A THEORY OF PERSONALISED NUDGING:

INTEGRATING HETEROGENEITY AND BEHAVIOURAL SCIENCE INTO POLITICAL DECISION-MAKING

S. Mills

PhD 2020

A THEORY OF PERSONALISED NUDGING:

INTEGRATING HETEROGENEITY AND BEHAVIOURAL SCIENCE INTO POLITICAL DECISION-MAKING

Stuart Mills

A thesis submitted in partial fulfilment of the requirements of

Manchester Metropolitan University for the degree of

Doctor of Philosophy

Department of Strategy, Enterprise and Sustainability

Manchester Metropolitan University

2020

Abstract:

While successful, behavioural nudges have often been one-size-fits-all, inducing different behaviours from different people despite both people being nudged in the same way. This is called the problem of heterogeneity, and one proposed solution is to personalise behavioural nudges.

One area where personalised nudges may be of pertinent interest is the online political advertising space. In recent years, concerns regarding the use of social media sites as part of highly targeted political campaigns have grown. For any personalised nudging programme, this is area of social significance.

This thesis investigates two strategies for personalising nudges using an experimental approach. Following an RCT experimental design (n = 962), the effect of impersonal nudges embedded into hypothetical political advertisements are first examined. The first part of this study finds limited evidence that impersonal nudges can influence decision making.

In the second part, two strategies for personalising nudges are used to investigate if personalisation renders nudging more effective in this domain. These strategies involve personalising the type of nudge shown to a participant (so-called delivery personalisation) and personalising the outcome which a participant is nudged towards (so-called choice personalisation).

Across all personalisation strategies (choice, delivery, and both combined), this thesis finds personalised nudges are statistically significantly more effective at influencing political decision-making than impersonal nudges and not nudging at all. Furthermore, data from the personalisation stage suggests further refinement of this experiment is possible, and so the effects of personalisation may be even greater than observed here.

Abbreviations

ANOVA	Analysis of Variance
ANS	Abbreviated Numeracy Scale
CFC	Consideration of Future Consequences
CO	Choice Only
CD	Choice and Delivery
DO	Delivery Only
FAFSA	Free Application for Federal Student Aid
GDMS	General Decision-Making Style
JNT	Johnson-Neyman Technique
NFC	Need For Cognition
OLS	Ordinary Least Squares
PTG	Personalised Treatment Group
RCT	Randomised Controlled Trial
SLMM	Simple Linear Moderation Model
UK	United Kingdom
US	United States
WMW	Wilcoxon-Mann-Whitley

Acknowledgements

The word *apocalypse* finds its origins in ancient Greek, with a literal translation being "uncovering" or a revelation of knowledge… according to *Wikipedia*. Adopting this very literal interpretation of the term, one may be tempted to describe a doctoral thesis as an apocalypse of sorts.

The notion of apocalypse has an intriguing place here – besides the tenuous link to revelation of knowledge – as I wrote this thesis, for the most part, during a global pandemic. I do not want to linger too much on this, mostly because I find it boring. I am self-aware enough to know that locking myself away in a room to write the best part of 100,000 words on a subject in minute detail was not because of a pandemic. It more comes with the territory of doing a PhD. But I think, as a note to myself if nothing else, it is important to remember the strange times during which this piece of work was born.

To really hammer home the idea of apocalypse, however, I suppose I must propose a candidate for quite what knowledge is being revealed here. Personalised nudging has an odd place in my heart. I remember reading page 1871 of Cass Sunstein's *Storrs Lectures* where he briefly discusses personalised nudges. I had had to leave the house because prospective buyers were viewing it, and so I sat in the car directly across the street as I read.

Anyway, I think like all nascent ideas, I had no clue where to really begin with it. I wrote some stuff, all of which was bad. I set the idea aside, occasionally returning to write more stuff… which was also bad. Eventually, I kind of gave up with the idea, and started working on other things. But I am nothing if not consistent, and my struggles with those other things caused me to give up on them too, and in a state without direction or optimism (an intellectual apocalypse in the more common sense of the word, perhaps?) I just started reading things.

I don't even remember how I discovered the idea of *hypernudge*, or what led me to Eyal Pe'er's then working paper on personalised nudges. I read that latter paper in July; it had perhaps only quietly arrived onto the internet a few months (if not weeks) earlier. But I loved the latter,

and with Karen Yeung's *hypernudge* concept, I suddenly found myself back to personalised nudging. That, more or less, is how this thesis began. This story is not so much a tale of *what* knowledge has been revealed, but more of *how* knowledge sometimes gets revealed. Perhaps, in these strange times, the how will provide us with more comfort than the what.

I should apologise at this point. I have rambled far too long, basically trying to disguise the fact that at the point I am meant to dispense some wisdom about life, the universe and everything I have little to say. I should, then, get into the main purpose of this part of the thesis.

I would like to thank, firstly, Richard Whittle and Kevin Albertson for their support, guidance and collaboration. I would also like to thank my parents, my sister, my girlfriend Bethany, Robbie who I love and Ben whom I will always love, and all those other relations who constitute in one way or another my clan. I would like to thank David Roberts, whose influence on my life trajectory I cannot overstate, and Jack Christian, for giving me my first research opportunity and for introducing me to David. I would like to thank Rebecca Weicht for all the challenging and entertaining discussions on the doctoral programme, and for proofing-read part of this thesis. In addition, I would like to thank Kevin Price for his interest, compassion and conversation regarding my work, my work, and pretty much everything else the progression of history seems to bring our way. Finally, I am grateful to Manchester Metropolitan University and the Royal Economics Society for supporting this research.

Declarations

The author declares no conflict of interest. All work presented in this thesis is the author’s own. The author acknowledges financial support from the Royal Economics Society and Manchester Metropolitan University.

Related Works and Publications

The following publications have been developed and published during the process of writing this thesis. The content of these publications has contributed, wholly or partially, to the content of this thesis:

> Mills, S (*forthcoming*) ‘Personalized Nudging’ *Behavioural Public Policy*. DOI: 10.1017/bpp.2020.7

The following works have been developed but not yet published during the process of writing this thesis. The content of these works has contributed, wholly or partially, to the content of this thesis:

> Mills, S (2020) ‘*Personalised Nudging and Political Decision-Making: An Experimental Approach’* Unpublished Manuscript.

Additional Works and Publications

The following works have been developed during the process of writing this thesis, but have not contributed to the content of this thesis:

> Mills, S (2020) ‘*Nudge/Sludge Symmetry*’ in review with *Behavioural Public Policy*

> Mills, S (2020) ‘*#DeleteFacebook: From Popular Protest to New Model of Platform Capitalism?*’ in review with *New Political Economy*

> Mills, S (2020) ‘*Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership’* Future Economies Research and Policy Paper #7. [Online] [Date

accessed: 06/10/2020]:

https://www.mmu.ac.uk/media/mmuacuk/content/documents/business-school/future-economies/Mills-2020.pdf

Mills, S, Whittle, R, Albertson, K (2020) '*Nudging Vaccination Engagement: A review of the experimental evidence in response to COVID-19'* in review with *Social Science and Medicine*

Mills, S, Whittle, R, Brown, G '*The Implications of a Crisis-Driven Societal Shift to Online Consumption*' in Vorley, T, and McCann '*Productivity and the Pandemic: The Way Forward*' Edward Elgar (*forthcoming*).

Sætra, H S, Mills, S (2020) '*Psychological Force, Liberty and Technology*' in review with *The Journal of Ethics*

Table of Contents

Table of Figures

List of Tables

This thesis is dedicated to the memory of David Roberts.

Chapter 1 – Introduction

> "When you act, you are" – Slavoj Žižek, on Hegel.
>
> "Melancholic and fascinated, such is our general situation in an era of involuntary transparency" – Jean Baudrillard, *Simulacra and Simulation (1981)*
>
> "Am I out of touch? No, it's the children who are wrong" – Seymour Skinner, Principal, Springfield Elementary, *The Simpsons, Season 5, episode 20.*

The story of behavioural theory begins with the realisation that humans often diverge from the rational, utility maximising model of human behaviour (Thaler, 2015). To many humans, rather than so-called *econs* (Thaler and Sunstein, 2008), this fact is tremendously evident. People rarely have perfect knowledge of any particular situation or decision, rarely have sufficient cognitive capacity to meaningfully analyse this information even if they did have it, and are rarely equipped with the tools to conduct an analysis which is sufficiently detailed and accurate (Simon, 1955).

The notion of nudging involves, at its heart, two aspects of behavioural theory. The first is that humans exhibit behaviour which diverges from the utility maximising model of human behaviour, and thus interventions which reduce such diversions may lead to better outcomes. The second is that people often diverge in a systemic fashion, allowing behaviours to be understood through an expansive taxonomy of biases (Gigerenzer, 2018; Thaler and Sunstein, 2008). Nudging as an outgrowth of behavioural theory holds that because people exhibit systemic biases which cause them to diverge from a utility maximising (or even utility *increasing*) path, small changes to so-called choice architecture (i.e. the landscape through which choices are posed and decisions are made) can be used to *nudge* them to make decisions which will leave them "better off" (Thaler and Sunstein, 2008, p. 5; Thaler and Sunstein, 2003, p. 175).

1.1 – Two Perspectives on Personalised Nudging

Behavioural nudges have become a popular feature of the policymaker toolkit (Oliver, 2019; Sanders, Snijders and Hallsworth, 2018; Halpern, 2015). The premise – to affect significant and predictable changes in behaviour without imposing significant economic (dis)incentives (Thaler and Sunstein, 2008) – has been a tempting and appealing prospect, prompting a slew of research and application (Sanders, Snijders and Hallsworth, 2018).

Along with this enthusiasm, however, has come frequent criticism (Rebonato, 2014; Gigerenzer, 2015). One common criticism is that many nudges take a one-size-fits-all approach (Ruggeri et al., forthcoming; Peer et al., 2019; Yeung, 2017; Sunstein, 2012; Carroll et al., 2009). This typically means that the same nudge is used for all members of a target population, despite the fact members of this population likely differ in meaningful ways, meaning the result of the nudge can produce very different – and potentially harmful – outcomes (Sunstein, 2012). This problem is known as the problem of heterogeneity (Sunstein, 2012), where heterogeneity describes individual differences between people.

Since the relative infancy of nudge theory, there have been those that have criticised nudging because of the problem of heterogeneity (Rizzo and Whitman, 2009; Carroll et al., 2009), as well as proposals to rectify the problem of heterogeneity (Sunstein, 2012, 2013a). One proposed solution is *personalised* nudging (Mills, forthcoming; Ruggeri et al., forthcoming; Peer et al. 2019; Sunstein, 2012, 2013a). By personalising nudges such that every individual, or every significantly different group of individuals, are nudged in such a way as would be expected to respect their heterogeneity, the problem of heterogeneity might be resolved.

This is one perspective on personalised nudging. Another comes from the perspective of information technology and digital nudging (Benartzi, 2017; Yeung, 2017; Weinmann, Schneider and vom Brocke, 2016; Thaler and Tucker, 2013). Here, the discussion follows that information technologies enable the collection of ever-more data about individuals, with *digital* choice environments expanding the range of choice architecture available to those who would seek to nudge others (i.e. choice architects). In addition, these environments are often highly personal spaces, resolving an implicit problem associated with personalised nudging: that of targeting (Benartzi, 2017). For advocates of digital nudging and digital choice environments, the growing ability to combine individual-level data with behavioural insights will produce a world inhabited by ever-more personalised nudges (Schöning, Matt and Hess, 2019; Yeung, 2017). This is not so much a response *to a problem*, but an embrace *of an opportunity*.

Both of these perspectives are valid, and often appear interconnected. For instance, it has often been noted that the past and present challenges of personalised nudging have been access to data and methodologies capable of analysing these data (Mills, forthcoming; Ruggeri et al., forthcoming; Porat and Strahilevitz, 2014; Sunstein, 2013a; Thaler and Tucker, 2013). As the information era develops, several of these challenges – notably targeting (Liu, 2020; Zuboff, 2019) – are being addressed. At present, however, there is a lack of a consistent theory of personalised nudging, as well as empirical examination of personalised nudging. When one says 'personalisation,' what is *actually being personalised?* When one states 'heterogeneity,' what *does this actually mean,* in the context of nudging? And when one talks of 'integrating heterogeneity,' what, *on a practical level*, does this entail? Furthermore, and perhaps most importantly of all, does personalisation actually deliver on the promise of resolving the problem of heterogeneity and producing more effective behavioural nudges? This is the central question which motivates this work.

1.2 – Beyond Targeting

In 2018, revelations about the role of the psychographics firm Cambridge Analytica and their role in several elections in 2016 came to light (Cadwalladr and Graham-Harrison, 2018). The firm was found to have harvested Facebook data from some 87 million users of the social media platform and combined these data with sophisticated data analytical procedures to micro-target political advertisements at users (Chang, 2018). They had, however, harvested these data without the permission of Facebook and – arguably more importantly – the *informed* consent of users (Lapowsky, 2019). The ensuing furore led to the collapse of the firm (Chang, 2018) and investigations by elected officials in both the UK and the US (Cadwalladr, 2019).

In developing such sophisticated micro-targeting techniques, Cambridge Analytica (and a whole slew of other technology companies at the time; Liu, 2020) had solved one of the key problems associated with personalised nudging, namely, knowing whom to nudge (Peer et al., 2019; Sunstein, 2012). But as Liu (2020) – a former start-up founder whose company specialised in micro-targeting advertisements on social media – argues, targeting is only half of the objective when trying to persuade someone. While targeting can identify an individual, knowing *what to target them with to affect behavioural change* can be rather different (Liu, 2020).

Personalised nudging emerges as a prime candidate for this 'what,' and in turn re-emphasises the timeliness of research into personalised nudging. Furthermore, it calls for research into personalised nudging in a very specific domain – namely, political decision-making. It is this domain which this thesis investigates.

1.3 – The Structure of this Thesis

This thesis is split into four *Sections*, with each Section containing multiple *Chapters*. This structure is adopted to assist the reader. Section 1 is titled *Background and Theory* and contains Chapters 2 and 3, which review the literature and propose a theory of personalised nudging, respectively.

In Chapter 2, the existing literature on personalisation is analysed. This discussion is broken into four parts. Firstly, the problem of heterogeneity is considered, with evidence from multiple studies which identify a problem in nudging arising due to heterogeneity discussed. This informs a discussion of how heterogeneity should be understood, and when heterogeneity should be rejected. Secondly, literature from the fields of marketing and consumer decision-making which investigate personalisation *without* nudging are considered. This research largely serves as a forerunner to the literature on personalised nudging which is currently emerging. Thirdly, said emergent literature is reviewed. Finally, the literature on digital nudging as it pertains to personalisation is examined, before a brief literature summary.

In Chapter 3, the central theory of this thesis is presented. This theory is dubbed the choice/delivery framework, and largely expands on the same framework proposed in Mills (forthcoming). The notions of choice personalisation and delivery personalisation are shown as clearly manifest in the literature, and the implications of this framework in relation to nudge theory are offered, as well as the hypotheses examined in this thesis.

Section 2 is titled *Methodology*, and – as the title suggests – concerns the method and methodological approach adopted in this thesis. This Section contains seven chapters. Chapter 4 considers the methods adopted by previous studies which are most relevant to this thesis. Chapter 5 presents an introduction to political advertising and decision-making before discussing the nudges examined in this thesis. Chapter 6 discusses the psychometric measures selected and proposes a psychometric map from which later results can be compared. Chapter 7 describes the process of constructing the political advertisements used in this thesis, and some of the experimental implications of these design choices, including the use of a randomised controlled trial (RCT) experimental design. Chapter 8 discusses data collection methods, as well as introducing two potential analytical approaches, matching and moderation analysis. Chapter 9 presents a power analysis of the proposed statistical tests and a discussion of sampling considerations and other factors which may impact the data collected. Chapter 10 provides a summary of the previous six chapters.

Section 3 is titled *Results* and consists of four chapters. Chapter 11 reports the findings of an initial pilot study, which is used to evaluate the experimental and analytical approach and adjust where necessary. These adjustments are offered in Chapter 12, and in Chapter 13, the findings of a second pilot study following these adjustments are presented. Chapter 14 provides the results of the main experiment in this thesis.

Finally, Section 4 is titled *Discussion and Conclusion*, and contains two chapters which discuss the results from Chapter 14 and provide concluding remarks, respectively. Chapter 15 discusses the results of Chapter 14 and the wider implications of this research. Firstly, the hypotheses proposed in this thesis are re-evaluated, with the first hypothesis being accepted and the second hypothesis being rejected. Secondly, explanations as to the relative performance of the personalisation strategies utilised in this thesis are offered, as well as proposed experimental adjustments to any future research. Thirdly, the results of this thesis are compared to those of previous research, with the apparent conclusion being that this research is broadly in-line with recent studies. Finally, a discussion of the wider implications of personalised nudging on society is offered. Chapter 16 concludes.

Section 1:

Background and Theory

Chapter 2 – Literature Review

2.1 – Introduction

Behavioural nudges have become important tools in the fields of public policymaking (Oliver, forthcoming; Sanders, Snijders and Hallsworth, 2018; Halpern, 2015), marketing (Akerlof and Shiller, 2017; Thaler, 2015) and widely used in the private-sector (Akerlof and Shiller, 2017; Lavi, 2017; Beggs, 2016; Sunstein, 2013a).

However, nudges are often criticised for their one-size-fits-all approach (Peer et al., 2019; Yeung, 2017; Sunstein, 2012; Carroll et al., 2009). For instance, Thunström, Gilbert and Jones-Ritten (2018) find that nudges which encourage saving behaviour can have a negative impact on individuals who already over-save. Sunstein (2012, 2013a) argues such phenomena occur because populations are heterogeneous – individuals and groups are different from one another, and these differences may result in different welfare outcomes from the same nudge (Sunstein, 2012).

Sunstein (2012) calls this, "the problem of heterogeneity" (Sunstein, 2012: 6) and argues, in many circumstances, it is desirable to respect heterogeneity. One solution may be to encourage active choices, but these can also be burdensome. An alternative, therefore, is personalisation and personalised nudging.

This chapter reviews the literature concerning the problem of heterogeneity, personalisation, and personalised nudging. In part 2.2, the problem of heterogeneity is examined. Following a brief synopsis of the difficulties of measuring heterogeneous populations, several behavioural studies which demonstrate unexpected and unintended results – and which show strong evidence of being explained by heterogeneity – are reviewed to evidence the problem of heterogeneity. These studies also inform the critique of Sunstein's (2012) relevancy principle, which seeks to conceptualise the broad concept of heterogeneity so that it may be useful within nudge theory.

In part 2.3, the existing body of literature on message personalisation in the fields of marketing and consumer decision-making is examined, before the relatively smaller body of literature on personalised nudging is considered. Significant evidence suggests that personalising or tailoring messages to match the cognitive styles of heterogeneous populations improves the likelihood the message will promote its intended consequence, bolstering the idea of personalised nudging.

The origin of this idea is then examined, turning to the original arguments of Sunstein (2012, 2013a) and the subsequent discussions about personalised nudging within the legal domain (Porat and Strahilevitz, 2014), before more contemporary empirical studies of personalised nudging are examined (Guo et al., 2020; Page, Castleman and Meyer, 2020; Peer et al., 2019; Schöning, Matt and Hess, 2019). As with the personalisation literature, these contemporary works show early evidence of the effectiveness of personalised nudging.

Finally, part 2.3 of this chapter considers the conceptions of personalisation and personalised nudging which have been developed in conjunction with the emergence of information technologies, automated systems and big data (Ruggeri et al., forthcoming; Benartzi, 2017; Yeung, 2017; Weinmann, Schneider and vom Brocke, 2016). In this review, it is argued that, while these sophisticated technological strategies facilitate personalised nudging, and will bring about *sophisticated*, personalised nudges, so-called crude personalised nudges (Porat and Strahilevitz, 2014; Sunstein, 2012, 2013a) demonstrate personalised nudging should be thought of as a *response* to the problem of heterogeneity, and not as an outgrowth of information technology.

This chapter concludes with part 2.4.

2.2 – Does Heterogeneity Matter?

It is worth considering whether heterogeneity matters, which necessitates an exploration of the effects of heterogeneity. A contentious argument around heterogeneity and nudging is that because nudges should not prevent a person from pursuing their own preferences, individuals should always be able to identify when a nudge is potentially harmful to their interests and adjust their behaviour accordingly. However, for this argument to be valid, it would simultaneously invalidate the purpose for nudging in the first instance; if decision-makers were evaluative enough to determine whether their specific preferences required them to exercise agency over the nudge, it seems reasonable to believe they also will be evaluative enough to make an optimal decision without requiring any nudge. Since most nudge theorists and behavioural economists reject the latter proposition,[1] it seems viable, if not necessarily intuitive, to suppose that whether or not individuals could avoid the potentially harmful consequences of nudges and go their own way, they do not do so.

Reasons for this behaviour may be speculated upon. For instance, a person who is relatively uninformed of, say, pension plans and so is inclined to follow a default option nudge will probably be uninformed of how their personal circumstances may mean the default plan is not right for them. This person, despite being significantly heterogeneous, may not go their own way, and thus the problem of heterogeneity emerges. In this part, evidence of the problem of heterogeneity will be presented. The question of what heterogeneity means will also be considered.

2.2.1 The Myth of the Average Person

The so-called problem of heterogeneity arises because people have individual characteristics, circumstances and preferences; they do not conform to a single set of specifications. While the belief in, say, assuming a population average is representative of a whole population can be traced to the emergence of population[2] data collection and social statistics in the 19th

[1] Indeed, Thaler and Sunstein's (2003) whole argument for nudges and behavioural interventions is that people often don't exhibit optimal behaviour.

[2] There is no standard method of evaluating a population down to a single representative individual in behavioural economics. The use of averages is a common strategy, particularly with social norm nudges (Schultz et al., 2007), but other less structured strategies can also be used (for instance, Butt et al. (2018) find choice architects try to estimate default plans based on what they think is best for those immediately available to them).

century – notably to the Belgium statistician Adolphe Quêtelet (Sposini, 2019) – the validity of this belief has also been challenged.

One notable example is the work of anthropologist Gilbert S. Daniels (1952) and his study of the average man within the U.S. Airforce. Daniels (1952) found that as the number of variables[3] used to discern the average man[4] increases, the number of observations (which is to say, people) which can be adequately represented by this average rapidly declines to zero. As Daniels (1952) writes as an introduction to his work, "The tendency to think in terms of the "average man" is a pitfall into which many persons blunder" (Daniels, 1952: 5).[5] This same tendency, and blunder, might be applied to the fields of behavioural science and nudge theory specifically[6] (Peer et al., 2019; Sunstein, 2013a).

2.2.2 The Heterogeneity Problem in Action

While the problem of heterogeneity in nudge theory has been a noted criticism of nudges for some time (Porat and Strahilevitz, 2014; Sunstein, 2013a; Johnson et al., 2012; Carroll et al., 2009), this discussion has largely been the reserve of theorists (Sunstein, 2013a). Only recently has the study of nudges sort to highlight the unintended and potentially harmful side-effects of some nudges (Thunström, Gilbert and Jones-Ritten, 2018; Beshears et al., 2016; Beshears et al., 2015b; Haggag and Paci, 2014; Schultz et al., 2007). This literature is explored to present evidence of the problem of heterogeneity within nudge theory. Furthermore, with an eye to heterogeneity, a critique of the effectiveness of nudging can also be made.[7]

---

[3] In this instance, bodily features (weight, height etc.).
[4] Daniels' (1952) work was restricted solely to men.
[5] Quêtelot's work, for example, led him to believe nations and races could be understood through comparison of their average man, and his work – according to Sposini (2019) – contributed somewhat to the legitimising of the fields of phrenology (skull measuring) and eugenics.
[6] For instance, averaging in the social norm nudge (Schultz et al., 2007).
[7] For instance, heterogeneity may lead many people to follow the nudge, and so from a non-heterogenous (homogenous) perspective the nudge can look highly effective. However, when considering some of those following the nudge are being negatively impacted due to heterogeneity, the effectiveness of the nudge (from a welfare perspective (Oliver, 2019; Sunstein, 2013b)) comes under question.

2.2.2.1 Thunström, Gilbert and Jones-Ritten (2018)

Thunström, Gilbert and Jones-Ritten (2018) investigate the use of a salience nudge to encourage people to reduce spending.[8] They define two groups within their target population, the so-called "tightwads" and "spendthrifts" (Thunström, Gilbert and Jones-Ritten, 2018: 268). As these names suggest, these are individuals who spend too little and individuals who spend too much, respectively.[9] Participants were offered the opportunity to buy locally produced honey, but were also informed (nudged) before making any purchase that said purchase would reduce their ability to buy alternative items in the future.

Thunström, Gilbert and Jones-Ritten (2018) find that for those who felt they spend too little (tightwads), this salience nudge was highly effective at discouraging them from spending more. However, for those who felt they spend too much, the salience nudge had no impact on their spending behaviour.[10] Thunström, Gilbert and Jones-Ritten (2018) therefore conclude that for tightwads who may have benefited from spending, the nudge actually *reduced* their welfare, while for those who would have benefited from following the nudge and not spending, the nudge did not enhance welfare.

Tightwad and spendthrift classifications represent an attempt to capture heterogeneity within the target population.[11] This is not necessarily typical of comparable nudges,[12] but having done so, Thunström, Gilbert and Jones-Ritten (2018) reveal additional insights about the nudge which would otherwise have been lost. For instance, without this heterogeneity information,

---

[8] Salience nudges typically highlight information which may otherwise be missed or not appropriately appreciated. Thunström, Gilbert and Jones-Ritten (2018) specifically use a reminder nudge.
[9] Thunström, Gilbert and Jones-Ritten (2018) define these categories around self-reported spending habits. In other words, those who believe they either spend too much or too little.
[10] Perhaps because spendthrifts already had a lot of experience facing opportunity cost information and were thus less sensitive to this information.
[11] This was a primary research objective of Thunström, Gilbert and Jones-Ritten (2018), and not simply a curious result. As they note, "[the] distributional effects of nudges are largely unknown" (Thunström, Gilbert and Jones-Ritten, 2018: 267)
[12] Examples include the UK Government's decision to encourage retirement saving by making workplace pensions opt-out rather than opt-in (Service, 2015), or Thaler and Benartzi's (2004) 'Save More Tomorrow' similarly designed to encourage retirement saving. In neither example is the heterogeneity of the target population considered. Evidence from Bourquin, Cribb and Emmerson (2020) and Beshears et al. (2016), respectively, suggest this may be an oversight.

the effect of the nudge may have been interpreted positively because the nudge did reduce overall spending across the sample.[13] However, because some heterogeneity within the sample has been collected, it is possible to understand for whom the nudge was effective, and re-evaluative whether the nudge was successful in enhancing welfare (Thaler and Sunstein, 2003, 2008) or was – in the words of Tor (forthcoming) – a “successful but undesirable” nudge “that should fail” (Tor, forthcoming: 1).

2.2.2.2 Butt et al. (2018)

Butt et al. (2018) argue *a priori* that heterogeneity may result in default employee retirement saving schemes not being suitable for many members. They adopt a mixed-methods approach, collecting various survey data[14] from over 1,000 employees of various Australian companies, before interviewing 28 executives from the same companies who are charged with establishing default retirement saving schemes for their respective companies.

Butt et al. (2018) find that schemes which are set as the default option do not reflect heterogeneity within the workforce. Of those who accepted the default plan, “the 18-34 years (youngest) age group is over-represented… as are women, singles, people with low education and low to middle income earners” (Butt et al., 2018: 553). This leads Butt et al. (2018) to argue that the over-representation of these groups suggests they are more susceptible to the default option.[15] Yet the default option is often not designed with these specific groups in mind and is instead designed to reflect – as best as possible – the whole workforce population.

Butt et al. (2018) offer another interesting observation, namely, that where attempts are made to tailor the default option scheme to individuals, they tend to rely only on those characteristics that can be easily observed or inferred: “Executives said that they designed their plan’s default

[13] Of course, this interpretation reveals something of a normative perspective on the relationship between welfare and saving. One might speculate about the welfare of those reliant on businesses now denied additional income, for instance.
[14] Survey data ranged from basic demographic information (age, sex, income level, education level) to information about risk preference, attitudes to decision making, and attitudes regarding the participant’s retirement scheme.
[15] Also see Beshears et al. (2015a)

investment strategy for passive members. At the same time, they acknowledged that they know little about them. Executives know the age, gender, plan account balance and insurance status of members, and can identify the martial status of some. By using the mandatory contribution rate as a guide, they can estimate members' incomes" (Butt et al., 2018: 553). Butt et al. (2018) argue that this reliance on easy to access information leads executives to ignore other important heterogeneous information such as risk preference, and further, through this ignorance, hinders their ability to interpret what the easy to access information implies about their staff. For instance, when investigating the risk preference of default plan members compared to the actual risk-level of their plans, Butt et al. (2018) find significant mismatching. This pattern was repeatedly found when they compared other survey metrics such as propensity to delegate or saving goals.

2.2.2.3 Rivers, Shenstone-Harris and Young (2017)

Rivers, Shenstone-Harris and Young (2017) investigate the long-term behavioural change in consumption resulting from a small charge being placed on plastic shopping bags. The use of a small charge is a famous nudge discussed by Thaler and Sunstein (2008), with the charge designed not to be significant enough as to constitute an economic disincentive, but salient enough as to remind consumers about the wastefulness of plastic bags and nudge them to use reusable alternatives.

Rivers, Shenstone-Harris and Young (2017) document several uses of the small charge nudge around the world and subsequent studies which attest to the success of the nudge.[16] However, they argue that, "many of these studies are flawed because they lack adequate temporal and geographic controls" (Rivers, Shenstone-Harris and Young, 2017: 153). To resolve these flaws, they use a natural experiment arising from the Canadian city of Toronto between the years of 2006 and 2013. In that time, Rivers, Shenstone-Harris and Young (2017) report, a

[16] Success is taken to be a reduction in the usage of disposal plastic bags which can be attributable to the nudge.

small charge was introduced on disposal plastic bags while not being introduced anywhere else in the country. This charge was eliminated following the election of a new mayoral administration. Using the rest of Canada as a control group, Rivers, Shenstone-Harris and Young (2017) evaluate whether the introduction of the charge in Toronto produced short- and long-term behavioural change in the consumption of plastic bags.

Overall, they find that the nudge did affect behavioural change, with an overall 3.4% increase in the use of reusable plastic bags.[17] However, given the detail of the data, they also find, "the levy was highly effective in encouraging people who already used reusable bags to use them more frequently, while having no effect on infrequent users" (Rivers, Shenstone-Harris and Young, 2017: 153).[18] Rivers, Shenstone-Harris and Young (2017) argue this result can be explained by the heterogeneity within the population. When the effectiveness of the nudge is analysed along household income, they find consistent evidence to suggest those with high household earnings change their behaviour following the nudge, while those in lower household income brackets show no significant adjustment in their behaviour.[19]

They link this heterogeneity result to the idea of nudge transparency, suggesting that those who have higher incomes are positioned to be more informed about the policy and the potential harms of disposable bag use, and thus are more susceptible to being nudged compared to those with lower incomes who may, for a variety of socio-economic reasons, not

[17] The data examined by Rivers, Shenstone-Harris and Young (2017) is survey data that captures household consumption on a Likert scale measuring frequency of use of *reusable* plastic bags. Therefore, they report increases in reusable plastic bags, not decreases in disposable plastic bags. This may be disadvantageous in some discussions. However, Rivers, Shenstone-Harris and Young (2017) argue the nature of the data (that it can be compared with households across the country unaffected by the charge) makes this a worthwhile compromise.

[18] This finding is possibly explained by the nature of the nudge itself. If, as Rivers, Shenstone-Harris and Young (2017) argue, the small charge nudge works by reminding consumers that disposable bags should be avoided, those who have already accepted this message (i.e. those who already use reusable bags) will benefit from being reminded, while those who have not accepted the message (i.e. those who have no impetus to use reusable bags) may find themselves, in a manner of speaking, protected from the nudge. This explanation is similar to that found by Thunström, Gilbert and Jones-Ritten (2018).

[19] Rivers, Shenstone-Harris and Young (2017) do note a curious result in the middle-income bracket ($40,000-$60,000), where the propensity to use reusable bags actually fell when the nudge was introduced. They do not elaborate on why this might be.

be as informed.[20] This explanation may thus reveal a slightly different perspective on the problem of heterogeneity. Where heterogeneity is often characterised as resulting in harm for some individuals (Thunström, Gilbert and Jones-Ritten, 2018; Beshears et al., 2016), the effectiveness of the nudge itself may also be subject to heterogeneous factors. For instance, should the hypothesis of Rivers, Shenstone-Harris and Young (2017) regarding lower-income households be correct, it may be a prudent observation that for some groups either a different nudge[21] or a different policy program entirely, is desirable.[22]

2.2.2.4 Beshears et al. (2016)

Beshears et al. (2016) present evidence of an unintended behavioural response which may be explained by the presence of heterogeneity in their study of present bias nudges designed to increase retirement saving. Beshears et al. (2016) re-examine Thaler and Benartzi's (2004) classic work on the present bias nudge and their proposal for a present bias retirement saving scheme known as 'Save More Tomorrow.' The present bias suggests individuals would rather receive gains immediately, while putting off losses until sometime in the future (O'Donoghue and Rabin, 2015; Laibson, 1994). Believing that people are reluctant to forgo present consumption in order to save, but more than willing to put off a 'loss'[23] until the future, Thaler and Benartzi (2004) utilise the present bias to nudge workers to commit to saving part of their *future* income. They report that the 'Save More Tomorrow' plan significantly increased the number of workers saving for their retirement.

Beshears et al. (2016) also find a positive impact from the present bias nudge. However, while they find that many participants pre-committed to saving, they also find that a notable number of participants reneged on that commitment. Beshears et al. (2016) argue that this behaviour

[20] Of course, other explanations may persist. For instance, those on lower incomes may see the purchase of reusable plastic bags, which are often more expensive in the short-term, as a luxury or excessive cost.
[21] Say, a small subsidy on reusable plastic bags which appeals to the economic interests of lower earners.
[22] Say, free public service information on the harms of disposable plastics to the environment.
[23] While not a loss, it can be useful when considering the present bias nudge to describe income forgone in the form of savings as a loss of income which could have been used for other activities.

was the result of a phenomenon known as information leakage (Sher and McKenzie, 2006; McKenzie and Nelson, 2003; Madrian and Shea, 2001).

Information leakage occurs when decision-makers infer additional information about a choice based on the framing and choice architecture of the choice. For example, Sher and McKenzie (2006) offer a thought experiment involving two statements, A and B. First, they establish that the likelihood of A being used to nudge is contingent on some condition C being met. Second, they suggest that the choice architect (the "nudger") and the decision-maker (the "nudged") have access to some common information.[24] Based on these criteria, Sher and McKenzie (2006) argue that despite the fact that, at any time both statements could be presented (regardless of the actuality of condition C), based on what statement is presented (A or B), the decision-maker will always be able to infer some additional information about the choice and condition C.[25]

Beshears et al. (2016) offer information leakage as an explanation of the failure of some workers to fulfil their commitment to begin saving. They argue that the use of the present bias unintentionally leaks a message to some participants that saving commitments can always be deferred into the future. As such, when the time came to begin saving, this message allowed some participants to renege on that commitment.

The role of heterogeneity in this study is revealed by the fact that only *some* participants reneged, while others didn't, with Beshears et al. (2016) seemingly unable to find an alternative explanation which would explain these observed behaviours. For instance, it may have been the case that only low-earners reneged on their commitment. However, even allowing this to be true – and rejecting the information leakage explanation – this would still

[24] For instance, that saving money is generally considered a good habit, and something to encourage others to do.

[25] For instance, if the nudge encourages saving, and the decision-maker knows that generally saving is considered a good habit, the decision-maker can infer that the choice architect wants them to save more.

be evidence of heterogeneity within the target population resulting in harm and/or undesirable behaviour from part of the target population.

2.2.2.5 Beshears et al. (2015b)

Not to be confused with Beshears et al. (2016), Beshears et al. (2015b) investigate the effect of providing individuals with information about the workplace pension saving behaviour of their peers, and in turn, nudging individuals using a social norm nudge. They investigate uptake of 401k programs by employees in a large American manufacturing company and distinguish between two groups of employees not contributing at a baseline contribution rate. The first group are those who are not enrolled in the scheme at all, and so are said to have a contribution rate of 0%.[26] The second group are those who are enrolled in the scheme but are contributing less than the typical contribution rate for that scheme, which Beshears et al. (2015b) state is 6%. This second group are said to be those contributing less than 6%, but more than 0%.

After providing these employees with information about their peers who were meeting the 6% baseline contribution rate, Beshears et al. (2015b) find significant increases in saving amongst those already enrolled in the program (the less than 6% group), confirming their hypothesis that a social norm nudge can be used to effectively increase employee retirement saving. However, for those not enrolled in the scheme (the 0% group), they find the social norm nudge produces a significant, negative reaction compared to the behaviour of a control group.[27] Beshears et al. (2015b) write, “discouragement from upward social comparisons seems to drive this [negative] reaction” (Beshears et al., 2015b: 1161).

Beshears et al. (2015b) dub this phenomenon, “oppositional reaction,” (Beshears et al., 2015b: 1166) and posit that individuals who are highly and negatively divergent from the social norm

[26] It is implied that these employees have no retirement provision, though it may be the case that some have some private provision. However, it is unlikely all have a private plan.
[27] Beshears et al. (2015b) argue this negativity manifests as a reluctance to even consider saving at all, as these employees feel they are already so far behind they cannot possibly catch up to the norm.

become discouraged from ever achieving the social norm when provided with this information about their divergence, so much so that they give up and begin to exhibit behaviour which is counter to expectations,[28] in this instance, not enrolling in the saving scheme.

Once more, the heterogeneity in the target population seems to explain the observed phenomenon. As with Thunström, Gilbert and Jones-Ritten (2018), by examining the social norm nudge with a measure of heterogeneity already established (non-savers vs. low-savers vs. baseline-savers), Beshears et al. (2015b) can describe more detailed behavioural manifestations. Without a measure of heterogeneity, the social norm nudge appears to be highly effective because – across the targeted population – employees are encouraged to save more. But by distinguishing between low-savers and non-savers, it can be discerned that the welfare effects of the nudge[29] may in fact be negative for non-savers; ironically, the group most in need of saving.

2.2.2.6 Haggag and Paci (2014)

Haggag and Paci (2014) investigate the use of the default option nudge on customer tipping of taxi drivers in New York City. In this natural experiment, customers who chose to pay using a credit card were presented with a payment system offering three (default) tip values. Customers could also, should they choose, manually enter any value they wanted as a tip. For journeys costing less than $15, the default tip values were set at $2, $3, and $4, while for journeys costing more than $15, default tip values were calculated at 20%, 25%, and 30% of the journey cost.

Haggag and Paci (2014) assume that because both customers who pay in cash and those that use credit cards regularly tip and have the same freedom to choose how much to tip, these groups are comparable.[30] They suggest that the default tips shown to credit card

[28] Namely, conformity and convergence towards the norm. See Bernheim (1994) and Schultz et al. (2007).
[29] Assuming increased retirement saving does indeed enhance welfare.
[30] There may be some arguments to suggest that these groups aren't comparable (Prelec and Simester, 2001). For instance, a person's propensity to use cash or card may be indicative of their financial situation. Furthermore, the method itself may induce different behaviour. For instance, a credit card payment may be

customers will nudge these customers into tipping, and thus credit card customers will tip more than their cash equivalents.[31] This initially appears to be the case, with credit card customers tipping higher values than cash customers. However, Haggag and Paci (2014) also find that, as the value of the default tip increases, the likelihood of a customer leaving no tip also significantly increases.[32]

To explain this phenomenon, Haggag and Paci (2014) introduce Brehm's (1966) theory of reactance. Following Brehm (1966), reactance is said to occur when an individual responds to a suggestion or an attempt at coercion by demonstrating behaviour which is opposite to the behaviour being desired. Haggag and Paci (2014) speculate that as the recommended value of the default tip increases, some customers respond to the prompt to leave a larger-than-expected tip by tipping significantly less than they would have had they not been prompted at all.[33]

Initially, such a finding does not seem explicable in terms of heterogeneity. For instance, it seems reasonable to believe that all customers have a maximum absolute tip value which the default tip could potentially exceed. Under this circumstance, it would be expected that participants revise the default down when they come to leave a tip.[34] However, even if this behaviour could be expected to occur in all people, the value at which this behaviour would be exhibited likely varies between people, and so this behaviour could be described in terms

more convenient than a cash payment, and in turn tipping may be more convenient. Equally, cash customers may see tipping as a chance to dispose of burdensome change, for example. Nevertheless, for the purpose of their study, Haggag and Paci (2014) assume such arguments are not significant.

[31] It is possible that greater tips are wholly attributable to the tendency for people to tip more if they are using a credit card (Prelec and Simester, 2001). However. Haggag and Paci (2014) find a significant jump in tip amounts, compared to a competitor, at the default tipping amounts, leading them to conclude it is the default, and not the medium of payment, which is responsible for the higher tips.

[32] As might be expected, this result is observed significantly more for journeys costing more than $15, as the absolute value of these tips can grow to significantly greater values than the set values shown to customers with journeys costing less than $15.

[33] While Haggag and Paci (2014) do not discuss it, it is worthwhile to consider whether at some tip values the value moves from an insignificant economic cost usually permissible under nudging, to a more significant economic cost typically associated with shoving (Oliver, 2015). If so, one might expect people to demonstrate more resistance to these interventions, prompting the observed backlash.

[34] Such a response would not necessarily be reactance either, merely seem to be reactance. If one tips a value less than a default because they cannot afford the default value or cannot justify it, they are not necessarily reacting to the default value, but instead behaving in accordance with their economic beliefs/limitations.

of heterogeneity.[35] Furthermore, assuming this is *not* true, the cost of individual journeys is a piece of heterogeneity information which the default tips do not respect, and as such an unintended behaviour due to *this* heterogeneity occurs. Finally, accepting the reactance hypothesis, reactance may also be heterogeneous within the population (Brehm, 1966).

2.2.2.7 Schultz et al. (2007)

Schultz et al. (2007) examine the use of a social norm nudge within the household energy market. Specifically, they investigate how a social norm nudge could be used to reduce household energy use. Schultz et al. (2007) argue that while many social norm interventions have been successful, social norms also have a tendency, "[to act] as a magnet for behaviour for individuals both above and below the average" (Schultz et al., 2007: 430).[36] They suggest that this tendency towards the norm means not only will positive results be observed when nudging those currently exhibiting undesirable behaviour (high energy usage), but negative results may be observed from those currently exhibiting desirable behaviour (low energy usage).

Schultz et al. (2007) investigate this phenomenon by providing households with social norm information regarding average household energy use and measuring the energy use of those households after the nudge has been implemented. Initially, they find that providing high energy use households with information regarding *average* energy use leads to a significant reduction in energy usage by these households. However, Schultz et al. (2007) also find that this information produces a significant increase in the energy use of households which were previously low energy use households (below the average). Thus, they argue that the tendency for social norm nudges to lead decision makers to converge on the norm is correct within an energy market context.

[35] Which is to say, the value a person is willing to tip, and/or the value they are able to tip, are heterogeneous.
[36] While Schultz et al. (2007) use the word "average," other authors (Beshears et al., 2015b; Allcott, 2011) suggest social norms act more as frames or reference points for decision makers, and thus it may be more accurate for Schultz et al. to state the word 'norm' rather than 'average.' Nevertheless, the principle of the statement remains the same.

Schultz et al. (2007) do believe social norms are useful and do produce some positive effects. For instance, they report that the reduction in energy use by high energy use households was slightly greater in both the short- and the long-term than the comparable increase by low usage households. Thus, the *net* effect of this norm was positive despite the nudge encouraging undesirable behaviour from some households. However, by providing all households with the same social norm information and not respecting the heterogeneity between households (energy usage), the size of the positive benefit produced by the nudge is reduced.

2.2.3 The Relevancy Principle

As shown, the problem of heterogeneity can be found across a variety of nudges, populations and policy goals. In many circumstances, the nature of the heterogeneity which explains the undesirable behavioural response also varies, from propensity to spend (Thunström, Gilbert and Jones-Ritten, 2018) to the duration of one's taxi journey (Haggag and Paci, 2014). Introducing heterogeneity, therefore, produces an important problem.

Namely, it is plausible that a great many results could be re-analysed with respect to a large number of measures of heterogeneity and some evidence of significant heterogeneous differences be found. Eye colour, for instance, would not be expected to influence spending behaviour, but given such data, it is possible such a statistical quirk might be found. Furthermore, allowing the number of heterogeneity criteria to increase, as Daniels (1952) has shown, quickly suggests that a single, one-size-fits-all approach will not adequately satisfy anyone within a target population.[37] This is to say, without a clear understanding of what heterogeneity means within a given context, appeals to heterogeneity could easily be used to undermine a great many policies, behavioural or otherwise.

Sunstein (2012) recognises this problem and makes efforts to limit any discussion of heterogeneity within nudges by stipulating that heterogeneity be *relevant to the context that*

[37] This may be a particular issue for social norm nudges (Beshears et al., 2015b; Schultz et al., 2007) that must generate a single standard to nudge a population.

the nudge is being used. For instance, leaving aside the many anthropological measures such as height and body-mass, propensity to spend and duration of one's taxi journey independently appear to reflect relevant heterogeneity in different contexts, namely reducing spending and tipping for taxi journeys, respectively. Within the context that heterogeneity is measured, propensity to spend and duration of one's taxi journey are relevant heterogeneity data.[38] Yet, it is harder to justify the relevance of these examples of heterogeneity information when the contexts are switched. Of course, there may be some reason to speculate at relevancy, and often choice architects (i.e. nudgers) may not know what information is relevant to their nudges until after-the-fact (Yeung, 2017; Rizzo and Whitman, 2009). But, on the whole, the use of these measures of heterogeneity is subject to their relevance within the context that the nudge is being used.

Sunstein (2012) offers further arguments to support the relevancy principle. He argues that acquiring heterogeneity information could be very costly because of the added level of detail which is required. Because of this, collecting any additional heterogeneity information beyond what is immediately required[39] to personalise the nudge reduces the potential net benefits produced by the nudge. Furthermore, because heterogeneity may often take the form of personal information and data, Sunstein (2012) argues relevancy is crucial to ensuring the privacy of individuals is not infringed. Following this argument, Sunstein (2012) suggests that acquiring heterogeneity information that is not relevant to the nudge being implemented unnecessarily violates the privacy of the target population and may therefore be unjustified.

Sunstein (2013a) offers some additional thoughts on the nature of heterogeneity which further inform the relevancy principle. Invoking Mill ([1859] 2015), Sunstein (2013a) acknowledges that allowing individuals to pursue their own interests can be important to identity formation and learning and suggests that a factor in determining the relevancy of heterogeneity should be whether ignoring heterogeneity would limit these opportunities. This argument is a

[38] See Thunström et al. (2018) and Haggag and Paci (2014).
[39] Which is to say relevant.

departure from the arguments of Sunstein (2012) which have generally focused on welfare maximisation and cost benefit comparison.[40]

Sunstein (2013a) also suggests, though does not explicitly state, that the type of nudge being used may also be an important factor when determining if heterogeneity is relevant or not. For instance, large variances in apparent discount rates have been found in investigations of the present bias and hyperbolic discounting (Frederick, Loewenstein and O'Donoghue, 2002), with Frederick, Loewenstein and O'Donoghue (2002) concluding it is unlikely there is any natural or common rate at which people discount future events.[41] While the range of possible discount rates can vary significantly between people,[42] lending greater importance to heterogeneity, the range of possible selections when, say, a default option is used is limited to the number of options available. In some circumstances, such as workplace pension schemes, these options may be as simple as opt-in or opt-out (Service, 2015). Therefore, because some nudges have more ability and reason to accommodate heterogeneity, while others do not, the nudge itself may be a contingent factor when evaluating the relevancy of heterogeneity.

The relevancy principle, then, reconciles two aspects of heterogeneity. By requiring any heterogeneity information to be relevant only to the context that the nudge is being used, it leaves choice architects (i.e. nudgers) free to think about heterogeneity in a way that is representative of the multiplicity of the concept. But equally, by demanding relevancy, it stops this multiplicity being used to overly complicate or unhelpfully undermine policies such as nudges. In this sense, the relevancy principle provides a helpful focus to the notion of respecting heterogeneity by allowing some *irrelevant* heterogeneity to not be respected.

[40] It could be argued, depending on how one defines it, that enabling identity formation and learning is also maximising welfare. However, it is worthwhile here to consider these ideas as distinct to the concept of welfare discussed previously.
[41] Sunstein (2013a) uses these as examples in his discussion of heterogeneity, writing, "different discount rates can reasonably be chosen by people who are in different life circumstances," (Sunstein, 2013a: 1870).
[42] And based on the work of Frederick, Loewenstein and O'Donoghue (2002) would be expected to do so.

Further exploration of the relevancy yields additional insights also. Where the costs of gathering heterogeneity information are high, relevancy may be a pressing principle. However, where the costs are reasonably low, relevancy could seem less significant, until one considers the potential privacy implications associated with personal information.[43] Obvious costs[44] may also not be a sufficient metric by which to assess relevancy, as some choices may be important to personal development, and thus justify respecting heterogeneity, even if an immediate cost-benefit analysis would not support this conclusion.[45] Finally, it may be easier to integrate heterogeneity into some nudges compared to others, suggesting that nudges may also mediate the relevancy principle.[46]

2.3 – Personalised Nudging as a Concept

Personalised behavioural nudges are offered as a means of integrating heterogeneity information about a target population of decision-makers into behavioural science, and as such, ameliorate[47] any harms or unintended behaviours which result from impersonal nudging.

The concept of heterogeneity has already been explored, with Sunstein's (2012) specification of "relevant" (Sunstein, 2012: 4) heterogeneity serving as a reasonable touchstone. This section considers the existing literature on personalisation, which can broadly be split into four

[43] It is worthwhile to recall that this privacy argument is made by Sunstein (2012), a vocal advocate of cost-benefit analysis in public policy. However, cost-benefit analysis is potentially difficult and subjective when concerned with notions such as individual data privacy.
[44] Such as the costs of collecting heterogeneity information via surveys or developing choice architecture to embed heterogeneity.
[45] One such example may be choosing from a restaurant menu. For those seeking to improve their diet, nudging diners towards healthy options via menu design may be a worthwhile strategy. However, meal selection remains a choice subject to tremendous personal taste, and while it might be costly to gather heterogeneity information on diners, and while the welfare-maximising outcome may always be to have a salad, there is scope to try and marry a nudge towards *generally* heathy options with personal taste which may diverge somewhat from the *healthiest* option.
[46] I.e. where it is easier to integrate heterogeneity into a nudge, the 'burden' of relevancy may be less so.
[47] The word ameliorate is used here as opposed to, say, eliminate, for two reasons. Firstly, it is unlikely any choice architect could reliably determine that all side-effects resulting from heterogeneity have been eliminated, simply because it may not be possible to measure all potential side-effects (Dolan and Galizzi, 2015). For instance, how could we possibly know all the inferences which result from information leakage (McKenzie and Nelson, 2003)? Secondly, there is likely a marginal cost to personalisation (Sunstein, 2012), and as such, it is likely the cost of any personalisation which would eliminate all the side-effects of heterogeneity (theoretically) would quickly exceed any calculable benefits received. For these reasons, it seems more appropriate to view personalisation as a *reduction* of sorts, rather than an antidote.

categories, reflecting the key strands of thought between authors and aiding the structure of this part of this chapter. These categories are i) personalisation without nudging; ii) personalised nudging; iii); empirical investigations of personalised nudges; and iv) personalised nudging as an outgrowth of technologies.[48]

2.3.1 Personalisation Without Nudging

Personalisation exists as a broad idea external to nudge theory and finds its roots largely in the fields of consumer psychology and marketing (Matz et al., 2017; Dubois, Rucker and Galinsky, 2016; Hirsh, Kang and Bodenhausen, 2012; Cesario, Higgins and Scholer, 2008; Cesario, Grant and Higgins, 2004; Moon, 2002). For this reason, it is first appropriate to consider the concept of personalisation as distinct from the sub-concept of personalised nudging, before exploring how nudge theory may integrate with personalisation.

Matz et al. (2017) investigate how social media advertisements for several products, from beauty products to mobile game applications, can be targeted to users who exhibit various psychological characteristics. Using a repository of Facebook profiles and matching personality data, they identified which Facebook 'likes' most frequently corresponded to users who exhibit highly introverted and highly extroverted personality types. Then, using Facebook's advertiser interface, Matz et al. (2017) targeted individual Facebook users with adverts corresponding to their expected personality profile given their Facebook likes.[49] Across several advertisements, Matz et al. (2017) find significantly higher click-through-rates (CTR) and purchase rates from those targeted with an advertisement matched to their personality profile.[50]

[48] These categories are somewhat arbitrary and have been designed to reflect the key ideas which reoccur within the literature. As such, one should not presume no cross-over between these categories, nor should one take these categories as definitive.
[49] As Matz et al. (2017) note: "As of now, Facebook advertising does not allow marketers to directly target users based on their psychological traits. However, it does so indirectly by offering the possibility to target users based on their Facebook likes" (Matz et al., 2017: 12715)
[50] CTR is the ratio of how many users saw the advertisement compared to how many clicked on the advertisement, while purchase rate is the ratio of how many users ultimately bought/downloaded the product compared to how many clicked on the advertisement. In both instances, a high ratio is desirable.

Dubois, Rucker and Galinsky (2016) investigate how matching perceptions of power between a communicator and an audience affects the persuasiveness of the communicator. They identify several scenarios in which an imbalance of power can easily be seen.[51] For instance, in a charitable advertisement, Dubois, Rucker and Galinsky (2016) argue the persuasiveness of the message will vary depending on whether a senior executive at the charity is delivering it, or an individual who benefits from the charity's work.[52]

Like others (Matz et al., 2017; Cesario, Grant and Higgins, 2004; Moon, 2002), they hypothesise that audiences are likely to be more persuaded by messages when the power-level of the communicator matches that of the audience. To test this hypothesis, Dubois, Rucker and Galinsky (2016) assigned participants roles as either communicators trying to persuade audience members to join a new gym-facility, or audience members listening to the communicators. To imbue a sense of power (or lack thereof), all participants were asked to complete a sentence scrambling task, with participants trying to form sentences from a scrambled set of words. Half of the participants were given a high-power set of words, while the other half were given a low-power set of words.[53] Following this task, communicators were asked to persuade audience members, before the persuasiveness of the communicator was measured via audience feedback.

Dubois, Rucker and Galinsky (2016) find that audience members were significantly more persuaded when the communicator's power-level matched that of the audience, compared to when power levels did not match. To check the robustness of the result, they considered whether the assignment of roles (communicator and audience member) itself imbued a sense of power imbalance but found no significant effect from this consideration. Furthermore, after

[51] Dubois, Rucker and Galinsky (2016) define power as, "asymmetric control over valued resources in social relationships" (Dubois, Rucker and Galinsky, 2016: 69).
[52] According to Dubois, Rucker and Galinsky (2016), this dynamic can also mean the message itself could be varied between actors with different levels of power, with a high-power actor unlikely to appeal to emotional 'warm' sentiments, but very likely to appeal with calculated, factual sentiments.
[53] I.e. sets of words which when combined would make statements imbuing a sense of high or low power.

measuring the feelings of power using a manipulation check,[54] Dubois, Rucker and Galinsky (2016) find that the gap in feeling of power scores between communicators and audience members was lower when the power-level matched than when the power-level did not match. This would be expected only if the power-level framing task successfully imbued a sense of power (either high or low) in participants.[55] The implication, from a nudging perspective, may be that matching participants with nudges which frame decisions in a manner congruent with how the participant makes decisions (i.e. personalisation) may make the nudge more effective.

Cesario, Grant and Higgins (2004) investigate a specific model of personality in their study of regulatory fit. Cesario, Higgins and Scholer (2008) define regulatory fit as, "a goal-pursuit theory that places special emphasis on the relation between the *motivational orientation* of the actor and the *manner in which that actor pursues the goal* (Cesario, Higgins and Scholer, 2007: 444-445, original emphasis). Cesario, Grant and Higgins (2004) relate the concept of regulatory fit to several hypothetical campaigns designed to persuade individuals to change their behaviour.[56] They suggest that if the messaging contained within the campaign[57] matches with the way people are motivated to act, then the messaging would be more effective at encouraging action than messaging that doesn't match because it would have a better regulatory fit.[58]

Cesario, Grant and Higgins (2004) frame the language of the campaigns around what they dub eager and vigilant language, with the former promoting the potential benefits/gains from accepting the subject (changing behaviour/adopting the policy), and the latter promoting the potential harms/losses of not accepting the subject (not changing behaviour/adopting the

[54] A typical device in surveys used to test whether the participant has become aware of the phenomena being examined; do they know they're being manipulated? As well as Dubois, Rucker and Galinsky (2016), Moon (2002) also employs a manipulation check.
[55] Dubois, Rucker and Galinsky (2016) also considered whether the medium of communication (in this case oral) had an effect by repeating the study using written persuasion messages. They find similar results in this this written study.
[56] These were 1) eating more fruit and vegetables; 2) accepting a new regulatory policy; and 3) acceptance of a new after-school program.
[57] Following Cesario, Higgins and Scholer (2008), the manner in which the actor pursues the goal.
[58] While not referred by Cesario, Grant and Higgins (2004), this effect might be tied to confirmation bias – the tendency to view favourably information which reinforces a person's pre-held beliefs (Nickerson, 1998).

policy).[59] After assessing the regulatory fit of participants using a variety of questionnaires,[60] Cesario, Grant and Higgins (2004) showed participants campaign material written to emphasise either eager or vigilant language, before finally measuring how persuaded participants felt by the material, such as how likely they were to subsequently eat more fruit.

They anticipated that those with an eager regulatory fit would be more persuaded by appeals to benefits/gains, and those with a vigilant regulatory fit would be more persuaded by appeals to harms/losses. As expected, where the framing of the material matched the regulatory fit of participants in each study, participants reported being more persuaded by the material. Again, the implication, from a nudging perspective, may be that matching the framing (if not the mechanism) of a nudge to that of a decision-maker (i.e. personalisation) may render the nudge more effective.

In a similar study to Cesario, Grant and Higgins (2004), Moon (2002) investigates message persuasion style and individual personality. While Cesario, Grant and Higgins (2004) investigate eager and vigilant personalities, Moon (2002) investigates one of the personality types found in the Big Five personality scale,[61] specifically extraversion. This personality type is selected over the four alternatives because, “not all of the dimensions [of the Big Five] are equally salient… The most ‘psychologically prominent’ factor is the dominance and submissiveness (“extraversion”) dimension” (Moon, 2002: 314). As this statement suggests, Moon (2002) splits the extraversion personality types into the two manifestations of extraversion – dominance (high extraversion) and submissiveness (low extraversion) – and

[59] It is interesting to note that the language used by Cesario, Grant and Higgins (2004) follows closely with the language used by Kahneman and Tversky (1979) in their study of risk-taking in the domain of gains and losses. Cesario, Grant and Higgins (2004) seem aware of the work of Kahneman and Tversky, citing their 1973 paper on the availability of information, but do not establish a connection between the use of gain/loss framing in regulatory fit and the use of the same framing in what would later develop to be the loss aversion nudge. Thus, it may be unreasonable to conclude that Cesario, Grant and Higgins (2004) adopt the gain/loss framing because of the behavioural phenomena previously identified regarding this framing. However, the closeness of these techniques may be a worthwhile observation.

[60] Cesario, Grant and Higgins (2004) measured regulatory fit differently depending on the context of the study. This is because a person’s motivations about, say, diet, may be very different to their motivations about, say, regulation.

[61] A commonly used personality scale that splits human personality into five personality types.

investigates the persuasiveness of messages which match these personality styles. Moon (2002) conducted two studies, each of which involved an automated computer program randomly showing participants messages framed either with dominant language, or submissive language.[62] Prior to conducting either study, Moon (2002) collected personality data from participants to measure their individual extraversion.

In the first study, participants were asked to rank a set of cars from best to worst using whatever criteria the participant saw fit to use. The ranking was then inputted into a computer program which was pre-programmed to always offer a slightly different ranking of the cars.[63] Therefore, regardless of how participants ranked the cars, the computer would always offer an alternative ranking that was identically transformed across participants. Finally, the computer would then provide the participant with messages explaining why it had ranked the cars differently, with either dominant or submissive language used throughout the messages.[64] After being given some time to process the messages, participants were then invited to re-rank the cars any way they wanted.

In the second study, participants were shown various entertainment content, such as news headlines, music, cartoon strips and health tips. While all participants were shown the same content, prior to being shown, some participants received a framing message using dominant language, and others received a message using submissive language. After processing the content, participants were then asked to evaluate each piece of content.[65]

[62] Moon (2002) writes, "In the dominant message condition, all of the messages contained strong language consisting of assertions and commands. This manipulation was consistent with the theoretical definition of dominance as being the tendency to command and direct others to take certain actions. Conversely, in the submissive message condition, all of the messages contained weaker language consisting of questions and suggestions" (Moon, 2002: 316).

[63] For instance, whatever car the participant ranked in the 4th position, the computer would always rank as the 1st position.

[64] These messages would contain information about the cars which the participant had been made aware of. However, the messages would use this information in a conversational style, rather than intermating to the participant that they had used the wrong criteria to rank the cars. For instance, a typical dominant message was, "The Dodge Neon is definitely ranked too low," but this would then be followed with the information, "The Neon is one of the most affordable cars on the road" (Moon, 2002: 316).

[65] For instance, participants were asked to state how funny the cartoon strip was, or how enjoyable the music was.

Consistent with the findings of other authors reviewed here, Moon (2002) finds significant evidence of more effective outcomes when a participant's personality type matches the framing of the message. In the first study, those whose personality type matched the framing of the message were significantly more likely to be persuaded by the messages and subsequently change their ranking to match that of the computer, compared to those whose personality type didn't match the messaging. Similarly, in the second study, those whose personality type matched the messaging were significantly more approving of the various entertainment content shown to them, compared to those whose personality type did not match the messaging. Moon (2002), therefore, concludes, "the matching of message style to the personality style of the recipient increases the effectiveness of messages" (Moon, 2002: 322).

So far, all the reviewed work has analysed individual personality using a dichotomous variable. For instance, dominance and submissiveness (Moon, 2002) or eagerness and vigilance (Cesario, Grant and Higgins, 2004). Hirsh, Kang and Bodenhausen (2012) argue this approach means much of the personalisation literature has failed to "systematically [relate psychological characteristics] to a comprehensive model of personality traits" (Hirsh, Kang and Bodenhausen, 2012: 578).[66]

Hirsh, Kang and Bodenhausen (2012) attempt to "systematically [relate psychological characteristics] to a comprehensive model of personality traits" in their study of personalised advertisements and the so-called 'Big Five' personality aspect scale, expanding on the work of Moon (2002) who only considers one aspect of the 'Big Five.' They showed participants one of five fictional advertisements for a mobile phone. The advertisement slogan was varied in each advert so that one slogan corresponded to each of the five personality types.[67] After

[66] Hirsh, Kang and Bodenhausen (2012) were writing prior to the work of Matz et al. (2017), and so are not talking directly to the latter's work, but to the general tendency of the personalisation literature which Matz et al. (2017) later continue. As such, the criticism remains valid.

[67] For the reader's immediate benefit, these are: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. An example slogan included – for extraversion – "with XPhone, you'll always be where the excitement is" (Hirsh, Kang and Bodenhausen, 2012).

randomly showing participants these five advertisements, participants were then asked to complete the Big Five personality aspect scale.

Across all five advertisements, Hirsh, Kang and Bodenhausen (2012) find statistically significantly higher product approval ratings for those whose personality trait matched the personality type embedded within the advertisement slogan compared to those who did not. Thus, the conclusion of Hirsh, Kang and Bodenhausen (2012) is similar to that of Moon (2002), namely that "tailoring messages to match recipients' personality characteristics appears to be a promising technique" (Hirsh, Kang and Bodenhausen, 2012: 581).

Following Hirsh, Kang and Bodenhausen (2012), Egelman and Peer (2015) investigate the 'Big Five' personality scale in their study of privacy and security systems. Egelman and Peer (2015) approach this subject from a slightly different perspective to previous literature. Instead of matching messages such as advertisements to personality types, they investigate the predictive power of previously used personality scales (notably the 'Big Five' personality scale) to critique whether such scales are best placed to capture heterogeneity in target populations.

Egelman and Peer (2015) use the same 'Big Five' scale used by Hirsh, Kang and Bodenhausen (2012)[68] and administer a privacy attitude scale[69] to participants to measure their desire for privacy.[70] Egelman and Peer (2015) then test how successfully the 'Big Five' personality types predict the desire for privacy of participants. Across all five personality types,

[68] Several scales which proport to capture the Big Five personality types exist, with Egelman and Peer (2015) and Hirsh, Kang and Bodenhausen (2012) using the ten-item personality index (TIPI), a scale that measures each personality type using two questions.

[69] Specifically, Egelman and Peer (2015) use four privacy scales, namely the Strahan-Gerbasi version of the Marlowe-Crowne Social Desirability Scale (SDS), the Affirmative Admissions Rate (AAR) scale, the privacy concerns scale (PCS) and the Internet Users Information Privacy Concerns (IUIPC) scale.

[70] Desire for privacy is inferred from two of the privacy scales examined (SDS and AAR). The SDS measures the propensity for a person to reveal information about themselves, and Egelman and Peer (2015) argue a person who demonstrates a low propensity to reveal their information greatly desires privacy (and vice versa). The AAR measures peoples' willingness to admit behaving immorally or unethically, and again Egelman and Peer (2015) argue willingness to reveal unsavoury information about oneself demonstrates a reduced desire for privacy.

they conclude the ‘Big Five’ personality scale is only a weak predictor of privacy behaviour, with no individual personality type being a consistent predictor.[71]

Egelman and Peer (2015) therefore argue that the ‘Big Five’ scale may be a useful scale in broad contexts but is relatively weaker when used in very specific contexts, such as privacy attitudes. They suggest it may be more effective to utilise specific personality or psychometric scales which capture personality traits which can be reasonably expected to relate to the specific context.

In a second experiment, Egelman and Peer (2015) evaluate the predictive power of several specific psychometric scales and contrast these results with the results found for the ‘Big Five’ personality types. They examine three scales: the General Decision Making Style (GMDS) scale, which captures ways of thinking *specifically* about decision-making; the Need for Cognition (NFC) scale, which captures propensity for cognitive tasks; and the Domain Specific Risk Attitude (DoSpeRT) scale, with measures risk attitudes. They continue to measure privacy attitudes using the PCS and IUIPC scale, though do not use inferential privacy scales such as the SDS or AAR.[72]

Supporting their hypothesis, Egelman and Peer (2015) find that these more specific psychometric scales are significantly better predictors of attitudes towards privacy than the more general ‘Big Five’ personality scale. They then enter into a discussion about the implications of these findings and offer two broad conclusions. Firstly, that the findings of previous work may be enhanced by embracing a context-specific measure of individual personality.[73] Secondly, building from this, Egelman and Peer (2015) speculate on how the effectiveness of decision-making tools such as nudges and choice architecture could be

[71] The best performing personality type was found to be openness, followed by conscientiousness, while all others failed to demonstrate predictive power.
[72] Egelman and Peer (2015) provide no explanation for this decision.
[73] For instance, while Hirsh, Kang and Bodenhausen (2012) report significant results in their work matching advertisements with personality types in the ‘Big Five,’ they also describe the effect size of these results as, “modest” (Hirsh, Kang and Bodenhausen, 2012: 580). The findings of Egelman and Peer (2015) may be one explanation of this modesty.

similarly enhanced by embracing, “a more ‘targeted’ nudging approach” (Egelman and Peer, 2015).[74] Drawing a parallel which is similar to the heterogeneity discussion found within the nudge literature, Egelman and Peer (2015) argue that – like they have shown in the personalisation literature – more personalised nudge strategies could lead to more effective nudges.

In a very recent study which seeks to tie nudging and behavioural interventions into this literature, Lipman (forthcoming) investigates small financial incentives and employee-health outcomes. Lipman (forthcoming) tasks participants with selecting one of four financial incentive schemes, each of which is expected to deliver the same reward, but in a different way.[75] Participants are told the incentive is be used by their employer to encourage them to improve their health by part-taking in exercise. Participants, however, are free to choose how they would like to be compensated using any criteria they see fit.

This freedom to crucial to Lipman’s (forthcoming) investigation. Following their selections, participants are then asked to complete several behavioural questions designed to measure behavioural characteristics such as their present bias and their risk preferences. Lipman (forthcoming) argues that a significant concentration of behavioural characteristics within a given incentive group[76] would suggest that people who exhibit the concentrated characteristic would be best nudged using that given incentive. However, contrary to expectations, Lipman (forthcoming) finds no significant difference in the behavioural make-up of any of the groups, leading Lipman (forthcoming) to conclude that, while the behavioural theory surrounding personalisation and heterogeneity would suggest individual differences should matter, the empirical evidence suggests this is not the case.

[74] One might speculate that this conclusion relates to the notion of relevancy and heterogeneity. Following Egelman and Peer (2015), less targeted (i.e. more general) measures of personality capture less relevant heterogeneity, owing to their relatively lower predictive power.
[75] For instance, a lump-sum reward versus a staggered pay-out.
[76] Participants were grouped after-the-fact into groups depending on which incentive scheme they had selected.

Lipman's (forthcoming) result appears to be an exception. Overall, there is compelling evidence that matching message frames with various individual criteria (i.e. personalisation) produces more effective messages.[77] Variation in these studies can be found. For instance, some authors investigate power dynamics (Dubois, Rucker and Galinsky, 2016) or dominance/submissiveness (Moon, 2002), while others investigate goal-setting (Cesario, Grant and Higgins, 2004) and introversion/extraversion (Matz et al., 2017). Yet, across these variations – and across the different contexts in which each of these authors conduct their research – compelling evidence of the effects of personalisation persists.[78]

While not contradicting this conclusion, Hirsh, Kang and Bodenhausen (2012) stand out as an example of personalisation which goes beyond a dichotomous measure of individual personality. To an extent, this may not be surprising; previous authors – most notably Moon (2002) – have investigated specific individual personality types captured by the Big Five personality scale examined by Hirsh, Kang and Bodenhausen (2012). Thus, if various studies find these personality types – when examined individually – demonstrate the potential of personalisation, then an examination of all five personality types is likely to produce a similar result.

This is not to downplay the significance of the work of Hirsh, Kang and Bodenhausen (2012); with perhaps the exception of Matz et al. (2017), there is little clear reason beyond particular intellectual curiosity to investigate some personality types and ignore others. While the Big Five personality scale may lack some detail and be inferior – in terms of specificity – to more specific measures of personality (Egelman and Peer, 2015), Hirsh, Kang and Bodenhausen (2012) can still attest to examining a reasonably broad and comprehensive range of personality types. In this sense, the corroborating evidence provided by Hirsh, Kang and

[77] As measured using several criteria.

[78] The term personalisation may imply these authors were intentionally matching participants to messages which appealed to their personality types, which in several instances (Dubois, Rucker and Galinsky, 2016; Hirsh, Kang and Bodenhausen, 2012; Cesario, Grant and Higgins, 2004; Moon, 2002) was not the case. The use of the term personalisation in this instance is simply meant to represent the idea of matching message frames with personality types.

Bodenhausen (2012) when personalisation is examined across a range of personality types strengthens the legitimacy of contemporary results stemming from analyses utilising more specific personality classifications.

In many ways, rather than criticising Hirsh, Kang and Bodenhausen (2012), Egelman and Peer (2015) build from their work to provide additional valuable insights. While acknowledging that Hirsh, Kang and Bodenhausen (2012) are right to investigate a complete range of personality types, Egelman and Peer (2015) demonstrate that the personality types investigated should be reasonably related to the context in which the investigation is being conducted.[79] Further, Egelman and Peer (2015) demonstrate the benefits of utilising more detailed measures of personality. Finally, Egelman and Peer (2015) begin to relate the ideas of personality and message-matching within the personalisation literature to concepts such as heterogeneity and choice architecture within the nudge literature.

2.3.2 Personalised Nudging

While Egelman and Peer (2015) establish a relationship between nudges and personalisation, they are neither the first to make such a proposition (Sunstein, 2012, 2013a; Thaler and Tucker, 2013; Porat and Strahilevitz, 2014) nor do they present an especially detailed analysis of what might be called a theory of personalised nudging. At best, they reformulate the problem of heterogeneity previously identified in the nudge literature as an opportunity, suggesting, “nudges should be tested on various different populations, and once a nudge is revealed to have higher potency among specific populations, a more ‘targeted’ nudging approach could be employed, and [would be] expected to produce better results” (Egelman and Peer, 2015). This section considers additional theories of personalised nudging.

[79] A perspective reminiscent of the relevancy principle.

Perhaps the most prominent exploration of the concept of personalised nudging is found in Sunstein's (2012)[80] discussion of personalised default options.[81] Sunstein (2012) contrasts the advantages and disadvantages of active choices, impersonal default options, and personalised default options. In doing so, he returns to the problem of heterogeneity. On the one hand, an active choice can resolve the heterogeneity problem by requiring each individual to actively choose whatever option they want, removing any need for pre-selection criteria.[82] On the other hand, Sunstein (2012) argues active choices can be burdensome,[83] with some people unwilling or unable to make an active decision in a variety of circumstances.[84] Sunstein's (2012) initial conclusion is that often the costs of not respecting heterogeneity and removing some active choices are less than the costs of respecting heterogeneity and mandating active choices.[85] Following an implicitly utilitarian approach (Sætra, 2019; Itai, Inoue and Kodama, 2016), Sunstein (2012) concludes that the use of impersonal default options is frequently desirable.

However, Sunstein (2012) argues that active choosing need not be the only way to resolve the problem of heterogeneity. He posits that collecting information about individuals would provide choice architects with enough of an understanding of the heterogeneity in the population to create personalised default options. These personalised default options would, "offer most (not all)[86] of the advantages of active choosing without the disadvantages"

[80] Sunstein (2013a) also considers personalisation and nudging, but largely reiterates the arguments made in Sunstein (2012) or focuses on ideas surrounding the problem of heterogeneity rather than personalisation.
[81] The default option nudge is a common behavioural nudge. The default option is said to be the option a person receives if they do nothing. Behavioural economists have shown that changing the option which is set as the default option can have a significant impact of the outcome which is selected (Brown and Krishna, 2004; Johnson and Goldstein, 2003; Madrian and Shea, 2001).
[82] Which is to say, because no nudge is being used, no assumptions which may ignore the heterogeneity in the population are being made.
[83] Sunstein (2012) writes, "if active choices were required in all contexts, people would quickly be overwhelmed" (Sunstein, 2012: 1).
[84] For instance, it may be hard for many people to choose from several dozen options available, each with a dozen specifications that need to be negotiated. This is often the case when selecting from various financial products (Sunstein, 2012).
[85] The term 'cost' is often used by Sunstein (2012) as a catch all term for what alternatively may be dubbed welfare or utility.
[86] Sunstein (2012) argues that personalised default options, "might be burdensome and expensive and might also raise some serious questions about privacy" (Sunstein, 2012: 1).

(Sunstein, 2012: 1) by respecting heterogeneity within the population while retaining the relative ease of a default option. Sunstein (2012) does not go so far as to argue personalised default options are always superior to either active choosing or impersonal defaults, choosing to ground his assessment firmly in terms of costs versus benefits. For instance, Sunstein (2012) invokes his relevancy principle, arguing, "when the relevant group is not diverse, and when an impersonal default rule will satisfy the informed members of that group, it is generally most sensible to select that default rule" (Sunstein, 2012: 41).[87] Yet Sunstein (2012) is also of the belief that personalised default options have significant advantages and potential, writing, "personalized default rules are the wave of the future. We should expect to see a significant increase in personalization as greater information becomes available about the informed choices of diverse people" (Sunstein, 2012: 41).[88]

However, Sunstein (2012) is less forthcoming about the practicalities of personalised nudging. On the question of heterogeneity, he establishes the important idea of relevancy which has already been discussed. He also argues that the form heterogeneity information takes, and the means of personalising defaults, could vary significantly. For instance, Sunstein (2012) argues pension schemes could be personalised using only demographic information,[89] but also recognises that tracking technologies could provide significantly more information to choice architects (i.e. nudgers), who in turn may be able implement more sophisticated personalised defaults.[90] Thus, heterogeneity information could be quite basic (e.g. demographic information) or complex (e.g. individual health data), with personalisation also being quite basic (e.g. setting pension contribution rates based on age) or complex (e.g.

[87] As Sunstein (2013a) quips, "with respect to one-size-fits-all approaches, universal scepticism is itself a one-size-fits-all approach, and a bad one" (Sunstein, 2013a: 1870).
[88] Sunstein (2012) broadly proposes a rule of thumb regarding active choices, impersonal defaults and personalised defaults where cost is not an issue. If heterogeneity is not significant, an impersonal default is best. However, where heterogeneity is an issue and it is possible to personalise the default, personalisation should be adopted. However, where it may not be possible to personalise, or where privacy is of concern, an active choice is the best option.
[89] What Porat and Strahilevitz (2014) call "crude" (Porat and Strahilevitz, 2014: 1465) personalised defaults.
[90] See Weinmann, Schneider and vom Brocke (2016) and Schöning, Matt and Hess (2019) on digital nudges, Thaler and Tucker (2013) on choice engines, Yeung (2017) on hypernudges and Benartzi (2017) on personalisation online.

adjusting personalised defaults throughout the day in accordance with bodily rhythms). Yet, these discussions lack significant detail, and wander rather into the realm of speculation, when compared to the empirical personalisation research examined previously, notably Matz et al. (2017) and Hirsh, Kang and Bodenhausen (2012).[91]

It is also important to note that Sunstein's (2012) analysis does not extend beyond personalising the default option nudge and remains grounded largely in the dynamics of default options versus active choices. Thus, Sunstein's (2012) work contributes to the literature regarding personalised nudging, but does not itself cover a *broad program* of personalised nudging.[92]

A similar criticism could be levied at the work of Porat and Strahilevitz (2014), who build from Sunstein's (2012) work to conceptualise how personalised default rules might be incorporated into contract law.[93] Further, it is less clear if Porat and Strahilevitz (2014) seek to contribute to a personalised nudging programme, or whether their contribution is largely one which adopts a legal perspective on default rules, which happen to share a platform with behavioural science.[94] Regardless, Porat and Strahilevitz (2014) do provide some additional insights worth considering.

Porat and Strahilevitz (2014) consider how individuals could be incentivised to disclose heterogeneity information, potentially resolving the problems of cost and privacy violation associated with heterogeneity outlined by Sunstein (2012). In their discussion of contract law, they consider the idea of a minoritarian or penalty default rule.[95] Porat and Strahilevitz (2014) explain, "the penalty default rule is not aimed at mimicking the contractual term most parties

---

[91] While Sunstein (2012) may be excused as these works are relatively recent, the concept of matching personality traits and psychometrics with messages has been shown to be quite established far before 2012.
[92] Though certain discussions such as privacy concerns and data access do extend beyond personalised default options. See Thaler and Tucker (2013) and Yeung (2017). Also see Sunstein (2013a), who is more willing to discuss personalised nudges as a general concept, albeit rather briefly.
[93] Porat and Strahilevitz (2014) do not disguise the influence Sunstein's work has had on their own thinking, writing, "We agree wholeheartedly, and regard his [Sunstein's] contribution to the literature as significant" (Porat and Strahilevitz, 2014: 1452)
[94] The latter is likely the case.
[95] See Ayres and Gertner (1989)

prefer but instead at penalizing the party who has private information that the other party does not have. Such a penalty is designed to incentivize the party with private information to reveal that information to the party without it" (Porat and Strahilevitz, 2014: 1428). To explain this idea further, Porat and Strahilevitz (2014) consider the legal concept of foreseeable losses, and argue that because an aggrieved party stands to lose out by not making unforeseeable losses foreseeable,[96] they are therefore incentivised to disclose all possible losses to the other party.

Porat and Strahilevitz (2014) relate this idea to personalisation. They argue that individuals may be incentivised to disclose information about themselves, because if they don't, they would be subjected to an impersonal default rule which may produce a less equitable outcome than a personalised default rule.[97] There is a potentially significant implication arising from this proposal. The cost of individual privacy itself is likely to be heterogeneous (Barton and Grüne-Yanoff, 2015).[98] Therefore, even when choice architects believe a personalised nudge will confer significant welfare benefits onto an individual, a given individual may still believe their privacy to be more valuable than these benefits. As Porat and Strahilevitz's (2014) argument emphasises the disincentives of impersonal defaults, the cost of privacy as determined by the choice architect is revealed.[99] Therefore, individuals have more information from which to decide whether or not to reveal information about themselves.[100]

Porat and Strahilevitz (2014) also offer some commentary on heterogeneity information itself. Not unlike others (Sunstein, 2012, 2013a; Thaler and Tucker, 2013; Yeung, 2017), Porat and Strahilevitz (2014) argue that the growth in data and digital technologies such as big data will propel the growth of personalised default options and will empower choice architects (i.e.

[96] Insofar as unforeseeable losses for one party are sometimes foreseeable to the other party.
[97] It is likely this altered framing may induce a different behavioural response to the decision to disclose private information. For some emerging research on how psychology and behavioural economics can be incorporated into models of privacy and disclosure, see Dinev, McConnell and Smith (2015).
[98] I.e. different people may attach different values to their individual privacy.
[99] I.e., the cost of privacy is equal to or less than the cost of impersonal defaults.
[100] This idea may follow the mechanism of information leakage described by McKenzie and Nelson (2003) and Sher and McKenzie (2006).

nudgers) relative to the decision-maker (i.e. the nudged). However, much like Sunstein (2012, 2013a), Porat and Strahilevitz (2014) also argue that personalisation can be achieved *without* the use technologies such as big data.[101] They argue that personalisation requires heterogeneity information, but as with Sunstein (2012), Porat and Strahilevitz (2014) acknowledge that basic heterogeneity information such as age or gender could be used to create personalised default options.[102] They call these basic personalised defaults "crude" personalised defaults (Porat and Strahilevitz, 2014: 1465), and distinguish crude personalisation from what might be called sophisticated personalisation strategies which utilise tracking software (Yeung, 2017), dynamic choice architecture (Weinnman, Schneider and vom Brocke, 2016; Benartzi, 2017; Yeung, 2017; Schöning, Matt and Hess, 2019) and big data technologies (Thaler and Tucker, 2013; Yeung, 2017).[103] This topic is returned to in part 2.3.4.

In a further, recent contribution to the personalised nudging discussion, Ruggeri et al. (forthcoming) argue personalised nudging may be well suited to the world of medicine. Following the arguments advanced by Sunstein (2012), Ruggeri et al. (forthcoming) argue that despite some nudges and other behavioural interventions being effective overall, often those who exhibit heterogeneous preferences are not considered by policymakers. In the case of medical care, they suggest this may be impermissible, and thus the use of personalisation may be justified simply by consideration that everyone ought to receive appropriate medical care.[104]

[101] A distinction Yeung (2017) does not make in her discussion of nudges and big data.
[102] See Butt et al. (2018), for instance.
[103] Porat and Strahilevitz (2014) do not suggest that sophisticated personalised nudges must utilise *all* of these technologies. Instead, crude personalisation is defined as the absence of these technologies. See, for instance, Butt et al. (2018), who find managers often rely on heterogeneity information which is immediately and easily available to them.
[104] This perspective contributes another aspect of the relevancy principle, one that supposes even when the costs of personalisation may outweigh the benefits, the social or contextual background of the personalisation may justify the expense. On the one hand, this may challenge Sunstein's (2012) cost/benefit perspective in regard to the relevancy principle. On the other, it may be argued that the social or contextual background can be *incorporated* into any cost/benefit analysis, and that doing so would reveal personalisation to the worthwhile.

Ruggeri et al. (forthcoming) further argue medicine represents a unique domain for personalised nudging, as a person's medical condition and thus required treatment is almost certainly different to that of other people. The importance of frequent check-ups, regularly taking medication, and diet can all be expected to be highly specific activities which could benefit from the use of personalised behavioural interventions.

2.3.3 Empirical Investigations of Personalised Nudges

It may be helpful to try and reconcile some of the strategies employed in the broad personalisation literature with the strategies outlined by those who have explicitly considered personalised nudges. Two key distinctions emerge between the two. Firstly, while the personalisation literature has focused on personality and psychometrics,[105] conceptual ideas surrounding personalised nudging – perhaps leaving aside Egelman and Peer (2015) – have broadly discussed *crude* criteria such as demographics, and *sophisticated* criteria arising from the spread of information technologies. As such, while both literatures consider how personalisation might manifest, each demonstrates divergent thought concerning how to measure heterogeneity.

Secondly, while authors within the personalisation literature have examined a broad range of contexts, all focus on the question of how the framing of the messages can be altered to match individual characteristics. By contrast, discussions of personalised nudges have thus far remained grounded within the default option nudge. Where there has been divergence from this ground, it has largely benefited a discussion other than personalised nudging,[106] or has mimicked but not necessarily synthesised the ideas raised in the personalisation literature.[107]

---

[105] As Egelman and Peer (2015) attest, personality captures how one thinks broadly, while psychometrics captures more specific ways of thinking.

[106] For instance, Sunstein's (2013a) consideration of discount rates in his discussion of heterogeneity.

[107] For instance, Thaler and Tucker (2013) and Busch (2017) consider personalised information disclosure, which very much borrows from the language of behavioural science and nudging, but insofar as it could be compared to the idea of personalising messaging frames found in the personalisation literature, the latter is far more developed conceptually and practically than the former.

While the work of Peer et al. (2019) is primarily empirical, and so offers important practical insights into personalised nudging, Peer et al. (2019) also offer (quite modestly) an important conceptual contribution to the idea of personalised nudging which may be attributed to a marriage between the idea of personalising framing in the personalisation literature, and the idea of personalising options/choices/outcomes in the personalised nudge literature. Peer et al. (2019) write,

> "it is possible that a stronger outcome could be achieved if existing nudges, which have already been shown to work on average, are deliberately given only to the specific groups of individuals on which they are expected, ex ante, to yield a positive effect, while other groups would receive different nudges or be treated differently.[108] In other words, personalization could be more effective if it is directed at selecting a nudge from a pool of existing nudges. In this we distinguish between personalization of *a certain nudge* vs. personalizing *the selection of the nudge*" (Peer et al., 2019: 4, original emphasis).[109]

Peer et al. (2019) look to test this latter type of personalised nudging using techniques developed in the personalisation literature.[110] They conduct two related studies. In the first

[108] Peer et al. (2019) may be eluding to the concept of shielding, where a person who would be harmed by a nudge is shielded from the nudge. For instance, one response to the heterogeneity identified by Thunström, Gilbert and Jones-Ritten (2018) might be to shield tightwads from the nudge entirely, as they already save enough and do not need *any* further encouragement to save. However, shielding may be a contentious issue. It may be reconciled into personalised nudging as a form of nudge selection (i.e. the range of nudges which could be used in any given context presumably includes the option to not nudge). However, Thaler and Sunstein (2003, 2008) argue some form of choice architecture is inevitable, and so it is potentially spurious to believe a person could be shielded from any influencing choice architecture. Equally, Hansen (2016) wonders, though offers no definitive conclusion, whether phenomena such as unintentional nudging, or intentionally not nudging, can actually be considered as part of nudge theory. Insofar as this remains a (pedantic but) unanswered question, shielding may be problematic. See Chapter 3 for more.

[109] To an extent, Beshears et al. (2015a) have also considered this rationale by arguing that some people may be predisposed to being nudged, while others may be more resistant. Furthermore, Benartzi (2017), Schöning, Matt and Hess (2019) and Ruggeri et al. (forthcoming) all offer commentary which reflects the idea given by Peer et al. (2019).

[110] Peer et al. (2019) claim theirs is the first of its kind, writing, "'nudge personalisation' has been advocated before, but its actual potency and feasibility has never been systematically investigated" (Peer et al., 2019: 1). To this author's knowledge, this statement is not false, though might be adjusted slightly, as nudge innovations which respond to heterogeneity have been tested prior to the work of Peer et al. (2019) – notably Beshears et al. (2016) – but never explicitly branded as personalised nudging.

study, participants are tasked with creating strong yet memorable passwords.[111] Building from the work of Egelman and Peer (2015), Peer et al. (2019) first ask participants to complete several psychometric tests before starting the password setting task.[112] Upon beginning the task, participants were randomly shown one of five nudges designed to improve the strength of passwords. There was also a control group.[113]

Peer et al. (2019) find significant evidence of interaction between several psychometric traits and the nudges, and argue this evidence supports their hypothesis that it may be possible to personalise the selection of the nudge itself. It is somewhat questionable how much of a result this is. For instance, Peer et al. (2019) have no *a priori* hypotheses about which psychometric traits would interact with which nudge strategies, and so do not and cannot test these hypotheses. Furthermore, they do not report comparisons between the nudge groups and the control group, so it remains unclear whether these nudges are effective when administered *impersonally*. Nevertheless, beyond being an overzealous statement, the results found by Peer et al. (2019) in their first study are of significance in their second study.

In their second study, Peer et al. (2019) use the findings from the first study to predict whether a specific nudge strategy will be more or less effective when used in conjunction with a given psychometric trait. By matching nudges with psychometrics with the goal of maximising the strength of participant-created passwords, they argue they are personalising the selection of the nudge. Peer et al. (2019) find that personalising the selection of the nudge leads participants to create significantly stronger passwords than impersonal nudging, or not

[111] If passwords were not memorable, participants would be able to very quickly create a strong password by randomly selecting characters.
[112] These are the General Decision-Making Style, Need for Cognition and Consideration for Future scales, as well as a numeracy scale. See Egelman and Peer (2015).
[113] On this, Peer et al. (2019) are less clear. As a working paper, developing findings may be forgiven, but neither the original draft discussed here (Peer et al., 2019) nor the most recent draft (Peer et al., 2020) considered in this thesis report substantial findings in relation to this control group. The likely purpose of the control group is to check whether nudges used impersonally are still effective. As discussed in Chapter 8, this is likely done but is not reported. One can infer Peer et al. (2019) find these nudges to be effective even when used impersonally as they are able to identify the Crack-Time nudge (a type of password nudge) as the best impersonal nudge.

nudging at all.[114] This result has two immediate consequences. Firstly, Peer et al. (2019) seem to demonstrate the benefits of personalised nudging only previously speculated and do so by introducing a new methodological approach – at least relative to previous work. Secondly, they provide strong evidence that several nudges can be involved in personalisation, and that personalising outcomes or choices may not be the only way of personalising nudges.

Schöning, Matt and Hess (2019) provide a comparable study. While more basic, methodologically speaking, than Peer et al. (2019), Schöning, Matt and Hess (2019) follow a similar rationale of matching "cognitive styles" (Schöning, Matt and Hess, 2019: 4395) to nudge strategies. In this sense, they seem to share the same view as Peer et al. (2019) that the selection of the nudge can be personalised. For instance, Schöning, Matt and Hess (2019) investigate how the layout of websites can be altered to improve the transmission of disclosure information, ultimately to encourage users to remove privacy restrictions. Thus, Schöning, Matt and Hess (2019) seem to be following the strategy used by Peer et al. (2019) of altering the type of nudge embedded within an advice message shown to users. This may, however, only be speculation regarding the intentions of Schöning, Matt and Hess (2019). As they write, "it is important to note that the presentation of choices is personalised, not the choice themselves"[115] (Schöning, Matt and Hess, 2019: 4397), which may be interpreted in support of personalisation via the selection of the nudge,[116] or may be interpreted as supporting an

[114] Peer et al. (2019) only personalise using two of the five nudges they originally examined. This is explained in what might be called the mapping procedure. The rationale for this procedure is as follows: multiple psychometric traits may predict a given nudge will be effective. To maximise effectiveness, it is necessary to determine which psychometric trait should be prioritised. This requires some form of mapping procedure, i.e. a way of determining the relative strength of prediction between psychometric traits. Peer et al. (2019) use a Monte Carlo simulation for their mapping procedure, and as a result, argue only two of the five nudges should be included in the second study. The nature of the simulation, or how this result arises, is not explained. This is discussed in more detail in Chapter 8.

[115] For some further context, Schöning, Matt and Hess (2019) argue that nudging must not reduce freedom of choice (Thaler and Sunstein, 2008), and therefore personalising choices would necessarily reduce freedom of choice. Therefore, personalised nudging must definitionally *not* personalise choices. The syllogism here seems robust, but robust only when taken on a disputable understanding of nudges. For instance, Sunstein (2012) seems to suggest that personalising choices means using heterogeneity information to select a personalised choice *from an existing range of options*, or to *expand the range of options to respect heterogeneity.* Thus, in this instance, no freedom of choice is lost.

[116] Schöning, Matt and Hess (2019) explicitly state that the choices themselves are not personalised, which would seem antagonistic to the idea of personalising outcomes/choices.

idea such as personalised default options.[117] As such, it seems more worthwhile to evaluate the actual procedure of investigation, rather than to speculate at the ideas of Schöning, Matt and Hess (2019).

Schöning, Matt and Hess (2019) investigate personalised nudging regarding privacy disclosure using two nudges and three "cognitive styles" (Schöning, Matt and Hess, 2019: 4395). They argue that the advent of digital technologies and the online space allow nudging to easily be embedded into the user interface (UI) of many websites. This medium also facilitates the use of several different communication styles, including text-based and image-based messaging. Thus, Schöning, Matt and Hess (2019) argue choice architects (i.e. nudgers) can readily nudge users by altering the UI.[118] They, therefore, define two nudges for use in their study: a visual nudge utilising imagery, and a verbal nudge utilising language.

Schöning, Matt and Hess (2019) asked participants to complete a series of questions to determine whether their cognitive style was either verbal or visual, before randomly assigning participants to either a verbal or visual UI nudge. As such, Schöning, Matt and Hess (2019) follow the method of several authors in the personalisation literature[119] who measure personality using a dichotomous variable before comparing groups whose cognitive styles match the message (or nudge) with those who do not match. Unlike the personalisation literature, however – and indeed unlike Peer et al. (2019) – Schöning, Matt and Hess (2019) measure the effectiveness of the nudge in several different ways. First, they give participants the option to disclosure private information about themselves after viewing the verbal/visual nudge explaining how any disclosed material would be used.[120] They hypothesise that those whose cognitive style matched the nudge would be more willing to reveal private information compared to those whose style didn't match. Secondly, Schöning, Matt and Hess (2019) measured how much participants trusted the verbal/visual nudge, hypothesising that trust

[117] For instance, changing which option is *presented* as the default option could very reasonably be described as, "[personalising] the presentation of choices."
[118] Also see Benartzi (2017).
[119] See part 2.3.1.
[120] Rather than simply recording willingness to disclose.

would be higher in the matching group. They also measured perceptions of privacy and risk, and respectively hypothesised that matching individuals would be less concerned about privacy and would see less risk in revealing private information.

Unlike previous studies,[121] Schöning, Matt and Hess (2019) report mixed evidence that personalisation is effective. When evaluating how willing matched participants were to disclose private information compared to unmatched participants, they find no significant difference between the groups.[122] Similarly, they identify no significant difference in trust levels between the groups. However, when measuring concerns regarding privacy and perceptions of risk, Schöning, Matt and Hess (2019) find those whose cognitive style matched the nudge were significantly less likely to express privacy concerns and significantly less likely to perceive giving away their private information as risky. Schöning, Matt and Hess (2019) suggest these mixed findings may be as a result of the way the task was evaluated. They argue, "it is relatively easy to express perceptions, but actually expressing behaviour always comes with a certain risk" (Schöning, Matt and Hess, 2019: 4401).[123] Another possible explanation is that there are outstanding costs to disclosing private information which they do not measure. For instance, participants may be satisfied that the risk is low, but may still want additional compensation for their disclosures.[124] It is also reasonable to suspect that various methodological shortcomings could account for these results.[125]

[121] Notably Peer et al. (2019).
[122] In fact, the matched group disclosed slightly less information.
[123] Schöning, Matt and Hess (2019) may have been wise to write "certain *additional* risk," as they are also measuring risk perceptions, and for this explanation to be correct, participants must be failing to notice an additional risk.
[124] For instance, Schöning, Matt and Hess (2019) asked participants to disclose "personal health information" (Schöning, Matt and Hess, 2019: 4400), but do not provide any additional information about what this constitutes. It is very reasonable to imagine an individual in good health may not be very guarded about their personal health information – because there is little to reveal – but may be of the general belief that health information is important and something that should be kept private, or revealed only when sufficiently incentivised (in a manner of speaking) to do so. For this individual, they may believe there is very little risk in them revealing their private information, but still not be willing to do so because of contextual factors.
[125] For instance, the authors had a sample size of 156 (i.e. N=78 for each group), and do not state how they measured trust, privacy concerns or risk. These criticisms will be discussed further in Chapter 4.

Despite these results, much like Peer et al. (2019), Schöning, Matt and Hess (2019) also contribute to a theoretical discussion of personalised nudging. However, unlike Peer et al. (2019), the assertions of Schöning, Matt and Hess (2019) seem antagonistic to previous ideas. They justify the use of personalisation in the same way Peer et al. (2019), Porat and Strahilevitz (2014) and Sunstein (2012, 2013a) do, namely that populations are heterogeneous and one-size-fits-all nudges may create problems for some individuals. However, Schöning, Matt and Hess (2019) proceed to define personalised nudging wholly around Weinmann, Schneider and vom Brocke's (2016) concept of a digital nudge.[126] Schöning, Matt and Hess (2019) write, "Personalised nudging is a form of digital nudging that takes into account users' individual characteristics and behaviour patterns" (Schöning, Matt and Hess, 2019: 4397). Within the context Schöning, Matt and Hess (2019) examine, digital nudging is used as a super-set into which all personalised nudges fall, as would be the case if the work of Peer et al. (2019) were to be operationalised. However, this definition of personalised nudging seems to ignore the ideas of Sunstein (2012, 2013a) or Porat and Strahilevitz (2014) (notably the latter) surrounding crude personalised nudges and sophisticated[127] personalised nudges. Furthermore, this definition is the exact opposite of Benartzi (2017), who defines digital nudging *around personalised nudging*.[128] This discussion is elaborated on more in part 2.3.4.1.

Page, Castleman and Meyer (2020) investigate personalised nudging in a rather different way. Interested in FAFSA completion rates,[129] they argue that FAFSA usually sees low uptake amongst those who would benefit from doing so,[130] but could be expected to increase if

[126] Weinmann, Schneider and vom Brocke (2016) write, "we define "digital nudging" as the use of user-interface design elements to guide people's behaviour in digital choice environments" (Weinmann, Schneider and vom Brocke, 2016: 1).
[127] E.g. digital nudges.
[128] "In order to take advantage of these digital nudges, I believe we need to tailor them for our new online environment" (Benartzi, 2017: 7).
[129] FAFSA is a state and federal financial assistance program for high school students applying for university in the United States.
[130] Page, Castleman and Meyer (2020) suggest a lack of uptake can ultimately discourage some from going to university, and present evidence from their own study to support this claim.

students were nudged into completing FAFSA. Page, Castleman and Meyer (2020) believe the nudge itself must be customised (i.e. personalised) given the large and potentially very heterogeneous population being nudged.[131] For instance, a reminder nudge which encourages a student to start their FAFSA application may be effective for someone who has not yet started it, but for someone who has started but has not yet completed it yet, the nudge may be ineffective.[132]

Using an automated text-message system linked to students' online FAFSA applications, Page, Castleman and Meyer (2020) personalise reminder nudges to reflect completion rates. For instance, a student who has not yet started their application may receive a text reminding them to start, while a student who is in the course of completing their application may receive a text to finish their application. It is interesting that this approach diverges from the approach taken from Peer et al. (2019) but is rather similar to the concept of personalised nudging developed by Sunstein (2012) of personalising the outcomes a person receives.

Controlling for school- and student-level effects, Page, Castleman and Meyer (2020) find that personalised reminder nudges significantly increased FAFSA application uptake and completion, and also link these personalised nudges to the significantly higher enrolment of students into university several months after the nudge was administered.[133] Once more, while a rather different approach to personalised nudging as seen previously, personalisation of nudges does appear to be an effective strategy.

Finally, a recent study by Guo et al. (2020) returns to the use of personalised nudging and password creation. Guo et al. (2020) argue that there are several common reasons why

---

[131] Heterogeneous in terms of educational outcomes, economic background, university acceptance, and so on.

[132] Page, Castleman and Meyer (2020: 5): "Information that is generic and not tailored to an individual's background and circumstances may seem less salient."

[133] The significance of these results falls as time elapses between the event and the nudge being administered. This may lead one to conclude that the result of higher university enrolment amongst the nudged students is not a *direct* result of the nudge, but an *indirect* result of the nudge. For instance, if a lack of FAFSA would prevent a student *who had not previously* planned on going to university going ultimately going, the nudge to complete the FAFSA application may have helped this student go to university, but only because it – in conjunction with the student ultimately choosing to go to university – enabled this outcome.

passwords may be weak, such as the need for passwords to be memorable, and thus simple. Building from this premise, they hypothesise that these password weaknesses may occur in individuals who exhibit distinct personality types. For instance, Guo et al. (2020) argue that people who are typically more open about themselves will also typically utilise common words in their passwords.[134]

Guo et al. (2020) thus suggest that if a person's personality type can be known, the password tip which is used to nudge the person into creating a stronger password can be personalised. This study, therefore, has a very similar premise to that of Peer et al. (2019). However, this is where the similarities end. Firstly, Guo et al. (2020) 'utilise' the Big Five personality scale, which Peer et al. (2019) – following Egelman and Peer (2015) – do not use. Utilise, only in the sense that the notion of five personality types is used to structure the thinking of Guo et al. (2020). Guo et al. (2020) do not *actually* test participants to determine their personality type. Instead, they assume *a priori* that specific password weaknesses (such as using a common word or repetitive character combinations) correspond to a particular personality type. They thus only ask participants to create a password, analyse the password for specific weaknesses, and nudge participants – given the detected weakness – in accordance with their *a priori* model.[135] Practically, there is clear benefit to doing this – in most password creation environments, one does not have the time or the willingness to complete a personality test. But within an experimental setting, actually administering the personality test would surely have been feasible, and thus the lack of this represents a weakness of Guo et al. (2020).

[134] This is, presumably, because openness is the opposite of secretive, and a secretive person may create passwords which are purposely more obscure. This, however, can only be speculated: Guo et al. (2020) offer little in the way of justification for their links between password weaknesses and personality types, much to the detriment of their work.
[135] To the credit of Guo et al. (2020), they allow for some error in their model by allowing each password weakness to correspond to two personality types, randomly choosing between these two, and then evaluating and updating the model as passwords are created. In this sense, Guo et al. (2020) adopt a trial-and-error approach which Ruggeri et al. (forthcoming) has suggested may be necessary in personalised nudging. However, by *a priori* restricting password weakness to *only* two personality types – rather than all five – they still embed within their model a degree of unsupported assumption about personality and password weakness.

This weakness is likely borne out in their results. Guo et al. (2020) do find that the personalising password tip nudges produced significantly stronger passwords, compared to two other commonly used password composition policies. However, the personalised password tips also required the participants to take significantly more time in creating passwords and was reported by participants to be significantly harder to use than alternative composition policies.[136]

2.3.4 Personalised Nudging

2.3.4.1 …as an Outgrowth of Technologies

The mistake of equating personalisation with the use of big data, so far as it is asserted here, is not an uncommon one. Indeed, several authors (Thaler and Tucker, 2013; Yeung, 2017) focus on personalised nudging not necessarily as a response to heterogeneity, but as an opportunity emerging from information technology (Benartzi, 2017). While the conflation of these concepts may not be quite as explicit as that of Schöning, Matt and Hess (2019),[137] this question of personalisation, technology and data remains a pertinent one worthy of exploration.

One possible place to begin this exploration is with the originators of the wider concept into which Schöning, Matt and Hess (2019) assign personalised nudging: that of digital nudging. Digital nudging is defined by Weinmann, Schneider and vom Brocke (2016) as, "the use of user interface design elements to guide people's choices or influence users' inputs in online decision-environments" (Weinmann, Schneider and vom Brocke, 2016: 4). Insofar as this discussion concerns only *digital* nudging, it is hard to dispute that digital nudging is wholly a

[136] These weaknesses can possibly be explained by the assumptions built into the model. Without grounding their personalisation model within behavioural theory (e.g. presenting evidence to suggest a given password weakness should be associated with a given personality type), Guo et al. (2020) are largely adopting a speculative approach. Furthermore, without a control group from which to evaluate the effectiveness of the password tips when used impersonally, it is difficult to conclude the effectiveness of these nudges was due to personalisation, and not simply due to nudging.

[137] I.e. defining personalised nudging as a subset of digital nudging.

response to the development of information technologies.[138] However, insofar as this discussion concerns *personalised* nudging, Schöning, Matt and Hess (2019) are right to argue that digital nudging is merely a groundwork on which a form (but not all forms) of personalised nudging may be built.[139]

Given the proposition of crude personalised nudges (Porat and Strahilevitz, 2014; Sunstein, 2012, 2013a), there is certainly a clear argument that personalised nudges do not necessarily have to be digital nudges. Equally, the expanded opportunities to nudge in an online, digital space (Weinmann, Schneider and vom Brocke, 2016) does extend the discussion of personalised nudges. Several authors (Thaler and Tucker, 2013; Yeung, 2017) have explored these ideas.

Thaler and Tucker (2013) argue that people quickly become overwhelmed when too much information is provided to them. They further state that easy access to information may be exacerbating these difficulties, while attempts to communicate important information such as contractual obligations often fail to address this problem.[140] Instead, to address the problem of information overload, Thaler and Tucker (2013) propose the idea of a "choice engine" (Thaler and Tucker, 2013: 44). Choice engines would be information technologies such as recommendation algorithms that would interpret much of the disclosure information available

[138] "The increasing adoption of digital technologies in large areas of our private and professional lives leads to a situation in which most decisions are made within – or are influenced by – digital choice environments. Already, when designing the user interfaces, like a Web site or a mobile app, we create a digital choice environment; for example, by the way defaults are set or workflows are organised predefines decisions [*sic*]" "Weinmann, Schneider and vom Brocke, 2016: 2).
[139] Weinmann, Schneider and vom Brocke (2016) make a single reference to personalisation, writing, "we propose five steps of a *digital nudging process* for online decision environments that takes into account specific affordances of information systems (e.g., personalisation, data availability, real-time tracking)" (Weinmann, Schneider and vom Brocke, 2016: 4). Thus, Weinmann, Schneider and vom Brocke (2016) also seem to imply that personalisation is an outgrowth of information technology, but it may also be important to appreciate that they mention personalisation without any evidence to suggest personalisation is significant part of their proposal.
[140] Which is to say, often it is not the language used, but the quantity of information communicated, which overwhelms individuals.

to individuals, and – using personal (heterogeneity) data – personalise the information that is ultimately disclosed to individuals.[141]

Where Thaler and Tucker (2013) develop the idea of choice engines, Yeung (2017) establishes the concept of the "hypernudge" (Yeung, 2017: 118) and in many ways expands the scope of what might be called personalised digital nudges. Yeung (2017) initially argues hypernudges are simply the combination of behavioural nudges and big data,[142] but reveals the assumed, *personalised*, nature of hypernudges when defining hypernudging: "Big Data-driven nudging is… nimble, unobtrusive[143] and highly potent,[144] providing the data subject with a highly personalised choice environment – hence I refer to these techniques as 'hypernudge'" (Yeung, 2017: 122).

As with Thaler and Tucker (2013), Yeung (2017) does not emphasise the personalisation aspect of hypernudges in their discussion; however, again in accordance with Thaler and Tucker (2013) – and Weinmann, Schneider and vom Brocke (2016) – Yeung (2017) recognises that information technology facilitates the personalisation of nudges. Further, the concept of "a highly personalised choice environment" seems very similar to the broad discussions of Weinmann, Schneider and vom Brocke (2016) and Schöning, Matt and Hess (2019) surrounding the integration of nudges and personalised nudges, respectively, with user-interface design.

[141] Thaler and Tucker (2013) also contribute to the conversation surrounding access to heterogeneity information, generally calling for more open data systems and transparency from governments and private firms.

[142] Yeung (2017) writes, "My central claim is that, despite the complexity and sophistication of their [big data systems] underlying algorithmic processes, these applications ultimately rely on a deceptively simple design-based mechanism of influence – 'nudge'… By characterising Big Data analytic techniques as a form of nudge, this provides an analytical lens for evaluating their persuasive, manipulative qualities and their legal and political dimensions" (Yeung, 2017: 119).

[143] This is presumably because hypernudges should also follow the definition of standard (non-hyper) nudges offered by Thaler and Sunstein (2008).

[144] The concept of potency is also mentioned by Peer et al. (2019), and may suggest that using personalisation to increase conformity with the nudge is separate from the idea of people following the nudge because it leads to a better outcome, as considered a normative standard in nudge theory (Oliver, 2019).

Benartzi (2017) offers a slightly different perspective. Writing about influencing behaviour in online spaces,[145] Benartzi (2017) argues that digital nudging is merely an extension of non-digital[146] behavioural science and nudging into a new medium. They further assert, however, that with a new medium comes new possibilities and argues that digital nudges and choice architecture have an opportunity to become extremely effective by embracing personalisation. In this sense, Benartzi (2017) does not argue that all personalised nudges are digital nudges; instead, Benartzi (2017) merely recognises the ease which with digital choice environments can be personalised, and advocates for this to happen. Thus, returning to Weinmann, Schneider and vom Brocke (2016), one may conclude the view of Schöning, Matt and Hess (2019) is a misunderstanding of previous discussions. Rather than personalised nudging *necessarily* being digital nudges, digital nudges may *potentially* be personalised nudges.

2.3.4.2 …In Relation to Previous Ideas

Digital nudging, choice engines and hypernudges exist primarily as outgrowths of information technology, rather than specifically as conceptions of personalised nudges. This is not to say that these 'sophisticated' personalised nudges[147] do not contribute to this discussion, with ideas regarding potency (Peer et al., 2019; Yeung, 2017), and user-interface design (Benartzi and Bhargava, 2020; Schöning, Matt and Hess, 2019; Weinmann, Schneider and vom Brocke, 2016) expanding on both early ideas of personalisation (Moon, 2002; Hirsh, Kang and Bodenhausen, 2012) and theories of personalised nudging (Ruggeri et al., forthcoming; Sunstein, 2012; 2013a). Further, Porat and Strahilevitz (2014) even find it necessary to distinguish between crude personalised defaults and personalised defaults which use big data because they also see the advantages of information technology in achieving ever-more precise personalisation.[148]

[145] Also see Bhargava and Benartzi (2020).
[146] One may be tempted to call it analogue.
[147] So-called here to contrast with Porat and Strahilevitz's (2014) crude personalised nudges.
[148] As does Sunstein (2012, 2013a), though Sunstein emphasises information technologies significantly less.

Yet information technology seems to have such a close relationship with the idea of personalisation not because it explicitly seeks to personalise users experiences,[149] but because the data necessary for the basic function of many online and digital services can often be used as heterogeneity information, and thus incorporated to solve problems arising from heterogeneity.[150] Personalised nudging driven by information technology may resolve the persistent challenge in some of the personalised nudging literature (Sunstein, 2012, 2013a; Porat and Strahilevitz, 2014) of a practical method for collecting heterogeneity information and personalising nudges. But previous research – both in personalised nudging and personalisation more generally – have also demonstrated the possibilities to personalise without using big data.[151] Therefore, while personalised nudging using big information technology is likely one (perhaps the major) future of the discipline (Ruggeri et al., forthcoming; Porat and Strahilevitz, 2014; Thaler and Tucker, 2013), it is important to note that personalised nudging does not emerge from of an outgrowth of information technology, but rather recourse to information technology emerges as one of several means of personalising nudges in response to the problem of heterogeneity (Sunstein, 2012, 2013a).

2.4 – Conclusion

The effectiveness of behavioural nudges can be undermined when individuals within target populations are significantly different. Often, a one-size-fits-all or impersonal nudge will still produce a significant benefit for many people (Sunstein, 2012), and may continue to produce

[149] The contemporary work of Zuboff (2019) on surveillance capitalism may disagree with this assertion, and such a comment would be fair as the business model of many information technology companies has subsequently come to revolve around targeted advertising and behavioural prediction. See Zuboff (2019).
[150] Sunstein (2012, 2013a) argues that some data such as demographic data often serve a purpose beyond capturing heterogeneity. For instance, gender may be known to an employer as part of equal opportunity employment requirements. Butt et al. (2018) also demonstrate this idea, and this idea could also be extended to incorporate information technology firms such as Facebook. While personal data may be used by Facebook to personalise content (Luckerson, 2015), Facebook's basic function of connecting people also requires this information (also see Zuboff (2019) for a discussion of Google's targeted advertising as an outgrowth of their desire to be the best search engine). The use of personal data beyond a primary functionality purpose is what Zuboff (2019) calls, "behavioural surplus" (Zuboff, 2019: 97)
[151] Rather than, as is often the case in the discussions surrounding, say, recommendation algorithms, *retrospectively* interpreting these systems in terms of nudge theory. Again, see Zuboff (2019).

a net benefit across the population even when considering the potential harm or loss of welfare suffered by heterogeneous individuals (Ruggeri et al., forthcoming; Sunstein, 2013a). In some cases, however, heterogeneity within a target population may be so great that aggregate benefit is eliminated. Even when this isn't the case, addressing the problem of heterogeneity still represents an opportunity to improve behavioural interventions (Ruggeri et al., forthcoming; Peer et al., 2019; Sunstein, 2012, 2013a). This review has shown that various unexpected and unintended results, from spending behaviour to policy setting, can be explained in terms of heterogeneity. This review has also examined possible solutions to the problem of heterogeneity, focusing centrally on strategies for personalising nudges.

Personalisation has been an emerging area of study in the fields of marketing and consumer decision-making, and several authors (Matz et al., 2017; Dubois, Rucker and Galinsky, 2016; Hirsh, Kang and Bodenhausen, 2012; Cesario, Grant and Higgins, 2004; Moon, 2002) have found evidence to suggest personalisation produces significantly more effective outcomes, as measured in several ways. With the exception of Egelman and Peer (2015), who briefly ponder about personalisation and nudging, none of these authors have sought to marry the field of message personalisation with behavioural science and nudge theory.

Some (Sunstein, 2012, 2013a; Porat and Strahilevitz, 2014) have approached the topic of personalisation primarily from a behavioural science perspective and have subsequently contributed greatly to the theory of personalised nudges but have also failed to facilitate a satisfying union. However, emerging research is demonstrating how personalisation and nudge theory can practically be combined. Peer et al. (2019) and Schöning, Matt and Hess (2019) show that using personalisation methods developed in the marketing and consumer decision-making literature (Hirsh, Kang and Bodenhausen, 2012) to personalise behavioural nudges produces significantly more effective nudge strategies. Peer et al. (2019) in particular also contribute to the conceptual discussion around personalised nudging, arguing that the *selection* of the nudge could be personalised *in addition to* Sunstein's (2012, 2013a) original conception of personalising choices/outcomes.

Personalisation, however, requires heterogeneity information, and this review has explored the arguments of Sunstein (2012) and the relevancy principle. Many individual differences could be found between any two individuals, and these differences *may* influence the respective preferences of these individuals.[152] However, many of these differences will probably not be relevant to the circumstances in question, say when making a decision. Relevancy, following Sunstein (2012), is a concept that should be applied to heterogeneity when personalising nudges. Several components of relevancy emerge from the literature: heterogeneous information that is relevant in one instance may not be relevant in another; even when heterogeneous information is believed to be relevant, accessing this information may violate a person's privacy. Therefore privacy is a component of relevancy; as is the cost of personalisation, for it is reasonable to believe that the costs of personalisation may sometimes outweigh the benefits of respecting heterogeneity; and personalisation may be more necessary for some nudges than others, meaning the need to respect heterogeneity and thus relevancy may also be dependent on the decision-maker's (i.e. nudgee's) desired outcome and the nudge in question, as well as the circumstance (Ruggeri et al., forthcoming).

The review of the literature gives some indication of what might be considered relevant heterogeneity information. Porat and Strahilevitz (2014) distinguish between crude personalised nudges which utilise heterogeneity information that is easy to access and simple to incorporate, such as age or gender, and more sophisticated personalisation strategies which utilise big data and information technologies. Several authors in the personalisation literature (Dubois, Rucker and Galinsky, 2016; Cesario, Grant and Higgins, 2004; Moon, 2002) examine personality traits using dichotomous variables, while Hirsh, Kang and Bodenhausen (2012), Egelman and Peer (2015) and Peer et al. (2019) use more sophisticated personality and psychometric tests (respectively) to investigate how heterogeneous cognition information could be utilised. Finally, Matz et al. (2017) investigate how social media data and other personal datasets could be used to infer cognitive styles and thus personalise advertising,

[152] Or *seem to influence* preferences.

following from the arguments of several authors (Yeung, 2017; Porat and Strahilevitz, 2014; Thaler and Tucker, 2013).

Finally, this review has considered personalisation and personalised nudging as an outgrowth of information technology. The basic requirement of heterogeneity information in order to personalise nudges has meant that technologies which provide access to and automate the use of large amounts of personal data have become closely associated with the idea of personalisation (Yeung, 2017; Thaler and Tucker, 2013). Furthermore, the relatively fluid canvas of online user-interfaces has led some (Bhargava and Benartzi, 2020; Schöning, Matt and Hess, 2019; Benartzi, 2017; Weinmann, Schneider and vom Brocke, 2016;) to imagine how nudges could be used in significantly more dynamic ways, prompting some to conflate the two into ideas such as personalised choice environments (Yeung, 2017). From this review, it is argued that information technologies do not hinder any programme of personalisation or personalised nudges. However, personalisation should not be thought of as an outgrowth of information technology, but as a response to the problem of heterogeneity found in impersonal nudges (Sunstein, 2012, 2013a).

Chapter 3 – Theory

3.1 – Introduction

Two concepts of personalisation emerge from the literature. Firstly, Sunstein's (2012, 2013a) concept of personalisation, which seeks to address the problem of heterogeneity by personalising the options/outcomes which decision-makers are nudged towards. This concept of personalised nudging, therefore, might be said to be considering *what to nudge.* Secondly, Peer et al. (2019) and Schöning, Matt and Hess (2019) offer a concept of personalisation which seeks to address the problem of heterogeneity by personalising the nudge strategy implemented, while not altering the option which is nudged towards. This concept of personalised nudging, therefore, might be said to be considering *how to nudge*.[153]

This chapter explores these two concepts in greater detail. Henceforth, the cumbersome '*what to nudge*' concept of personalisation is called *choice personalisation*, emphasising that under this procedure, the nudge *strategy* remains impersonal, while the *outcome* supported by the nudge (i.e. *nudged towards*) is personalised. Furthermore, the equally cumbersome '*how to nudge*' concept of personalisation is called *delivery personalisation*, emphasising that under this procedure, the nudge *strategy* (i.e. the type of nudge selected) is personalised, while the *outcome* supported by the nudge remains impersonal.

[153] Ruggeri et al. (forthcoming) and Benartzi (2017) also adopt similar ideas, albeit implicitly. For instance, consider this extended quote from Benartzi (2017: 51):

> "the McDonald's Chinese language site is full of information – reflecting the Chinese preference for higher levels of visual complexity – while the German site is very plain. Such aesthetic adjustments are currently done by hand, but it's easy to imagine a future in which each Internet user has his or her own "aesthetic algorithm," customizing the appearance of every site they see. Just as Pandora recommends music based on what I like, and Netflix sends me suggestions based on my viewing history, so might our browser automatically "format" Web sites in accordance with our visual preferences. Life is too short for ugly screens."

Note that Benartzi (2017) makes an equivalency between two notions of personalisation which are not equivalent. First, there is website design, or how *any information is presented.* Second, there is content recommendation, or *what content is presented regardless of design*.

The structure of this chapter is as follows. Firstly, definitions of *choice* and *delivery* personalisation are presented, with any objections emerging from these definitions unpacked and addressed. These definitions follow from Mills (forthcoming), as does much of the discussion in this chapter. Secondly, by way of situating the choice/delivery framework into the existing literature, a brief re-analysis of some of the literature considered in Chapter 2 is presented with the intention of demonstrating that the choice/delivery framework *compliments* many existing studies examining personalisation. Finally, a brief discussion is offered regarding the use of choice and delivery personalisation as separate personalisation strategies and combined. It is here the two main hypotheses of this thesis are presented.

3.2 – Choice and Delivery Personalisation

Defined in terms of heterogeneity, definitions of choice and delivery personalisation are offered below:

- *Choice personalisation* utilises various heterogeneity data to determine what is the best outcome to nudge a decision-maker towards when the method of nudging *has already been determined.* For instance, if a default nudge is being used to increase pension saving, one individual might have a higher contribution product set as the default because they frequently under save, while another might have a lower contribution product set because they frequently over save (Porat and Strahilevitz, 2014; Sunstein, 2013a). Choice personalisation, therefore, is personalisation *within nudges.*
- *Delivery personalisation* utilises various heterogeneity data to determine what is the most effective method of nudging an individual. For instance, some individuals might be impatient and respond well to default nudges, while others might greatly value the opinions of their peers and respond better to social norm nudges (Peer et al., 2019; Schöning, Matt and Hess, 2019; Beshears et al., 2015a). Delivery personalisation, therefore, is personalisation *across nudges.*

Immediately, there are two items to note from these definitions. Firstly, the definitions are such that *delivery* personalisation is assumed to *precede* choice personalisation. Secondly, various language is used – notably the phrases “best outcome” and “most effective” – which require some consideration.

3.2.1 Delivery Before Choice?

In the above definitions, delivery seems to precede choice. In other words, choice personalisation is defined as being contingent on some decision regarding the delivery of the nudge (the nudge strategy or the method of nudging) having already been determined (either personally or impersonally), while no such contingency is placed on the definition of delivery personalisation.

One potential reason for defining these terms in this way can be found in Schöning, Matt and Hess (2019), who argue choice personalisation itself *cannot* be a type of nudging, because choice personalisation necessarily changes the options available to decision-makers, thus infringing on freedom of choice. As a result, any personalisation of choices – according to Schöning, Matt and Hess (2019) – should not be considered a type of nudging.[154] While it may be initially tempting to embrace this argument to counter the question at hand, not only would such an embrace seem to undermine half of the choice/delivery framework outlined above, but the argument itself contains two noteworthy issues.

Firstly, Thaler and Sunstein (2008) are rather open to the concept of nudging by reducing the number of options available to people (and the salience of those options) insofar as the smaller range of options enables people to better evaluate which they would prefer. This follows from Simon’s (1955) bounded rationality critique of decision-making, whereby people rationally consider only a small amount of information available to them. Hansen (2016) expands on the

[154] Schöning, Matt and Hess (2019): “Personalised nudging is a form of digital nudging that takes into account users’ individual characteristics and behaviour patterns. It is important to note that the presentation of choices is personalised, not the choices themselves, i.e. *freedom of choice is ensured*” (Schöning, Matt and Hess, 2019: 4397, emphasis added).

question of nudging by reducing options, and argues many people – including, at times, Thaler and Sunstein themselves – mis-interpret the idea of preserving freedom of choice as maximal choice (i.e. making *all possible* options available) or equal salience of options (i.e. making *all options equally noticeable*). In this instance, Schöning, Matt and Hess (2019) seem to have fallen into this trap.

Secondly, Schöning, Matt and Hess' (2019) argument leaves a tremendous amount of discussion unaccounted for. Besides disregarding Sunstein's (2012) arguments about personalising choice, Schöning, Matt and Hess (2019) seem willing to allow personalisation to respond to some heterogeneity in the population – namely, differences in decision-making style associated with delivery personalisation – but ignore other heterogeneity in the population – namely, differences in outcomes given individual circumstances. This is problematic in two ways. Firstly, Sunstein (2012) argues the primary criterion for disregarding heterogeneity is relevance, yet Schöning, Matt and Hess (2019) other no compelling reason why heterogeneity of outcomes is irrelevant to personalisation.[155] Secondly, accepting heterogeneity of outcomes is irrelevant creates a myriad of philosophical problems associated with objectively determining outcomes to nudge towards. These objections usually follow that choice architects (i.e. nudgers) who nudged populations (i.e. nudge impersonally) must ignore individual preferences (Rizzo and Whitman, 2009). This is another way of characterising the problem of heterogeneity.[156]

---

[155] One argument may be that, because they do not believe in personalising choices, heterogeneity information which could be used to personalise choices is irrelevant. Yet, as argued above, it seems incorrect to disqualify choice personalisation, and by extension, to disqualify various heterogeneity information. Furthermore, such an argument requires the assumption that choice architects can reasonably know that certain heterogeneity will *not* be relevant to delivery personalisation. Such an assumption may be unjustified. For instance, heterogeneity of outcomes might infer decision-making style, i.e. a person's observable circumstances are a result of their unobservable methods of decision-making.

[156] It is interesting to note the role of paternalism in nudging. One of course can object to the claims of nudging being paternalistic, as some do (Rebonato, 2014; Rizzo and Whitman, 2009; Mitchell, 2005), but the primary reason why nudging claims to be paternalistic is centrally to justify the techniques as part of a program for improving outcomes (Thaler and Sunstein, 2008, 2003). Even if one disagrees with the reality of this claim, it is more difficult to deny the ambition behind this claim. Accepting this, consider the definitions of paternalism given by Thaler and Sunstein (2008, 2003). In their 2003 paper, they write, "In our understanding, a policy counts as "paternalistic" if it is selected with the goal of influencing the choice of affected parties in a way that will make those parties better off" (Thaler and Sunstein, 2003: 175). In their 2008 book, however, this

If not for the argument of Schöning, Matt and Hess (2019), then, why might delivery personalisation be said to precede choice personalisation? The argument put forth here is that delivery precedes choice because the *type of nudge determines the sensible type of options available*.

Consider, for instance, UK workplace pension schemes. The United Kingdom's Pensions Act 2008 introduced legislation that sort to require employees to 'opt-out' of their workplace pension scheme, rather than 'opt-in' – a change known as automatic enrolment, and in the language of nudges, can be described as changing the default option (Service, 2015). This change to UK pension law was to occur gradually, beginning in 2012, and being completed by 2017 (Pensions Act, 2008).

Much discussion of automatic enrolment has focused specifically on the nudge in question, with little analysis of the range of options available *as a consequence of the nudge* (Service, 2015).[157] Having chosen to use a default option nudge, the government implicitly limited the options available to decision-makers (i.e. employees) to either opting out of a workplace pension scheme, or staying in that scheme. Of course, alternative schemes may have existed for employees, as well as various features might have been adjustable by employees once part of a scheme, but these are choices which only occur *following* the initial decision to stay with the scheme or leave the scheme. In short, the type of nudge chosen (i.e. the delivery) impacts the type of options available (i.e. the choice).

By way of further explanation, consider an alternative retirement savings nudge, *Save More Tomorrow* developed by Thaler and Benartzi (2004). Thaler and Benartzi (2004) utilise a present bias nudge to encourage employees to commit to saving more for retirement. The

---

definition has changed: "In our understanding, a policy is 'paternalistic' if it tries to influence choices in a way that will make choosers better off, *as judged by themselves*" (Thaler and Sunstein, 2008: 5, original emphasis). Demanding that better off be judged by the person making a decision implicitly imbues a tendency towards personalisation into nudging, as for a nudge to be paternalistic, it must endeavour to render better off often very different people.

[157] To aid smaller businesses who may not have the sufficient resources to establish and manage a workplace pension scheme, the UK government established NEST, a workplace pensions service which businesses could register with in order to offer workplace pension schemes.

present bias holds that people value the future less than the present, and Thaler and Benartzi (2004) use this to develop a retirement savings plan whereby employees commit to save more of their earnings in the future. As with automatic enrolment, *Save More Tomorrow* had already chosen the method of nudging, namely the present bias nudge, and as a consequence, the type of options available to employees are limited – employees can choose to save more today, or save more sometime in the future. Again, the type of nudge has influenced the choices that can be nudged towards.

Finally, consider Thaler and Sunstein's (2008) famous cafeteria analogy. Thaler and Sunstein (2008) argue that rearranging the layout of foodstuffs within a cafeteria can nudge people towards different choices. For instance, by placing healthy snacks such as fruit at the front of the cafeteria, and unhealthy snacks such as chocolate at the back of the cafeteria, people can be encouraged to choose healthier options. As evidenced with nudges that encourage retirement saving,[158] there are a myriad of ways of nudging healthier lifestyles. Yet, by choosing to use a convenience nudge to make healthier snacks more convenient and unhealthy snacks less convenient, Thaler and Sunstein (2008) are limited in the type of choices that can be nudged towards, namely healthy snacks versus unhealthy snacks. To emphasise this point, there are many ways a person could be nudged to be healthier, from social norm nudges about obesity, to gamification of exercise, to reminder nudges to get up and have a walk. Each different nudge, while trying to achieve the same broad objective – namely, healthier lifestyles – engenders different types of options to be nudged towards (i.e. healthy snack versus unhealthy snack; walk versus sitting down; overweight versus underweight).

There are two further notions to consider. Firstly, the type of nudge chosen impacts the *type* of options available but does not necessarily impact the *range* of options available within a given type. For instance, automatic enrolment reduces choices down to opt-out versus opt-in

[158] In this short discussion, two strategies – default options and present bias nudges – have already been presented.

(the type), but does not restrict the ability of a choice architect to select which option, from a *range of schemes*, to automatically enrol an employee in. Similarly, the *Save More Tomorrow* plan reduces choices down to 'save today' versus 'save in the future' (again, the type), but does not restrict the ability of the choice architect to select which date, from a *range of dates in the future*, to offer as a potential future starting date. Finally, the decision to make healthy snacks more convenient than unhealthy snacks reduces choices down to 'healthy' versus 'unhealthy' (once more, the type), but does not restrict the ability of choice architects to select which snack, *from a range of healthy snacks*, to make convenient for a decision-maker.

As such, by distinguishing between the *type* of choice and the *range* of choice, the idea that the delivery component of nudging precedes the choice component of nudging does not invalidate the choice component. This is vital when considering *personalisation*. Beshears et al. (2016), for instance, personalise the *Save More Tomorrow* programme by personalising the future date which employees begin saving on. As such, even when the delivery of the nudge has already been selected (and selected *impersonally*), thus limiting the *type of choice*, it is still possible to *personalise the choice from a range of options*.

3.2.2 "Best Outcome" and "Most Effective"

In the above definitions, two terms are used which may evoke controversy. These are the terms "best outcome" when discussing choice personalisation, and "most effective" when discussing delivery personalisation. Exploring these terms and the conditions which surround them is not easy, and involves traversing several layers of conflict, some which spawn directly from the complication of *personalisation*, and some which emerge from the basics of nudge theory.

A good place to begin may be by asking a question: why do the objectives of choice and delivery personalisation differ? The answer is because these terms – especially the term "most effective" – endeavour to respect the language and ideas found in the existing literature.

Notably, Yeung (2017) and Peer et al. (2019) introduce the notion of *potency* into this conversation.

The idea of potency is not clearly defined by either author(s), but some understanding of the term can be inferred from the context in which these terms are used. For instance, Peer et al. (2019) contrast the term potency with that of feasibility, distinguishing between the practical ability to personalise nudges (feasibility) and the effectiveness of personalised nudges at nudging people (potency). Similarly, Yeung's (2017) exact wording is, "*highly* potent" (Yeung, 2017: 122, emphasis added), a phrase which implies degrees of potency much like there might be variation in the number of people seemingly being nudged.[159] Potency, as can best be understood from these uses, seems to follow a definition such as:

> *the number of decision-makers who choose an option supported by a nudge (i.e. follow the nudge) compared to the number of decision-makers who choose an alternative option (i.e. do not follow the nudge).*

Thus, the term "most effective" seeks to capture the idea of potency. For instance, an impersonal nudge suffers from the problem of heterogeneity. In some instances, heterogeneity will be so great that the nudge, or the option nudged towards, will be rejected in favour of some other option (Sunstein, 2012). As such, because of heterogeneity, the *relative* potency of an impersonal nudge would be expected to be less than the potency of a personalised nudge, when defined as a ratio of the number of people who follow the nudge versus the number who do not (Peer et al., 2019; Yeung, 2017).[160] By defining delivery personalisation around the term "most effective," therefore, the notion that personalisation should maximise potency is established.

[159] Yeung (2017) also uses this phrase as an extra qualification of their hypernudging definition, which includes practical or feasible (to use the language of Peer et al. (2019)) qualifications. This would suggest potency, in Yeung's (2017) definition, does not describe the *feasibility* of personalised nudging, but some additional characteristic of personalised nudging.

[160] Assuming personalisation resolves the problem of heterogeneity.

Yet this still does not explain why "most effective" is used when considering delivery personalisation, and "best outcome" is used when considering choice personalisation. Indeed, personalising choice so as to promote a good outcome might lead more people to 'following' the nudge, even if the nudge exerted no effect on their decision. This would seem to contribute to potency too.

The difference in terms between choice and delivery personalisation can seem somewhat arbitrary. This is not the case, though one may find themselves treading on treacherous ground if the concept of potency is reserved solely for a discussion of delivery personalisation. The difference, instead, stems from a desire to emphasise the *differences* between choice and delivery personalisation. Delivery personalisation, for instance, says little about the choices which a decision-maker is nudged towards, in much the same way that choice personalisation says little about the method of nudging as determined by delivery personalisation. The use of the term "best outcome" to describe choice personalisation, therefore, purposely emphasises *outcomes* to demonstrate the great importance of outcomes in choice personalisation, and the relatively low importance of outcomes in delivery personalisation. For this same reason, "most effective" emphasises the importance of potency to delivery personalisation, as argued by Yeung (2017) and Peer et al. (2019).[161]

Yet, for such an argument to stand, "most effective" and "best outcome" must be understood to generally be the same. To evidence this, consider the concept of libertarian paternalism developed by Thaler and Sunstein (2003). Broadly, libertarian paternalism argues people should be nudged towards outcomes which are better for them (paternalism) while maintaining freedom of choice (libertarianism). Oliver (2019) posits that nudges could be defined normatively around libertarian paternalism, which is to say, for a behavioural intervention to

[161] One might even argue that effectiveness, as defined here, can be determined relatively objectively. If a great many people are following a nudge, irrespective of the outcome they are receiving, the nudge can be considered effective. However, what the "best outcome" is is significantly harder to determine objectively (hence why Thaler and Sunstein (2003) ultimately change their definition of paternalism). Only in a hypothesised world of perfect personalised nudging, where all decision-makers are nudged towards their *subjectively* determined best outcome, might a degree of consensus around the effectiveness of nudges and outcomes be achieved.

count as a nudge, it should preserve freedom of choice, and promote an outcome which will benefit the decision-maker (Thaler and Sunstein, 2003, 2008).[162]

From this perspective, a *highly potent* nudge is also a nudge which nudges towards an outcome which will improve the welfare of the decision-maker. As such, the notion of “effective” takes on both the qualities of potency (i.e. lots of people following the nudge) and welfare (i.e. people *benefiting* from following the nudge). Thus, the use of the term “most effective,” from a libertarian paternalist perspective, can be understood as the notion of nudging towards improved welfare with an emphasis on *potency*, while the term “best outcome” can be understood as the notion of nudging towards improved welfare with an emphasis on specific *outcomes*. In this sense, the terms may be equivalent.

One might note, however, that Thaler and Sunstein (2003) argue that policies should promote outcomes that will leave decision-makers “better off” (Thaler and Sunstein, 2003: 175), and place no special emphasis on choosing optimal outcomes to nudge towards, or utility maximising outcomes. Yet, the use of the term “best outcome” when defining choice personalisation emerges as a necessary condition: “best outcome,” from a personalisation perspective, means the best outcome for a given individual on the basis of which outcome will minimise disutility arising from heterogeneity (Sunstein, 2012, 2013a). The use of the language “better off” by Thaler and Sunstein (2003) is better understood when considering impersonal nudging through the lens of utilitarianism (Sætra, 2019; Itai, Inoue and Kodama, 2016), where a policy is good if overall welfare or happiness are increased, while personalised nudging is less easily appraised under a utilitarian doctrine.[163]

3.3 – Re-analysis with the Choice/Delivery Framework

[162] Such a normative definition is controversial. For instance, Hansen (2016) argues that libertarian paternalism and the existing definition of a nudge are insufficient to deal with contemporary nudge theory problems, while Schubert (2017, 2015) argues nudges are simply tools, and whether something promotes a ‘good’ or a ‘bad’ outcome matters little when determining if something is or is not a nudge. Even Sunstein (2017a) seems mixed in their assessment of the normative status of nudging.

[163] Again, Thaler and Sunstein’s (2008) adjustment to their definition of paternalism imbues nudging with a tendency towards personalisation, and away from a utilitarian evaluation of welfare.

For all the arguments made thus far, in an attempt to clarify and address issues which may arise from the choice/delivery framework, an outstanding query must surely be whether the choice/delivery framework is consistent with research which has come before. In many ways, it likely is. For instance, in a paragraph highly influential to the ideas developed here, Peer et al. (2019) write, "we distinguish between the personalization of a *certain nudge* (e.g., adding the recipient's first name to the nudge's message) vs. personalizing *the selection of the nudge* (e.g., assigning different kinds or versions of nudges to different individuals)" (Peer et al., 2019: 3). In this statement, whether intentional or not, Peer et al. (2019) seem to acknowledge the dynamics of nudge personalisation conceptualised within the choice/delivery framework.

3.3.1 Peer et al. (2019)

The use of delivery personalisation by Peer et al. (2019) is rather evident from their statement given above. Yet, for prudency, it is worthwhile examining the procedure which they undertake. Immediately, one can determine that Peer et al. (2019) *cannot* have been using choice personalisation, as participants are only nudged towards a single outcome – stronger passwords. It should be evident that personalisation cannot genuinely occur when the range of options to choose from numbers one.[164] Thus, for Peer et al. (2019) to have any claim at all to personalisation, they must undertake delivery personalisation.

Such evidence by elimination is, to an extent, lacking insight. A closer consideration of the procedure undertaken by Peer et al. (2019) reveals that they clearly investigated how a range of different methods of nudging could be used as part of a personalisation strategy to nudge participants into creating stronger passwords. This was done by collecting heterogeneity information about decision-making styles. As such, Peer et al. (2019) clearly use various heterogeneity data to determine the most effective method of nudging, and thus undertake delivery personalisation.

[164] One is reminded of the famous Henry Ford quote: "You can have any color so long as it's black."

3.3.2 Schöning, Matt and Hess (2019)

The case for Schöning, Matt and Hess (2019) is not as immediately obvious as that of Peer et al. (2019). This is not for lack of trying on the part of Schöning, Matt and Hess (2019); as above, they clearly believe that *only* delivery personalisation is possible, and that the prospect of choice personalisation renders the nudge being personalised something else entirely. Despite this assertion being flawed (again, as outlined above), it should be taken as a valuable clue as to the direction they follow. However, as with Peer et al. (2019), it is necessary to go beyond what Schöning, Matt and Hess (2019) state and examine their actual procedure.

It is here that complications arise. As will be discussed in Chapter 4 and Chapter 8, Schöning, Matt and Hess (2019) adopt a *matching* approach; a procedure which has a dubious claim to personalisation. They impersonally show participants verbal and visual nudges, and then assess whether individuals had verbal or visual preferences. By examining the effectiveness of the nudge when the nudge and the preference matched, versus when they did not, Schöning, Matt and Hess (2019) are able to assess the impact *personalisation would have had*. As will be seen, this is not an uncommon strategy within the literature.

For the immediate purpose, however, the work by Schöning, Matt and Hess (2019) seems to use delivery personalisation. Ignoring the problems created by the *matching* approach in claiming to *actually* be personalising nudges, as with Peer et al. (2019), Schöning, Matt and Hess (2019) are nudging people to reveal more private information but doing so using a range of nudges. Therefore, they are not personalising choice, but are personalising delivery.

3.3.3 Page, Castleman and Meyer (2020)

Page Castleman and Meyer (2020) adopt the term “customized nudging” (Page, Castleman and Meyer, 2020: 3), though there is no indication that the term customized marks any difference from that of personalisation. As with Peer et al. (2019) and Schöning, Matt and Hess (2019), Page, Castleman and Meyer (2020) offer something of a theory of personalisation (or customization), writing, “a potentially important distinction [between decisions] is what *kind* of information is likely to be most salient to individuals […] for instance, information about the benefits of pursuing education […] may not resonate with individuals if they already have some basic understanding of the benefits" and continuing, "information that is generic and not tailored to an individual's background and circumstances may seem less salient" (Page, Castleman and Meyer, 2020: 5).

It is immediately obvious that Page, Castleman and Meyer (2020) are grappling with the same notions of heterogeneity which have been tackled by Peer et al. (2019) and Sunstein (2012). Yet, unlike Peer et al. (2019), Page, Castleman and Meyer (2020) focus their attention on what may impact the *salience* of information delivered via a *reminder* nudge, and do not elaborate on alternative nudges. Furthermore, they talk about *tailoring* information, not nudges, which resonates much more with the theories of personalisation considered by Sunstein (2012) and Porat and Strahilevitz (2014) than by Peer et al. (2019) and Schöning, Matt and Hess (2019). Thus, there is an initial suspicion that Page, Castleman and Meyer's (2020) work may be classified as choice personalisation.

Again, this becomes evident in an examination of method. Page, Castleman and Meyer (2020) investigate how customized (personalised) nudging can be used to increase uptake of FAFSA, a federal benefit available to US students applying to university. They utilised a reminder nudge administered via a text-message to students. The contents of the reminder nudge were customized based on the completion status of the student's FAFSA to nudge them towards various behaviours.[165] As such, it seems likely the work by Page, Castleman and Meyer (2020) can be classified as choice personalisation, because the method of nudging was already selected (text-message reminder nudges), while the outcome nudged towards was varied in accordance with various heterogeneity data (namely, application status).

3.3.4 Beshears et al. 2016

[165] For instance, those that hadn't started their application were reminded to do so, those who had only partially completed their application were reminded to complete it, and those who had completed were reminded to ensure various tangential considerations were under control.

Beshears et al. (2016) do not classify their work as personalised nudging, but an analysis using the choice/delivery framework soon creates a compelling case for their work to be classified as such. Yet, because Beshears et al. (2016) do not approach their research from the perspective of personalisation, they offer no explicit theory of personalisation from which to base an initial assessment.

They do, however, identify a clear example of what Sunstein (2012) calls the problem of heterogeneity. By examining Thaler and Benartzi's (2004) *Save More Tomorrow* program, which – as above – encourages employees to save for retirement by committing to saving more in the future, Beshears et al. (2016) find that the implicit message "saving can be deferred into the future" leads some employees to renege on their commitment to save when the pre-selected date arrives. In response, Beshears et al. (2016) formulate a concept called the fresh-start nudge, which sets the future commitment date as a date which is *personally* important to each employee (e.g. a birthday or wedding anniversary).

Here, two key pieces of evidence emerge. Firstly, the fresh-start nudge uses various heterogeneity data to respond to a problem created by heterogeneity, and thus insofar as Sunstein (2012) formulates it, Beshears et al. (2016) do engage in personalised nudging. Secondly, by retaining the present bias nudge used by Thaler and Benartzi (2004), they do not seem to be personalising delivery, and by considering that the fresh-start nudge uses heterogeneity data to select, from any date in the future, the optimal date for an employee to begin saving, it seems likely that Beshears et al. (2016) are engaging in choice personalisation.

3.3.5 Guo et al., 2020

Finally, Guo et al. (2020) investigate how personalised password tip nudges can be used to increase password security, much like Peer et al. (2019). Topic is not the only similarity between these studies. As with Peer et al. (2019), Guo et al. (2020) utilise inferred information about participants personalities to choose, from a range of password tip nudges, which nudge would be most effective at improving password security. Once more, it can be seen that various nudges are being used to nudge towards a single outcome, namely effective passwords. This study, therefore, seems to follow delivery personalisation.

3.3.6 Re-analysis Conclusion

From this brief re-analysis of some of the literature, it can be seen that the choice/delivery framework rather easily categorises existing work on personalised nudging. Furthermore, by re-analysing using the choice/delivery framework, components which exist only in the most abstract of personalisation theory – such as heterogeneity – become more obvious.

3.4 – Choice and Delivery in Isolation and in Tandem

One outstanding question is whether choice and delivery personalisation can be used in conjunction, or whether they represent distinctly different forms of personalisation which must remain separate. The literature on personalised nudging would seem to suggest the latter, with no study seemingly combining both, and no conceptual understanding of personalised nudging – even that put forth by Peer et al. (2019) – bringing both together.

Yet, there is no clear reason why choice and delivery personalisation must operate separately. There may be a myriad of contextual reasons – for instance, a given choice may not lend itself to using both, there may be inhibitive costs associated with one or the other, or personalisation itself may not be a worthwhile endeavour.[166] Yet, from a *theory* perspective, there seems little reason to not conceive of a personalisation process which first uses heterogeneity data to personalise the delivery of the nudge, before using the same or additional heterogeneity data to personalise the choice which a decision-maker is nudged towards.

Furthermore, such an evolution of personalised nudging seems absolutely necessary to conform with future imaginings of nudging as found in the literature. Yeung (2017), for

[166] For instance, Sunstein (2013a) argues nudges which likely involve a lot of heterogeneity such as the present bias nudge may better candidates for personalisation than those with less associated heterogeneity.

instance, writes of hypernudges, "technologies thus operate as self-contained cybernetic systems, with the entire tripartite regulatory cycle continuously implemented via a recursive feedback loop which allows dynamic adjustment of both the standard-setting and behaviour modification phases of the regulatory cycle, enabling an individual's choice architecture to be continuously reconfigured in real time" (Yeung, 2017: 122). Clearly, the ambition for behavioural nudging and interventions within the information technology era vastly exceeds the hypothesis (and in many ways presumes to be correct) that the two base-components of personalisation, choice and delivery, might be used in tandem. Furthermore, contextual evidence, where it exists, supports this hypothesis. Luckerson (2015), for instance, notes that the Facebook algorithm prioritises content which adopts a medium which an individual user is expected to interact with. For instance, a user who regularly engages with photos, but generally scrolls past text, will be shown more photo-based content. Matched with the Facebook advertising algorithm which shows users targeted advertising (Zuboff, 2019), a picture begins to emerge of a "continuously reconfigured", "[individual] choice architecture", to use the language of Yeung (2017), one that combines both personalisation in delivery (i.e. selecting the best medium) and personalisation in choice (i.e. selecting the best product).

This example, of course, is not an example of personalised *nudging*, both from the perspective that Facebook is not explicitly *nudging* users, and from the libertarian paternalist perspective that advertising may not be a welfare-bolstering outcome.[167] Yet, choice and delivery are much more methods of *personalisation*, with personalised nudging representing these methods applied to nudge theory. As such, the question of using choice and delivery personalisation in conjunction is not a question which is the reserve of nudge theory; rather, it is a distinct

[167] Following Sunstein (2017a), what Facebook does with its algorithms may be considered nudging. But this argument is based on a fraught argument used by Sunstein (2017a) that a GPS also nudges users; an argument which quickly reveals itself to suggest that most anything can be considered a nudge. Equally, following Schubert's (2017, 2015) arguments that nudges are merely tools, and Beggs' (2016) and Lavi's (2017) arguments that nudges can be used for non-paternal reasons, once more an argument could be made that Facebook, within this example, is nudging – simply not following libertarian paternalism.

question which has implications for personalised nudging, just as it has implications for personalisation generally.

As such, despite the lack of empirical evidence that choice and delivery personalisation can be used together, the expectation of those with futuristic attitudes, coupled with some commentary on real-world personalisation systems, and *in absence* of any compelling argument to the contrary, it seems more than reasonable to anticipate that choice and delivery personalisation could be used in tandem.

What might be the expected result of such a formulation? Without a direct examination, such a conclusion can only be speculated at. As seen with the empirical literature into personalised nudging, personalised nudges – be them choice personalised or delivery personalised – are often *more effective* than impersonal nudges. Using Sunstein's (2012) notion of heterogeneity, the reason for this increase in effectiveness can be ascribed to a reduction in any issues created by heterogeneity, which is to say, by personalisation *respecting more* heterogeneity. A fair hypothesis, therefore, may be that using both choice and delivery personalisation would respect *even more* heterogeneity than simply using choice or delivery, and thus nudges which are personalised using both methods would be expected to be *even* more effective.

From this discussion, two hypotheses are thus proposed:

> **Hypothesis 1:** *Personalised nudges will be statistically significantly more effective at influencing political decision-making than impersonal nudges, which in turn will be more effective than not nudging.*
>
> **Hypothesis 2:** *Choice and Delivery personalised nudges will be statistically significantly more effective at influencing political decision-making than delivery or choice personalised nudges alone.*

3.4.1 A Note on "Shielding"

It is feasible that, in some circumstances, a person may be found to be better off if they were not nudged at all. For instance, in Thunström, Gilbert and Jones-Ritten (2018), over-savers were found to be negatively impacted by nudges which encouraged saving. One conclusion, using personalisation, may be to nudge over-savers towards lower contribution rate products. But these people *already over-save!* Surely, a more optimal solution would be found by not nudging at all. Of course, further research would be required to examine this hypothesis, but it opens up an intriguing possibility. As does another example, that of reactance. Reactance is a psychological phenomenon described by Brehm (1966) whereby individuals, by virtue of being told to do one thing, engage in the opposite behaviour. Sunstein (2017b) has identified reactance as a potential reason why nudges may fail. For individuals who might be expected to exhibit reactance, the presence of any nudge may prompt deleterious behaviour. As such, when personalising nudges for these individuals, it may be appropriate to consider not nudging at all.

The notion that personalised nudging might result in a person not being nudged at all is a concept hereby dubbed *shielding*. Shielding is not a focus of this research, but it is an idea which emerges out of discussions surrounding personalised nudging. Further, it is an idea which necessarily needs to be embedded within the theory proposed here, less the theory remain incomplete, and requiring some adjustment. It is the opinion of this author that shielding is not inconsistent with the theory offered here.

Firstly, it is prudent to consider were shielding fits within the choice/delivery framework. Shielding almost certainly represents a type of delivery personalisation. When imagining various types of nudges which could be used to nudge an individual, an additional implicit option – *not nudging* – is always available for selection. The apparent omission of shielding from the discussion thus far is the result of the unstated assumption that regardless of whether a nudge is being personalised or used impersonally, and regardless of whether choice or delivery personalisation is being used, a *nudge is always being used*. But dropping this assumption, shielding emerges, and it emerges within the choice/delivery framework as a form of delivery personalisation.

Secondly, shielding is not inconsistent with the definitions of choice and delivery personalisation. If shielding is a form of delivery personalisation, it would be expected that shielding would impact the *types of choices* available, but not the *range of choices* (which are adjusted using choice personalisation). Indeed, shielding does this. For instance, if a default option is being used, the type of choices are opt-in versus opt-out. Shielding removes the nudge, and thus the choice reverts to an *active* choice. This reversion says nothing of the *range of choices* available, which don't change at all. This consistency with theory is just further evidence that shielding is a type of delivery personalisation. Furthermore, when considering the language of "most effective" and "best outcome," it can be seen that neither phrase *demands* the use of a nudge. Indeed, the notion of shielding emerges from accepting that circumstances exist whereby the use of a nudge may produce harmful outcomes, and therefore not be effective. In some ways, it may even be argued that shielding is a necessary consideration to ensure the holistic consistency of the choice and delivery definitions.

Furthermore, shielding is not simply an emergent quirk of personalised nudging. As Sunstein (2012) argues, a perfectly valid alternative solution to the problem of heterogeneity are active choices – in other words, not nudging at all. Thus, the consistency between nudge theory and personalised nudging is maintained.

Section 2:

Methodology

Chapter 4 – Review of Previous Methods

4.1 – Introduction

There is very little prior research on personalised nudging, and while the use of psychographics in political campaigns – notably the Cambridge Analytica scandal which emerged in 2018 (Cadwalladr and Graham-Harrison, 2018) – has been extensively documented in recent years (Chen and Potenza, 2018; Resnick, 2018; Rokka and Airoldi, 2018; Wade, 2018), no research has sort to apply the behavioural theory of personalised nudging to political decision-making.

This is not to say that valuable prior research, applying either wholly or partially ideas pertaining to personalised nudging to alternative domains, does not exist, as evidenced by the review of the literature in Chapter 2. Notably, studies by Peer et al. (2019), Schöning, Matt and Hess (2019) and Page, Castleman and Meyer (2020). Methodologically, some additional, if tangential, studies are also of benefit to this analysis, notably Hirsh, Kang and Bodenhausen (2012) and Moon (2002).

Not unexpectedly, methodological weaknesses exist in all these studies. Some appear as oversights on the part of the respective authors, while others emerge when these studies are recast against the backdrop of the personalised nudging theory developed thus far in this thesis. Insofar as an optimised methodology must be developed – and is desirable – understanding these weaknesses is advantageous. More advantageous, however, are the numerous insights a consultation of previous work can provide. Via a critical, methodological review of the studies outlined in the paragraph above, some direction on an effective methodological approach to investigating personalised nudging can be ascertained.

Each paper discussed in this chapter will be split into five categories, or points of discussion: *Summary*, where the approach of the respective authors is summarised; *Psychometric Selection*, where the type of heterogeneity measurement, and the rationale for that measure's selection, is considered; *Nudge Selection*, where the selection of nudges, and the rationale for their selection, is considered; *Mapping Procedure,* where the process of creating a model from which to personalise interventions is considered; and *Additional Comments,* where miscellaneous commentary on respective papers, if necessary, can be given. This chapter concludes with a summary of these findings.

4.2 – Peer et al. (2019)

4.2.1 Summary

The work conducted by Peer et al. (2019) examining personalised nudging in the domain of cybersecurity is the strongest forerunner to the research undertaken in this thesis. Peer et al. (2019) conduct a two-stage project utilising incentivised survey experiments using Amazon's micro-tasking platform, *Mechanical Turk* (MTurk hereinafter).

In the first stage, they collect various psychometric data about respondents, before giving respondents a password-setting task which utilises impersonal nudges.[168] These data are then used to construct a model for personalising nudges based on psychometrics. This is done via (what will be dubbed here) a mapping procedure, whereby psychometric traits are mapped onto nudges.[169]

In the second stage, Peer et al. (2019) first collect psychometric data from respondents, before utilising these data within their personalisation model (i.e. the mapping procedure) to personalise which nudge is shown to an individual respondent, in order to maximise the strength of passwords which are created (i.e. maximising the effectiveness of the nudges). By comparing the passwords which the first (impersonal) group created with those created in the

[168] Which is to say, randomly assigns a nudge to a participant, and is thus, in effect, *impersonal* nudging.
[169] The mapping criteria could, in theory, vary depending on the purpose of personalisation. Consistent with the definition of delivery personalisation outlined in Chapter 3, Peer et al. (2019) utilise a mapping procedure which maximises the effectiveness (i.e. potency) of the nudge. From a normative nudging perspective (Oliver, 2019), which is to say from a libertarian paternalist perspective (Thaler and Sunstein, 2003), this mapping criterion is the expected mapping criterion.

second (personalised) group, Peer et al. (2019) find the latter group's passwords to be significantly stronger.

4.2.2 Psychometric Selection

The first stage consisted of a sample of 2,074 participants who were asked to complete two psychometric tests, the General Decision-Making Style (GDMS) and the Need for Cognition (NFC) scale. Because their research concerns nudging and decision-making, Peer et al. (2019) argue – following Egelman and Peer (2015) – psychometric tests which are designed to measure decision-making traits are more appropriate than broader psychometric tests such as the so-called 'Big Five' personality scale used by Hirsh, Kang and Bodenhausen (2012) and Moon (2002). Once these psychometric tests had been completed, Peer et al. (2019) asked participants to part-take in a password setting task. It was only after participants completed the password setting task that Peer et al. (2019) asked participants to complete two more psychometric tests, the Consideration of Future Consequences (CFC) scale, and the abbreviated numeracy scale (ANS).[170]

Peer et al. (2019) do not explain why these tests are performed after the password setting task, nor why the GDMS and NFC are performed prior. It may be speculated that they did not want to lead or frame the thinking of participants prior to the password setting task. However, if this is so, explanation as to why the GDMS and NFC tests were performed prior remains lacking. Furthermore, only the CFC would seem to have an obvious framing effect, perhaps prompting some participants to think about the longevity of their password, and thus impacting the results. The delay in administering the ANS, however, remains unexplained, and given any explanation for a delay in the CFC scale is speculative, the decision by Peer et al. (2019) to order their procedure in the way that they have still requires explanation. Indeed, in absence

[170] Peer et al. (2019) do not refer to their numeracy scale as the ANS, but they do cite Peters et al. (2006), who develop the ANS.

of an explanation, any reason for replicating the order of psychometric administration would be speculative at best.

What is more evident, and argued explicitly by Peer et al. (2019), is the rationale for the selection of the various psychometric tests they use. Contrasting specifically with Hirsh, Kang and Bodenhausen (2012; see below), Peer et al. (2019) argue that the general personality types of the Big Five personality scale used by Hirsh, Kang and Bodenhausen (2012) cannot capture the various cognitive traits which would be expected to contribute to decision-making in sufficient detail. However, such an assertion does immediately prompt the question of whether the rationale behind the selection of the GDMS, NFC, CFC and ANS is sensible.[171]

Beyond the name of the GDMS, this scale – developed by Scott and Bruce (1995) – attempts to measure the cognitive patterns through which individuals choose between a set of options. The GDMS is a scale dedicated to measuring typical or "habitual" (Scott and Bruce, 1995: 818) decision-making styles, and so seems suitable for an investigation of decision making.[172]

The NFC scale also seems reasonable. Originally developed by Cohen, Stotland and Wolfe (1955), the scale seeks to measure a person's propensity to perform cognitively taxing tasks. In turning to the nudge literature, the notion of laziness or a path of least resistance frequently emerges (Benartzi, 2017), most notably in discussions surrounding default options and the status quo bias (Madrian and O'Shea, 2001). The idea, then, that propensity to engage in cognitively taxing tasks may moderate the effectiveness of various nudges seems rather reasonable.

In a similar fashion to the NFC scale, the CFC scale can also find justification for its inclusion via a comparison with existing behavioural phenomena and nudges. Developed by Strathman et al. (1994), the CFC scale attempts to measure a person's propensity to behave in such a way as to consider future consequences, versus more immediate consequences. As such, the

[171] For a fuller review of the literature surrounding each of these scales, see Chapter 6.
[172] Scott and Bruce (1995) note that the term 'decision-making style' can be used interchangeably with the term cognitive style, hence the use of both terms above.

CFC may be closely linked to the present bias (O'Donoghue and Rabin, 2015), a behavioural bias that suggests people place greater importance on immediacy (i.e. the *present*) compared to the future.

Finally, the rationale for including the ANS is set out by those who develop the scale, namely Peters et al. (2006), in their introductory line: "many judgements and decisions rely heavily on understanding basic mathematical concepts" (Peters et al., 2006: 407). Peer et al. (2019) acknowledge the relevance of numeracy on decision-making, and further link numeracy to the notion of risk-taking, a pertinent consideration given their research into cyber security.

In summary, while the order with which the various psychometric tests are administered remains unexplained, and subject only to speculation, the selection of the tests themselves seem, at least in this brief review, sensible within the context of decision-making, particularly decisions with a degree of risk, and decisions with a degree temporality.

4.2.3 Nudge Selection

Peer et al. (2019) select strategies which seem to qualify as nudges, following Thaler and Sunstein (2008), though the choices of nudges examined by Peer et al. (2019) remain somewhat particular. This is in no small part because of the domain in which they situate their research, and the experimental conditions/limitations which are imposed.

To begin, it is helpful to explain what nudges Peer et al. (2019) use in their research. In the first round of data collection, they offer five different nudges to participants: an insertion nudge, which provides participants with an example of what makes a good password; a meter nudge, which indicates to participants how strong/weak their password is, in the form of a progress meter; a crack-time nudge, which estimates and informs participants how long it would take to crack their password; a social norm nudge, comparing the strength of the participant's password to the password strength of other users; and a correct horse battery staple (CHBS) nudge, which offers participants a specific tip on how to generate a good password, in this instance, concatenation of words. These nudges are selected based on their prominence within the cyber security literature.

While the validity of these strategies at encouraging good password-setting is not under question, the discussion offered by Peer et al. (2019) remains lacking, as they do not explicitly relate these strategies to any underlying behavioural phenomena, or a more conventional nudge strategy. With the exception of the social norm nudge, Peer et al. (2019) fail to explicitly recognise that, for instance, a meter nudge may invoke loss aversion (Kahneman and Tversky, 1979), or a crack-time nudge the present bias (O'Donoghue and Rabin, 2015; Laibson, 1994).[173] While these interventions still do operate as nudges, the lack of relation back to the literature leaves several areas under-explained.

Centrally, the relationship between psychometrics and nudges becomes muddled, as Peer et al. (2019) relate *their nudges* to psychometrics, and fail to offer broad hypotheses for how general types of nudges (i.e. the present bias, social norms, the default option effect) relate to psychometrics. Insofar as they seek to only ground their research within the domain of cyber security, this oversight may be permissible, but insofar as this work is being considered as a contribution to an analytical method, this absence is unfortunate.

4.2.4 Mapping Procedure

Perhaps the most significant contribution offered by Peer et al. (2019) to the present discussion is their procedure for analysing data collected from their first, impersonal, nudging stage, in order to develop a model of how to personalise the nudges shown to participants

[173] By way of further explanation, the meter nudge may be thought of as something of a probability. While no password is totally secure, it may be reasonable to expect a password which scores, say, 100% on a password meter is more secure than one with a score of, say, 50%. If the 100% score is *evaluated* as a certainty, then the 50% score has a component of loss associated with it. This loss takes the form of a risk of the password being cracked. For the crack-time nudge, while this may also be evaluated as a sort of probability, it may be much harder to calculate the probability of a password being cracked, given an estimate of how long it would take to *be cracked*. As such, the time component would seem more reminiscent of the present bias, where the nudge is moderated by one's propensity to value the short over the long.

given their psychometric profiles.[174] This procedure is dubbed here the *mapping procedure*, and is a term which will be used throughout this and further discussions.

Following the first round of data collection, Peer et al. (2019) report 1,824 participant responses which are used in their analysis. They aggregate the results of their various psychometric scales – which each consist of multiple questions – and test for the internal reliability of this aggregation using Cronbach's alpha. They find no issues with internal reliability. These aggregated figures are then used in a regression analysis procedure.

Specifically, Peer et al. (2019) use *moderation* regression analysis, followed by the Johnson-Neyman technique (JNT hereinafter), which is also known as "floodlight analysis" (Hayes, 2018: 254; Spiller et al., 2013).[175] Referring to Hayes (2018), who offers an authoritative overview of moderated regression and floodlight analysis, Hayes (2018) notes that moderated regression may be a suitable means of analysis when there is reason to believe the effect of an independent variable on a dependent variable is *moderated* by another independent variable, namely a moderator variable.[176]

The research question Peer et al. (2019) investigate would seem to fit these criteria, holding an implicit hypothesis that the effectiveness of a nudge in inducing stronger passwords is moderated by individual differences in decision-making, which are captured by the various psychometric scales. While Peer et al. (2019) neither state this hypothesis outright, nor their use of *moderated regression*, they do write, "This [the JNT] allowed us to examine the moderation effects of each trait on each nudge's effectiveness" (Peer et al., 2019: 9), which is consistent with the use of the JNT following a moderated regression (Hayes, 2018).[177]

---

[174] I.e., delivery personalisation. See Chapter 3.

[175] For further discussion on the JNT and floodlight analysis, see Chapter 8. In brief here, the JNT allows one to evaluate the significance of a moderating (interacting) variable across a continuous range of values.

[176] By 'moderated,' Hayes (2018) is referring to a change in effect size due to the presence of a moderator variable. The validity of moderation is often evaluated quantitatively and qualitatively, with the former relationship tested for statistical significance, and the latter tested based on what might be expected from the literature. See Hayes (2018) for more. Moderated regression is also discussed in more detail in Chapter 8.

[177] A review of the supplementary material provided by Peer et al. (2019) does not offer further clarity here.

Presently, however, several questions remain outstanding. Chiefly: how is the effectiveness of the nudge being measured?; what variable is taking the form of the moderator variable?; what mathematical form does the moderated regression take?; and what purpose does the JNT serve? These questions will be addressed in this order, with some reference to Hayes (2018) where appropriate. Furthermore, for a more complete explanation of moderated regression and floodlight analysis (the JNT), see Chapter 8.

Peer et al. (2019) explain that the effectiveness of the nudge is measured by the strength of passwords being created by participants, which in turn is determined by using a neural network to estimate how many guesses it would take to crack the password. The number of guesses is then log-transformed, and this forms the variable which is said to capture the effectiveness of the nudge interventions.[178]

To answer the second and third questions, Hayes (2018) is a more useful source of information, as Peer et al. (2019) do not explicitly outline their statistical procedure. Hayes (2018) argues that mathematically the moderator variable, and the variable that is moderated, can be interchangeable. This is because – to address the question of the mathematical form of a moderated regression – the moderat*or* variable, and the moderat*ed* variable, are multiplied together as an interaction term within an Ordinary Least Squares (OLS) regression. This interaction term can also be called the *moderation* term (Hayes, 2018), and as the product of moderator and moderated is the same regardless of which variable is which, these labels are mathematically interchangeable. Of course, the actual interchangeability of these variables, as Hayes (2018) explains, is dependent on the subject under research.[179]

[178] This, however, does not necessarily determine that the nudges were effective. Peer et al. (2019) are aware of this, and beyond the five subgroups for each of the five nudges examined, they also collect data from a sixth, control subgroup. This control subgroup is simply asked to create a password and receives no nudge which might influence their judgement. The original draft of this study (Peer et al., 2019), nor the most recent draft which is available (Peer et al., 2020) do not offer comparisons of these nudges *when used impersonally* to the control group. However, Peer et al. (2019) do seem to have performed this analysis, as they are able to conclude the Crack-Time nudge is the best impersonal nudge.

[179] Hence why Hayes (2018) devotes much time to emphasising not only the quantitative significance surrounding a moderation effect, but also the validity of the *qualitative* underpinnings of that effect.

Furthermore, as Johnson and Neyman (1936) argue, the JNT should only be used when the moderator variable is a continuous variable.

Peer et al. (2019) clearly explain that they treat psychometric traits as moderation variables (see above). Qualitatively, with reference to the concept of heterogeneity and nudging developed by Sunstein (2012), it makes sense for the psychometric traits to be the moderator variables. Quantitatively, the various nudge subgroups are likely coded as dummy or dichotomous variables, and so would not meet the criteria for use in the JNT as set out by Johnson and Neyman (1936), while the psychometric trait variables, being aggregates of multiple questions using Likert scales, would qualify as continuous variables.[180]

Finally, Peer et al. (2019) use the JNT as a means of probing "regions of significance," (Peer et al., 2019: 8), a term which has also been used extensively by Hayes (2018) and by Preacher, Curran and Bauer (2006).[181] As Peer et al. (2019) note, "regions of significance [are] ranges of [psychometric] trait values where the effect of each nudge on password strength is statistically significant" (Peer et al., 2019: 8-9). The presence of ranges is a result of the continuous moderator variable, as some values of the moderator may exhibit a significant moderation effect on the dependent variable, while other values may not (Hayes, 2018). This is somewhat intuitive; a person who scores highly on one psychometric is, based on that psychometric, different from someone who scores very low on that same psychometric. It would thus be somewhat deceptive to equate the relationship these disparate individuals' experience with any nudge to be identical simply because the psychometric *overall* has a significant moderating effect.[182]

[180] More so if the aggregation process follows an averaging of responses, rather than a summation. Peer et al. (2019) do not state which method they used, though it is likely they used averaging.

[181] Peer et al. (2019) draw on the use of the JNT by Preacher, Curran and Bauer (2006) rather than Hayes (2018). The term 'regions of significance' can, in fact, be found in Johnson and Neyman's (1936) original paper on the topic.

[182] Peer et al. (2019) perform some initial probing of the effect different scores have in their supplementary material. While these results do not form a significant part of their final contribution, they do find some initial evidence of differences seemingly resulting from arbitrarily 'high' scorers and arbitrarily 'low' scorers, reinforcing this intuitive notion.

Following the moderated regression analysis, Peer et al. (2019) identify significant moderation effects from all psychometric tests for the CHBS and crack-time nudges; significant moderation effects from the CFC and GDMS scales for the meter nudge; some significant moderation effects from the NFC scale for the social norm nudge; and no significant moderation for the insertion nudge. Following the JNT, they then identify the corresponding regions of significance for each nudge.[183]

Having identified these scores, Peer et al. (2019) seem close to having built a mapping model to personalise the nudges shown to participants, which they do in a second stage of data collection. However, an outstanding problem remains, namely, if a person could be nudged effectively with two or more nudges – given their psychometric profile – some procedure must determine which nudge to use. Without resolving this problem, one may be personalising nudges, but not *optimally personalising*. The lack of clarity in this step is the biggest methodological weakness of Peer et al. (2019).

They first establish that they seek to *maximise* the effectiveness of the nudge, noting, "we computed for each participant the nudge that would be expected to produce the highest effect on the password strength" (Peer et al., 2019: 12). The problem facing Peer et al. (2019), however, is not so much understanding the criteria by which a single nudge is selected from a pool of nudges, but rather in explaining what computation they did once establishing these criteria. Peer et al. (2019) give some insights into their procedure, noting that they used a Monte-Carlo simulation to compute which nudge was best in a given situation. However, they do not explain the parameters of their simulation, nor do they explain why they used a simulation, when alternative strategies may be available (see Chapter 8). Furthermore, through using this strategy, Peer et al. (2019) note, "Our simulations estimated that the crack-time nudge would be optimal for 85% of the sample, whereas the meter nudge would be optimal for 15% of the sample" (Peer et al., 2019: 12-13). In other words, this simulation

[183] I.e., the values of a given psychometric variable that would be expected to significantly moderate the effectiveness of the nudge.

method greatly reduced the range of nudges they use in the second stage of their analysis, prompting an observer to question whether such a reduced range of nudges is sufficient to adequately personalise interventions.[184]

Regardless of questions surrounding the use of a Monte-Carlo simulation, Peer et al. (2019) now argue they are able to adequately personalise nudges, given a person's psychometric profile, such that password strength is maximised.

4.2.5 Additional Comments

While the data collection and mapping procedures capture most of the significant aspects of the approach to personalisation adopted by Peer et al. (2019), it is still worthwhile to explore, ultimately, how Peer et al. (2019) assess the effectiveness of their personalisation strategy. This is achieved by collecting data from a second group of participants, again using an incentivised experiment hosted on Amazon's MTurk. Of the 1,146 participants initially recruited, Peer et al. (2019) retain 931 participants following the failure of some 215 to correctly respond to an attention check question.

Unlike in the first group, this second group were asked to complete all four psychometric tests (GDMS, NFC, CFC and ANS) prior to being shown any nudge. Recall in the previous round, participants were asked to complete some psychometrics prior, and some following, the nudge task. The rationale for conducting all the tests prior to nudging participants in the second group is quite sensible – the psychometric data is needed in order to personalise the delivery of the nudge. This does, however, raise two points of consideration. Firstly, the slight difference in procedure may call into question the direct comparability of the first and second groups. Secondly, and additionally, the speculated rationale for having administered tests at different times in the first group was to avoid any priming or framing effects those tests might have

[184] Note, they may be. However, it is difficult to rectify this criticism without a more thorough explanation on behalf of Peer et al. (2019) as to why their simulation approach was more appropriate than alternative approaches, such as ranking coefficients. Without such an explanation, the conclusion that only two nudges would be sufficient is mired in doubts surrounding methodology, and any truth to this claim remains unverified.

contributed; having performed all the psychometric tests before nudging in this second group, it may be speculated that either a) those priming effects are now a potential source of variance in response between the first and second cohorts, or b) the risk of priming effects is insignificant enough so as to be ignored in this second stage, prompting the question of why tests were administered in the order that they were in the first. Peer et al. (2019) are not forthcoming with explanations of these criticisms.

Nevertheless, as above, it seems wholly reasonable to collect all psychometric data prior to nudging, as all psychometric data are needed to personalise the delivery of the nudge. As above, Peer et al. (2019) only select from two nudges in this second stage, the crack-time nudge, and the meter nudge. It may be inferred that they examine the various psychometric results produced by a given participant, refer to the regions of significance information produced by the JNT to identify which nudge/nudges are viable for that given participant, before deferring to the Monte Carlo simulation results to ultimately choose between the two nudges, when either are deemed appropriate.[185]

Once personalised nudges have been selected, Peer et al. (2019) invite participants to part-take in the same password setting task as that undertaken by participants in the first stage. These passwords are then appraised for their strength. Peer et al (2019) test for differences between the personalised group (the second group), the impersonal group (members of the first group who were nudged), and the control group (members of the first group who were not nudged). These comparisons are done using a t-test and a Wilcoxon-Mann-Whitley U-test (WMW-test).

4.3 – Schöning, Matt and Hess (2019)

[185] It is important to note that this is an inference, and Peer et al. (2019) remain rather scant on the exacting details of their process. Issues thus arise; for instance, without them explaining the nature of their simulation model, it cannot be known whether the simulation model alone is relied upon, or whether the output of the JNT is also consulted. Furthermore, it is possible that neither nudge would be deemed suitable for a given participant (and the likelihood of this occurring increases as the range of nudges to choose from decreases – another issue with reducing the range of nudges, as Peer et al. (2019) do).

4.3.1 Summary

For what is, in principle, a similar investigation to that conducted by Peer et al. (2019), Schöning, Matt and Hess (2019) take a very different methodological approach (though, their approach is rather similar to those of Hirsh, Kang and Bodenhausen (2012) and Moon (2002). See below).

Where Peer et al. (2019) situate their investigation of personalised nudging in the domain of cybersecurity, Schöning, Matt and Hess (2019) investigate the use of personalised nudging in the domain of data privacy, specifically, personal health information (PHI hereinafter). Schöning, Matt and Hess (2019), much like Peer et al. (2019), view online, digital infrastructures as an opportunity to embed nudges into a process which they very much view as a decision, complete with risks and rewards. In fact, Schöning, Matt and Hess (2019) go so far as to define personalised nudging as a *subset* of digital nudging.[186]

They conducted a survey experiment which received a total of 156 respondents. Participants were recruited via a university database (though the exact selection criteria are omitted) and via Facebook (though the conditions of this sampling are not elaborated upon). The experiment was not incentivised, and was modelled around, "a health bonus programme [used] by a health insurance company," (Schöning, Matt and Hess, 2019: 4399) which Schöning, Matt and Hess (2019) explain is typically provided via a mobile app.

Measuring two "cognitive styles" (Schöning, Matt and Hess, 2019: 4398), namely verbal and visual, Schöning, Matt and Hess (2019) investigated levels of trust, perceptions of risk, concerns for privacy, and willingness to disclosure PHI between groups who were nudged using a nudge which matched their cognitive style, versus those whose nudge didn't match. Following a t-test, they find mixed results, suggesting personalisation may significantly lower

[186] This is disputable given the choice/delivery framework developed here. For more, see Chapters 2 and 3.

privacy concerns and risk perceptions, but not significantly increase levels of trust or willingness to disclose PHI.

4.3.2 Psychometric Selection

Schöning, Matt and Hess (2019) select two "cognitive styles" for their investigation.[187] They investigate dichotomous verbal/visual styles, with the former suggesting text and language-based information is preferred, and the latter suggesting visual and image-based information is preferred. The justification for this selection is somewhat unclear. Schöning, Matt and Hess (2019) are quite forthcoming with some semblance of a justification, suggesting the verbal/visual styles were selected because they are, "a widely agreed upon distinction in psychology, marketing, and education" (Schöning, Matt and Hess, 2019). A wide agreement regarding the existence of the verbal/visual distinction, however, is not a sufficient justification for the *selection* of the verbal/visual styles in their investigation.

Given this, one is inclined to infer the selection criteria. Such inference is quite forthcoming, given the definition Schöning, Matt and Hess (2019) provide for personalised nudging, and the discursive landscape in which they place their research. As above, Schöning, Matt and Hess (2019) consider personalised nudging a type of digital nudging, and borrowing from Weinmann, Schneider and Brocke (2016), define digital nudging as, "the use of user-interface design elements to guide people's behavior [sic] in digital choice environments" (Schöning, Matt and Hess, 2019: 4396; Weinmann, Schneider and Brocke, 2016: 433). As such, Schöning, Matt and Hess (2019) are *definitionally* limited in their investigation of personalised nudging to an exploration of user-interface (UI) adjustments. Given this limitation, the choice

[187] The use of terminology in the literature (i.e. psychometric vs. cognitive style vs. decision-making style) remains varied, but seemingly inconsequential. Typically, decision-making styles may be described as psychometrics (Peer et al. 2019; Egelman and Peer, 2015) but have also been considered an entity in und themselves (Scott and Bruce, 1995), as well as a sub-component of cognitive styles (Dewberry, Juanchich and Narendran, 2013; Kozhevnikov, 2007). In each instance, the principle of the term (i.e. a numerical measure of some cognitive phenomenon) remains the same, and so the language used by respective authors is adopted when discussing their work, but these terms are considered essentially interchangeable throughout.

of measuring verbal/visual cognitive styles makes sense; UIs consist of various verbal and visual components which UI designers can manipulate.

As such, the selection of the verbal/visual cognitive styles can be justified within the constraints established by Schöning, Matt and Hess (2019). Yet, the rationale for selection still remains rather weak. For instance, following the arguments of Peer et al. (2019) and Egelman and Peer (2015), specificity in psychometric selection may be important to improve predictive power. Insofar as they discuss the verbal/visual cognitive styles, Schöning, Matt and Hess (2019) do not address how these measures may be adapted (or be outright suitable for) the PHI context they are investigating.[188]

4.3.3 Nudge Selection

Much like Peer et al. (2019), Schöning, Matt and Hess (2019) suffer from an ill-defined nudge framework, though differences persist. With Peer et al. (2019), the assertion was that various graphical adjustments to their password-setting task would function as nudges, and as above, the main criticism levered was that they failed to explicitly relate most of these nudge ideas to underlying behavioural phenomena.

With Schöning, Matt and Hess (2019), they define the nudges they investigate around the concept of verbal/visual. The first nudge, the *visual* nudge, "[displayed] the information about data usage with icons" while the second, *verbal*, nudge, "[displayed] the information with bullet points" (Schöning, Matt and Hess, 2019: 4399). Schöning, Matt and Hess (2019) offer little justification to support their nudge design, and – as with Peer et al. (2019) – generally fail to relate their nudges to underlying behavioural phenomena. Though, this may be forgiven, as

[188]Schöning, Matt and Hess (2019) do not even state precisely what questions make up the verbal/visual scale. They may indicate a source of this material, appearing to reference Kirby, Moore and Schofield (1988) and Solomon and Felder (1991). However, this remains unclear, as Schöning, Matt and Hess (2019) write, "for determining respective cognitive styles, i.e. if someone classified [sic] a visual or a verbal type, we employed three semantic differentials with questions drawn from the literature" (Schöning, Matt and Hess, 2019: 4399) before proceeding to cite the above authors. As such, it is unclear whether Schöning, Matt and Hess (2019) first classify participants, and then differentiate further, or differentiate candidates using the "three semantic differentials" found in the literature.

they seem to be following a rather intuitive rationale; a visual style will likely need to involve images or icons, while a verbal style will likely need to involve text.

This is not to say outstanding questions cannot be asked. Namely, given the same information regarding privacy must be conveyed by each nudge,[189] one may question how significantly different swapping bullet points for icons may be? This criticism may not just be speculative; given Schöning, Matt and Hess (2019) report generally mixed results regarding differences in decision-making between verbal and visual participants, there is good cause to question whether these changes were sufficient.

4.3.4 Mapping Procedure

Schöning, Matt and Hess (2019) adopt a mapping procedure that is very different to that of Peer et al. (2019) but is – when contrasted with other literature considered here – not particularly unusual. This is not to say that Schöning, Matt and Hess (2019) share no ground with Peer et al. (2019) in terms of mapping; in fact, with Peer et al. (2019) in mind, it is evident that the method adopted by Schöning, Matt and Hess (2019) could easily be reformulated so as to follow the method of Peer et al. (2019). What's more, the method adopted by Schöning, Matt and Hess (2019) is entirely feasible within the method taken by Peer et al. (2019). These existing methodological links are interesting, and worthy of some discussion. Immediately, however, the specific method adopted by Schöning, Matt and Hess (2019) should be considered.

The name '*mapping procedure*' is something of a misnomer when considering the work of Schöning, Matt and Hess (2019) – it is likely more appropriate to use the name *matching procedure*.[190] This is because they do not adopt a two-stage data collection method as Peer

[189] Schöning, Matt and Hess (2019) report information from GDPR, which mandates various disclosures which must be reported.

[190] Hirsh, Kang and Bodenhausen (2012) have also used the phrase "message-person congruence" (Hirsh, Kang and Bodenhausen, 2012: 578). The use of the term 'matching' as given here, however, rather than this alternative phrase, is that the alternative places unwarranted emphasis on messaging, which will not always be the nudge under examination. Furthermore, Moon (2002), whose process is very similar to Hirsh, Kang and

et al. (2019) do, and instead utilise only one round of data collection.[191] Schöning, Matt and Hess (2019) first showed each of their 156 participants PHI privacy disclosures, with participants randomly shown these disclosures under a visual frame, or under a verbal frame. They then measured a participant's trust, privacy concerns and risk perceptions surrounding privacy disclosures, before testing to see if participants had verbal or visual cognitive styles. Finally, participants were asked to provide PHI.

Several items of note emerge from this process. Most intriguing is the measurement of verbal/visual styles. While Schöning, Matt and Hess (2019) give examples of the questions asked of participants (e.g. "I am rather the verbal/visual type" (Schöning, Matt and Hess, 2019: 4400)), it remains unclear precisely what scale, if any, they are drawing from. Furthermore, each question could be answered on a 6-point Likert scale, and so some degree of aggregation is necessary, as with Peer et al. (2019) and their psychometric scales. Schöning, Matt and Hess (2019) aggregate using a mean average but offer no Cronbach's alpha to indicate whether this process maintained internal validity.[192] Additionally, with these questions now aggregated, they simply assign participants who had an average score less than 3.5 (based on a 6-item, 1 through 7 Likert scale) as visual, while those with a score greater than 3.5 were verbal. While the structure of the questions asked enable this procedure,[193] and while there may be practical reasons for doing so, such a decision may be unnecessarily simplifying. For instance, it seems likely that many people exhibit a range of *both* verbal and visual

Bodenhausen's (2012) and Schöning, Matt and Hess' (2019) uses the terms "matching" (Moon, 2002: 322) and "matches" (Moon, 2002: 313). The use of the term here, therefore, finds precedence in Moon (2002).

[191] The implications of this are to be elaborated on more below. In sum, the second round serves, for Peer et al. (2019), as something of a prediction round, whereby the mapping procedure can be tested by assessing the effectiveness of its predictions. For Schöning, Matt and Hess (2019), no such round exists, which prompts some epistemological concerns.

[192] There are two items to note. Firstly, Schöning, Matt and Hess (2019) do provide *some* Cronbach alpha results later in their paper, but these appear related to the various questions measuring trust, privacy and risk, rather than verbal/visual cognition. Secondly, only three questions are used to measure verbal/visual cognition, so it might be argued that Cronbach's alpha serves little purpose here (as relatively little aggregation is happening). This, however, seems insufficient, as a) Cronbach's alpha could still be tested; b) if so few questions are necessary, could verbal/visual not simply be measured using a single question? and c) if not, are three questions sufficient to measure verbal/visual cognition?

[193] By which, all questions had low scores indicating visual, and all high scores indicating verbal, and as such, an average would still encode this information.

preferences, even if one dominates. By way of their assignment, Schöning, Matt and Hess (2019) allow this detail to be lost. Furthermore, questions must be raised about the validity of this measure, assuming a normal distribution of scores. Given this assumption, many people will have scores around 3.5 for both verbal and visual, and so it must be considered whether it is reliable to simply split the sample based on achieving an above or below average score.

Another item of note is the order in which Schöning, Matt and Hess (2019) collect their data. Assume, for a moment, that data pertaining to trust, privacy concerns and risk propensity is actually akin to the psychometric data collected by Peer et al. (2019).[194] Under this assumption, Schöning, Matt and Hess (2019) collect these psychometric data *after* nudging participants. As will become clear in a moment when discussing the *matching* procedure, the use of a nudge before collecting additional data greatly diminishes the claims of Schöning, Matt and Hess (2019) that they were *personalising* nudges, both from the perspective of *predicting* which nudges would be best, and from the perspective of actually adjusting nudges to match preferences.[195]

However, the assumption that underpins this critique does not hold – in fact, to proceed with such an assumption would reveal a great misunderstanding of their work. Schöning, Matt and Hess (2019) do not treat trust, privacy concerns and risk propensity as psychometrics which moderate the effectiveness of the verbal/visual nudge, but rather regard these measures as *additional* dependent variables – along with willingness to disclosure PHI – with which to scrutinise the effectiveness of personalisation.

[194] They are not, but there is more to be said of this notion. See below.

[195] There is a fair argument that these are the same thing. They may be considered different if one believes that sometimes it *can be known* which nudges correspond to which preferences, and thus no prediction is occurring. This belief may be questionable, and so the argument that these perspectives are the same seems valid. The benefit of framing these perspectives as such largely comes from methodological perspectives. For instance, Peer et al. (2019) use a mapping process from which to *predict* how to personalise nudges; Schöning, Matt and Hess (2019) utilise a matching process whereby they presume to know that, say, a visual person will respond better to a visual nudge. In the language of reasoning, the former, predictive approach is first *inductive*, before becoming *deductive*, while the latter, matching approach is only *deductive*.

In this sense, and as already alluded, Schöning, Matt and Hess (2019) do not use a *mapping* procedure where they search out relationships between nudges, psychometrics, and a measure of effectiveness; instead, they use a *matching* procedure, whereby they hypothesise a relationship between a nudge and a psychometric, and test this hypothesis using a dependent variable. For instance, Schöning, Matt and Hess (2019) hypothesise that visual individuals will respond better to visual nudges, and verbal people will respond better to verbal nudges. They then randomly assign people to either a verbal or visual nudge, and then *after-the-fact* determine if a) a person was a verbal/visual person and b) if that person was matched or not to a corresponding verbal/visual nudge. In a comparison of groups, therefore, Schöning, Matt and Hess (2019) do not contrast verbal participants with visual participants, but instead, what might be called matched participants with unmatched participants (or *personalised participants* with *impersonal participants*).

Using a t-test, they are able test, for each of the four dependent variables, whether these groups were statistically significantly different, and through an examination of the means, determine which group seemed most effected by the nudge. There are, centrally, three core criticisms to be made of this *matching* approach, compared to the previously examined *mapping* approach. Firstly, as discussed above, the act of personalising *after-the-fact* makes any claim to personalising the nudges rather dubious. At best, the matching procedure allows researchers to make statements to the effect of, “if nudges were personalised, they would have X effect.” This, in the opinion of this author, does not constitute *actually personalising nudges.*

Secondly, in this specific instance, Schöning, Matt and Hess (2019) did not contrast their ‘personalised’ and ‘impersonal’ groups against a third, control group. They may implicitly assume that the unmatched group functions as the control group, but given members of this group are by definition being nudged in an impersonal way (which, presumably, means a less effective way), it is unfair to consider this group a representative control group. Furthermore, by consciously assembling groups into matched and unmatched *after-the-fact*, Schöning, Matt and Hess (2019) may be manufacturing significance. For instance, consider Peer et al. (2019). In their first stage of data collection, participants were impersonally nudged. Within this group will be many that could be determined to be, after-the-fact, being nudged sub-optimally, and a *few* being nudged optimally, albeit unintentionally. The unmatched group assembled by Schöning, Matt and Hess (2019), however, definitionally contains no one who was unintentionally nudged optimally. As such, this assembly of the groups imbues a degree of methodological bias which may exaggerate any significance.[196]

Thirdly, the *matching* procedure actually exists as a sub-procedure within the *mapping* procedure. This will be expanded on shortly – the immediate point to be made here is that, methodologically, one should not see *mapping* versus *matching* as a binary *choice*; rather, an advantage that emerges only through an analysis of the *matching* procedure is the clear, holistic approach of the *mapping* procedure.

4.3.5 Additional Comments

In the case of Schöning, Matt and Hess (2019) in particular, this revelation is quite significant. One methodological contribution they offer – which, to their detriment Peer et al. (2019) do not – is a conceptual model of how nudges (in the model, 'stimulus'), psychometrics (in the model 'perception') and effectiveness (in the model 'behaviour') interact. Recall, for instance, an above criticism of Peer et al. (2019) was a lack of a clear statement relating – even as a hypothesis – nudges and psychometrics. Schöning, Matt and Hess (2019) should face no such criticism, as their conceptual model (visually shown as a flow-chart on page 4398) clearly shows that the effectiveness of the nudge (behaviour) is affected by the nudge (stimulus) via psychometrics (perception). The language chosen here is selected so as to best reflect the

[196] There are two arguments one could make when playing Devil's advocate. Firstly, it may not be possible to have a 'true' control group within the domain Schöning, Matt and Hess (2019) examine. This follows from Thaler and Sunstein's (2008) argument that choice architecture and nudges cannot be avoided – as long as a decision must be made, framing surrounding that decision must exist. Secondly, assuming Schöning, Matt and Hess (2019) had limited resources or reason to adopt alternative methods, the matching procedure reveals itself to be a rather sensible and easy method to adopt. Even if one can criticise it, these practical considerations should not be dismissed.

ideas found within their paper, but the conceptual model given by Schöning, Matt and Hess (2019) seems frightfully close in conception to the notion of psychometrics moderating nudges examined by Peer et al. (2019).

It is for this reason that one might be tempted to describe trust, privacy concerns and risk perceptions as *psychometric* measures, rather than additional dependent variables as Schöning, Matt and Hess (2019) do. What's more, given the data available to Schöning, Matt and Hess (2019), it seems entirely possible that a mapping procedure following Peer et al. (2019) could be replicated by Schöning, Matt and Hess (2019) *with no change in how the data are gathered.* In this sense, it first becomes evident that the *matching* procedure is contained within the *mapping* procedure. This conclusion might also be arrived at in reverse; by considering the first stage data collected by Peer et al. (2019), it is feasible that these authors could have hypothesised various relationships between nudges and psychometrics within a conceptual model, and assigned first stage participants to matched/unmatched groups *after-the-fact*, in accordance with their hypotheses.

In conclusion, by contrasting the methods of Peer et al. (2019) and Schöning, Matt and Hess (2019), both points of weakness and points of complement can be identified, and a stronger methodological approach begins to emerge.

4.4 – Page, Castleman and Meyer (2020)

4.4.1 Summary

Page, Castleman and Meyer (2020) consider how personalised (or "customized") reminder nudges can be used in conjunction with FAFSA completions – FAFSA being a US state and federal programme designed to provide financial support to students graduating from high school and entering higher education. They argue that FAFSA is, like other public policies, "often hindered by complicated application processes that make it difficult for people who are eligible for public benefits to access them" (Page, Castleman and Meyer, 2020: 3). To increase FAFSA uptake, therefore, they suggest behavioural nudges could be introduced. However, Page, Castleman and Meyer (2020) argue, "information that is generic and not tailored to an individual's background and circumstances may seem less salient" (Page, Castleman and Meyer, 2020: 8), and thus hypothesise that customizing or personalising any nudges used in conjunction with FAFSA may prove more effective than impersonal nudging.

Page, Castleman and Meyer (2020) take advantage of an automated text message distribution service available to some 66 high schools, with a reach of as many as 17,000 students, to introduce a reminder nudge in the form of a text message to encourage students to engage with and complete their FAFSA application.[197] For half of these schools, however, Page, Castleman and Meyer (2020) also customize (personalise) these text messages by linking the automated service to a database of students' FAFSA applications. This allows them to send application-appropriate messages to students in the personalised treatment group, and generic reminders to students in the impersonal control group.

Generally, Page, Castleman and Meyer (2020) find consistent statistical evidence that personalised nudging produced a significant increase in student engagement with and completion of FAFSA, as well as enrolment in higher education, compared to the control group.

4.4.2 Psychometric Selection

It would be rather disingenuous to talk of psychometric selection in regard to Page, Castleman and Meyer (2020), as they make no endeavour to collect psychometric data (despite acknowledging that people may differ in how salient they find the same piece of information). To be sure, this lack of collection was likely outside of their control, given their research concerns several thousand *children*, for many of whom FAFSA and higher education will be their first significant financial decision.[198] Even if Page, Castleman and Meyer (2020) could

[197] Cadena and Schoar (2011) have demonstrated the effectiveness of reminder nudges at encouraging loan repayments, while Altmann and Traxler (2014) have shown reminder nudges can increase attendance of dental appointments.
[198] This is to say there is an ethical component (working with children) and a methodological component (are participants suitable for psychometric testing?).

and did collect psychometric data, therefore, it may be speculated that the validity of such data would be clouded by a lack of experience from participants; distorted given the time in life the testing would be occurring in (e.g. at a time of applying for university or a job, considering moving away from home, taking on a significant financial commitment; all big decisions which may influence judgment); and compromised by the character of participants (e.g. participants may still be developing physically and cognitively).

It is also not clear that collecting any psychometric data within this study would necessarily improve the quality of the research; indeed, when contrasted with the procedure undertaken by Page, Castleman and Meyer (2020), it seems likely psychometric data would have complicated the procedure. Furthermore, in consultation with the definitions of choice and delivery personalisation seen in Chapter 3 – and accepting the proposition laid there that Page, Castleman and Meyer (2020) use choice personalisation – there is no compulsion to use *psychometric* data, only some measure of heterogeneity. As such, one must consider what measure of heterogeneity Page, Castleman and Meyer (2020) use.

Insofar as Page, Castle and Meyer (2020) can be fit into the general schema constructed in this thesis, they identify heterogeneity broadly within the FAFSA application process. This is to say, at any given moment during the course of their experiment, the automated system used by Page, Castleman and Meyer (2020) could identify the completion status of a given student's application. As can be expected, at any given time, many students were at different points in completing their applications, with some having not begun their FAFSA, and some having completed their FAFSA. Stage-of-completion, therefore, appears as the heterogeneity data Page, Castleman and Meyer (2020) use to personalise the reminder nudge.

This need not be the only data they could use, however. The case for why conventional psychometric testing cannot be used has been made above, but Page, Castleman and Meyer (2020) outline in great detail the various demographic data they are able to acquire about participants, as well as school-level data.[199] One might speculate as to whether these data could have been used as a means of personalising interventions. Certainly, a hypothesis may be constructed around, say, family income and attentiveness to financial matters, which may moderate the effectiveness of any nudge, to use the language of Peer et al (2019).

Page, Castleman and Meyer (2020) do not pursue a hypothesis like this, though their use of stage-of-completion data does not seem inconsistent with their general theory of personalisation. They write, “a potentially important distinction is… what *kind* of information is likely to be most salient to individuals… for instance, information about the benefits of pursuing higher education… may not resonate with individuals if they already have some basic understanding of the benefits” (Page, Castleman and Meyer, 2020: 8, original emphasis).

By this same logic, it is easy to imagine that an impersonal reminder text to two students, one who has nearly completed their FAFSA, and one who has not started, would resonate much more with the latter if it were reminding them *to start* the application, and much more with the *former* if it were reminding them to *finish* their application. The reverse, however, would seem to communicate the wrong kind of information (to use the language of Page, Castleman and Meyer (2020)) and thus, “may seem less salient” (Page, Castleman and Meyer, 2020: 8). The immediate conclusion, therefore, is that while Page, Castleman and Meyer (2020) utilise a rather different approach to heterogeneity as seen in Peer et al. (2019) and Schöning, Matt and Hess (2019) (and as will be seen with Hirsh, Kang and Bodenhausen (2012) and Moon (2002)), Page, Castleman and Meyer (2020) still utilise a concept of heterogeneity that is consistent with *their* theory of personalisation, and consistent with the theory of personalisation developed in Chapter 3.

[199] One might even argue FAFSA information itself, such as academic grades or family income, could have also been incorporated into this analysis. There are three clear objections to this notion, however. Firstly, this supposes these data were available, which would not be the case for someone who never starts their FAFSA or does not complete the necessary sections. Secondly, if starting an FAFSA application is a precursor to accessing this data, it would not be possible for these data to be used to personalise the FAFSA experience (especially the initial nudge *to start an FAFSA application*). Thirdly, even ignoring the first two objections, Page, Castleman and Meyer (2020) may not have had permission (or ethical legitimacy) to use these data, given the type of data, and the subjects being investigated.

4.4.3 Nudge Selection

With further reference to Chapter 3, it is argued there that Page, Castleman and Meyer's (2020) "customized" nudges constitute *choice personalisation*, as the method of nudging – a reminder nudge – is not personalised, but the outcomes/choices/options nudged towards are personalised. Given this, and as with the previous subsection, it is quite inappropriate to critique their *selection of nudges*, as only one nudge has been selected.[200] Rather, it seems appropriate to consider the outcomes/choices/options Page, Castleman and Meyer (2020) instead select from.

Page, Castleman and Meyer (2020) detail on page 11 of their study four categories or classifiers which relate to their measure of heterogeneity – these are "*FAFSA not yet started"*; "*FAFSA submitted, not yet complete"*; *"FAFSA complete"*; and "*FAFSA complete, selected for income verification*." From these classifications, they write, "Students' FAFSA status information was updated in districts' data systems every 1 or 2 weeks. As this information was updated, OneLogos [the automatic messaging service utilised by Page, Castleman and Meyer (2020)] updated the message stream that students received" (Page, Castleman and Meyer, 2020: 11). In other words, when a student advanced from one classifier to another – say, from having no FAFSA to starting an application – the message sent to the student would also adjust.

The exact wording of these messages is provided by Page, Castleman and Meyer (2020) in their appendices. A review of Appendix A reveals that students in the control group received *some* of the messages received by students in the treatment group. Thus, the control group seems to function as an *impersonal nudge* group, while the treatment group receives *additional* text messages which are personalised. In all, the treatment group receives 3

[200] A conversation could be had, as part of a methodological critique, as to whether a reminder nudge was the best choice, or what rationale was used to support the selection of a reminder nudge. On this first point, it seems largely moot given the study at hand, as any investigation to resolve this query would, most likely, constitute an entirely different study in and of itself. On the second point, the pre-existence of communication infrastructure designed around disseminating text messages seems reasonable justification for the selection of a text-based reminder nudge.

personalised messages, despite their being four classifiers, because all students – irrespective of grouping – who have not started their application receive the same generic message.[201]

4.4.4 Mapping Procedure

Page, Castleman and Meyer (2020) adopt a rather different analytical approach compared to either Peer et al. (2019) and Schöning, Matt and Hess (2019), opting for a standard regression analysis (i.e. OLS). A departure from previous methods is to be expected here, as Peer et al. (2019) and Schöning, Matt and Hess (2019) investigate delivery personalisation, while Page, Castleman and Meyer (2020) investigate choice personalisation. Given the latter investigate a different aspect of personalised nudging, it is not surprising that they adopt a different analytical technique.

As above, Page, Castleman and Meyer (2020) utilise regression analysis, specifically OLS. It may be helpful at the immediate moment to ignore their dependent variable and return to this question shortly. Instead, immediate attention is paid to the independent variables which they use. Much like Peer et al. (2019) and Schöning, Matt and Hess (2019), Page, Castleman and Meyer (2020) demarcate their control and treatment groups (those who do not receive personalised reminder messages, and those who do, respectively) with a simple dummy variable. They include additional control variables, specifically variables controlling for individual student demographics, and variables controlling for school-level variation. Page, Castleman and Meyer (2020) do not give an indication of whether such controls were necessary following an analysis of distributions across the control and treatment groups and may have simply been included for prudency. Furthermore, as a single school could only belong to one group (i.e. all members of school A would be in a control group, all members of

[201] Generic in terms of information, though not content. A review of the content of messages found in Appendix A reveals that all students receive texts messages which include their name and their high school. In a rather washed-out sense of the word, therefore, all messages could be said to be personalised or customized. Insofar as the topic is discussed here, such personal inclusions do not count as personalisation. However, it should be acknowledged that these inclusions may have induced some effect, and thus the control group may not represent a true control group (where no personalisation is offered at all).

school B would be in a treatment group), it is possible that such a comparison of distributions across the control and treatment groups on a school-level would not be meaningful. Nevertheless, after controlling for individual- and school-level differences, Page, Castleman and Meyer (2020) determine whether personalisation was a) statistically significant and b) positive through an interpretation of their dummy variable and its associated coefficient.

Returning now to the dependent variable, Page, Castleman and Meyer (2020) share some similarities to Schöning, Matt and Hess (2019) in that they investigate multiple dependent variables, contrasting various regression results to inform their conclusions.[202] Page, Castleman and Meyer (2020) explore three broad dependent variables in greater minutiae than is necessary to discuss here. From a methodology perspective, they consider how personalised reminder nudges impacted student engagement with the FAFSA process, the completion rate of FAFSA, and the enrolment rate following FAFSA.[203] Student engagement is measured by the frequency of engagement with the reminder service (i.e. responding to messages, or messaging the service with questions, arranging appointments with expert advisers, etc.), while completion rate and enrolment rate are calculated as a percentage of students who A) complete FAFSA and b) who enrol at a university, respectively.

Page, Castleman and Meyer (2020) find statistically significant and positive effects on each of their dependent variables arising from the personalising reminder messages, compared to their control group. However, the significance and the size of the effect decreased from engagement, to completion, to enrolment. An explanation for why this occurred seems reasonably forthcoming – as time elapsed between being nudged, the effect of the nudge

[202] This is by no means obvious upon a first, or indeed second or third read. Page, Castleman and Meyer (2020) write "Y_{ijk} [dependent variable] is the *outcome of interest*" (Page, Castleman and Meyer, 2020: 13, emphasis added) but do not elaborate further on what "outcome of interest" means within the investigation's context.
[203] On the minutiae: Page, Castleman and Meyer (2020) break down their sample to investigate the significance of the treatment on students who interact more/less with the reminder service, those applying for courses of different length, and those enrolling at different times, for instance. These analyses in the context with which Page, Castleman and Meyer (2020) undertake their research – educational policy – seem worthwhile but are largely just repetitions of a general method outlined in the main body here, hence the decision to not focus on them.

decreased.[204] Nevertheless, the findings reported by Page, Castleman and Meyer (2020) are broadly consistent with the significant, positive effects identified by Peer et al. (2019) and Schöning, Matt and Hess (2019).

4.5 – Hirsh, Kang and Bodenhausen (2012)

4.5.1 Summary

In several ways, the work by Hirsh, Kang and Bodenhausen (2012) serves as an example of the intersection of methodologies and methodological arguments already explored in this chapter. The reason for this is quite clear – on the one hand, Hirsh, Kang and Bodenhausen (2012) follow from previous work (including Moon (2002), see below) in utilising a matching approach, rather than a mapping procedure.

On the other hand, Hirsh, Kang and Bodenhausen (2012) contribute important criticism of previous literature (see, again, Moon (2002), but also this chapter's discussion of Schöning, Matt and Hess (2019)), notably their criticism of the simplistic, dichotomous approach the likes of Moon (2002) and Schöning, Matt and Hess (2019) take. Hirsh, Kang and Bodenhausen (2012) write, "existing research has examined congruence effects primarily by separating message recipients into one of two psychological categories… utilizing a model of personality based on dimensional variation could allow for more fine-grained personalization of persuasive messages based on an individual's relative standing on a given trait dimension" (Hirsh, Kang and Bodenhausen, 2012: 578-579). Such an argument serves as the foundation of Egelman and Peer's (2015) criticism of the Big Five personality scale, with both this study

[204] Given these results, and some intuition, it is possibly inappropriate to directly attribute, as Page, Castleman and Meyer (2020) do, something like enrolment rate to the reminder nudge. Two broad reasons come to mind. Firstly, the decision to pursue higher education is a complex one involving many factors, and one should be mindful of this when attributing any effect. Secondly, while the nudge may have initially been significant, future significance may be better understood as something like a snowball effect, with students more likely to enrol given their access to FAFSA, and their access to FAFSA being a function of the nudge. Thus, it may be more realistic to think of this nudge as indirectly impacting, say, enrolment rates – a principle which Page, Castleman and Meyer's (2020) method would somewhat undermine. This criticism, however, is not a significant one.

and Hirsh, Kang and Bodenhausen (2012) contributing greatly to the methodological approach taken, ultimately, by Peer et al. (2019).

By way of summary in the context of this chapter, such points are sufficient. Hirsh, Kang and Bodenhausen (2012) do situate their work within the marketing literature, framing their investigation around the advertisement of a new smartphone, but they emphasise that their contribution is not to the field in which their study is framed – as opposed to Peer et al. (2019), who greatly situate their research within the cybersecurity sector, or Page, Castleman and Meyer (2020), who situate their work in the education sector. As such, less of an emphasis is offered here. Hirsh, Kang and Bodenhausen (2012) utilise a survey-experiment approach, with a sample of 324 participants recruited via Amazon's MTurk.

4.5.2 Psychometric Selection

Hirsh, Kang and Bodenhausen (2012) spend relatively little time justifying their psychometric selections, and comparatively more time critiquing the more simplistic approach of others. They write, "Although message-person congruence effects have been examined in relation to a variety of psychological characteristics, they have not yet been systematically related to a comprehensive model of personality traits. Such integration, however, would advance the message-framing literature by opening the door to exploring new ways to make persuasive messages more personalized and effective" (Hirsh, Kang and Bodenhausen, 2012: 578). As such, Hirsh, Kang and Bodenhausen (2012) select psychometrics based on the scale's comprehension and multidimensionality. As they continue, "examining message-person congruence effects within a comprehensive model of personality would allow for a multidimensional assessment of recipients' characteristics with a single measurement instrument" (Hirsh, Kang and Bodenhausen, 2012: 578).

It is perhaps no surprise, therefore, that they opt to utilise the Big Five personality scale, which they note is well-researched, generally regarded as robust and capable of measuring several (i.e. five) different aspects of personality. Within the context that Hirsh, Kang and

Bodenhausen (2012) conduct their research – namely, attempting to move beyond analysis using a simple dichotomous measure – the criticisms levied by Egelman and Peer (2015) that the Big Five personality scale is not specific enough and is inferior to other psychometric scales can be somewhat disregarded. Of course, this is a fair criticism given the evidence Egelman and Peer (2015) present, but as Hirsh, Kang and Bodenhausen (2012) emphasise, their immediate concerns are with the *comprehensiveness* of the scale. Given specificity at times can be anathema to comprehensiveness (or, to use an alternative word, *generality*), to take Hirsh, Kang and Bodenhausen's (2012) approach and subject it to direct criticism from Egelman and Peer (2015) may be unfair.[205]

4.5.3 Nudge Selection

Hirsh, Kang and Bodenhausen (2012) investigate *personalisation*, but do not investigate *personalised nudging*. In fact, they do not draw upon nudge theory at all, and thus it would be most inappropriate to characterise their work as such. The correct characterisation would be an analysis of their *message* selection, as Hirsh, Kang and Bodenhausen (2012) do adjust the advertising slogan shown to participants, in accordance with each of the five personality types covered by the Big Five personality scale.

Some parallels between what Hirsh, Kang and Bodenhausen (2012) do, and nudge theory, can be drawn, however. For instance, Hirsh, Kang and Bodenhausen (2012) offer a slogan which appeals to neuroticism (which they characterise as "especially sensitive to threats and uncertainty"; Hirsh, Kang and Bodenhausen, 2012: 579) "Stay safe and secure with the XPhone", a slogan which – cast under the guise of nudge theory – could easily be understood as designed to appeal to a person's sense of risk or loss aversion. As above, it is wrong to characterise this and other slogans developed by Hirsh, Kang and Bodenhausen (2012) as

[205] It is important to note that Egelman and Peer's (2015) work is informed directly by Hirsh, Kang and Bodenhausen (2012), and the former very much proceed with an interest in improving the conceptual procedure advanced by Hirsh, Kang and Bodenhausen (2012). Therefore, one should be reluctant to see these studies as antagonistic and would be better served conceiving as Egelman and Peer's (2015) work as a direct attempt to build on and improve Hirsh, Kang and Bodenhausen's (2012) work.

nudging, as they do not utilise behavioural insights or a nudging framework, but mechanically, the integration of nudging into a process such as adjusting slogans seems eminently reasonable.

4.5.4 Mapping Procedure

As above, Hirsh, Kang and Bodenhausen (2012) follow previous literature (see Moon, 2002) and precede other literature (Schöning, Matt and Hess, 2019) in their use of a *matching* procedure, or as they dub it, “message-person congruence” (Hirsh, Kang and Bodenhausen, 2012: 578). Yet, in comparison to, say, Schöning, Matt and Hess (2019), Hirsh, Kang and Bodenhausen (2012) take a slightly more sophisticated approach.

Primarily, they use regression analysis (again, OLS), constructing a dependent variable measuring the effectiveness of an advertisement by averaging six answers to slightly different questions probing effectiveness and persuasiveness.[206] They demonstrate the validity of this aggregation (i.e. averaging) with a test for Cronbach’s alpha, reporting a high score for internal validity. Statistically, therefore, this construction is adequate, yet with reference to Page, Castleman and Meyer (2020) and Schöning, Matt and Hess (2019), both of whom investigate the effectiveness of personalisation against several dependent variables, it may be reasonable to consider whether averaging these responses was necessary, or whether Hirsh, Kang and Bodenhausen (2012) might have used each response as a different dependent variable and assessed the consistency of their results across multiple regressions.[207]

Nevertheless, Hirsh, Kang and Bodenhausen (2012) proceed with an average effectiveness score for each type of advertisement as their dependent variable, with a small adjustment (see below). It is worth taking a moment, before proceeding, to discuss the experimental procedure which they undertake.

[206] For instance, agreement statements such as, “this is an effective advertisement”; “I would purchase this product after seeing this advertisement” and so on.
[207] It is not obvious that this alternative approach would be better, and would certainly require more work than Hirsh, Kang and Bodenhausen’s (2012) approach requires.

Based on the phrasing given by Hirsh, Kang and Bodenhausen (2012), it would seem they did *not* adopt a random allocation procedure (i.e. randomly assigning nudges/messages to individuals), but instead showed *all* advertisements to every participant. With reference to Peer et al. (2019), this method may create potential for priming effects, as a participant may respond differently after seeing a similar advertisement multiple times (even if the personality framing varies). Hirsh, Kang and Bodenhausen (2012) report evidence which may indicate as such, with a very high correlation between all the effectiveness scores. Of course, the reason for adopting this approach may also be methodological; Hirsh, Kang and Bodenhausen (2012) use a relatively small sample of 324, meaning if they randomly assigned messages to participants and stratified based on matches after-the-fact, there may be too few natural matches for a reasonable comparison across matched-unmatched groups.[208] This being the case, the risk of priming effects may be outweighed by methodological necessity.

For independent variables, Hirsh, Kang and Bodenhausen (2012) administer the Big Five personality scale after recording participant feedback on the advertisements.[209] Each participant was thus measured for each of the five personality types, and these five scores associated with each individual were utilised as independent variables within their regression models.

As above, they note a small adjustment was made for their dependent variable. Hirsh, Kang and Bodenhausen (2012) report high correlation between their five measures of advertisement effectiveness, and argue such high correlation demonstrates that these advertisements – despite being framed differently – must have still been reasonably similar. To isolate the

[208] This may also be an issue facing Schöning, Matt and Hess (2019). Schöning, Matt and Hess (2019) had an even smaller sample of 156. However, for Schöning, Matt and Hess (2019), the probability of matching was 50%, so the matching group can be estimated to contain around 78 participants, compared to an unmatched group of another 78. For Hirsh, Kang and Bodenhausen (2012), the probability of matching is 20% (1 in 5), and based on a sample of 324, the estimated matching group based on random allocation would be around 65 compared to an unmatched group of 259. As such, the random allocation method may not have been suitable for Hirsh, Kang and Bodenhausen (2012) – despite a relatively larger sample size – because of the relatively lower probability of matching.

[209] Hirsh, Kang and Bodenhausen (2012) do not state what version of the Big Five they use.

variance attributable to the framing, rather than to aesthetic similarities,[210] they utilise some additional regression analysis. For the sake of discussion here, assume five variables measuring the effectiveness of the five personality type-advertisements, X1 through X5. Hirsh, Kang and Bodenhausen (2012) regressed, say, X2, X3, X4 and X5 on X1, before extracting the residual. This residual was said to capture the variance that cannot be attributed to similarities in the other advertisements, and as such, Hirsh, Kang and Bodenhausen (2012) do not use average effectiveness of each personality type as the dependent variable in their main regression analyses, but the *residuals* produced by these five initial regressions.

By way of a further departure, Hirsh, Kang and Bodenhausen (2012) do not construct a dummy variable to demarcate groups. By instead using the personality scores for each personality type produced for each individual (i.e. each individual had a score for each of the five personality types), they do not have to (somewhat arbitrarily) assign participants to, say, an extraversion group, or a neuroticism group. This resolves an issue present in the method of Schöning, Matt and Hess (2019), namely, Schöning, Matt and Hess (2019) determined whether a person was verbal or visual simply based on whether a person came above or below the average of their verbal/visual scale. As argued above, this approach could be called arbitrary and is subject to some criticism. By simply allowing the responses of participants to 'speak for themselves' (in a manner of speaking), Hirsh, Kang and Bodenhausen (2012) avoid this issue.

Over five regressions, Hirsh, Kang and Bodenhausen (2012) find that individual scores in a particular personality type statistically significantly and positively predict the effectiveness of the corresponding advertisement (i.e. the extraversion trait significantly and positively predicts the effectiveness of the extraversion-framed advertisement). Furthermore, in every regression, *only* the corresponding personality type is found to be significant, and *often* it is

[210] "We constructed five advertisements featuring a picture of a cell phone and a few lines of text; *the text was manipulated* so that each advertisement highlighted the motivational concerns associated with one of the five personality dimensions" (Hirsh, Kang and Bodenhausen, 2012, emphasis added).

only the corresponding personality type which has a positive effect (i.e. positive coefficient). Based on these results, Hirsh, Kang and Bodenhausen (2012) conclude that personalising advertisements to match personality types produces statistically significantly positive effects on advertisement effectiveness, which is consistent with all other studies discussed thus far. As with Schöning, Matt and Hess (2019) – and the matching procedure generally – it is questionable whether the method undertaken by Hirsh, Kang and Bodenhausen (2012) should be called personalisation, because they only demonstrate the *potential* effectiveness of personalisation, and do not purposely personalise themselves. Overlooking this (potentially semantic) criticism, however, Hirsh, Kang and Bodenhausen (2012) do produce results which are consistent and do so using a methodology which offers several insights which can (and in the case of Peer et al. (2019) do) inform other studies discussed here.

4.5.5 Additional Comments

One important contribution to this discussion offered by Hirsh, Kang and Bodenhausen (2012) is that of imagery or visualisation. Peer et al. (2019) adjust the password tips shown to participants, while Page, Castleman and Meyer (2020) adjust the content of a text message sent to participants. It is only Schöning, Matt and Hess (2019) who come close incorporating some sort of visual element into their intervention. Even here, however, the visual discussion is between the use of icons to demarcate blocks of text, or bullet points. Hirsh, Kang and Bodenhausen (2012) are the only set of authors thus far to integrate a significant visual element into their research, even if *the visual element is not significantly changing*.

By noting that similarities in the imagery used may have been impacting the effectiveness scores which participants produced, and treating for this effect by trying to isolate the variance resulting from the framing, Hirsh, Kang and Bodenhausen (2012) provide an insightful guide to how any intervention utilising visuals might proceed (again, even if the adjustment in the visuals is not a significant component of the intervention). Thus, despite being quite similar in methodological approach and result to previous studies discussed, Hirsh, Kang and

Bodenhausen (2012) provide practical methodological insights which previous authors have not needed to consider.

4.6 – Moon (2002)

4.6.1 Summary

Moon (2002) is, in many ways, a forerunner to the work of Hirsh, Kang and Bodenhausen (2012). Indeed, the substantive difference in approach between the former and the latter is quite clearly explained by Hirsh, Kang and Bodenhausen (2012),[211] and reiterated here, namely, Moon (2002) investigates the role of personality and responses to messaging using only a simple measure of personality. This simple measure is a single aspect of the Big Five personality scale, *extraversion*, which Moon (2002) deconstructs further into dominant personality types (dominance) and submissive personality types (submissiveness). Following Hirsh, Kang and Bodenhausen (2012), such a limited scope may not be sufficiently broad to truly capture personality; yet, following Egelman and Peer (2015), Hirsh, Kang and Bodenhausen's (2012) broadness may not be sufficiently specific. It is unfair, thus, to necessarily critique Moon (2002) with (at the time of writing) nearly two decades of hindsight and experimental development to draw upon.

The rationale for drawing on Moon (2002), then, given more contemporary literature exists which has developed the methodology of personalisation studies greatly, is largely to demonstrate this development and evolution of methodologies. Just as the contrast of the methods adopted by Peer et al. (2019) and Schöning, Matt and Hess (2019) reveal areas of methodological compliment and conflict, by returning to a relatively early study such as Moon's (2002), the evolutionary story beginning with Moon (2002) and ultimately arriving at Peer et

[211] Hirsh, Kang and Bodenhausen (2012) do not directly refer to Moon (2002), but they do discuss their work as building upon "existing research [which has] examined congruence effects" (Hirsh, Kang and Bodenhausen, 2012: 578), a group in which Moon (2002) can mostly certainly be considered.

al. (2019) and Schöning, Matt and Hess (2019) via Hirsh, Kang and Bodenhausen (2012) can be seen.

Much like Hirsh, Kang and Bodenhausen (2012), Moon (2012) necessarily investigates personalisation within a specific domain (the consumer goods market) but does not place their research within the consumer goods literature. Rather, Moon (2002) seeks to contribute directly to the personalisation and personality literature, and so throughout, her work places minimal emphasis on the domain in which her experiments are situated. Moon (2002) utilises an experimental task in combination with data collection via an incentivised survey, and presents results arising from a sample of 48 participants.

4.6.2 Psychometric Selection

Moon (2002) initially argues that the Big Five personality scale is a sufficient psychometric or personality scale to draw upon as it is the most extensively researched personality scale. Moon (2002) offers no reason why she chooses to only select a single characteristic captured by the Big Five personality scale – extraversion – though a reason may be speculated when considering Moon's (2002) analytical procedure, namely, a matching approach is used, and thus a psychometric or personality measure which can be easily converted into a dummy variable to indicate matching and unmatching groups may have been desirable. As Hirsh, Kang and Bodenhausen (2012) show, alternative methodologies can be developed which integrate all five aspects of the Big Five.

Nevertheless, Moon (2002) does justify her selection of the extraversion characteristic *in particular*, writing, "the most 'psychologically prominent' factor is the dominance and submissiveness ('extraversion') dimension of personality, which has been found to provide information, relative to other factors, about what an individual is 'really like.' It is thus more useful in understanding and predicting another individual's behaviour" (Moon, 2002: 314). Given this, it might be argued that despite the subsequent development of a methodology by Hirsh, Kang and Bodenhausen (2012) which could accommodate all aspects of the Big Five, this may not be necessary, as each aspect does not offer an equal contribution to the ultimate picture of individual personality the scale produces. Certainly, this is Moon's (2002) implicit assertion.

4.6.3 Nudge Selection

As with Hirsh, Kang and Bodenhausen (2012), it would be incorrect to characterise Moon (2002) as selecting a nudge, as at no point does Moon (2002) invoke nudge theory. Equally – again as with Hirsh, Kang and Bodenhausen (2012) – the message framing employed by Moon (2002) can be re-interpreted through the lens of nudge theory, and thus some semblance of a parallel established. For instance, in Moon's (2002) first experiment investigating consumer preferences in automobiles, participants are offered statements such as, "The *Elantra* does not come very well-equipped, which will be annoying if you like air-conditioning or power steering" (Moon, 2002: 316), a statement which could easily be understood as an appeal to, say, loss aversion.

The issue with the re-interpretation offered here, compared to the similar procedure offered in relation to Hirsh, Kang and Bodenhausen (2012), is that Moon's (2002) framing does not vary the component of the statement which would be associated with the nudge. For instance, in the above statement, the loss aversion component emerges from the phrase, "which will be annoying if you like air-conditioning or power steering." Yet, in both the dominant framing (given above) and the submissive framing,[212] the potential for loss is emphasised. As such, while some endeavour can be made to relate the procedure adopted by Moon (2002) to nudge theory, one must be cautious in doing so, and aware of any limitations which arise in turn.

However, a consideration of Moon's (2002) experiments through the lens of nudge theory is not a fruitless endeavour, and in doing so, potentially reveals some methodological oversights on the part of Moon (2002). In experiment 1, for instance, Moon (2002) asks participants to

[212] Moon (2002): "Perhaps you should put the *Hyundai Elantra* lower in your ranking? The *Elantra* does not come very well-equipped, which may be annoying if you like air-conditioning or power steering" (Moon, 2002: 316).

rank 12 cars in order of preference. An automated program then presents participants with an alternative ranking, which is generated using a standard transformation of the participants' original ranking.[213] The computer justifies its ranking with various statements, such as that quoted above. These statements are worded to appeal to either a dominant or submissive personality type. However, by the computer presenting an explicit ranking, one might wonder whether this serves as a kind nudge, such as establishing a default option. Madrian and Shea (2001) consider such a hypothesis in their study of default options and retirement savings behaviour. As such, when participants are invited to change their ranking, given the messaging and the computer ranking, any change that occurs *may* be attributed to appeals to various personality framing, but *may* also be attributed to some nudging effect.

Of course, this is just a hypothesis, and Moon (2002) conducts manipulation checks for both of her experiments, finding participants successfully identify dominant and submissive framing. This hypothesis may continue to stand, however, even in the presence of significant results, as the combination of effects may be producing statistical significance. In turn, one might wonder whether, rather than offering participants the computer's ranking, Moon (2002) should have not simply had the computer pass commentary on the *participant's* ranking.[214]

Regardless of this relatively small quandary, Moon (2002) constructs her messages following a reasonably consistent and logical formulation. In the first experiment, all statements begin with a firm (dominant) or softer (submissive) comment on ranking, before an additional statement referring to various features about the cars which seems to justify the statement.[215] In the second experiment, which relates to recommended entertainment content, messages

[213] For instance, whatever car the participants ranked as number 4 is ranked by the computer as number 1.
[214] For instance, a statement such as, "Number 4 is ranked too low" does not establish a new, recommended ranking for number 4, but does convey a message which is attempting to persuade the participant to change their ranking. Such a statement could be framed in accordance with the dominant and submissive framing without much issue: "Number 4 is ranked too low" follows a dominant frame, while "Perhaps number 4 is ranked too low?" follows a submissive frame. By showing the computer's ranking, and then explaining that the computer believes number 4 to be ranked too low, the participant is not just informed of the computer's opinion regarding the *relative* ranking of number 4 (i.e. too low) but the *absolute* ranking of number 4, which may influence the re-ranking undertaken by the participant just as much as the message framing itself.
[215] Participants were provided with various information about these cars during the experiment.

consist of an initial statement introducing the type of media content (i.e. news report, music, a joke, health advice etc.), followed by a statement either confidently assuring (dominant) or cautiously suggesting (submissive) the participant will enjoy the content.

4.6.4 Mapping Procedure

As touched on above, Moon (2002) utilises a *matching* procedure which has subsequently been reproduced rather faithfully by Schöning, Matt and Hess (2019), and shares characteristics with the matching procedure adopted by Hirsh, Kang and Bodenhausen (2012). Moon's (2002) treatment is utilised over two experiments, each utilising 48 participants.[216] Immediately, one should note the tremendously small sample size, even when compared to Schöning, Matt and Hess (2019). Of course, the relatively small sample may be due to factors external to Moon's (2002) research and beyond her control; nevertheless, the limited number of participants necessarily limits the statistical analysis which can be robustly utilised. This, though an unfortunate justification, may serve as justification for Moon's (2002) use of the matching procedure and relatively simple personality framework (compared to, say, Hirsh, Kang and Bodenhausen (2012)).

Participants were invited by Moon (2002) to complete a personality test several weeks prior to either experiment. Moon (2002) notes that this was done to avoid any association between the experiment and the personality test, suggesting she may have had concerns about the role of framing the personality test may play in participant behaviour. Like Peer et al. (2019), however, this is never stated as such, and can only be speculated on.

Upon completion of these personality tests, Moon (2002) aggregated participant scores, reporting a high Cronbach's alpha. However, she then proceeds to assign participants to either a dominant or submissive group using a rather questionable method. Recall that Schöning,

[216] It is unclear whether the same participants are used in both experiments, or whether 96 different participants are used across the experiments. The former seems most likely, as there is little reason to maintain the same sample size across experiments which are only going to be qualitatively compared, rather quantitatively contrasted.

Matt and Hess (2019) tested their participants using a verbal/visual test, before splitting the group based on whether a participant scored higher or lower than the midpoint. As above, this method is criticised because detail is lost, and those near the margins (i.e. near the midpoint) may not be faithfully represented by this division, and the selection is arbitrary. Moon (2002) adopts a similar method to Schöning, Matt and Hess (2019), splitting participants into dominant and submissive groups based on whether a participant scored above or below the *median* score (respectively). From a sampling perspective, this seems to have been done to ensure an equal sample of 24 participants in each group. But besides suffering from the arbitrariness which Schöning, Matt and Hess (2019) also suffer, the use of the median score embeds sampling bias into the groupings. For instance, participants in the sample may randomly contain many more dominant, or many more submissive, personalities, which would skew the group median compared to the expected median from the population and result in some participants potentially being misclassified. Some comparisons of the median with the mean could potentially strengthen the rationale for this procedure, but problems surrounding representativeness at the margins – which Schöning, Matt and Hess (2019) share – still persist. Furthermore, such a method of constructing groups must return a critical eye to the sample size; the use of median, mean or any dividing value becomes more acceptable as the sample comes to resemble the population, which is to say, when the sample size is very large. The risk that the sample group used by Moon (2002) is not representative of population is very high given the sample size is so small. It seems only fair, therefore, to conclude that the method of constructing groups adopted by Moon (2002) is very inadequate.

Regardless, Moon (2002) proceeds to conduct 2 experiments. The first experiment asks participants to rank 12 cars based on whatever criteria the participants see fit. Moon (2002) notes that all cars are reasonably comparable in terms of price and utility. Once participants have submitted their ranking, they are shown a computer's ranking, as well as various messages explaining why the computer's ranking differs with their own. The computer is programmed so as to always alter the participants' rankings in the same way, regardless of how individual participants rank the cars. The messages shown to participants will either follow a dominant tone, or a submissive tone. The tone is randomly determined for each participant. Finally, participants are offered the chance to adjust their ranking, given the computer's ranking and messages. As the computer's ranking is a standardised transformation of the participant's ranking, any changes which participants make can be quantified and compared across participants. Moon (2002) uses this variable as a measure of the message's persuasion. This is one of the dependent variables used in experiment 1. The second dependent variable is constructed from a series of survey questions which ask respondents how persuasive they found the computer to be and is verified using Cronbach's alpha. As such, Moon (2002) argues these two dependent variables capture persuasiveness as an expression of actual behaviour (adjusted rankings) and perception of persuasiveness.

Experiment 2 follows a somewhat similar strategy to experiment 1. In experiment 2, participants were presented with four pieces of entertainment content, including a news report, a song, a joke, and a health-tip. Before being shown each piece of the content, participants received a message from the computer summarising the computer's confidence that the participant would enjoy the content. Again, these messages took on dominant or submissive framing, and were randomly assigned to participants. Participants were then invited to complete a survey assessing how much they enjoyed each piece of content, and for each piece, a dependent variable was constructed. Again, these variables all showed a high Cronbach's alpha. A fifth dependent variable was also constructed, based on survey responses to participants' general perceptions of the computer's competence.

Following a matching procedure near identical to that of Schöning, Matt and Hess (2019), a dummy variable was constructed to distinguish between those whose messaging matched their personality type, and those whose messaging did not match. Differences in the various dependent variables between these groups were then examined using a t-test for both experiments, as well as ANOVA in experiment 2. Moon (2002) reports consistent evidence of statistical difference between matching and non-matching groups for *most* dependent variables. Where statistical significance was identified, the results consistently supported the hypothesis that messages which matched the personality types of participants would be more persuasive/received more favourably than those that did not match. For the two results which were not significant (the news content, and the health-tip content), the means still suggested matching produced more favourable responses to messaging. These results are consistent with all other literature examined.

4.6.5 Additional Comments

Few additional comments need be made regarding Moon (2002). The consistency in methodology between Moon (2002) and Schöning, Matt and Hess (2019), adjusted slightly by Hirsh, Kang and Bodenhausen (2012), indicates a relative methodological robustness in the matching procedure. However, as argued above, the matching procedure only dubiously personalises interventions, suffers from arbitrariness, and can be incorporated into the more rigorous mapping procedure developed in part by Peer et al. (2019).

4.7 – Conclusion

As an aid to the conclusions drawn from this section, details of the reviewed studies are summarised in Table 1:

Table 1: Summary Methodology Details for Reviewed Studies[217]

Study	Choice or Delivery?	Sample Size	Data Collection Method	Mapping or Matching?	Psychometrics	Nudges	Statistical Treatment	Significant?
Peer et al. (2019)	Delivery	2755	Incentivised Survey Experiment	Mapping	GDMS, CFC, NFC, ANS	Password Meter, Crack-Time, Social Norm, CHBS, Insertion	Cronbach's alpha, Moderated Regression, Johnson-Neyman Technique, two-tailed t-test, ANOVA, WMW-test	Yes
Schöning, Matt and Hess (2019)	Delivery	156	Survey Experiment	Matching	Verbal/Visual	Bullet points, Icons	Cronbach's alpha, Factor Analysis, two-tailed t-test	Yes
Page, Castleman and Meyer (2020)	Choice	17731	Partnered Project	Matching	Progress Score	Reminder Text Message	OLS Regression	Yes
Hirsh, Kang and Bodenhausen (2012)	Delivery	324	Incentivised Survey Experiment	Matching	Big Five Personality Scale	Message Framing	Cronbach's alpha, OLS Regression, two-tailed t-test	Yes
Moon (2002)	Delivery	48	Incentivised Survey Experiment	Matching	Dominance and Submissiveness (Extraversion)	Message Framing	Cronbach's alpha, ANOVA, two-tailed t-test	Yes

[217] Note: for all studies to be presented in a tabular format, some liberties have been taken with the reporting of some details. For transparency: 1) Page, Castleman and Meyer (2020) may not neatly align with the notion of mapping *or* matching; 2) Page, Castleman and Meyer (2020) do not refer to their data collection method as a "partnered project" but do establish that their data was collected as part of a collaboration with several outside institutions; 3) Hirsh, Kang and Bodenhausen (2012) and Moon (2002) do not liken their messages to nudges, and as such, "Message Framing" should not be misconstrued as a type of nudge; 4) "Progress Score" is not a psychometric, nor do Page, Castleman and Meyer (2020) refer to psychometrics, though the phrase "Progress Score" broadly captures the heterogeneity Page, Castleman and Meyer (2020) used to personalise their interventions; 5) while all studies produced significant results, significance was not consistent across all tests performed by all authors, and levels of significance varied on multiple occasions.

Immediately, several commonalities emerge from this side-by-side comparison. All but Page, Castleman and Meyer (2020) collect data via survey experiment, with, out of these, only Schöning, Matt and Hess (2019) *not* using an incentivised survey experiment. All but Page, Castleman and Meyer (2020) utilise common statistical treatments including Cronbach's alpha and t-test, while some – including Page, Castleman and Meyer (2020) – utilise some type of regression analysis, notably OLS. All but Peer et al. (2019) adopt a matching approach in their analysis, and all but Page, Castleman and Meyer (2020) investigate delivery personalisation. Finally, all find statistically significant results.

Equally, various disparities in approach can also be identified. There is a large variance in the sample size employed by these studies, with a tendency towards relatively small (>500) samples. There is also little consistency in terms of psychometrics or nudges which have been analysed. Both points, to an extent, have simple explanations. On the matter of sample size, beyond various exogeneous factors which may have impacted the capacity of the researchers, sample may have been influenced by the planned statistical analysis. On the matter of psychometric and nudge selection, given all the studies situated their research in rather different domains, it is to be expected that the types of psychometrics to be analysed and the variety of nudges to be examined would vary across these domains.

The advantage of the detailed analysis offered in this chapter, and the summary provided above, is that a methodological route for the investigation of the hypotheses established in Chapter 3 can now be determined. From this route, further details can be established, from psychometrics to nudges to advertisement design to statistical procedure and logic. The following chapters in this section discuss these considerations, making frequent reference to the literature discussed above as part of the methodological justifications for various choices made in this research design. On occasion, explicit reference is made to Table 1, however, readers are invited to use Table 1 as a helpful summary-resource even when explicit reference is not made.

Chapter 5 – An Introduction to Political Advertising and Nudge Selection

5.1 – Introduction

This chapter establishes the four behavioural nudges to be examined in this thesis. These are the status quo nudge, the present bias nudge, the loss aversion nudge and the social norm nudge. Firstly, a brief discussion of political advertising is offered. This is provided here because it is necessary to understand the context in which any nudges would be used, so as to interrogate the suitability of any selected nudges. Such interrogation follows a brief review of the literature pertaining to each nudge.

It is helpful to begin this chapter by returning to the research question and corresponding hypotheses pertaining to this thesis:

Research Question: *Can personalised nudges be used to significantly influence political decision-making?*

Hypothesis 1: *Personalised nudges will be statistically significantly more effective at influencing political decision-making than impersonal nudges, which in turn will be more effective than not nudging.*

Hypothesis 2: *Choice and Delivery personalised nudges will be statistically significantly more effective at influencing political decision-making than delivery or choice personalised nudges alone.*

The purpose of this exercise is to establish a sensible starting point to begin discussing the method used in this thesis, and the many moving parts involved. By reviewing the above question and hypotheses three worthwhile starting points emerge, namely: what is the political component investigated in this thesis?; what nudges are investigated?; and how do these relate?

5.2 – A Brief Introduction to Political Advertising

In trying to investigate political decision-making, one must first formulate a way of invoking a political decision. This thesis focuses on political advertising, more specifically, investigating how individuals respond to various political advertisements. Of course, one might consider an alternative – and possibly more realistic – political decision to be casting a ballot at the ballot box, rather than evaluating a political advertisement.[218] Yet this alternative is problematic for several reasons, most notably because few – if any – legitimate democratic exercises would accommodate attempts to influence said exercise. This is not to say that attempts are not made, but for the most part they are made *away* from the ballot box – in speeches, televised debates, newspaper columns and *political advertising, which are increasingly found on social media* (Bakir et al., 2019). Furthermore, while this project will be concerned only with citizens who *experience* democratic institutions,[219] it should be acknowledged that *political* influence can still be exerted – often in the form of propaganda – in non-democratic countries.[220] Experiences, therefore, of political advertising are likely more numerous than of *democratic exercise*, which remains limited to those eligible to part-take in supposedly democratic countries.

A further advantage of focusing on political advertising is the contemporary nature of the medium in relation to the mechanism under investigation here, namely personalised nudging. In recent years, controversial stories have emerged regarding political advertising being targeted at very specific individuals via social media networks which command – for the sake

[218] See, for instance, Downs' (1957) economic theory of voting. Here, Downs (1957) argues that not merely the *choice of who to vote for*, but the choice *to vote at all,* is subject to a rational evaluation of the costs and benefits.
[219] US citizens. See Chapter 7.
[220] For a contemporary account of propaganda and influence in democratic and non-democratic countries, see Robinson (2019).

of linguistic consistency – heterogeneity data about individuals (Zuboff, 2019). In the emerging rationale of modern political advertising, while influence may be minimised at the ballot box, influence can be exercised over individuals in a variety of ways prior to their voting.

5.3 – Selecting Nudges

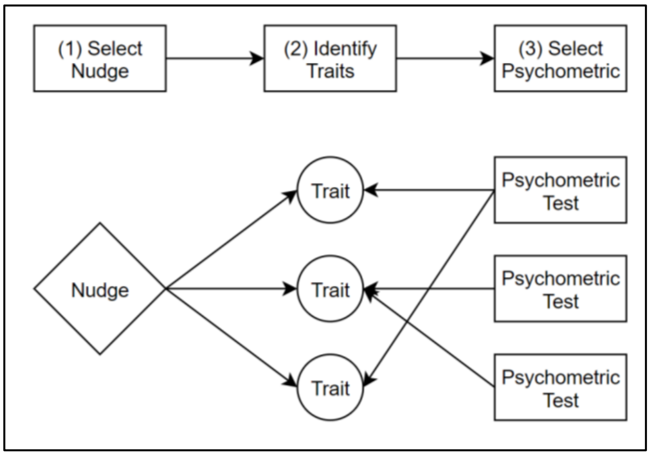
This thesis returns to a discussion of political advertising in Chapter 7, where choices over the construction of the political advertisements used in this research are discussed. At present, it is more opportune to turn to a discussion of nudges. This is because nudges are embedded within the political advertisements used in this research, and as such it is prudent to understand the nature and rationale for the selection of nudges prior to any fuller discussion of particular political advertisements.

When considering the existing literature, little indication of which nudges to investigate can be found. There are two reasons for this. Firstly, much of the literature does not investigate nudging *per se*. Secondly, where nudges are clearly used, these nudges often result from the psychometric tests particular authors use. For instance, see Schöning, Matt and Hess (2019) with their verbal/visual scale and corresponding verbal/visual nudges.

It is argued here that it is more prudent to first select nudges to examine, identify the psychological traits associated with those nudges, and then select psychometric scales which are designed to measure propensity of those traits. This approach has been taken, in varying degrees, with Peer et al. (2019), Guo et al. (2020) and Lipman (2020), with the former two selecting ‘nudges’ which respond to problems typically seen in password design, while the latter selects behavioural interventions which can be incorporated into financial incentives structures (e.g. loss aversion or the present bias).

As such, nudges are first selected, the psychological traits associated with these nudges identified, and psychometric scales which are expected to measure these traits thus chosen.

Figure 1: Selecting Nudges



This method of selection, visualised in Figure 1, also offers a guide for how to construct the hypothesised psychometric map constructed later (see Chapter 6).

Four nudges are selected for this research. These are 1) the status quo nudge; 2) the present bias nudge; 3) loss aversion nudge; and 4) the social norm nudge. These nudges are selected primarily on the basis that they can be incorporated into the medium for assessing political decision-making, namely, political advertising, and secondarily on their prominence in the literature. In 5.3.1 to 5.3.4, each nudge is discussed, with the psychological traits underpinning each nudge clearly identified. A demonstration of how the nudge can be incorporated into political advertising is then offered.

5.3.1 The Status Quo Bias

Investigation of the status quo bias largely begins with Samuelson and Zeckhauser (1988), who investigate the impact of status quos on subsequent decisions. They identify a significant bias towards the status quo (i.e. the status quo bias), writing, "Subjects in our experiments adhered to status quo choices more frequently than would be predicted by the canonical model" (Samuelson and Zeckhauser, 1988: 8), with the "canonical model" being the rational model of decision-making, whereby agents are expected to evaluate prospects independent of previous choices. The status quo bias, therefore, can be understood as a tendency for decision-makers to prefer whatever option is the status quo. Samuelson and Zeckhauser

(1988) hypothesise that this bias may occur because people seek to reduce perceived risk by selecting options which they are more familiar with, or they believe to have a more certain understanding of. By framing choices around a status quo, the status quo bias may be used to nudge decision-makers.

Closely related to the status quo bias is the default option nudge, a nudge which could be characterised as the nudge-transformation of the status quo bias. For instance, in their investigation of the default option nudge, Madrian and Shea (2001) make frequent reference to the status quo bias and Samuelson and Zeckhauser (1988). Madrian and Shea (2001) argue that decision-makers may prefer defaults because they see them as a sort of recommendation, a hypothesis which subsequent research into information leakage (Sher and McKenzie, 2006; Tannenbaum and Ditto, 2011) supports. Johnson et al. (2012) argue defaults and the status quo may appeal because people are impatient or seek to avoid making decisions.[221] Kahneman, Knetsch and Thaler (1991) also suggest the status quo bias may be driven by a desire to avoid risk.

Insofar as the default option nudge establishes a status quo – as the default option is often characterised as the choice an individual would receive if they did nothing (Thaler and Sunstein, 2008) – it may be tempting to equate the default option nudge with the status quo bias. However, this may be unhelpful for the purposes immediately at hand, for two reasons.

Firstly, the default option nudge changes a very specific piece of choice architecture, namely, whatever option is set as the default. While the changing of the default may be considered to be establishing a status quo of sorts, this specific adjustment is by no means the only way of establishing a status quo. Indeed, regardless of the default option, all decision-makers come to any decision with an established status quo. As such, while the default option nudge is very

[221] Also see Akerlof (1991), Madrian and Shea (2001) and O'Donoghue and Rabin (2001) for their respective work on procrastination and decision-making.

closely related to the status quo bias, it should be understood as *a* status quo nudge, rather than *the* status quo nudge.[222]

This distinction is advantageous given the second objection, namely that the use of a default option nudge as a means of appealing to the status quo does not seem sensible within the context of political decision-making. To be clear, a status quo exists in political decision-making, just as a status quo exists in all decision-making. In this instance, the status quo decision is to not vote, as this is the outcome a person will receive if they do nothing. But it does not seem substantially correct, by extension, to assert that a default option exists in political decision-making. Instead, political decision-making is better described as an active choice (Sunstein, 2012), whereby a person must indicate their preference. This is evidenced by the fact that in the voting process, people are presented with political choice to not vote via the ruining of their ballot (Lijphart, 1997). In fact, it seems rather anathema to the notion of democracy to impose a default option. Such an argument extends to the domain of political advertising; if the ultimate decision an advertisement is trying to influence a person about is an active choice, it seems wholly inconsistent to suppose the advertisement may nudge using a default option.[223]

As such, using a default option nudge as a status quo nudge is not sensible, as the second objection outlines, but equally, the default option nudge is feasibly *not the only* status quo nudge which could be used, as the first objection outlines. Combining these objections,

---

[222] Samuelson and Zeckhauser (1988) do investigate the effects of the status quo by using an experimental design likely comparable to the default option nudge. For instance, they suggest to participants that some investments have already been accounted for, and should the participant choose to do nothing, these accountings will be the fate of their investments. On the one hand, therefore, Samuelson and Zeckhauser (1988) can be seen to be describing the status quo in similar language to the description of the default option nudge given by Thaler and Sunstein (2008). On the other, however, one might wonder what a status quo is, if not the expected result of doing nothing? Defining, say, the default option nudge in these terms is not inaccurate, merely *incomplete*; such a description defines the default option nudge far more broadly than it need be.

[223] One might imagine, for instance, an advertisement which read, “now that you have seen this advertisement, if you do not indicate otherwise, you will vote for this candidate.” Such a statement is doubly absurd, as the advertisement has no licence in a free society to impose such a loss of agency on a decision-maker, nor does the advert have such a licence in a democratic society governed by free expression and political anonymity.

therefore, reveals a route to adequately utilising the status quo bias. If a status quo nudge is simply imagined as any choice architecture which appeals to a status quo, it seems eminently possible to imagine a statement which could be incorporated into a political advertisement which does this.[224] This endeavour is actually aided, rather than hindered, by the political domain, as many political (electoral) decisions are a choice between incumbent candidates and challenger candidates. As such, it would seem reasonable to imagine political advertising to evoke a candidate's incumbency.[225]

By way of a political slogan which nudges decision-makers by appealing to the status quo, the following is offered:

"Let's Keep Going!"

The emphasis on continuation clearly suggests that the candidate associated with this slogan is the incumbent, and as such, this slogan represents a status quo nudge as it is appealing to the status quo. Finally, given the format (i.e. a political slogan), this status quo nudge is clearly incorporated into a political advertisement, unlike the default option nudge.

5.3.2 The Present Bias

The present bias represents a more recent manifestation of a much older notion of value varying depending on whether it is received today or sometime in the future (O'Donoghue and Rabin, 2015). While O'Donoghue and Rabin (2015) attribute the foundational thinking of the contemporary study of present bias to Liabson (1994), it is in fact O'Donoghue and Rabin

[224] Furthermore, such a characterisation is not inconsistent with the definition of the default option nudge, which can very easily be understood as also appealing to a status quo, albeit by *establishing* a new status quo.
[225] Or, indeed, a candidate's insurgency. The actual electoral advantage of incumbency is mixed. See, for instance, Ade, Freier and Odendahl (2014). The point here is not to argue that appeals to incumbency universally convey an advantage; rather, it is to assert that incumbency represents a status quo which can sensibly be invoked.

(1999a) who distinguish between Laibson's (1994, 1997) analysis of time-dependent discounting and the behavioural notion of bias.[226] O'Donoghue and Rabin (1999a) write, "when presented a choice between doing seven hours of an unpleasant activity on April 1 versus eight hours on April 15, if asked on February 1 virtually everyone would prefer the seven hours on April 1. But come April 1, given the same choice, most of us are apt to put off the work until April 15. We call such tendencies *present-biased preferences*: When considering trade-offs between two future moments, present-biased preferences give stronger relative weight to the earlier moment as it gets closer" (O'Donoghue and Rabin, 1999a: 103, original emphasis).[227]

Closely related to the present bias is Laibson's (1994, 1997) work on hyperbolic discounting, a term used to describe the mathematical form discounting takes.[228] However, as O'Donoghue and Rabin (2015) note, the hyperbolic form can distract from the behavioural implications of the present bias by seeking to model relative discounting rather than emphasising the preference for the present. As O'Donoghue and Rabin (2015) write, "*Present Bias is About Now*" (O'Donoghue and Rabin, 2015: 274, original emphasis).

O'Donoghue and Rabin (2015) and Prelec (2004) have attributed the present bias to impatience, a very reasonable explanation given the characteristic feature of the bias, namely desire for most immediate consumption. O'Donoghue and Rabin (1999b) have also linked the present bias to procrastination, building from the rationale established by O'Donoghue and Rabin (1999a). On the one hand, the procrastination explanation emerges from the notion of prolonging unpleasant tasks. On the other hand, as O'Donoghue and Rabin (2015) emphasise, the present bias is about now, and so what appears as procrastination can also

[226] "We have contrived the term "present-biased preferences" as a more descriptive term for the underlying human characteristics that hyperbolic discounting represents" (O'Donoghue and Rabin, 1999a: 103, footnote 2).
[227] Also see Prelec and Loewenstein's (1991) discussion of the immediacy effect, which suggests that future discounting is more dramatic if one is asked to forgo consumption which would otherwise have been received immediately (Camerer and Loewenstein, 2004).
[228] Being hyperbolic, if one were to receive a gift, the relative discounting of that gift when received tomorrow rather than today would be significantly greater than the relative discounting when received the day after tomorrow, rather than tomorrow.

be understood as a preference for non-unpleasant tasks now.[229] Risk may also contribute to the present bias, yet many who model hyperbolic discounting and present bias assume risk-neutrality,[230] resulting in relatively little appreciation of the risk aversion as a mechanism driving the present bias (Benhabib, Bisin and Schotter, 2010; O'Donoghue and Rabin, 1999b).[231] However, as O'Donoghue and Rabin (1999b) note, without the assumption of risk-neutrality, risk aversion may have a role to play. Andreoni and Sprenger (2012) go further in their analysis of risk and time preferences, presenting results which, "cannot be explained by… hyperbolic discounting, or preferences for resolution of uncertainty, but seem consistent with a direct preference for certainty" (Andreoni and Sprenger, 2012: 3357). Thus, they conclude that certainty and an aversion to risk may contribute to the present bias.

Is a present bias nudge suitable for use in a political decision-making context? Returning to Downs' (1957) economic conception of voting, he argues that great uncertainty arises for a voter when making a political decision.[232] Such uncertainty can be attributed to a time-lapse between a person voting for a candidate, and that candidate implementing the policies which initially led to the person's vote.[233] Invoking Andreoni and Sprenger (2012) and their work on uncertainty, therefore, it can be speculated that a present bias of sorts might be manifest in political decision-making. Such speculation need not be mired in the language of (un)certainty either; it is perfectly feasible to imagine a candidate who emphasises their intentions to get

[229] Any implications here seem largely insubstantial. At best, one might engage in a semantic debate about the use of the word procrastination. Yet, insofar as a preference for non-unpleasantness now is the same as putting off unpleasant tasks until later (i.e. procrastination), very little changes. The impetus to demonstrate this flipside is wholly so that the discussion of procrastination might be couched in the same language and perspective as O'Donoghue and Rabin (2015) adopt when discussing the present bias.

[230] "Most of the experimental literature assumes risk neutrality" (Benhabib, Bisin and Schotter, 2010: 218).

[231] Risk-neutrality assumes that there is no risk associated with delaying consumption. For instance, O'Donoghue and Rabin (1999b) write, "Temporal incentives can impose risk on the agent… Because we wish to focus solely on the procrastination issue, we will assume that the agent is risk-neutral." Yet, as they acknowledge, risks do arise from delay, and so a preference for the present may also feasibly be explained by aversion to risk.

[232] Downs (1957: 13): "though we can find out something about how rational governments operate by analysing them in a "certain" world, we learn much more by facing uncertainty and the problems it creates."

[233] If, indeed, the candidate acts on their promises at all. This is just one of the many sources of uncertainty Downs (1957) identifies in political decision-making.

policies done sooner rather than later would appeal to an impatient mindset described by O'Donoghue and Rabin (2015) and Prelec (2004).

Building on the general notion of the present bias manifesting in relative appeals to the present over the future (O'Donoghue and Rabin, 1999a), the following political slogan is imagined:

"Fighting for You Today, Not Tomorrow."

The above slogan incorporates the present bias nudge in a very simple – if not subtle – way. By establishing a time framing (i.e. "today" vs. "tomorrow"), this present bias nudge nudges decision-makers towards the candidate associated with the present (i.e. "today") rather than the future (i.e. "tomorrow"). Importantly, this nudge can be incorporated into a political advertisement in *the same way* as the status quo nudge is incorporated, and as will be seen, the other nudges examined are.

5.3.3 Loss Aversion

Loss aversion is a behavioural phenomenon identified first by Kahneman and Tversky (1979). It is succinctly described as the tendency for losses to loom larger than gains, which is to say, a loss confers greater disutility upon an individual than the *utility* conferred onto a person who received a proportionate gain.

Loss aversion is closely related to risk aversion, so much so that Kahneman and Tversky (1979) directly draw upon the notion of risk aversion in their formulation of decision making under risk (i.e. Prospect Theory). There is intuitive sense to relate the concepts – if one is averse to losses, surely such aversion extends to risk-taking which presents the possibility of a loss? Yet, Kahneman and Tversky (1979) find this argument only *partially* true, specifically only in the domain of gains. They find that when there is a choice between a certain gain, and a gain with a risk component, decision-makers tend to prefer the certain gain (risk aversion).

In the domain of losses, however, Kahneman and Tversky (1979) find people become risk-loving when faced with a certain loss versus only a *possible* loss. Only aversion to losses can explain the apparent changing attitudes to risk-taking across the domains of gains and losses.

Benartzi and Thaler (1995) offer an additional mechanism which may drive loss aversion. In their work on loss aversion and stock markets, they point out that decisions with associated risks resolve over extended periods of time (say, several years). When allowed to resolve over said periods, risky returns will often yield greater returns than much safer investments (as these carry a lower premium). Yet, when evaluated in the short-term, the safer investment may appear preferable because of the lower risk. This version of loss aversion, which Benartzi and Thaler (1995) dub *myopic* loss aversion, thus suggests that individuals who are more sensitive to the time horizons of their decisions may exhibit loss aversion less than those who are not as sensitive.[234] Several studies (Fellner and Sutter, 2009; Langer and Weber, 2008; Benartzi and Thaler, 1995) confirm the concepts established by Benartzi and Thaler (1995).

Finally, loss aversion may be explained by a tendency to procrastinate. Akerlof (1991) writes, “Procrastination occurs when present costs are unduly salient in comparison with future costs, leading individuals to postpone tasks until tomorrow without foreseeing that when tomorrow comes, the required action will be delayed yet again” (Akerlof, 1991: 1). Given this definition, it may be reasonable to suppose that a person who exhibits the tendency to procrastinate does so as they are averse to the costs associated with making any decision or engaging with a particular activity. When faced with the choice of accepting a cost (embracing a loss) today or forgoing a cost (avoiding a loss) until tomorrow, it would be expected that a loss averse person would thus demonstrate procrastination-tendencies.

While a political candidate can try and appeal to voters by promoting the benefits (i.e. gains) they will bring to the voter, they can similarly emphasise the potential dangers (i.e. losses)

[234] Thaler and Benartzi (1995: 75, original emphasis): “The longer the investor intends to hold the asset, the more attractive the risky asset will appear, so long the investment is not evaluated frequently. Put another way, two factors contribute to an investor being unwilling to bear the risks associated with holding equities, loss aversion and a short evaluation period. We refer to this combination as *myopic loss aversion*.”

which their opposing candidate could bring about (Hemphill and Shapiro, 2019; Allcott and Gentzkow, 2017; Hopp and Vargo, 2017). From this perspective, it seems more than reasonable to suppose that a loss aversion nudge – which highlights the potential losses which might befall an individual if they *don't* support a candidate – could be used as part of a political advertising strategy to influence political decision-making. The following slogan is offered:

"Let's Not Go Backwards."

By emphasising the regression associated with the alternative candidate (i.e. "backwards"), this slogan is designed to emphasise the potential loss associated with that candidate, and thus nudge decision-makers towards the candidate who is not associated with loss (i.e. the candidate using the slogan). The use of the word backwards is also pertinent when considering Tverksy and Kahneman's (1992) work on reference frames. Tversky and Kahneman (1992) note that gains and losses are relative to a frame of reference, above which one must regard utility as being gained, and below which one must regard utility as being lost. The word backwards speaks illicitly to a reference frame, namely that regardless of the relative utility position of the decision-maker, backwards always refers to regression.

5.3.4 Social Norms

Social norms can generally be understood as the standards or expectations which a person holds – informed by their peers and wider environment – which in turn influence how they themselves act (Sunstein, 1996). Sunstein (1996) argues that norms represent powerful forces in the organisation of society and speaks in great detail regarding the power (and danger) of establishing new norms, or reshaping older norms, to engender a change in behaviour.

In more recent years, the power of social norm nudges – which seek to change norms or establish new norms – has been demonstrated in a variety of areas, from energy usage (Allcott, 2011; Allcott and Rogers, 2014) to charitable giving (Bartke et al., 2017) to healthy eating (Huitink et al., 2020) and voting (Gerber et al., 2014).

Sunstein (1996) argues such nudges work because they invoke shame within a decision-maker. Yet, this seems far too simple a proposition. Schultz et al. (2007), in their investigation of social norms and household energy use, describe a magnetic effect associated with social norms: "Because a social-norms marketing campaign provides specific descriptive normative information that can serve as a point of comparison for an individual's own behaviour, the descriptive norm acts as a magnet for behavior [*sic*] for individuals both above and below the average" (Schultz et al., 2007: 430).[235] Such a finding could find a hypothesised explanation in *shame*; a person may feel shame in their excessive consumption but may also feel shame in their lack of consumption.[236] But shame is just one explanation which could be offered here. Equally, it might be imagined that a person reduces their consumption as the norm makes them aware that they may continue to live adequately with less consumption, while another may increase their consumption for the norm offers them a legitimising reason to overcome their frugality. Here, shame plays no role.

Returning to Schultz et al. (2007), it seems more appropriate to say that social norms work because they, "can serve as a point of comparison for an individual's own behaviour," and speculate that those more conscious to this comparison may be more influenced by the social norm. This does not invalidate Sunstein's (1996) argument regarding shame; rather, it better situates shame as one response which might be expected *given a person's awareness of others*.

[235] Also see Bernheim and Exley's (2015) "*gravity effect*" (Bernheim and Exley, 2015: 3, original emphasis).
[236] See, for instance, Veblen's (2012 [1899]) concept of conspicuous consumption as a means of demonstrating one's class and *mere ability to consume.* Bernheim (1994) has drawn a similar comparison with social norms and Veblen's work. See footnote [3] of Bernheim (1994).

It is important to note, however, that awareness of others does not by extension mean one is inclined to *follow* or *mimic* the actions and decisions of others. An individual may be fully aware of others and still pursue the so-called road less travelled (Bernheim, 1994). Following Bernheim (1994), the utility a decision-maker derives from a decision might be said to have two components: the *intrinsic* utility associated with the decision itself, and the utility conferred onto an individual given their decision. For those whom the later represents a significant degree of their utility, they can be described as conformists. For others, the former might dominate the utility function, and thus may be wholly willing to ignore the status their decision may confer. As such, a person's tendency to conform or to be autonomous also represents mechanisms which may explain the effectiveness of social norms.

Finally, it is important to recognise that the awareness of others represents information which may contribute to a decision even if an individual cares little for the what Bernheim (1994) identifies as the status component. Moon (2010), for instance, argues people are more inclined to follow recommendations when they know relatively little about an area as the collective wisdom of others may make up for this shortfall.[237] Bernheim (1994) addresses this argument more generally, arguing conformity may arise due to the positive externalities generated by conforming.[238] As such, a person who is less inclined to evaluate information about a decision (i.e. establish expertise) may be more susceptible to following a social norm.

Voting is a typical example of an individual decision which is greatly entangled with the decisions of other individuals. Unt, Solvak and Vassil (2017) argue that voting has an inherent social nature and produces a sense of civic duty, with Gerber et al. (2014) finding that whether a person votes or not can be an important factor in shaping the attitudes directed towards that

[237] This argument broadly follows Friestad and Wright's (1994) persuasion knowledge model, which argues a person's expert knowledge (topic knowledge) within a context in which someone is trying to persuade them will moderate the effectiveness of the persuasive technique. It follows from this model that a person with less knowledge may be more susceptible to persuasion, while someone with more knowledge may be more inclined to ignore the persuasion attempts.

[238] For instance, Bernheim (1994) cites Katz and Shapiro's (1986) work on networks as an exemplar of this point. Here, Katz and Shapiro (1986) argue that the value received by each member of a network from the network grows as the network expands. As such, one may follow the crowd purely *because* there is a crowd.

individual. This social nature of voting creates opportunities to nudge individuals via social norms. For instance, given many people may be supporting a candidate, emphasising this apparent popularity could be used – within a political advertisement – to nudge a decision-maker. This is demonstrated with the following slogan:

"Trusted by Voters."

The slogan represents a social norm nudge by establishing that *other* voters place trust in the candidate. In this sense, this slogan is dissimilar to social norms which, for instance, report the monetary value of a neighbour's household consumption, establishing a reference point from which the decision-maker can evaluate their own consumption. However, such a reference point seems neither necessary nor appropriate within the political advertising medium; instead, this slogan simply indicates to decision-makers that other voters have already placed their trust in the candidate, and thus that it is acceptable for them to do the same.

5.3.5 Summary

As above, it is prudent to first select the nudges to be examined, and then identify the psychological traits associated with those nudges, before selecting the psychometric scales to be used in this research. Having considered the four nudges selected, various psychological traits have been identified, each requiring measurement via some psychometric scale.

Immediately, however, the details of the nudges discussed, psychological traits identified, and slogans proposed are summarised in Table 2:

Table 2: Nudges, Psychological Traits and Slogans

Nudge	Psychological Traits	Example Slogan
Status Quo Bias	Risk Decision-impatience Cognition	*"Let's Keep Going"*
Present Bias	Time Immediacy Procrastination Risk	*"Fighting for You Today, Not Tomorrow"*
Loss Aversion	Risk Time Procrastination	*"Let's Not Go Backwards"*
Social Norm	Conformity Cognition	*"Trusted by Voters"*

In total, seven broad psychological traits are identified here. These are described as follows:

- *Risk* – the propensity to engage in decisions where multiple outcomes are possible, and the final outcome is uncertain. Risk may be understood on a scale from risk-loving (high engagement) to risk-averse (low engagement).
- *Decision-impatience* – the propensity to make decisions quickly. The reverse may be understood as *decision-patience*. See, for instance, Kahneman (2011, 2003).[239]
- *Cognition* – the propensity to engage in tasks and decisions which require a high (low) level of thinking.
- *Time* – the propensity to evaluate decisions with several time horizons in mind. Time does not necessarily capture time-*preference* but is instead indicative of lesser or greater consideration of the temporal-nature of a given decision.
- *Immediacy* – the propensity to prefer immediate outcomes over delayed outcomes. Immediacy can be understood as a form of impatience, but regarding *outcomes* rather

[239] In this literature, Kahneman (2011, 2003) distinguishes between two 'systems' of thinking: system 1, described as, "Fast", "Effortless" and "Emotional", and system 2, described as, "Slow", "Controlled" and "Effortful" (Kahneman, 2003: 1451). In the language of 'decision-patience' and 'decision-impatience' used above, these terms should be taken as corresponding, *generall*y, to systems 2 and 1 respectively.

than the decision-making process itself. See, for instance, O'Donoghue and Rabin (2015) and Prelec (2004). The reverse, *postponement,* can be understood as a form of *patience* with preferences for delayed outcomes.[240]

- *Procrastination* – the propensity to put off or delay making a decision.[241] Procrastination differs from *patience* in that pertains to a propensity to *not* make decisions, while *patience* pertains to the speed at which a person *willingly* makes a decision.[242]
- *Conformity* – the propensity to engage in the same or similar activities to those who have come before (Bernheim, 1994). The reverse, *autonomy,* can be understood as a propensity to engage in different or dissimilar activities to those who have come before.

[240] See Kahneman (2003), who discusses the time-preference component of systems 1 and 2.

[241] Akerlof (1991: 1): "Procrastination occurs when present costs are unduly salient in comparison with future costs, leading individuals to postpone tasks until tomorrow without foreseeing that when tomorrow comes, the required action will be delayed yet again."

[242] See, for instance, Akerlof (1991) who notes the apparent similarly between patience and procrastination. Akerlof (1991) notes that procrastination models might be used to offset deleterious impatient behaviour, but ultimately argues that procrastination is better understood as occurring when a person wishes to put off incurring a cost associated with acting, rather than because a person who procrastinates is necessarily evaluating the circumstances of a decision more carefully than an impatient individual (namely, exhibiting *patience*).

Chapter 6 – Psychometric Selection and the Psychometric Map

6.1 – Introduction

Having selected the nudges to be used in this thesis, and identified the psychological traits associated with each nudge, appropriate psychometrics can now be selected so as to capture these various traits. Additionally, returning to Table 1 offers some guidance as to which psychometric scales may be appropriate to draw upon.

This chapter discusses three psychometric scales, consisting in total of seven psychometric measures, which are used in this investigation. These are the General Decision-Making Style (GDMS), the Need for Cognition (NFC) scale, and the Consideration of Future Consequences (CFC) scale. The suitability of each scale is assessed against the psychological traits identified from the literature discussed in the previous chapter. Two more scales – the Big Five personality scale, and the Abbreviated Numeracy Scale (ANS) – are also discussed, but not utilised in this thesis. Combining information on nudges, psychological traits and psychometric scales provided in this chapter and Chapter 5, this chapter concludes by providing a hypothesised psychometric map, which can be used to interrogate empirical results.

6.2 – General Decision-Making Style

The General Decision-Making Style (GDMS) is developed by Scott and Bruce (1995) as a means of broadly measuring decision-making style, which they state relates to the habits and characteristics individuals exhibit when making decisions. The initial robustness of the GDMS across several different contexts is demonstrated by Scott and Bruce (1995), but has been further demonstrated by Loo (2000), Thunholm (2004) and Spicer and Sadler-Smith (2005) in the context of general decision-making, del Campo et al. (2016) in the context of *heuristic-based* decision-making, and Peer et al. (2019) in the context of nudging.[243]

From the literature, Scott and Bruce (1995) propose an initial scale consisting of some 37 items contributing to 4 variables: *rational*, "characterized by a thorough search for and logical evaluation of alternatives"; *intuitive*, "characterized by a reliance on hunches and feelings"; *dependent*, "characterized by a search for advice and direction from others"; and *avoidant*, "characterized by attempts to avoid decision making" (Scott and Bruce, 1995: 820). Through an analysis of their initial dataset, they reduce their scale from 37 items to 25. They also conclude, based on a factor-analysis, that a 5 five-factor solution, rather than their hypothesised 4-factor solution, would be superior. Scott and Bruce subsequently propose a fifth variable, *spontaneity*, described as, "the amount of time devoted to decision making" (Scott and Bruce, 1995: 823).

Is the GDMS appropriate for an investigation into political decision-making? Likely, it is, as the GDMS is designed to capture styles of decision-making, and political choices are a type of decision-making (Downs, 1957), with variables that appear well adapted to capture several of the psychological traits identified as being associated with the selected nudges. The GDMS is disadvantaged in that it is designed to remain very broad (Scott and Bruce, 1995). However, as Egelman and Peer (2015) demonstrate, the GDMS remains a stronger predictor of decision-making than an alternative and popular scale in the literature (Hirsh, Kang and Bodenhausen, 2012; Moon, 2002), the Big Five personality index.

6.2.1 Risk-Taking

There is evidence linking variables within the GDMS with risk. For instance, Gambetti and Giusberti (2019) find that rational and avoidant styles predict risk perceptions within an investment environment. They suggest this finding could be related to desires for control, with

[243] The effectiveness of the GDMS has also been shown across languages, with Girard, Reeve and Bonaccio (2016) demonstrating its robustness in French, Alacreu-Crespo et al. (2019) in Spanish, Gambetti et al. (2008) in Italian and Fischer, Soyez and Gurtner (2015) in German.

the rational style leading people to have more understanding and thus a sense of control over the risks they are undertaking, and the avoidant style leading people to avoid decision-making and in turn attempt to distance themselves from risks. In their review of control and the GDMS, Thunholm (2004) supports the arguments of Gambetii and Buisberti (2019). In their study of medical decision-making, Bavolar and Orosova (2015) also report this relationship between risk and the avoidant style.

6.2.2 Decision-Impatience

To this author's knowledge, no study has established a link between decision-impatience and any aspect of the GDMS. As such, any proposed link must be speculative. However, propositions can be made. For instance, the intuitive variable – as characterised by Scott and Bruce (1995) – could be hypothesised to positively predict decision-impatience, as reliance on one's own intuition may grant licence to a decision-maker to spend less time searching for additional information. As such, it may also be expected that the rational variable would negatively predict decision-impatience. Finally, spontaneity, by its very characterisation, seems a likely candidate to positively predict decision-impatience.

6.2.3 Conformity

Again, little research has linked conformity and the GDMS. However, unlike above, *some* research has explored concepts such as emotional intelligence and manipulative tendencies, both of which may be indicative of susceptibility to feelings such as shame (Sunstein, 1996) or self-awareness which contribute to conformity (Bernheim, 1994).

Geisler and Allwood (2017) find the dependent style to be significant when decision-makers are faced with negative emotions, suggesting that – in accordance with Sunstein's (1996) assertions surrounding conformity and shame – those who are dependent would be expected to demonstrate more conformity. This seems intuitively sensible – a person who seeks out advice and guidance from others can be reasonably expected to care about the opinions of others. This idea is also consistent with the ideas advanced by Moon (2010) and Bernheim (1994) that individuals may conform simply for the advantages conforming confers.[244]

Geisler and Allwood (2017) also find some evidence that spontaneity may be related to, “a tendency to have an amorally manipulative social orientation” (Geisler and Allwood, 2017: 424). Extending this result into the current discussion, it might thus be speculated that spontaneity is negatively related to conformist behaviour as amorality and manipulative actions would seem to demonstrate a diminished social importance of emotions such as shame. Even ignoring this postulate, an intuitive assertion may be that a person who acts *spontaneously* does so without giving great importance to the actions of others.

6.2.4 Time

Carelli, Wiberg and Wiberg (2011) find that future prospects tend to lead decision-makers to adopt avoidant and dependent styles. They argue avoidance may be invoked because decision-makers wish to avoid a future they see as negative, while the uncertainty associated with the future may prompt some to seek support from others (dependency). By contrast, those who viewed the future *positively* tended to exhibit a rational decision style. Finally, they find those who are more focused on the present tend to demonstrate the spontaneous and intuitive decision styles. Carelli, Wiberg and Wiberg (2011) explain this may be because these people are used to relying on feelings and hunches which occur in the moment.

6.2.5 Immediacy and Procrastination

Geisler and Allwood (2017) largely corroborate the findings reported by Carelli, Wiberg and Wiberg (2011). However, they go further and relate time preferences to the ideas of procrastination (deferment of a decision) and immediacy (eagerness for outcomes). Consistent with Carelli, Wiberg and Wiberg (2011), Geisler and Allwood (2017) find that the

[244] For instance, Moon’s (2010) assertion that often non-experts will follow the road most travelled while experts will tread their own path is consistent with the notion of dependency – why would an expert seek out the guidance of others?

spontaneous decision style positively predicts immediacy in decision-making, while the avoidant style positively predicts procrastination tendencies. Again, these results are unsurprising, with the decision style associated with in-the-moment decision-making (spontaneity) being associated with immediacy, and the style associated with avoiding decisions (avoidance) being associated with procrastination.

6.2.6 Cognition

As with decision-impatience, no literature appears to examine the cognitive strain of decisions in conjunction with the GDMS. As such, the GDMS may not be appropriate for measuring this psychological trait. Nevertheless, relationships may be hypothesised. For instance, the rational decision-making style characterises a decision-maker as one who surveys their options and thinks logically, and thus this decision-maker may feel very comfortable with highly cognitive decisions. By contrast, intuitive and spontaneous decision-makers, who rely on in-the-moment, instinctive decision-making, may do so because they dislike the cognitive burden associated with more rational decision-making.

6.3 – Need for Cognition

As seen above, there is good evidence that the GDMS may be able to capture many of the psychological traits which are expected to be associated with the selected nudges. However, for two of these traits – decision-impatience and cognition – possible links can only be hypothesised. As such, the GDMS alone is likely insufficient. Fortunately, existing literature suggests two additional psychometric scales which might be used to supplement the shortcomings of the GDMS. Specifically, Peer et a. (2019) also utilise the Need for Cognition (NFC) scale, and the Consideration of Future Consequences (CFC) scale.[245] The latter will be discussed in part 6.4. At present, discussion turns to the NFC scale.

[245] Peer et al. (2019) also utilise a numeracy scale. The relevance and suitability of this scale in this thesis is discussed in more detail below.

The NFC scale was developed by Cacioppo and Petty (1982), though as these authors note, the notion of a cognitive need existed prior to their work (Cohen, Stotland and Wolfe, 1955). Cacioppo and Petty (1982) state the NFC scale is designed to assess, “the tendency for an individual to engage in and enjoy thinking” (Cacioppo and Petty, 1982: 116). Cacioppo and Petty (1982) initially propose a 45-item scale to measure cognition, though over the course of several trials they narrow the NFC scale down to 34 items. Cacioppo, Petty and Kao (1984) subsequently reduce the scale to only 18-items. Unlike the GDMS, the NFC scale is wholly contained within these 18-items and does not take the form of multiple variables (e.g. rational, intuitive etc.).

By design, the NFC scale measures cognition, with a high-NFC score indicating a person who likes engaging with tasks which require a lot of thinking, while a low-NFC score indicates a person who does not like tasks which require a lot of thinking. Given the need for some measurement of cognition following the lack of measurement from the GDMS, the NFC is therefore welcome. Following Peer et al. (2019), furthermore, the NFC has been shown to be appropriate in a study of nudges. Finally, the NFC has been found to be indicative of several other psychological traits previously discussed.

As with the GDMS, it should be considered whether the NFC scale is suitable for political decision-making. This consideration should go beyond the need for a robust measure of cognition as found given the shortcomings of the GDMS. Political decision-making often consists of a myriad of factors (Boiney and Paletz, 1991) which require cognitive effort to appropriately resolve in a political decision-making process. For instance, Gomez and Wilson (2006) have found that politically sophisticated voters – those with more knowledge of issues and capacities to consider said issues – are motivated by different interests compared to less sophisticated voters, while Coulter (2008) has found cognition to be a mediating factor in how voters assess the positive and negative frames of advertisements. Using the NFC scale, more contemporary studies (Sohlberg, 2019; O’Hara, Walter and Christopher, 2009) have found similar results. There is not only a reasonable argument as to the relevance of the NFC scale in political decision-making, therefore, but also evidence demonstrating the scale's effectiveness within investigations of political decision-making.

6.3.1 Additional Psychological Links

Some evidence exists which draws a link between the NFC scale and propensity to take risks. In their study of healthcare decision-making, Hadj-Abo et al. (2020) find that those with high-NFC tended to be more careful in their healthcare decisions, and thus may be indicative of a low risk propensity. Lin, Yen and Chuang (2006) find similar in their study of risk and consumer choice. They find that high-NFC participants were less likely to be influenced into undertaking riskier activities than those low in NFC. Following both Hadj-Abo et al. (2020) and Lin, Yen and Chuang (2006), those who had high-NFC were more likely to evaluate and understand all available information, allowing them to better understand the relative risks associated with their choices. Estelami (2020) has elaborated further on this argument, suggesting that high cognition leads to *over* consideration of information, which increases sensitivity to risk.

Srivastava and Sharma (2012) investigate the NFC scale in the context of consumer decision-making and find that NFC is a significant factor in the speed of consumer decision-making, with those with low-NFC acting faster than those with high-NFC. Furthermore, Das et al. (2003), in their investigation of the NFC scale and online consumer purchasing, find that individuals with high-NFC consider more information and take longer processing information than those with low-NFC. These results aren't surprising, given that high-NFC is characterised by *more* thinking, which one would naturally expect to take longer. As a result, the NFC scale may also be able to capture decision-impatience.

Finally, following Bernheim's (1994) suggestion that conformity may arise due to the benefits which arise from following the crowd, it may be speculated that a conformist individual may have a low-NFC as they are satisfied relying on collective wisdom or the wisdom of others, while a non-conformist individual may have a high-NFC. Lee (2014), for instance, finds that individuals with low-NFC tend to be more persuaded by the opinions of others in political decision-making, while those with high-NFC do not demonstrate this behaviour.

6.4 – Consideration of Future Consequences

While the need for cognition may somewhat capture it, decision-impatience remains an outstanding psychological feature which requires some form of measurement. Fortunately, a third psychometric test – consideration of future consequences (CFC) – appears suitable for this task.[246] The CFC scale is 12-item scale developed by Strathman et al. (1994) and is designed to capture, "the extent to which people consider distant versus immediate consequences of potential behaviours" (Strathman et al., 1994: 742).

Such a definition would seem to invite commentary on time preferences and immediacy – indeed, Strathman et al. (1994) very much consider these aspects too. However, the CFC scale measures *consideration* with respect to the future outcomes of decisions, both in the short- and long-term. As such, a low-CFC is interpreted as a demonstration of decision-impatience, as a person who makes decisions quickly is also likely to demonstrate less consideration of their decisions. Equally, a high-CFC score is typified by demonstrating great consideration over their decisions, and as such, would be expected to make decisions slower (i.e. decision-*patience*).

As previously, it must be considered whether the CFC scale is appropriate within the context of political decision-making. Given political decision-making has an inherent temporal dimension (the time between engaging with the political process and an outcome resulting from that engagement; Downs, 1957), it is reasonable integrate a measure of temporality into this thesis. For instance, Fowler and Kam (2006) argue that the inherent "delayed gratification" (Fowler and Kam, 2006: 113) associated with political decision-making means patience is an

[246] See Peer et al. (2019).

important factor in engagement with the political process. As the CFC scale attempts to capture the temporal thinking of decision makers, it seems appropriate for this thesis.

6.4.1 Additional Psychological Links

As above, the CFC scale can also be linked to time preference and immediacy. Orbell and Kyriakaki (2008) find evidence – consistent with the expectations of the CFC scale (Strathman et al., 1994) – that those who have low-CFC tend to respond to emphasis on immediate outcomes, while those with high-CFC tend to respond to outcomes which are emphasised to occur over a longer-term timespan.

Similar findings are reported by Orbell, Perugini and Rakow (2004), who find that low-CFC individuals, “were more persuaded when positive consequences were short term and negative consequences were long term. The opposite was true for high CFC individuals” (Orbell, Perugini and Rakow, 204: 388). O’Connor et al. (2009) also report this relationship in their study of healthcare decision-making.

While the Orbell, Perugini and Rakow (2004) or O’Connor et al. (2009) do not make such a link, such a result may also indicate a relationship between the CFC scale and procrastination, as low-CFC individuals may put off negative consequences (costs) while high-CFC individuals embrace them. Evidence for this link is provided by My Lien Rebetez et al. (2016) and Sirois (2004), who both find a negative relationship between CFC and procrastination. My Lien Rebetez et al. (2016) suggest greater consideration about the future may led individuals to see the fruitlessness of delay and thus prompt action in the present.

Finally, the CFC scale may be indicative of conformist behaviours and awareness of others. For instance, Griffin and O’Cass (2010) find that individuals who have high-CFC are more likely to obey the speed limit, and Ebreo and Vining (2001) and Lindsay and Stratham (1997) have found individuals with high-CFC are more inclined to participate in recycling initiatives. Griffin and O’Cass (2010) argue such behaviour is manifest because high-CFC individuals are more conscious of the potential outcomes of their behaviour, including the outcomes of non-compliance and non-conformity.

6.5 – Numeracy Scale

Despite being used by Peer et al. (2019) in their study of personalised nudging and cybersecurity, this thesis will not use a numeracy scale. There are two reasons for this.

Firstly, as Peer et al. (2019) acknowledge, there are several reasons why numerical competence may be important to cybersecurity and password design. For instance, an understanding that an 8-character password will be harder to crack than a 7-character password – while adding little additional inconvenience to the password-holder – is reliant on a simple understanding of exponentials.

However, it is not so obvious that numerical understanding is relevant in political decision-making. A *rational* theory of voting, such as that proposed by Downs (1957), may contribute to an argument that individuals must understand the expected payoffs of any political decision. However, such a theory of voting is one of several,[247] and subject to tremendous criticism. For instance, people often make political decisions which are driven by emotional, social or moral reasons (Borah, 2019; Hamlin and Jennings, 2011; Boiney and Paletz, 1991). Even when political decisions are expected to occur along economic lines, this may only occur in a small proportion of a population (Gomez and Wilson, 2006). The major justification for imagining political decision-making as a rational cost-benefit analysis, and thus using a numeracy scale, is based on an unrepresentative model of political decision-making. Beyond this view, there is little reason to suspect that competency with numeracy may guide political decision-making.

Secondly, unlike the scales already considered, numeracy scales often consist of multiple-choice questions with only a single correct answer. This is the case with the abbreviated numeracy scale (ANS) utilised by Peer et al. (2019), as well as various longer numeracy scales

[247] For a review, see Boiney and Paletz (1991).

from which the ANS descends (Weller et al., 2012). This means it may be unfair to compare a numeracy psychometric scale – which, even when averaged, is constructed from a series of right/wrong questions – with another psychometric scale – when, despite being averaged, is constructed from a genuine scale of response.

6.6 – The Big Five Personality Scale

From the review of previous literature, one scale which emerges on several occasions is the Big Five personality scale. See, for instance, Moon (2002), Hirsh, Kang and Bodenhausen (2012) and Guo et al. (2020). This scale is an extensively studied set of questions which seek to profile personalities from five personality traits: openness, conscientiousness, extraversion, agreeableness and neuroticism. Given the scale's previous usage, it may be tempting to include the Big Five personality scale in this project.

However, following Egelman and Peer (2015), the Big Five personality scale is omitted from this project. Egelman and Peer (2015) argue that scales should match the context in which they are used, while the Big Five personality scale is designed to be extremely broad. Indeed, in the two studies which quantitatively utilise the scale – Moon (2002) and Hirsh, Kang and Bodenhausen (2012) – the former only considers one trait – extraversion – because of its apparent explanatory power *compared to the other four traits*, while the latter considers all five traits, but *only because previous research has failed to do so*. The prior arguments for using this scale, then, are weak.

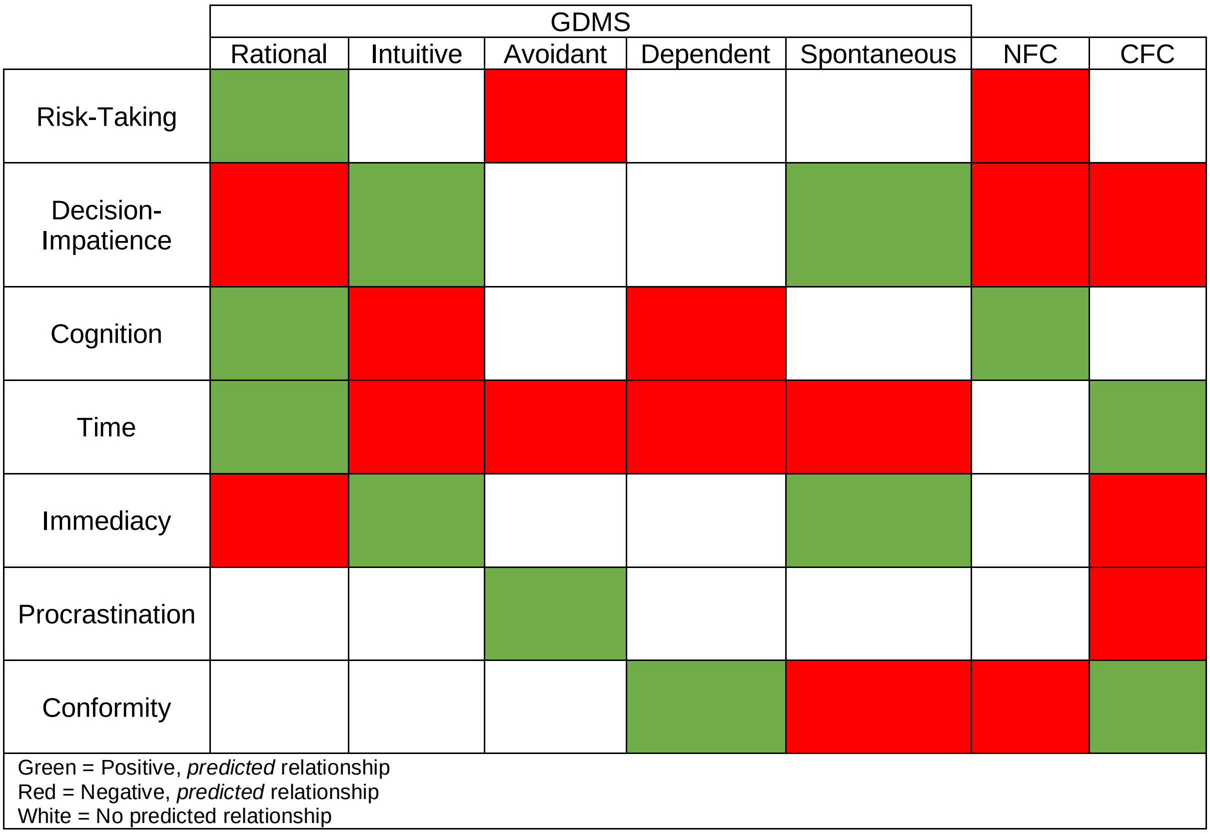
Following their argument, Egelman and Peer (2015) further demonstrate that the predictive power of the Big Five personality scale is weak compared to the GDMS, and subsequently argue that in investigations of decision-making, the GDMS represents a superior scale. For this reason, the Big Five personality scale is not used in this thesis, and the GDMS is used.

6.7 – Hypothesised Psychometric Map

Having first established the nudges which will be investigated in this thesis, and having identified the psychological traits which are expected to be associated with these nudges, three psychometric scales, consisting of a total of seven variables, are selected which are expected to capture each of the seven psychological traits identified. Returning to Figure 1, therefore, a psychometric map relating these psychometric scales – which can be measured – with nudges – which can be observed – via psychological traits – which are predicted – can now be offered.

Firstly, however, the expected relationship between psychometric scales and psychological traits is visualised in Figure 2:

Figure 2: Predicted Relationships Between Psychometric Scales and Psychological Traits



Several items must be noted with Figure 2. Firstly, green cells represent a predicted *positive* relationship between the psychometric scale and psychological trait. For instance, a person who scores highly in the avoidant style is predicted to demonstrate high procrastination. Red cells, on the other hand, represent a predicted *negative* relationship between the psychometric scale and psychological trait. For instance, a person who scores highly in the spontaneous style is predicted to demonstrate low conformity. As such, Figure 2 represents something of a correlogram, albeit one based on indicative relationships and predictions found in the literature, rather than demonstrated via data analysis.

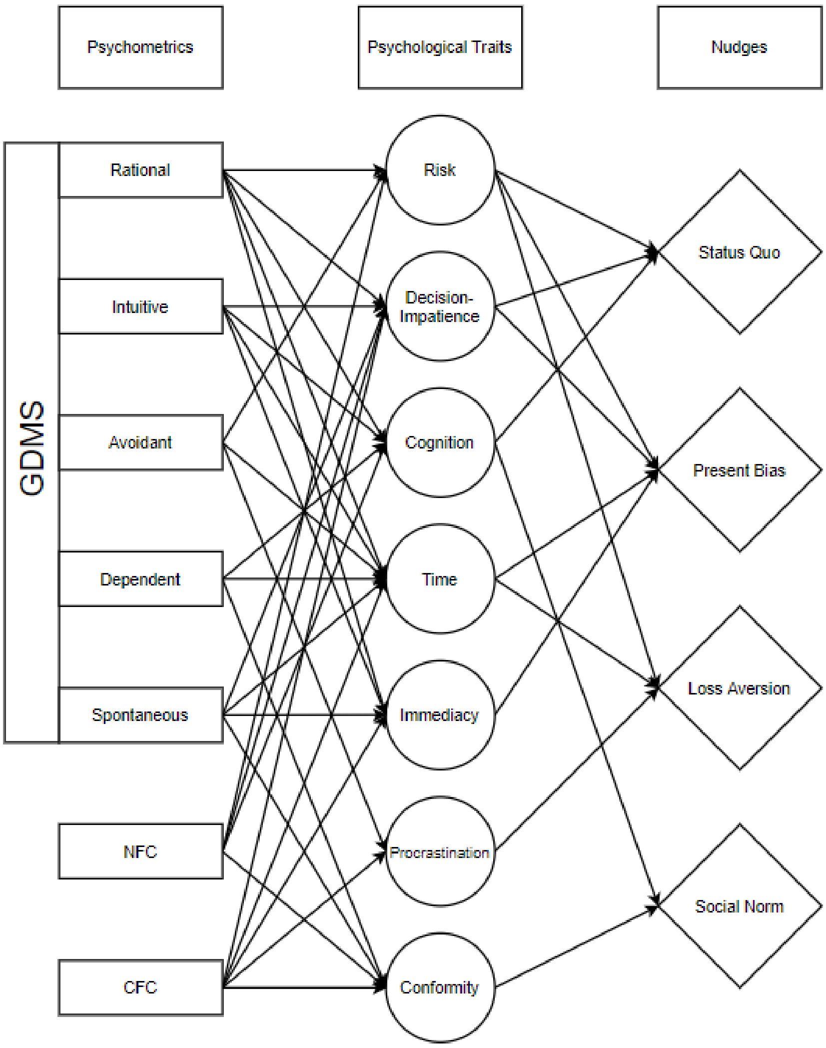
Secondly, the two-tone nature of Figure 2 may suggest that all predicted relationships are strong.[248] This may be a misrepresentation. The visualisation is merely meant to demonstrate the *direction* of the relationship (i.e. positive or negative) and should not be used to infer the *strength* of the relationship (e.g. strongly positive, weakly negative, and so on). This is a limitation of this visualisation, but an inevitable one, given Figure 2 is derived from reported findings in the literature, and not specific data which has been analysed. Figure 2, therefore, is merely a visual aid, and is not a visualisation which should be taken to capture statistical precision.

Finally, several cells within Figure 2 remain blank. This is because evidence to suggest a relationship between a psychometric scale and a psychological trait has not been identified. As above, this should not be taken as there being *no relationship*, only that the evidence at present does not support the assertion that there is a relationship. Indeed, it is more correct to interpret Figure 2 as *suggesting* – where a cell is coloured – that a significant relationship would be expected, and – where a cell is blank – that either no relationship, or an insignificant relationship, is expected. These expectations, of course, may be misguided when interrogated with data. This, to an extent, will be undertaken in proceeding chapters.

For now, attention should turn to the hypothesised psychometric map, which is shown as Figure 3:

[248] In the language of correlation, highly (un)correlated.

Figure 3: Hypothesised Psychometric Map



Such a 'map' brings together the various nudge-trait-psychometric elements discussed thus far. The construction of the map differs from Figure 1; where in Figure 1 nudge selection was the first step, in Figure 3 there is a clear progression from psychometric measurement, via psychological traits, to nudges. This primarily reflects practical reality, with the psychometric measurements expected to *moderate* the effectiveness of nudges (see Chapter 8), rather than as previously understood *when nudges were used as the basis for selecting psychometrics* (see Figure 1).

The use of psychometric maps is not a novel contribution here. Schöning, Matt and Hess (2019) and Guo et al. (2020) both illustrate their personalised nudging procedures via rudimentary psychometric maps. The purpose of doing so, both for these authors and for this thesis, is to visualise expected relationships between psychometrics and nudges which can inform a *matching procedure* analysis. For instance, using Figure 2 and Figure 3, it is hypothesised that those who score high in the rational decision style will be more risk-taking, and so will be less susceptible to the status quo nudge. From a data sample, therefore, high- and low-scorers for the rational decision style can now be identified and divided – likely in a somewhat arbitrary fashion such as above/below an average or a midpoint (Schöning, Matt and Hess, 2019; Moon, 2002) – before the effectiveness of the status quo nudge across both groups is tested.

In fact, many such hypotheses can now be formulated and considered via a matching procedure approach. This approach carries with it all the issues already identified with the matching procedure (e.g. 'personalising' after-the-fact, arbitrary value selection) but introduces a new concern in turn. Namely, by identifying and testing all possible relationships using the matching procedure, one might fall into a pick-and-choose mentality, with 'personalisation' sometimes seeming effective, and sometimes seeming ineffective. Indeed, beyond overwhelming evidence of (in)significant effects, it would seem rather difficult to confidently conclude 'personalisation' is or is not effective using a matching procedure alone.[249] This is not to say that the matching procedure cannot be of interest and indicative – on both counts, this remains to be seen – only that when a matching procedure is not selected *a priori*, but is instead arrived at via this selection process and understood using a psychometric map, the subjective and arbitrary selections which seem apparent in previous studies using the matching procedure become amplified.

Nevertheless,

Table 3 details the 44 such hypotheses which can arise from Figure 2 and Figure 3 and the wider discussion of the literature offered here. For the readers aid, each item in

[249] For evidence of such difficulties, see Chapters 11 and 13.

Table 3 is best read as:

"*If* **A** *is high, then* **B** *is expected to be* ***C****, and so the effectiveness of* **D** *is expected to be* **E**.*"*

Table 3: Matching Statements (Hypotheses)

Psychometric (**A**)	Trait (**B**)	High/Low (**C**)	Nudge (**D**)	High/Low (**E**)
Rational	Risk	High	Status Quo	Low
Rational	Decision	Low	Status Quo	Low
Rational	Cognition	High	Status Quo	Low
Rational	Risk	High	Present Bias	Low
Rational	Risk	High	Loss Aversion	Low
Rational	Cognition	High	Social Norm	Low
Rational	Time	High	Present Bias	Low
Rational	Time	High	Loss Aversion	Low
Rational	Immediacy	Low	Present Bias	Low
Intuitive	Decision	High	Status Quo	High
Intuitive	Cognition	Low	Status Quo	High
Intuitive	Cognition	Low	Social Norm	High
Intuitive	Time	Low	Present Bias	High
Intuitive	Time	Low	Loss Aversion	High
Intuitive	Immediacy	High	Present Bias	High
Avoidant	Risk	Low	Status Quo	High
Avoidant	Risk	Low	Present Bias	High
Avoidant	Risk	Low	Loss Aversion	High
Avoidant	Time	Low	Present Bias	High
Avoidant	Time	Low	Loss Aversion	High
Avoidant	Procrastination	High	Present Bias	Low
Avoidant	Procrastination	High	Loss Aversion	High
Dependent	Cognition	Low	Status Quo	High
Dependent	Cognition	Low	Social Norm	High
Dependent	Conformity	High	Social Norm	High
Spontaneous	Decision	High	Status Quo	High
Spontaneous	Time	Low	Present Bias	High
Spontaneous	Time	Low	Loss Aversion	High
Spontaneous	Immediacy	High	Present Bias	High
Spontaneous	Conformity	Low	Social Norm	Low
NFC	Risk	Low	Status Quo	High
NFC	Risk	Low	Present Bias	High
NFC	Risk	Low	Loss Aversion	High
NFC	Decision	Low	Status Quo	Low
NFC	Cognition	High	Status Quo	Low
NFC	Cognition	High	Social Norm	Low
NFC	Conformity	Low	Social Norm	Low
CFC	Decision	Low	Status Quo	Low
CFC	Time	High	Present Bias	Low
CFC	Time	High	Loss Aversion	Low
CFC	Immediacy	Low	Present Bias	Low
CFC	Procrastination	Low	Present Bias	High
CFC	Procrastination	Low	Loss Aversion	Low
CFC	Conformity	High	Social Norm	High

It is a testament that, when the results found in the literature are arranged this way, there is a remarkable degree of consistency in expectation. Two contradictions do emerge. Firstly, a high NFC would seem to produce a high effectiveness with the status quo nudge when acting via risk, but a low effectiveness when acting via cognition or conformity. Secondly, the CFC would seem to produce a high effectiveness with the present bias nudge when acting via procrastination, but a low effectiveness when acting via immediacy or time preference.

Yet, these apparent contradictions demonstrate a reasonable expectation regarding the hypotheses shown in Table 3 and the hypothesised psychometric map presented as Figure 3, namely that these predicted relationships arise from a wide literature, and such relationships are likely to be much sparser when investigated with data. This, it is expected, will allow a much more meaningful picture to emerge, with apparent contradictions resolved.

The question of data collection will be addressed shortly. Immediately, however, discussion returns to the format of the political advertisements.

Chapter 7 – Constructing Political Advertisements and Experimental Implications

7.1 – Introduction

This chapter develops preliminary political advertisements to be used in the first pilot study. To begin, the concept of *dynamic* choice architecture is developed. This concept bridges the behavioural concept of choice architecture with the visual design considerations of advertisements to offer a schema for producing the political advertisements to be used in this thesis. These advertisements are then offered, followed by a discussion of various design choices.

Attention then turns to the experimental implications of these political advertisements. Broadly, there are two key experimental implications. The first concerns the sample population, as all design is often grounded in a social or cultural aesthetic. In this instance, an American style is adopted, and thus an American sample is selected. The second concerns the role of differing aesthetics. Where advertisements differ in ways which are not of experimental interest, the effect of any difference between a control and a treatment is obscured. This is the case in this experiment. The solution proposed here is to adopt a randomised controlled trial (RCT) design. The use of RCTs within the behavioural science literature is offered in this chapter, as well as a consideration of a (potential) alternative solution proposed by Hirsh, Kang and Bodenhausen (2012).

7.2 – Dynamic Choice Architecture

Political decision-making can be a very complex process (Borah, 2019; Stone, 2012; Hamlin and Jennings, 2011). The advertisements which support and encourage political decision-making, therefore, come to reflect the complexity of an electorate through their designs (Kehle and Naimi, 2019). From a behavioural perspective, *design elements, components and aesthetics* represent choice architecture (Benartzi, 2017), especially in online and digital advertising (Yeung, 2017; Weinmann, Schneider and vom Brocke, 2016). It is this language which is utilised here when considering the design of political advertisements.

Classically, choice architecture is characterised as the various framing conditions under which a proposition is given, and a decision is made (Thaler, Sunstein and Balz, 2014). Nudges exist, in relation to choice architecture, as small, often singular changes which influence decision-making (Thaler and Sunstein, 2008). With the emergence of digital landscapes, a growing body of work on digital nudging and digital choice architecture (Benartzi and Bhargava, 2020; Benartzi, 2017; Yeung, 2017; Weinmann, Schnieder and vom Brocke, 2016) frequently posits that choice architecture is much more dynamic (Yeung, 2017) than the classical account would suggest, with different choice environments being built or designed (Benartzi and Bhargava, 2020; Thaler and Tucker, 2013) depending on different individual preferences.[250]

This offers a rather useful, if quite reductionist, view of design. An initial proposition is put forth: an advertisement consists as the sum of its constituent parts. Depending on the advertisement, these parts may include a product name, a logo, a background image, a colour scheme and so on.[251] This list is illustrative, not exhaustive. In other words, an advertisement can be thought of as:

Advertisement = Product + Logo + Colour Scheme…

In a more mathematical language, an advertisement A is assembled from components within a set $\{X_1, X_2, \dots X_i\}$ such that $A = X_1 + X_2 + \cdots + X_i$. Each component X represents an element of choice architecture which, *in theory*, can be changed so as to nudge a decision-maker.

[250] Here, the intersection between digital nudging and personalisation is rather obvious.
[251] Furthermore, depending on the medium, this list can be expanded. For instance, a video advertisement may include a soundtrack as a constituent part.

In fact, each component might be thought of existing in a state of being either 'switched off' – whereby that component is not changed, and no nudge is intentionally embedded – or 'switched on' – whereby that component is changed and a nudge is intentionally embedded.[252] For instance, recall Schöning, Matt and Hess (2019) who nudge decision-makers within a digital space by changing whether images are used, or whether bullet points are used.

It is through this lens that Yeung's (2017) discussion of "dynamically reconfiguring the user's [decision-maker's] informational choice context in ways intentionally designed to influence her decisions" (Yeung, 2017: 122) can be understood; dynamism may refer to the 'switching' on or off of different choice architectural components depending the circumstances of the decision-maker and the goals of the choice architect.[253] Furthermore, for a thesis concerned with *personalised nudging*, with an experiment conducted online, this approach to choice architecture and design seems the most appropriate. As Benartzi (2017) notes in their authoritative work on the choice architecture of online spaces: "The logical endpoint [of digital choice architecture] is an internet in which the best Web sites and apps customize their appearance based on our demographic background. Are we an educated senior citizen from Poland? Then take away all the colors and give us plenty of text and links. Are we a young Thai man? Then give us lots of bright color and imagery" (Benartzi, 2017: 50).[254]

7.3 – Political Advertisements

[252] It is important to note that Thaler and Sunstein (2008) and Sunstein (2013, 2017) argue that choice architecture cannot be avoided. It is for this reason the notion of *intentionality* is included.

[253] There is much more to be said of this notion than is appropriate to include here. For instance, one might be inclined to assume the effectiveness of an advertisement A, measured as the proportion of people who choose whatever option A is nudging them towards, would *always* be greater when two components are used to nudge, rather than one. This, of course, assumes that the effect of 'switching' one component on is the same as 'switching' any other on – an assumption which is likely unsupported. An investigation of this notion, however, is beyond the scope of this thesis.

[254] Benartzi (2017), here, is drawing on the work of Reinecke and Gajos (2014), who investigate the aesthetic design choices of websites and corresponding user engagement. Benartzi (2017) is clear, later on, that the simple colour scheme and complex nexus developed by Reinecke and Gajos (2014) can be more complex: "There are, of course, countless variables that go into human attention, from font size to the color [*sic*] palette of a Web site. (A lot of A/B testing is about fine tuning these details)" Benartzi (2017: 68). This notion, once more, speaks to the concept of dynamic choice architecture proposed here.

This dynamic, *choice architectural* approach to advertisement design is utilised here. Following the proposition, the political campaign slogan is taken to be one component of the whole political advertisement, and the *only* component which is 'switched' on, which is to say, the only element of the advertisement which is used to nudge participants. All other components of the advertisement, including the political candidates, background imagery, accompanying graphics and colour scheme remain 'switched' off, which is to say, no nudge is intentionally embedded within these components.[255]

The proposed political advertisements to be used in this project as shown in Table 4:

[255] Kehle and Naimi (2019) demonstrate the ability to dynamically reconfigure advertisements to appeal to different kinds of voters (e.g. right or left, young or old) by varying several components within their advertisements. See Figure 4, below. In a relatively recent study, Praino and Stockemer (2018) demonstrate how changing the facial features of political candidates within advertisements can influence voters, extending the notion of dynamic reconfiguration even to the *physical image* of the candidate shown.

Table 4: Political Advertisements

Candidate	Control	Status Quo Nudge	Present Bias Nudge	Loss Aversion Nudge	Social Norm Nudge
A					
B					
Slogan	"Working for You"	"Let's Keep Going"	"Fighting for You Today, not Tomorrow"	"Let's not go Backwards"	"Trusted by Voters"

Several features of these advertisements warrant further elaboration.

1. *The Control Advertisement* – Firstly, a control advertisement is offered. The relevance of this advertisement will be discussed in greater detail below. Immediately, this discussion concerns the selection of the slogan “Working for You.” In accordance with the dynamic, choice architectural approach described above, this slogan is designed to represent a slogan which has no nudge embedded within it. A common argument within nudge theory, of course, is that no choice architecture is neutral (Sunstein, 2013, 2017a), and as such, the desire for this slogan to carry no nudging influence is an ambition more than a proposition which can be verified. The alternative would be to have no slogan at all, but this raises a comparability issue. Namely, without a slogan, participants may favour the nudge slogan not because of the nudge, but simply because the presence of *any* slogan provides them with more information about their decision. This would, then, obscure the ability to discern the effectiveness of the nudge.
2. *Background imagery* – The background imagery used with candidates A and B is designed to appeal to neutral spaces. Both images are outdoors, and contain buildings, some greenery and some water, with none of these aspects in such a concentration so as to intentionally imply something about the candidates, e.g. the use of a lot of greenery may imply something to do with an environmental agenda. Finally, the backgrounds are blurred in an effort to draw attention away from the background images and towards the foreground which contains the nudge.
3. *Candidate portraits* – Two stock images of middle-aged, white men in suits are used to visualise the candidates in this project. These images are selected based on their broad demographic similarities, as well as their body language. For instance, both candidates have their arms together, are orientated towards the viewer in the same way, and have a generally pleasant facial expression. The choice of white, male candidates was made to reflect the present reality of US politics: 78% of the 116th US Congress is white, and 76% is male (Bialik, 2019; Desilver, 2018).

4. *Candidate names* – Generic names are chosen so as to not be unusual given the demographics of the selected candidates. The names ‘Steve Jones’ and ‘Jack Smith’ are chosen as the names ‘Steve’, ‘Jack’, ‘Jones’ and ‘Smith’ are quite typical examples of very common names. In addition to these names, a banner with the word ‘Vote’ is added to ensure participants recognise these advertisements as *political* advertisements.
5. *Graphics* – a small graphic consisting of three arrows is included to make the advertisement feel more genuine. The arrows go from left to right in the same direction that English and other western languages are typically read, and thus are meant to indicate *progression* rather than *regression*. The colour scheme used with this graphic is red, white and blue, a very typical colour scheme used in American political advertisements (Kehle and Naimi, 2019). These colours are also the colours of the US flag and capture to two main colours of the two main US political parties. In addition to the arrow graphic, a transparent curve graphic is used, again to make the advertisement feel more genuine.
6. *Font* – candidate names are displayed in bold block text following design choices used by Kehle and Naimi (2019). However, the campaign slogans use a ‘handwriting’ font so as to encourage participants to associate the words of the slogans with the candidates shown in the advertisements.

Despite these design choices and the various reasons which accompany them, two outstanding points remain to be addressed. Firstly, given the designs which have been adopted, what target population will the sample be drawn from? Secondly, given the advertisements for Candidate A and Candidate B differ *by more than just the nudge*, how will the conflating effect of these differences – so-called *aesthetic* effects – be isolated and removed?

7.4 – A Note on Sampling

This discussion of sampling centrally concerns the target population from which a data sample will be drawn, rather than the question of sample *sizes*, which is considered in Chapter 9. The inspiration for the political advertisements shown in Table 4 comes from Kehle and Naimi (2019) and their work on big data and customised political advertisements:

Figure 4: Example Advertisements from Kehle and Naimi (2019)

Kehle and Naimi (2019) design their advertisements for an American audience. It makes sense, having taken influence from these advertisements, for this experiment to therefore sample an American audience. There are several additional advantages to selecting an American audience. Firstly, Americans have experience of the electoral process with an established history of democracy. Someone who has limited experience participating or even seeing political campaign material may evaluate the advertisement very differently to someone who has more familiarity with said material. Secondly, data are collected in the same year as the 2020 US presidential election, during which some state elections also are planned. Therefore, an American audience can be expected to *already* be evaluating election material. Thirdly, most users on the MTurk platform report to be from America, meaning the ability to recruit participants for this experiment will not be hindered by the selection of an American audience.

7.5 – Aesthetic Differences and Randomisation

One important item of consideration is the aesthetic differences in advertisements for Candidates A and B. The presence of these differences prompts the question of how such differences can be controlled for so as to not inflate or diminish the observed effect of the nudge. A more immediate question, however, is why two advertisements – which produce the potential for aesthetic effects – are even used to begin with?

7.5.1 Why Use Two Advertisements?

To be sure, one resolution to the problem of aesthetic effects would be to only use a single advertisement, or to use multiple advertisements but only use a single candidate. The use of two advertisements supporting two candidates, however, is largely for realism. A basic tenant of the democratic process is the notion of a *choice* between two or more candidates. Where only a single candidate is running for an elected position, the act of election itself becomes moot.[256] In an effort to encourage participants to treat their choices as choices made within a genuine democratic setting, two different candidates are provided.

The act of providing two different candidates itself creates the opportunity for aesthetic effects to occur. While efforts have been made to control for demographics, body language and expression, it is a wholly plausible possibility that participants find favour with one of these candidates based on appearance alone (Praino and Stockemer, 2018; Praino, 2018; Lawson

[256] Rebonato (2014), writing on the falsity of choice, argues that when people see no meaningfulness in the exercise of choice, the claim of freedom of choice itself becomes "nominal" (Rebonato, 2014: 360).

et al., 2010; Little et al., 2007). Furthermore, the inclusion of different candidate pictures may not be enough to make participants feel as though this is a genuine choice. One can imagine two advertisements, absolutely identical apart from the slogan and the candidate picture, engendering one to believe these advertisements are – despite their differences – more or less the same. Indeed, political advertisements between different candidates often convey similar ideas in different ways. For instance, Figure 4 shows the same candidate discussing the same policy, but advertisements targeted at democrats, republicans and independents vary quite significantly in terms of aesthetics. In other words, to encourage a feeling of genuine choice between candidates, it seems prudent to introduce differences in the advertisements beyond merely changing the candidate and the slogan.

7.5.2 Resolving Aesthetic Effects

Now that the question of why aesthetic differences are introduced has been addressed, attention turns to the more demanding question of how the resulting aesthetic *effects* can be resolved. As a term, aesthetic effects is taken here to mean any influencing effect a difference between the advertisements *other than the nudge* has. As such, the difference between a treatment advertisement and a control advertisement without controlling for aesthetic effects can be described as:

Equation 1

$$Y_1 - Y_0 = (x_1 - x_0) + (S_1 - S_0) \tag{1}$$

where $Y_1 - Y_0$ is the difference in observed effectiveness scores, $(x_1 - x_0)$ is the implied effect attributable to the nudge, and $(S_1 - S_0)$ is the implied effect attributable to presence of other differences, such as aesthetic differences.[257] Two routes to resolving the presence of aesthetic

[257] In the randomised controlled trial literature, from which this equation draws, $(S_1 - S_0)$ captures a large range of differences, including the demographic differences of respondents. The equation shown above is simplified and should be taken as representing a single decision-maker. As such, there is no demographic

effects emerges from this model. Firstly, an adjustment of some kind can be made to the observed effect so as to remove the inflating or diminishing effect of aesthetics. Secondly, an adjustment of some kind can be made to the experimental design itself so as to reduce $(S_1 - S_0)$ to zero.[258]

7.5.2.1 Hirsh, Kang and Bodenhausen's (2012) Approach

In the personalisation literature, an example of the first strategy can be found. Hirsh, Kang and Bodenhausen (2012) create an advertisement for each of the Big Five personality types, varying these advertisements only by their appeal to a given personality type. Each participant is then asked to rate the effectiveness of every advertisement. Hirsh, Kang and Bodenhausen (2012) initially find that the effectiveness ratings of the advertisements are very similar and suggest this is because the aesthetics of the advertisements were the same. To draw out the effect due to the treatment,[259] they regress the effectiveness of four of the advertisements onto the fifth, arguing that the outstanding variance (the residual) captures the effectiveness of the personality type attributable to the fifth advertisement.[260]

This strategy of resolving aesthetic effects is appealing as it is relatively simple. However, it is also problematic for both specific and very general methodological reasons. Specifically, the experiment design which would be necessary to implement Hirsh, Kang and Bodenhausen's (2012) adjustment would reduce the experiment to merely asking participants to rate the same advertisement which is subtly changed each time. This is not congruent with a typical political decision-making experience, and in fact would reduce this experiment to just asking respondents to rate slogans. At that point, the need for aesthetics of any kind evaporates entirely. In short, Hirsh, Kang and Bodenhausen's (2012) procedure may not be suitable for

---

variance and the term $(S_1 - S_0)$ captures only aesthetic effects. Where Equation 1 describes the aggregate decisions of multiple people, this narrow understanding is not correct.

[258] At the very least, to afford a responsible argument that $(S_1 - S_0)$ is zero and thus assume as much.

[259] This is to say, the authors argue $Y_1 - Y_0$ may not be significantly different from zero, but $(x_1 - x_0)$ may be significantly different from zero.

[260] This process was done for each of the five advertisements.

this *specific* experiment. Generally, Hirsh, Kang and Bodenhausen's (2012) procedure may be problematic in that it requires the experimenter to adjust the *data after they have been collected*, rather than *the experiment before data are collected.* For instance, it may be worthwhile to consider whether Hirsh, Kang and Bodenhausen (2012) would have undertaken the process of regression and residual extraction had the absolute effectiveness scores initially collected produced results congruent with their expectations.[261] While not to disparage Hirsh, Kang and Bodenhausen (2012), insofar as one anticipates the presence of an aesthetic effect which could be obscuring an effect of primary interest, adjusting for this obscuring effect after-the-fact grants experimenters licence to engage in a variety of activities under the guise of isolating and removing an effect which – in the process – may render any investigation into the effect of primary interest questionable.

Given the specific issue the experiment proposed here would have should the only direct example of treating for aesthetic effects be implemented, and given the general concerns which can be raised by adjusting for aesthetic effects after-the-fact, it would seem most advantageous to engage with an adjustment *to the experimental* design prior to any data collection.

So far, the consequences of aesthetic effects have been described as having either an inflating or diminishing effect on the treatment effect. Partly, this is because the role of aesthetic effects cannot be known *a priori*. More importantly, where aesthetic effects are significant,[262] it is because the experimental design produces circumstances where aesthetic effects *always* inflate or *always* diminish the treatment effect. If the experiment design can be adjusted such that half the observations gathered experience an inflationary effect, while half experience a diminishing effect, the *net* aesthetic effect across the sample can be assumed to be zero. This is the principle behind a randomised controlled trial (Deaton and Cartwright, 2017).

[261] This is not to say that they did not, nor is it to attack authors who are unable to defend their approach at present. Rather, such a question is merely designed to demonstrate how attempts to treat aesthetic effects after data have been collected can produce such critiques of process.

[262] Which is to say, where without any adjustment, $(S_1 - S_0)$ is expected to be non-zero.

7.5.2.2 Randomised Controlled Trials

Randomised controlled trials (RCT) have become a popular method in behavioural science (Madrian, 2014). Ho and Imai (2008) investigate how randomising the position of political candidates on a voting ballot influences voter outcomes, compared with a control ballot of alphabetised candidates. They find that, for non-partisan voters, regardless of who was first on the voting ballot, that candidate received more votes than their rivals. Both Redfern et al. (2016) and Just and Price (2013) use an RCT experiment design to investigate the use of small incentives to encourage healthier eating in children, while Fryer (2011) uses an RCT to investigate the use of different educational-support interventions amongst children living within different US cities. Most famously, the UK's Behavioural Insights Team (BIT) have made RCT experiments the cornerstone of their research approach (Haynes et al., 2013). For instance, BIT projects have used RCT designs in experiments looking at messaging to reduced over-prescription of anti-biotics within the UK (Hallsworth et al., 2016), reminder nudges to improve hospital appointment attendance (Hallsworth et al., 2015) and engagement with UK pension provisions (Glazebrook, Larkin and Costa, 2017).

By randomising which advertisement is the treatment advertisement (i.e. contains the nudge) and which advertisement is the control, it can be assumed that over a large enough sample the effect of aesthetic differences will be zero.[263] Returning to above, randomisation rather than *allocation* is used as it often cannot be known *a priori* under what circumstances aesthetic differences are expected to have an inflationary effect versus a diminishing effect (Banerjee, Chassang and Snowberg, 2016). Furthermore, randomisation eliminates a source of experimenter bias (Deaton and Cartwright, 2017). As experimenters may indeed have *a priori*

[263] A large sample is required because randomisation over a small sample may not conform to an expected distribution. For instance, randomisation into one of two groups is expected to produce a 50/50 split amongst participants. However, over a small sample (say, N = 10), randomisation may produce a 60/40 or 70/30 quite frequently. In such a situation, the assumption that $(S_1 - S_0) = 0$ is much weaker. The Australian Behavioural Economics Team (BETA) note that in a review of published RCT's, some 55% of studies used samples of N = 1,000-10,000, followed by some 29% that used samples N = 100-1,000. Few (4%) used less than 100 observations, and none (unsurprisingly), used less than 10 (BETA, 2016).

beliefs about the effectiveness the treatment and the role of aesthetic effects, allocation may lead the experimenter, consciously or not, to construct control and treatment groups which reinforce these expectations, and in turn obscure the true effect size (Deaton and Cartwright, 2017; Banerjee, Chassang and Snowberg, 2016).

RCTs are not beyond criticism. For instance, Deaton and Cartwright (2017) argue that because RCTs are often considered the 'gold standard' approach to experimental design, some researchers may use RCT designs despite alternative approaches being better for their specific research project. Haynes et al. (2013) have also criticised RCTs from a time and cost perspective, noting that RCTs typically require larger sample sizes compared to other (though potentially less methodologically robust) approaches. The time and the cost of these larger samples can sometimes be inhibitive.

Nevertheless, an RCT approach appears advantageous here. Indeed, following Deaton and Cartwright's (2017) criticism of RCTs, the most appropriate alternative methods of tackling aesthetic effects (namely Hirsh, Kang and Bodenhausen's (2012) approach and the use of a single advertisement) have been considered and, per the arguments provided above, determined problematic for several reasons. As such, in using an RCT and varying which advertisement contains the nudge and which is the control, the theoretical aesthetic effect arising from aesthetic differences between advertisements can be assumed to be zero, and thus the difference in recorded effectiveness can be determined to be the average treatment effect without any adjustments to the data post-collection.

The outstanding questions to which attention now turns are the experiment design, method of data collection and analysis.

Chapter 8 – Data Collection and Matching and Moderation Analysis

8.1 – Introduction

This chapter details data collection methods, as well as specific methods for statistical analysis of the data which are collected. The first part of this chapter establishes the idea of a primer group or stage. This stage of experimentation is used to collect data which a) form the basis of prediction in the second stage, and b) can be used as a contrast with data collected in the second stage. Primarily, this part of the chapter is concerned with point a), and offers two methods of analysis, the advantages and disadvantages of which are discussed at length: matching analysis, and moderation analysis.

The second part of this chapter considers an issue which arises when making predictions about personalised nudges; when multiple means of personalisation are predicted to be effective, the optimal method must be determined. A novel method of 'ranking coefficients' is offered here, and this method is contrasted with the simulation method offered by Peer et al. (2019).

The third part of this chapter considers the second stage of data collection, where predictions based on the data gathered in the primer stage are implemented, and data collected. This part details practical information about experimental implementation of predictions, as well as outlining the delivery only (DO), choice only (CO) and choice and delivery (CD) groups which are to be analysed as part of the investigation into hypotheses 1 and 2. This chapter concludes by addressing point b), which concerns how these various groups will be compared and the statistical procedure to be undertaken in this thesis.

As shown in Table 1, many previous studies utilise incentivised-survey experiments. Following Peer et al. (2019) and Hirsh, Kang and Bodenhausen (2012), distribution of the survey-experiment is done using Amazon's Mechanical Turk (MTurk) micro-tasking platform, a service commonly used for behavioural research (Peer et al., 2017; Chan and Holosko, 2015;

Mason and Suri, 2011).[264] Schöning, Matt and Hess (2019) also recruit respondents online, but using Facebook's survey service. This platform, however, is undesirable as the social media nature of Facebook may encourage users to share survey links, producing a sampling bias. The survey-experiment itself is constructed using Qualtrics, a survey creation and hosting service.[265]

As with the use of the MTurk platform, incentivised research is common practice within the personalised nudging literature (Peer et al., 2019; Hirsh, Kang and Bodenhausen, 2012; Moon, 2002) and the wider behavioural science literature (Chan and Holosko, 2015). Much research has been conducted into the effects of incentivised data collection. Singer and Ye (2012) find that incentivisation increases response rate; that increased incentives increase response rates but with ever reduced increments; and that no significant difference in data quality is found between incentivised and non-incentivised data.[266] Furthermore, Singer and Cooper (2008) find that incentivised participants exhibit similar willingness to take risks in experiments as non-incentivised participants. The expected effects of incentivisation, then, would seem to be an increase in the rate of response but not change in the quality of response. Given the RCT design and the method of analysis (moderated regression) discussed below, a reasonably large sample size, without degrading the quality of data, is desirable, and so incentivisation appears appropriate.

Following Peer et al. (2019), data collection occurs in two stages. The first stage, referred to hereinafter as the primer group or primer stage, collects various psychometric data before impersonally nudging respondents. The second stage, henceforth referred to as the treatment group or treatment stage, collects various psychometric data before showing respondents personalised nudges. This is done by exploiting relationships identified between nudges and

[264] Peer et al. (2017) investigate alternative micro-tasking platforms, arguing that the maturity of MTurk means users may now be too experienced to provide reliable behavioural data. However, in their research, they find no alternative service performs better than MTurk in terms of response rate or data reliability.
[265] Qualtrics is also a commonly used service, but this is not the primary reason for its selection. Primarily, Qualtrics is used because of its functionality. See below for more.
[266] Also see Singer (2002), Singer and Kula (2002) and Cantor, O'Hare and O'Connor (2008).

psychometrics in the primer group. This approach is adopted because, as outlined above, it is questionable whether a matching approach with its deductive approach can be described as *actual personalisation*. As such, an inductive approach must be taken, and thus, an initial dataset must be assembled from which the means of personalising can be identified, before a second dataset is constructed capturing the results of *actually* personalising.

8.2 – The Primer Group

The structure of the primer group survey is as follows:

1. Participants are first asked a series of demographic questions. These include their age, their gender, their education level and their political identity. Education level and political identity questions consist of five categories for education (*"what is the highest level of education you have completed"*: none; high school; bachelor's degree; master's degree; PhD) and political identity (*"what is your political affiliation"*: left-wing; left-leaning; centre; right-leaning; right-wing). These questions are used to check that, across a range of potentially relevant criteria, that various sample groups are comparable. For anonymity reasons, no names or geographical data are collected, beyond geographical specifications available through MTurk.
2. Participants then complete the three psychometric scales discussed Chapter 6. Firstly, the GDMS is completed in the order avoidant, dependent, intuitive, rational and spontaneous. Next, the standard ordering of the 18-item NFC scale is given. Finally, the standard ordering of the 12-item CFC scale is given. As above, there is some indication that previous studies have purposely selected the order of item delivery, possibly to avoid framing effects (Peer et al., 2019; Moon, 2002). However, little evidence is offered to support a compelling narrative regarding the mitigation of framing effects via ordering or indeed the presence of framing effects at all. The order used here is designed to encourage participant completion of the survey-experiment,

with the 25-item total GDMS followed by the 18-item NFC followed by the 12-CFC. In other words, the scales get shorter as participants progress.

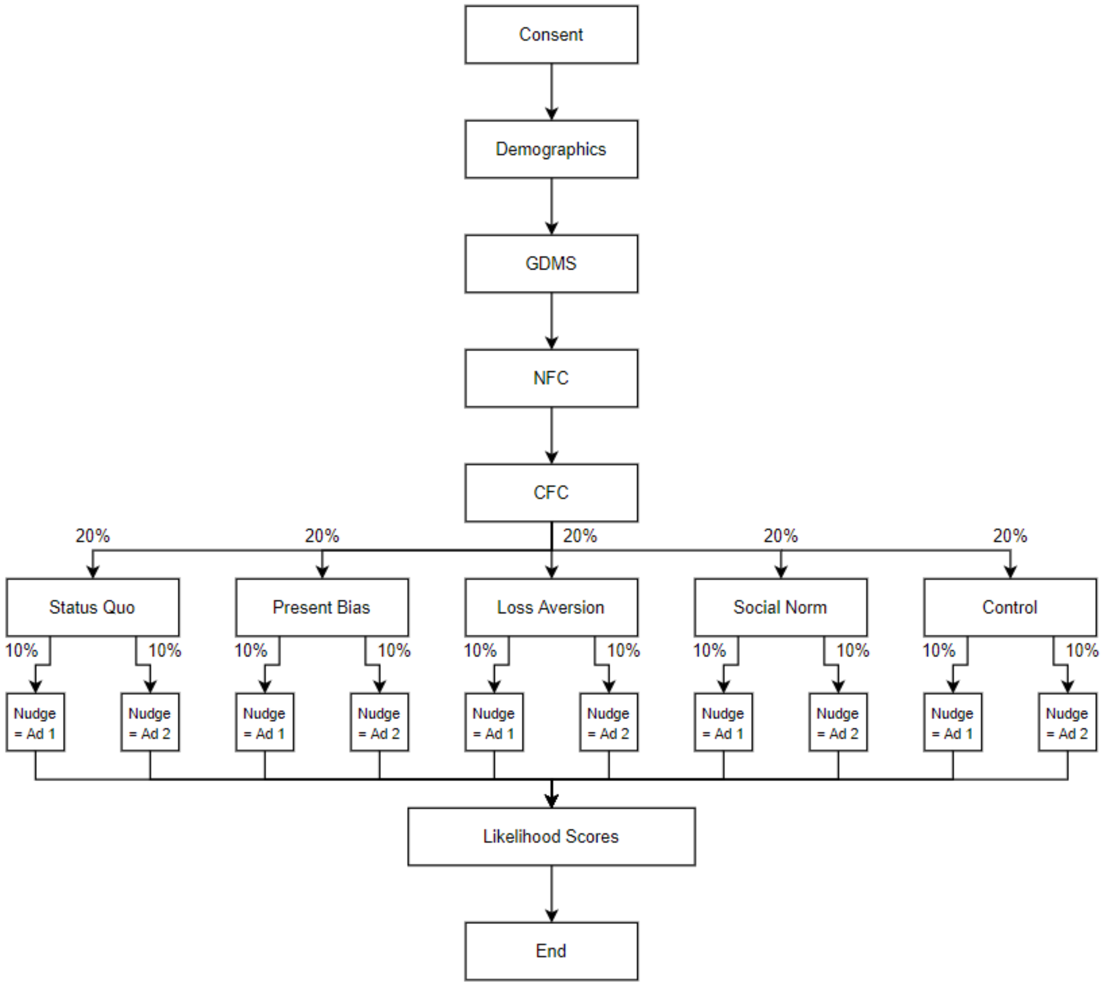
3. Upon completion of the psychometric scales, participants are randomly assigned into one of five groups: a status quo nudge group; a present bias nudge group; a loss aversion nudge group; a social norm nudge group; and a control group. Each of these groups is further divided into two subgroups: a subgroup where Candidate A is used in the nudge-advertisement and Candidate B in the control; and a subgroup where Candidate B is used in the nudge-advertisement and Candidate A in the control. In all subgroups, the advertisement containing the nudge is shown above the control advertisement. Distribution of participants into these subgroups is set evenly, so each subgroup will – on average – contain 10% of respondents (1 in 10), and each group – on average – will contain 20% of respondents (1 in 5). While the distribution of respondents is determined by the author, the author remains blind to the allocation of participants. As such, this is a blind randomised control trial.
4. Once assigned to a subgroup, participants are given the following prompt:

   > "*Imagine these candidates are running in an upcoming election which you can vote in. Based on these advertisements, please indicate how likely you would be to vote for each candidate:*"

   Participants can then indicate their likelihood of voting for Candidate A – labelled for the participant as whoever is featured in the (nudge) advertisement at the top of the page – or Candidate B – labelled for the participant as whoever is featured in the (control) advertisement at the bottom of the page. Participants may answer from 0 to 100.

A visualisation of the survey flow for the primer group survey-experiment is shown in Figure 5:

Figure 5: Primer Group Survey Flow



8.2.1 Effectiveness Variable

All previous studies construct or directly measure a variable which can be said to capture the effectiveness of whatever nudge or intervention is being used. In all cases, this variable becomes the variable of central interest in that it captures the effectiveness of the nudge and, often by proxy, the effectiveness of the personalisation method utilised. A similar procedure is undertaken here.

Given the design of the survey-experiment, the construction of the effectiveness variable (so-called because it is taken to account for the effectiveness of the nudge) is rather simple. Each respondent is required to indicate their likelihood of voting for *both* candidates A and B (as displayed for the participant), which can be labelled $L(A)$ and $L(B)$ respectively. The effectiveness *of the nudge*, therefore, can be characterised as:

Equation 2

$$Effectiveness = L(A) - L(B) \tag{2}$$

In theory, this construction of the effectiveness variable will render all effectiveness scores in the control group (i.e. the group which sees two control advertisements) zero. Therefore, a nudge which positively influences participants should have an (average) effectiveness which is greater than zero, and a nudge which negatively influences participants should have an (average) effectiveness which is less than zero.[267] Note that the name 'effectiveness' is not indicative of whether the nudge is *effective* in a statistical sense, which is to say significant. Insofar as an (in)effective nudge differs from the control group, *effectiveness* is merely the name for a variable that captures this difference.

Of course, the (average) effectiveness of the control group may not be zero because, despite the RCT design, there may still be significant aesthetic effects. This is because $L(A) - L(B)$ is equal to $Y_1 - Y_0$ in Equation 1. Allowing $(x_1 - x_0) = 0$,[268] any difference between zero and $L(A) - L(B)$ must be attributable to other effects, i.e. $(S_1 - S_0) \neq 0$. However, such a conclusion can only be drawn when the variation from zero in the control group is significant; the control group will likely vary slightly (but insignificantly) from zero simply due to a finite sample size.

Regardless, the relative relations between the effectiveness score of the control group and that of the nudge groups still applies; where the nudge has a higher (average) effectiveness score than the control group, this would indicate the nudge having a positive influence, and where the nudge has a lower (average) effectiveness score, this would indicate a negative

[267] A positive influence is taken to be indicating a greater likelihood of voting for the candidate whose advertisement contains the nudge compared to the alternative candidate. A negative influence is the reverse.
[268] Because no nudges are used in the control group.

influence. However, accepting that the control group may have a non-zero effectiveness score means that it is not correct to interpret *any* positive score as an indication of positive nudging influence, nor is it correct to interpret *any* negative score as an indication of a negative nudging influence. Rather, such interpretations must be established *relative* to the effectiveness score of the control group.

8.2.2 Testing for an Aesthetic Effect

Assuming no aesthetic effects are acting on the data (either inflating or diminishing the effectiveness scores associated with the nudges), the difference in the effectiveness scores of each of the two subgroups within a given group in the sample should be zero. This holds when the group in question is the control group (which is expected to have an *overall, group* average effectiveness score of zero) or a treatment group (which are expected to have *overall, group* averages with are non-zero).

As such, a simple test for differences in the averages of these subgroups can be performed using a two-tailed t-test, with a significant difference suggesting the presence of significant aesthetic effects.[269] This test will be performed for each of the five pairs of subgroups. Following the RCT design, the hypothesis is that no statistically significant difference will be found.[270]

8.2.3 Comparability

As above, various demographic data are collected to ensure comparability between the control group and the four treatment groups. The impact of non-comparable groups is the requirement

[269] A two-tailed t-test is an inferential statistical test which examines the difference between the means of two groups. The t-test assumes these groups are normally distributed and have the same (or very similar) variances. The null hypothesis of a two-tailed t-test is that the means of both groups are equal. The alternative hypothesis is that they are not equal. Given the hypotheses outlined in this thesis, therefore, a one-tailed t-test may also be used. A one-tailed t-test has the same assumptions and null hypothesis as the two-tailed t-test, but the alternative hypothesis is that one group is larger than another. In this sense, the two-tailed t-test is a simply a test of difference in group means, while the one-tailed t-test is a test of difference in group means with a directional component.

[270] A two-tailed t-test is appropriate here as there is no *a priori* expectation regarding aesthetics.

to control for the factor which is not comparable. For instance, if the control group is significantly younger than a treatment group, the age variable will have to be controlled for in whatever subsequent analysis is performed. This, in turn, changes what analyses can be performed. For instance, often a t-test will be a suitable test, but in having to control for a demographic disparity, a t-test will need to be replaced with an OLS regression. Of course, if the groups are comparable, which is to say, if the distributions of the four demographic variables (age, sex, education and political identity) are not significantly different, no such adjustment will be necessary. A chi-squared test for differences in distribution is used to test the comparability of the data. Given a random assignment and a large sample size, it is expected that the groups will be comparable.

8.2.4 Testing Whether Impersonal Nudges Are Effective

The final piece of analysis to be conducted prior to any investigation of psychometric variables and nudge effectiveness is a simple examination of whether, without any conflating factors, the four nudges are effective at influencing behaviour in a positive direction. Again, positive is taken to be a higher likelihood score (effectiveness) for the nudge advertisement compared to the control advertisement.

The statistical procedure for investigating such a proposition follows the methods outlined for examining the presence of significant aesthetic effects, as well as by others in the literature (Lipman, 2020; Peer et al., 2019; Schöning, Matt and Hess, 2019; Hirsh, Kang and Bodenhausen, 2012; Moon, 2002). Namely, each nudge group is compared to the control group using a dummy variable which takes the value of 1 if an observation is in the nudge group, and a value of 0 for all other observations. A test for difference in the average effectiveness of these groups using a t-test can then be performed, with a significant difference indicating the nudge *is effective at influencing behaviour*.

8.2.5 A Note on Normality and Homogeneity of Variance

An assumption of the t-test is that the dependent variable (effectiveness) is normally distributed (Kim and Park, 2019). This can be examined though histogram plots and several statistical tests, notably the Shapiro-Wilk's test and the Kolmogorov-Smirnov test. For this thesis, the Shapiro-Wilk's test is adopted as it has been shown to be the most powerful (Razali and Wah, 2011). While normality is expected, provision should be made to account for the possibility of a non-normally distributed dependent variable. The usual solution is to utilise a non-parametric equivalent of the t-test, the Wilcoxon-Mann-Whitney U-test, or the WMW-test (Wilcoxon, 1945; Mann and Whitney, 1947; Fay and Proschan, 2010). The trade-off for non-normality comes in the form of statistical power, with the WMW-test having less statistical power than the t-test (Fay and Proschan, 2010). However, given non-normality is often a problem which cannot be easily overcome if a few initial strategies fail,[271] this compromise on statistical power is often a worthwhile trade.

A second assumption is homogeneity of variance (Kim and Park, 2019). Levene's test for homogeneity of variance is an effective test of this assumption (Levene, 1960). Levene's test is also less sensitive to violated normality than the alternative Bartlett's test (Snedecor and Cochran, 1989). Where homogeneity of variance is violated, Welch's t-test can be used (Welch, 1947).[272]

8.2.6 Identifying Relationships Between Psychometrics and Nudges

The basis on which the delivery personalisation of nudges will occur is the identified relationships between the seven psychometric scales and the four nudges. It is therefore paramount – and indeed, the *modus operandi* of the primer group – to identify these relationships and thus proceed to examine the effectiveness of nudges when delivery

[271] Such strategies usually include log-transformations and square-transformations. These adjustments may produce their own challenges, largely centred around interpreting means and coefficients. Furthermore, where multiple groups are under examination, the transformation of one group may not produce normality in another, prompting further discursive and comparability challenges.

[272] These assumptions and solutions also apply to ANOVA, which is discussed below.

personalisation is used. In the personalisation literature, two methods of identifying these relationships can be found.[273]

The first is the *deductive* matching approach, whereby groups of participants are constructed *after-the-fact* in accordance with *a priori* expectations about psychometric profiles and nudge effectiveness (Schöning, Matt and Hess, 2019; Hirsh, Kang and Bodenhausen, 2012; Moon, 2002). The second is the *inductive* moderation approach, whereby participants are analysed using a moderated regression model to identify significant moderation effects (i.e. relationships between psychometrics and nudges; Peer et al., 2019).

One may argue that this second method is actually two methods (Hayes, 2018). In conjunction with moderated regression, Peer et al. (2019) utilise the Johnson-Neyman technique (JNT; Johnson and Neyman, 1936) – also known as *floodlight* analysis (Hayes, 2018; Spiller et al., 2013) – to identify *regions of significance*: precise values between which the moderation effect is expected to be statistically significant (see part 8.2.6.3). Alternatively, Hayes (2018) discusses the use of the pick-a-points method – also known as *spotlight* analysis – whereby – once a statistically significant moderation effect is identified – various values of the moderator are tested for statistical significance.

Each of these three methods has advantages and disadvantages, and as argued above, can often be used *in conjunction* with one another to elucidate a greater understanding of the ongoing dynamics within the data. Below, each method, and the statistical procedure involved in each method, is discussed, before a summary discussion is offered. First, however, a note on psychometric variables is offered.

8.2.6.1 A Note on Psychometric Variables

In this experiment, seven psychometric scales are used: five of the components of the GDMS scale, the NFC scale, and the CFC scale. Each scale consists of multiple questions which

[273] An introduction to these methods is provided immediately, while more in-depth discussions, with terms clearly defined, are offered below.

must be aggregated into psychometric *variables* to be used in statistical analyses. Aggregation can take two forms, summation and averaging, with averaging being the most common approach and so the one taken here.

The issue aggregation raises, however, is that in aggregating, a single variable must capture a sufficient amount of the variance captured by the non-aggregate components, less the single variable be insufficient to represent these data. To verify the validity of this single variable, therefore, Cronbach's alpha test for internal consistency is used (Cronbach, 1951). Testing for validity using Cronbach's alpha is commonplace within the individual difference literature (Appelt et al., 2011; Cho, 2016) and has been used within the personalisation literature also (Peer et al., 2019). Given all the scales examined are well-established and have been consistently shown to produce high Cronbach's alpha scores (typically higher than around 0.6-0.7; Nunnally and Bernstein, 1994), no issue is expected in repeating these examinations here.

8.2.6.2 The Matching Approach

The principle of the matching approach is best demonstrated by Schöning, Matt and Hess (2019) and Moon (2002). The approach begins by imagining a 2 × 2 grid in which all respondents can be categorised:

Figure 6: Example 2 × 2 Matching Approach Grid

	Group 1	Group 2
Low	Match (D = 1)	No Match (D = 0)
High	No Match (D = 0)	Match (D = 1)

As shown in Figure 6, the 2 × 2 grid is assembled along two-axes. The first axis demarcates whether a respondent falls into some treatment group ‘Group 1’ or ‘Group 2’. The second axis demarcates whether a respondent scores ‘high’ or ‘low’ on some psychometric scale. Assuming *a priori* that the treatment in Group 1 will be more effective with ‘low’ scorers and the treatment in Group 2 will be more effective with ‘high’ scorers, those respondents which match the conditions of these assumptions are organised into a ‘matching’ group, indicated using a dummy variable with the value 1. Those respondents who do not match the conditions of these assumptions are also organised, however into an ‘non-matching’ group, indicated by the dummy variable with the value of 0.

A difference in the average effectiveness scores of these matching and non-matching groups can then be tested using a t-test. Where a significant difference is found, and the average effectiveness of the matching group is found to be greater than that of the non-matching group, researchers can conclude that their *a priori* expectations about the relationship between the treatments and the psychometric were correct. Furthermore, as Moon (2002) and Schöning, Matt and Hess (2019) do, researchers can propose that, *in the future*, matching respondents in accordance to the relationship subsequently identified will produce more effective outcomes. It is from this perspective that these authors use the matching approach to contribute to discussions of personalisation without, during their experiments, *actually* personalising interventions.

The key advantage of this form of the matching approach is it is very simple. Depending survey-experiment design, participants can automatically be sorted into one of the two treatment groups, while the researcher can determine the threshold for a high or low score once the data are collected and easily categorise participants appropriately. Furthermore, having constructed the match/non-match dummy variable, the analysis to determine the presence of a significant relationship is also very straightforward.

The matching approach, however, has several drawbacks. Firstly, the high/low scorers must be determined arbitrarily. Of course, within a specific research project their may be good reason to suspect responses above or below a given value should matter and be appropriately labelled. But in general, there is no fundamental reason to determine a score of, say, 2.2 out of 5 to be low and 2.8 out of five to high.

At best, on an n-point Likert scale, one might be tempted to construct the high/low demarcation around the midpoint, defined as $\frac{n+1}{2}$. This would initially appear to not be an arbitrary selection as the demarcation corresponds to the measurement scale of the psychometric itself. Of course, this is a false proposition, as the measurement scale for the psychometric is itself arbitrarily selected.

An alternative approach may be to demarcate high/low scorers based on the *data*, say defining a low scorer as someone who scores less than the mean or the median, and a high scorer as someone who scores more than the mean or the median. This does not resolve the problem of arbitrary selection, however. For instance, if the behaviour of 90% of the population is observed to be statistically similar, it is not clear what rationale might correspond to the selection of the mean (which is unlikely to capture 90% of the population) or median (which splits the population 50/50).[274] Indeed, even discounting this criticism, another emerges. Namely, the use of a sample mean or median requires one to assume the sample mean or median is similar to that of the population, less the demarcation of high/low can only be taken to describe that given sample. Such an assumption is not required of the midpoint demarcation.

Regardless of how the high/low demarcation is determined, the selection remains arbitrary. Acknowledging this, a further adjustment may be to examine the difference by matched/non-matched groups under *several* different demarcations, e.g. the mean, median and midpoint.[275] This pacifies somewhat the criticism that can be levied at the method by acknowledging the

---

[274] Note, such an argument rests on the lack of an *expectation* that 90% of the population is statistically similar. Where this is an *a priori* expectation, such an expectation could be tested, and such an expectation would serve as a reasonably fair reason for defining high/low around being above/below the 90th percentile.

[275] This represents a tendency towards expanding the range of examined values for significance, a tendency which will return in due course. So too, however, will the problem of arbitrary value selection.

arbitrariness of the demarcations and taking efforts to address it; but it in turn introduces a new form of arbitrariness. Namely, it is not clear how significant relationships might be determined under multiple constructions of high/low scorers. For instance, it would seem rather compelling if, regardless of construction, a significant difference in a consistent direction (i.e. positive/negative) was identified. But it is not clear what might the interpretation be if only two constructions produced such a result, or indeed only one. Or, further, the interpretation of the results is unclear if multiple instances of significance are identified, but the direction of the effect varies across constructions. When one is seeking to confidently identify a relationship between a treatment effect and a psychometric, constructed around a single value or within a bounded range of values, the matching approach can only be treated as exploratory of relationships, and not determinant.

A somewhat different matching approach is used by Hirsh, Kang and Bodenhausen (2012). This approach differs both in experimental design and analytical procedure. They investigate advertisements and the Big Five personality scale. Five advertisements are constructed – one for each of the five personality types captured by the scale. Each advertisement is shown to every participant, who then scores the effectiveness of each advertisement. Rather than constructing matching and non-matching groups along an arbitrary determinant of high/low scoring, Hirsh, Kang and Bodenhausen (2012) then simply regress the five psychometric variables for each of the Big Five personality types onto each of the advertisements:

Equation 3

$$Ad_n = \beta_1 + \beta_2 Agree + \beta_3 Consci + \beta_4 Extra + \beta_5 Neuro + \beta_6 Open + \varepsilon \tag{3}$$

where Ad_n is the effectiveness of advertisement n, which corresponds to one of the five personality type advertisements. In taking this approach, Hirsh, Kang and Bodenhausen (2012) are able to show that only when the psychometric and the advertisement match is the psychometric statistically significant.

This approach improves on the previous matching approach as it avoids arbitrary group construction and allows all of the values of the data to be considered. The major drawback of this approach, however, is that the regression analysis can only determine the direction of any effect (i.e. is the coefficient positive or negative?) and does not delineate at what values of the psychometric the significant effect may be most notable.

Both approaches are also unsuitable for this project as each requires some degree of adjustment to the experiment design. In the simple 2 × 2 matching approach, both groups which are being compared are treatment groups. While the current survey-experiment design would enable a comparison of treatment groups, it remains unclear how a comparison of treatment groups would be beneficial given the measurement of multiple psychometrics. The alternative may be to compare the control group with a treatment group in a 2 × 2 matching approach; however, it remains unclear what a person who 'matches' with the control group would represent. The control group is *not* a treatment group, and thus any matching group which can be constructed would not be congruent with the idea of matching *treatments*.

The same problem of survey design is present in the regression-based approach. In order to use regression analysis, participants would need to be shown all four nudge advertisements. As above, insofar as the survey-experiment is designed to resemble a real-world choice between political candidates, this survey design would undermine this endeavour.

Given this, an alternative matching approach – henceforth dubbed matching analysis for disambiguation – can be determined for this project.[276] Two key limitations remain with matching analysis: values must be selected arbitrarily, and relationships between nudges and psychometric variables are assumed to be dichotomous, existing either above or below these arbitrarily selected values. However, the matching analysis approach proposed here does not require an adaption of the survey-experiment design as outlined above.

[276] This approach shares similarities with Lipman (2020).

The approach broadly has three steps:

1. First, the significance of the difference between the control group and the treatment group needs to be examined. This analysis will already be conducted. The importance of this examination on the matching analysis, however, is that it informs the interpretation of the third stage of the matching analysis. A statistically significant difference *overall* would be expected to produce statistically significant results when the treatment group is split into high/low subgroups, just as a statistically insignificant difference would be expected to produce statistically insignificant results. This initial analysis, therefore, frames expectations later on, and allows contrary results to be identified.
2. Second, the treatment group is divided into high/low subgroups according to average psychometric scores. This must be done for each psychometric scale. Furthermore, as the demarcations of high and low are arbitrary, several constructions of high/low are used. These are the mean, median and midpoint as discussed above. With these subgroups constructed, a two-tailed t-test[277] is used to compare the effectiveness of the nudge between the high subgroup and the low subgroup. Where no statistically significant difference is found, no relationship between the psychometric and the nudge is determined. Where a significant difference is found, a relationship *may* exist.[278]
3. Third, a two-tailed t-test is used to investigate a significant difference in averages between a given treatment *sub*group and the control group. Where high/low subgroups associated with the same psychometric variable are statistically significantly different from each other, and one or both subgroups are statistically significantly different from the control group, there is good reason to conclude that a statistically significant

[277] A two-tailed t-test is used as matching analysis is exploring for relationships and therefore makes no *a priori* assumptions about the sign of the effect.
[278] Interpretation of these results, given the multiple constructions of the subgroups, still remains arbitrary, and a degree of discrepancy and prudence is required.

relationship exists between the nudge and the psychometric. Where neither subgroup is significantly different from one another, and both subgroups are statistically (in)significantly different from the control group, no relationship is the most likely conclusion.

In Chapter 11 and Chapter 13, this matching analysis is performed on two pilot studies, and the effectiveness of this approach evaluated. Where matching analysis is found ineffective, this approach is not used in the primary analysis presented in Chapter 14, with the alternative moderation analysis approach utilised.

8.2.6.3 Moderation Analysis

An alternative approach to matching analysis is *moderation* analysis, utilised by Peer et al. (2019) in their work on delivery personalisation. The term *moderation* describes an interaction between variables. As Saunders (1956) describes moderation: "There are many examples of situations in which the predictive validity of some psychological measure varies systematically in accord with some other independent psychological variable" (Saunders, 1956: 209). In recent years, Hayes' (2018) work on moderation has come to dominate. Hayes (2018), in more methodological language, describes moderation as follows: "The effect of X on some variable Y is moderated by W if its size, sign or strength depends on or can be predicted by W. In that case, W is said to be a *moderator* of X's effect on Y or that W and X *interact* in their influence on Y" (Hayes, 2018: 220, original emphasis).

As Peer et al. (2019) argue, moderation can be used to describe the relationship between a person's psychometric profile and their susceptibility to being nudged. Indeed, when the phrase, 'identify a relationship between a psychometric variable and a nudge' – a phrase which has been used several times in one form or another already – is used, this is implicit of an acknowledged belief in the moderated interaction of psychometric variables and nudges. Following Hayes and Rockwood (2017), the hypothesis that the relationship between psychometric variables and nudges as understood within the discussion of delivery personalisation and in Peer et al. (2019) is a moderated one seems justified. Hayes and Rockwood (2017) write, "moderation analysis is used to address *when*, or under *what circumstances*, or for *what types of people* [an] effect exists or does not and in what magnitude" (Hayes and Rockwood, 2017: 47, original emphasis). Certainly, given this description, the hypothesised relationship between nudges and psychometric variables qualifies as one to be described in terms of moderation.

Statistically, simple moderation effects are examined using OLS regression with the inclusion of an interaction term between the independent (sometimes called *focal*) variable X and the moderator variable W.[279] Hayes (2018) outlines the regression model associated with a simple linear moderated model (SLMM):

Equation 4

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 XW + \varepsilon \tag{4}$$

where Y is a dependent variable, X is a factor or dummy variable, W is a continuous variable which is taken to moderate the effect of X, and XW is an interaction term between X and W (i.e. the product of X and W).[280] Unlike in a standard OLS regression model without the interaction term, β_1 and β_2 take on slightly different interpretations. Hayes (2018): "these regression coefficients estimate the effect of X when W = 0 and the effect of W when X = 0, respectively" (Hayes, 2018: 239).[281] This is important to note because, in some instances such

[279] As Hayes (2018) notes and documents extensively in their appendices, moderated regressions can be extremely complicated. There is no theoretical limit on the number of interaction terms, or the number of terms interacting within a single interaction term. Furthermore, moderation effects can exist *within* other models. The model shown here, and the model focused on here, is the simplest and most commonly discussed construction of a moderated regression.

[280] Mathematically, X could be described as the moderator and W the focal variable (Hayes, 2018). However, it is most common to treat the dummy variable as the focal variable and the continuous variable as the moderator as these are the specifications for probing moderation effects post-estimation (Hayes, 2018; Preacher, Curran and Bauer, 2006).

[281] Hayes (2018) demonstrates this mathematically, though this is a rather simple conclusion. When either X or W = 0, the interaction term and respective variable falls out of the model, and the model reduces to either an X on Y or a W on Y form.

as this thesis, the value of W cannot equal 0 (as the psychometric scale ranges from 1 to 5), and so the coefficient of X (the dummy variable indicating control or treatment in this specific instance) cannot be interpreted in isolation using a moderated regression model.

However, with the inclusion of the interaction term, this limitation becomes less significant.[282] In seeking to identify a statistically significant (moderated) relationship between the psychometric variable and the nudge – to use the language of the this experiment – one needs only to identify and interpret a significant *moderation* effect, which is to say, the p-value associated with the interaction term need only be statistically significant to determine a significant relationship between the nudge and the psychometric. The effect size of the moderation effect, known as the *conditional effect* (Hayes, 2018; Preacher, Curran and Bauer, 2006), is then given by:

Equation 5

$$\theta_{X \to Y} = \beta_1 + \beta_3 W \tag{5}$$

As eluded to in Equation 5, the conditional effect can be understood as the effect of X on Y given the presence of a moderating effect W. This statement becomes useful when probing the interaction term following the model's estimation.[283] Two methods for probing the interaction exist in the literature. These are the pick-a-points method, sometimes known as *spotlight* analysis, and the Johnson-Neyman technique, sometimes known as *floodlight* analysis (Hayes, 2018; Spiller et al., 2013). These alternative names illuminate (pardon the pun) the difference between these techniques: spotlight analysis investigates the interaction term using specific values of interest, while floodlight analysis *solves for* specific values of

[282] Hayes (2018) notes that some researchers may choose to zero centre their moderator variable so as to remove this limitation and does not offer any distinct opinion on how worthwhile this is. Rather, the only comment is a reminder that such an adjustment should not change the substantial meaning of the model, only the values of the coefficients.
[283] Note that Preacher, Curran and Bauer (2006) use slightly different notation to that used in Equation 5, which follows Hayes' (2018) notation. These statements, however, are identical in meaning.

interest given the specifications of the model (Hayes, 2018; Preacher, Curran and Bauer, 2006).

The trade-off between spotlight analysis and floodlight analysis returns to previous discussions. Hayes (2018) and Preacher, Curran and Bauer (2006) argue spotlight analysis can be a useful tool for analysing a moderation effect when there is a justified interest in a particular value of the moderator. For instance, if previous research has found consistent evidence that people who score a three on a given psychometric scale exhibit a particular behaviour, using spotlight analysis to investigate the moderation effect when W = 3 makes sense. Hayes (2018) also notes that spotlight analysis is somewhat simpler to grasp than floodlight analysis, and can be interpreted easier too. The statistical significance of the effect can then be calculated by calculating the t-statistic associated with the value and in turn the t-statistic's associated p-value:

Equation 6

$$t = \frac{\theta_{X \to Y}}{SE_{(\theta_{X \to Y})}} \tag{6}$$

where $\theta_{X \to Y}$ corresponds to Equation 5.[284]

However, spotlight analysis suffers the same arbitrary selection problem as the matching approach does. This is to say, when there is no obvious reason to select a given value of the moderator, spotlight analysis becomes a very arbitrary way of probing an interaction (Hayes, 2018; Preacher, Curran and Bauer, 2006). In such a situation, it is often desirable to explore *all possible* values of the moderator. In principle, this is floodlight analysis, or the Johnson-Neyman technique.[285] In practice, floodlight analysis does *not* involve identifying p-values for

[284] Equation 6 is taken from Preacher, Curran and Bauer (2006), however, is adapted to use Hayes' (2018) more recent notation. This change in notation *does not* change the substance of the equation.
[285] Johnson and Neyman (1936) even note, "It [the JNT] is shown that the problem of matched groups may be generalized so that both a more detailed analysis of experimental data and a greater accuracy of results is obtained" (Johnson and Neyman, 1936: 209).

every possible value of the moderator. Rather, the significance level is pre-selected (usually the 5% level), and the t-statistic calculation is used to *solve for values of the moderator* (Johnson and Neyman, 1936). This is shown in an adapted t-statistic equation:

Equation 7

$$W = \frac{(t_{crit}SE) - \beta_1}{\beta_3} \tag{7}$$

However, while this rearrangement to isolate the moderator value W is correct, it does not reveal the full picture. When the calculation for the standard error (SE) is accounted for, Equation 7 actually takes a *quadratic* form (Hayes, 2018). This derivation is not provided by Hayes (2018), but is demonstrated below by combining Equation 5 and Equation 6 before rearranging:

$$t_{crit} = \frac{\theta_{X \to Y}}{SE_{(\theta_{X \to Y})}} \tag{6}$$

$$t_{crit} = \frac{\beta_1 + \beta_3 W}{SE_{(\theta_{X \to Y})}} \tag{6.1}$$

$$t_{crit} = \frac{\beta_1 + \beta_3 W}{\sqrt{SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2}} \tag{6.2}$$

$$t_{crit}\sqrt{SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2} = \beta_1 + \beta_3 W \tag{6.3}$$

$$t_{crit}^2\left(SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2\right) = (\beta_1 + \beta_3 W)(t_{crit}\sqrt{SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2}) \tag{6.4}$$

$$t_{crit}^2(SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2) = (\beta_1 + \beta_3 W)(\beta_1 + \beta_3 W) \tag{6.5}$$

$$t_{crit}^2\left(SE_{\beta_1}^2 + (2W)COV_{\beta_1\beta_3} + W^2 SE_{\beta_3}^2\right) = \beta_1^2 + 2\beta_1\beta_3 W + \beta_3^2 W^2 \tag{6.6}$$

$$t_{crit}^2 SE_{\beta_1}^2 + t_{crit}^2(2W)COV_{\beta_1\beta_3} + t_{crit}^2 W^2 SE_{\beta_3}^2 = \beta_1^2 + 2\beta_1\beta_3 W + \beta_3^2 W^2 \tag{6.7}$$

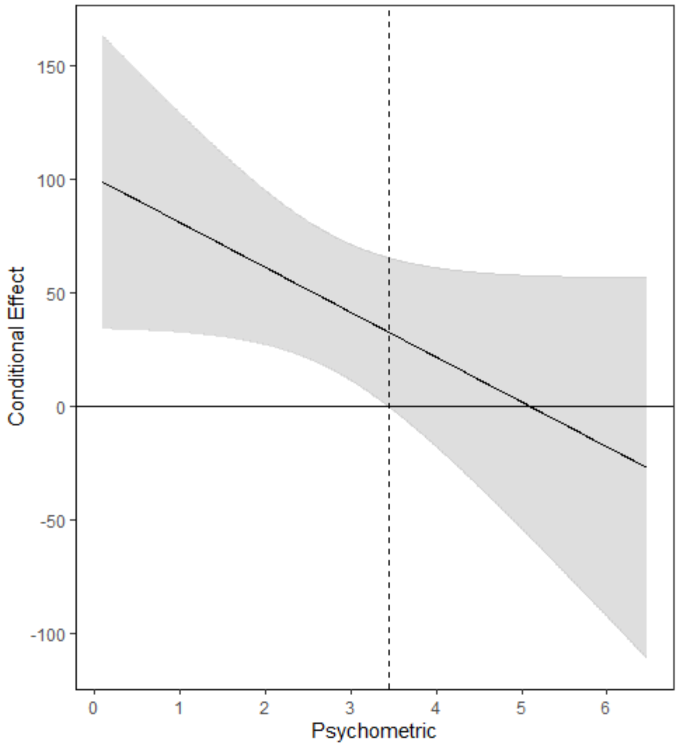
$$W^2\left(t_{crit}^2 SE_{\beta_3}^2 - \beta_3^2\right) + W\left(2t_{crit}^2 COV_{\beta_1\beta_3} - 2\beta_1\beta_3\right) + \left(t_{crit}^2 SE_{\beta_1}^2 - \beta_1^2\right) = 0 \tag{6.8}$$

where $\left(t_{crit}^2 SE_{\beta_3}^2 - \beta_3^2\right)$, $(2t_{crit}^2 COV_{\beta_1\beta_3} - 2\beta_1\beta_3)$ and $(t_{crit}^2 SE_{\beta_1}^2 - \beta_1^2)$ are constants (Carden, Holtzman and Strube, 2017). W can thus be solved for. The consequences of doing so, as Hayes (2018) and Preacher, Curran and Bauer (2006) *do note*, is that the Johnson-Neyman technique always produces *two* values for the moderator W which correspond to a t_{crit} value associated with a given critical confidence level. These two values represent the bounds of a region of significance,[286] between which all values of the moderator correspond to a statistically significant moderation effect. However, the JN technique may produce values which – while mathematically valid – are not values which the moderator can actually take. For instance, where a psychometric scale ranges from 1 to 5, a value of, say, -26.77 is not a value which can be practically used. Therefore, while the JN technique provides greater precision in determining moderator values than can be achieved with spotlight analysis, this precision can often be unnecessary.

Nevertheless, where regions of significance are identified, a common output accompanying the JN technique is a graphical representation. An example is shown in Figure 7:

[286] Or *regions* of significance.

Figure 7: Example Johnson-Neyman Plot



The JN plot shows values of the moderator on the x-axis, while the conditional effect (i.e. $\beta_1 + \beta_3 W$) is shown on the y-axis.[287] The grey area represents the upper and lower confidence intervals, while the dotted line represents the boundary between a region of significance and insignificance. A region of significance exists at the point that both the upper and lower confidence intervals are either greater than, or less than, zero on the y-axis.[288] Thus, the interpretation of Figure 7 would be that a region of significance exists for values less than 3.45. Of course, a second moderator value is produced by the JN technique which demarcates the other boundary for this region, but as this value is not a value the moderator can take, it is merely of mathematical interest, not practical.

Methodologically, the difference in approach between floodlight analysis and spotlight analysis (and, for that matter, the matching approach) is that floodlight analysis adopts an inductive approach to *identify* values of significance and thus interest, while spotlight analysis adopts a

[287] The y-axis can also be thought of as showing the *total* effect of X on Y, after accounting for moderation.
[288] In other words, the point at which the effect of X on Y can be confidently (i.e. at the 95% level) be said to differ from zero.

deductive approach to examine the significance or lack thereof of values *already determined* (by some criteria) to be of interest.

Where no such values are obviously of interest, as is the case here, and where an inductive approach may be desirable – again, as is the case here – floodlight analysis appears ideal. The trade-off, of course, is complication. Not only is floodlight analysis theoretically more complicated, but it is practically more difficult too. Hayes (2018) has made significant progress in easing the practical use of moderation analysis and the JN technique with their PROCESS macro for SPSS and SAS (Hayes, 2012). However, this macro lacks the graphical potential which is often reported as part of a floodlight analysis (Preacher, Curran and Bauer, 2006). Bachl (2015) has produced a series of functions in R which are capable of running the moderated regression, calculating the regions of significance, and plotting these results. However, Bachl (2015) has yet to update their package, and the present release is not functional with the latest version of R and the associated libraries. Brambor, Clark and Golder (2006) also offer code for the JN technique in their paper, this time for STATA. However, this software is, again, out-of-date, with the online repository housing the code – at the time of writing – inaccessible. Finally, Carden, Holtzman and Strube (2017) offer an Excel based program called CAHOST which is capable of calculating the moderated regression, undertaking the JN technique, and plotting the output. CAHOST, however, is limited to only 1,000 observations, while the plot output is difficult to use. Peer et al. (2019) do not document how they produced their analytical output.

The solution used here is a combination of programs. Firstly, STATA 13.1 is used to examine the presence of statistically significant moderation effects. Where a statistically significant effect is identified, data are entered into CAHOST, which is used to calculate regions of significance and data to be plotted. ggplot2 in R is then used to plot these data and edit the graphical output. As such, despite being technically more demanding than alternative methods, this disadvantage of the moderated regression and JN technique approach is not insurmountable. Furthermore, this approach resolves the problem of arbitrary selection associated with both spotlight analysis and matching analysis, while providing greater precision (in the form of regions of significance) than these approaches as well.

Of course, before any of this can be conducted, the moderated regression model which is to be estimated needs to be established. Fortunately, assuming no confounding differences arising from demographics, the SLMM model is sufficient for the proposed experiment. This is a further advantage of moderated regression over the matching procedure – namely, no adjustments to the experimental design are necessary in order to utilise moderated regression.[289]

The equation form of the estimated moderated regression model used here is:

Equation 8

$$effectiveness_i = \beta_0 + \beta_1 D_i + \beta_2 Psy_\lambda + \beta_3 D_i Psy_\lambda + \varepsilon_i \tag{8}$$

where $effectiveness_i$ is the effectiveness score of treatment group i compared to the control group (i.e. $Y_i - Y_0$), D_i is a dummy variable – or *focal* variable – taking the value of one if an observation is in treatment group i, and a value of zero for all other observations,[290] Psy_λ is a continuous variable – or *moderator* variable – containing the psychometric scores for psychometric λ, and $D_i Psy_\lambda$ is the interaction or moderation term. β_0 is an intercept and ε_i is an error term.

Where demographic differences between groups may need to be controlled for, this produces a relatively straightforward adaption of Equation 8:

Equation 9

$$effectiveness_i = \beta_0 + \beta_1 D_i + \beta_2 Psy_\lambda + \beta_3 D_i Psy_\lambda + \beta_n Demographics + \varepsilon_i \tag{9}$$

[289] In contrast to the matching approach, which itself has had to be adapted to this experimental design in the form of matching analysis simply to be usable here.
[290] I.e. if the observation is in the control group.

However, such an adjustment does reintroduce the technical challenges associated with floodlight analysis. CAHOST, for instance, can only accommodate SLMM designs (which is to say, 2 independent variables). As such, an alternative approach relying more heavily on Hayes' (2012) PROCESS tool may become necessary. Of course, this all assumes a statistically significant difference is found and will need to be controlled for – where this is not the case, no adjustments to Equation 8 will be necessary.

8.3 – Prioritising Relationships

Given the advantages of moderation analysis with the JN technique compared to the alternative methods of identifying relationships between nudges and psychometric variables and given the precedent of this method also having been used by Peer et al. (2019), moderated regression is selected as the primary method of analysis of the primer group. As noted above, however, it is feasible to engage with the matching approach while *also* utilising moderated regression, provided the matching approach adopted is adjusted so as to not alter the survey-experiment design. This adapted approach is offered here as matching analysis, and is investigated as a secondary method of analysing the relationships existing within the primer group and during the two pilot studies.[291]

An outstanding question pertains to the use of nudge-psychometric relationships once they are identified. The goal of the primer group is to construct an evidence-based psychometric map which can be used to personalise the delivery of nudges to the treatment group in the second round of data collection. However, an aspect of the mapping process which is missed in diagrammatic presentations of the psychometric map, such as that shown in Figure 3, is what might be called the nudge *preference*.

[291] One may be tempted to argue that the use of both methods enable comparability and thus the most robust conclusions. This is perhaps assertible; however, the many methodological weaknesses of the matching analysis mean it is liable to miss or misreport conclusions found using moderation analysis. See Chapter 13 for evidence of the limitations of matching analysis in comparison to moderation analysis.

The issue of nudge preference arises in several ways:

1. A region of significance may exist between psychometric A and nudge B, and a region of significance may exist between psychometric A and nudge C. Thus, if a respondent were to score for psychometric A a value which falls within both regions of significance, which nudge – B or C – should be delivered to this participant?
2. A region of significance exists between Psychometric A and nudge B, as well as between Psychometric B and nudge C. If a respondent were to score within these regions of significance *for both psychometrics*, which nudge – B or C – should be delivered to this participant?
3. Various regions of significance exist between various psychometric variables and various nudges. However, a participant does not score within a region of significance for *any* psychometric variable. Which nudge – from all available – should be delivered to this participant?

While not stated quite so situationally as expressed here, these problems are ones Peer et al. (2019) also tackle in their work. Their 'solution' is to use a Monte Carlo simulation to estimate which nudges should be delivered under various conditions to maximise the effectiveness of the nudges *overall*. This approach is inadequate for several reasons. Firstly, they provide no specifications for their simulation model, leaving the methodological space void. Secondly, the model is ultimately extremely simplifying, eliminating entirely two of their original nudges and suggesting 85% of participants should be nudged using only one nudge. Thirdly, even where a simulation model is appropriate, Peer et al. (2019) provide no explanation for why a Monte Carlo simulation over, say, a suite of machine learning techniques is selected.[292] Finally, it is not clear why a simulation model is used at all when a much simpler solution to the problem

[292] Note that this critique is not to say that machine learning techniques are definitely better than Monte Carlo simulation. Indeed, the literature on this specific question is extremely sparse, constituted entirely by Peer et al. (2019) and this discussion. As such, further research would need to be conducted before such a claim could be made.

at hand presents itself within the analyses which are already untaken to get to this methodological junction.

This simpler solution is merely to rank the effects of nudges. Recall that the conditional effect (i.e. $\beta_1 + \beta_3 W$) is defined as the effect of X (the nudge) on Y (effectiveness) *given* moderation by W (the psychometric). Insofar as the selection of nudges in the above scenarios is done so as to maximise the effectiveness of the nudge, one can simply rank the modulus of the coefficients of the interaction terms (i.e. β_3) prioritise delivering the nudge associated with the largest coefficient. The use of the modulus arises from a problem associated with an alternative approach, namely ranking the *conditional effect*. If one were to calculate the conditional effect for a positive moderation effect, and compare it to a negative moderation effect, a 1-unit change in the value of W would increase the conditional effect of the positive relationship and decrease the conditional effect of the negative relationship. To make these results comparable – which is to say, to determine which moderation effect and thus which nudge would produce a greater effect following a 1-unit shift *into a region of significance*, one must make these relationships comparable. This is easily done by taking the modulus.

This method resolves scenarios 1 and 2. Scenario 3 is resolved using an even simpler approach: where participants demonstrate no affinity for any particular nudge, they should be presented with the nudge which, *overall*, is found to be most effective compared to the control group (i.e. the best impersonal nudge).

Again, the data required for establishing nudge preferences are already produced in the planned analysis. The interaction term coefficients are produced as part of the moderated regression, while the most effective nudge is determined when the effectiveness of the nudges without any psychometric effects are examined against the control group. Two alternative solutions to this problem also emerge from the literature – both of which remain comparatively simpler than the simulation approach taken by Peer et al. (2020).

The first is to rank β_1 coefficients, i.e. the coefficient associated with the nudge without moderation effects. The logic of this solution is quite evident – if one is seeking to select the *most effective* nudge, then the nudge with the greatest effect as given by the coefficient seems ideal. Recall, however, that the *conditional effect* is defined as the effect of X (the nudge) on Y (effectiveness) *given* a moderating effect by W (the psychometric). Where a statistically significant moderation effect exists, it makes no sense to select nudges based on their effectiveness when not moderated.

The second solution as discussed by Hayes (2018) is to use a comparison of r-squared (so-called change in r-squared). By estimating the amount of variance explained (i.e. r-squared) by a regression without the interaction term, and comparing this to the amount of variance explained when the interaction term is included, an estimate of the impact of the interaction term on variance can be determined (i.e. $R_2^2 - R_1^2$). This estimate can be used, according to Hayes (2018), as a means of evaluating moderation effects, and Hayes' (2012) PROCESS macro provides this estimate as a default output. However, Hayes and Darlington (2017) are critical of the reductive effect such an approach has on the interpretation of moderation models. Indeed, this is evidenced here: just because one moderation effect may explain more variance in a given moderated regression model than another in its respective model does not mean the nudge associated with the former is necessarily *more effective* than the later. In fact, while potentially a handy trick for appraising moderated regression in some circumstances,[293] such an approach fails to offer any solution to the scenarios set out above.

8.4 – Personalised Treatment Group

Given the inductive approach taken here, a flow of the survey-experiment design for the personalised treatment group (PTG) cannot be provided *a priori*.[294] Indications of how this

[293] For instance, where both the focal and moderator variable are significant and r-squared is already high without the moderator term, an analysis of change in r-squared may be useful in deciding if the inclusion of the moderation term is useful, or whether the more parsimonious model is better.
[294] A survey flow the PT group is provided in Chapter 14.

survey-experiment will be constructed, however, can be offered. As with the primer group, the survey-experiment will be incentivised, hosted on Qualtrics and distributed via MTurk.

The major difference between the primer group survey design and the personalised treatment group survey design is the presence of personalisation. This manifests in two ways. Firstly, PTG participants are, *at the outset*, randomly assigned into one of three subgroups. These subgroups are delivery personalisation only (DO), choice personalisation only (CO) and choice and delivery personalisation (CD). These subgroups can then be compared to examine the hypotheses established in Chapter 3.

8.4.1 Delivery Personalisation Only

The DO subgroup will first be asked to complete the three psychometric scales, in the same order as the primer group (i.e. GDMS, NFC, CFC).[295] Unlike the primer group, however, once these questions are completed, based on their responses, participants are shown a nudge-advertisement which is predicted to be most optimal (in terms of effectiveness) from the primer group data.

This process is surprisingly complicated. Qualtrics, on a UI level, is extremely limited in its functionality. In order to automate the personalised delivery of nudges, it must be possible within Qualtrics to automatically calculate the average score for a given psychometric. If this is possible, the standard 'Branch' feature within Qualtrics can be used to assign participants to their delivery-personalised nudge. A review of the literature would suggest, however, that automated averaging is not possible. Peer et al. (2019), for instance, require participants to complete the psychometrics before inviting those same participants back to complete the nudge-task at a later date.

[295] This assumes all psychometrics will contribute something to delivery personalisation. It may be the case that, upon analysis, some psychometrics are not indicative of any moderated relationship with any nudge. As such, they may not need to be included in the survey-experiment. The same is true of nudges: where a nudge is ineffective or has no moderated relationships and is less effective than an alternative nudge, it need not be included.

However, this averaging limitation only exists at the UI level, and by utilising the 'Embedded Data', 'Recode Values and 'Math Operations' tools available within Qualtrics, as well as some basic JavaScript, it is completely possible to automatically average scores. These averages can then be used in conjunction with the 'Branch' function to deliver the nudges.[296] Automating the survey-experiment in this way, it is hoped, will encourage greater completion of the survey-experiment.

Nudge-advertisements shown to participants in the DO group will not be altered from those shown in the primer group. However, participants will still be randomly assigned to a group where the delivery-personalised nudge-advertisement features Candidate A and a control advertisement featuring Candidate B, or a group where the delivery-personalised nudge-advertisement features Candidate B and a control advertisement featuring Candidate A. As such, the RCT design remains, and any nudging effect observed can be ascribed as the effect of delivery personalisation *without* an aesthetic effect.

As above, participants who fail to produce any psychometric scores associated with a moderation effect will be shown the nudge-advertisement which is found, within the primer group, to be *overall* the most effective. Crucially, regardless of what nudge is delivered to a given participant, no semblance of choice personalisation is used; neither Candidate A or B appeal to outcomes which have been specifically chosen for the participant, and with the only difference being which *impersonal* outcome is being supported by *personalised* nudge delivery.

8.4.2 Choice Personalisation Only

Participants in the CO group are first asked to select, from four voter issues, which issue is in the abstract most important to them. These four issues are: 1) the economy; 2) education; 3) healthcare; 4) national security. These issues are selected as they are identified in Gallup's

[296] It is possible that Peer et al. (2019) were unable to take this approach as they use a more complex procedure to determine nudge preferences. This, then, is yet another advantage provided by the much simpler ranked-coefficient approach.

2020 poll of top voter issues as being most important to American citizens, with over 80% of respondents agreeing each issue is either extremely important or very important (Hrynowski, 2020).[297] Upon selecting one of these four options, participants are then shown a political advertisement with the slogan changed to embed *their most important issue*. However, the nudge and candidate used remains randomised. Example slogans are shown in Table 5:

Table 5: Choice Personalised Slogans

Nudge	Generic Slogan	Policy	Choice Personalised
Status Quo	"Let's Keep Going"	The Economy	"Let's Keep Building a Better Economy"
		Education	"Let's Keep Building a Better Education System"
		Healthcare	"Let's Keep Building a Better Healthcare System"
		National Security	"Let's Keep Building a Secure Nation"
Present Bias	"Fighting for You Today, not Tomorrow"	The Economy	"Fighting for a Better Economy Today, not Tomorrow."
		Education	"Fighting for a Better Education System Today, not Tomorrow."
		Healthcare	"Fighting for a Better Healthcare System Today, not Tomorrow."
		National Security	"Fighting for a Secure Nation Today, not Tomorrow."
Loss Aversion	"Let's not go Backwards"	The Economy	"Let's not let the Economy go Backwards."
		Education	"Let's not let Education go Backwards."
		Healthcare	"Let's not let Healthcare go Backwards."
		National Security	"Let's not let National Security go Backwards."
Social Norm	"Trusted by Voters"	The Economy	"Trusted by Voters to look after the Economy"
		Education	"Trusted by Voters to look after Education."
		Healthcare	"Trusted by Voters to look after Healthcare."
		National Security	"Trusted by Voters to look after National Security."

By randomising the nudge delivery but changing the slogans to embed policies participants have revealed to be important to them, these advertisements personalise choice (by nudging

[297] The economy (84%), education (83%), healthcare (81%), national security (80%). No other issue is found to be above 80%, with the closest being gun policy and immigration, both at 74%.

towards the candidate which would appear to be the desirable outcome for the participant) but do not personalise delivery (by still using a random (i.e. impersonal) nudge). This approach follows methodologically from Matz et al. (2017), who personalise advertisements based on user-revealed Facebook likes, Guo et al. (2020), who offer password tips based on previously constructed passwords, and Page, Castleman and Meyer (2020), who design personalised text messages based simply on however complete a person's FAFSA application is. By simply asking respondents to reveal those outcomes which matter to them, one can easily produce an advertisement which subsequently appeals to those preferences. To ensure the transparency of this strategy was not revealed, respondents are informed at the start and the end of the survey that the advertisements they are shown are randomly allocated. CO participants also complete a set of psychometric questions after they are asked which issue matters most to them, but before they are shown an advertisement, further separating these actions.

8.4.3 Choice and Delivery Personalisation

Following Chapter 3, the hypotheses in brief arrived at from theory were: a) personalised nudging would be more effective than impersonal nudging and not nudging at all, and b) choice and delivery personalisation, when used in conjunction, would be more effective than either used in isolation. The CO and DO groups provide a basis for evaluating this first hypothesis, but a group must be constructed which utilises both choice and delivery personalisation. This is the CD subgroup.

The CD subgroup is very much an amalgam of the survey-experiment designs used for the CO and DO subgroups. Upon completing the demographic questions, participants face the same policy question as in the CO subgroup, namely which of the four policies matters most to them. From here, the survey splits into four branches, with each participant following one branch depending on their answer to the policy question. Participants will answer the same psychometric questions in the same order and with the same branching structure as in the DO subgroup. Finally, upon completing the psychometric questions, participants will be shown a nudge advertisement and a control advertisement. The nudge used in the advertisement will be personalised based on the predictions borne from the primer group data, while the slogan used will be personalised based on the policy preference specified by the participant.

8.5 – Comparisons

With data on the nudge effectiveness for the primer group, the CO subgroup, the DO subgroup and the CD subgroup, several statistical comparisons can be used to evaluate whether the hypotheses of this project can be accepted or should be rejected.

Following Peer et al. (2019) an ANOVA is used to compare the CO and DO groups, respectively, to the primer group (impersonal nudging) and the control group. A positive, statistically significant difference in effectiveness between the personalisation subgroups and the primer and control groups would support the hypothesis that personalisation makes nudges more effective.[298] Tukey's test can add further detail to this analysis, as well as t-tests of various pairs. In instances of statistical exploration, such as are involved in matching analysis, the use of two-tailed t-tests is advantageous, as these tests test for any statistical difference in the means of group, regardless of the sign of the effect (i.e. positive or negative effect). However, when testing the proposed hypotheses, a one-tailed t-test may prove more instructive, as this test offers a statistical examination of not simply *difference in means*, but also the *sign* associated with the effect. As personalisation is expected to produce a statistically significant and *positive* effect, a one-tailed t-test is suitable here.

Finally, the CD subgroup needs to be compared to both the CO and DO subgroups to evaluate the second hypothesis that using choice and delivery in conjunction produces more effective nudges than using them in isolation. This, again, should utilise an ANOVA, Tukey's test and a t-test.

[298] Where the assumptions of an ANOVA are violated, a Kruskal-Wallis test (non-parametic ANOVA) or Welch's test (heterogeneous variance) can be used.

Chapter 9 – Sampling, Power Testing and Other Factors

9.1 – Introduction

This chapter provides a power analysis of the statistical tests discussed in the previous chapters, drawing on effect estimates from previous research. This analysis takes the form of *a priori* power analysis, and is used to inform sample *size* selection. Following this analysis, a short discussion of three exogeneous factors which may impact the data collected in this investigation is offered. These factors are the COVID-19 pandemic, the rise of populism, and the 2020 presidential election in the US.

9.2 – Sampling and Power Testing

While sample *population* has been addressed, the question of sample *size* remains outstanding. There are several indicative pieces of evidence to inform sample size selection. For instance, when discussing the Johnson-Neyman technique, Cronbach and Snow (1977) argue a minimum of 100 observations are required for moderation analysis. Additionally, when considering the technical limitations of the CAHOST program for calculating regions of significance, a *maximum* of 1,000 observations can be used.[299] Furthermore, following the review of RCT designs by BETA (2016), most RCT designs utilise samples between 1,000-10,000 observations in size, and the vast majority (84%) use samples between 100 and 10,000 observations in size. Finally, one must appreciate budget constraints associated with incentivised research which also limit the final sample size.

These pieces of information, however, are merely indicative. A more immediate approach to choosing an adequate sample size is to calculate the statistical power associated with the three means of analysis outlined above: moderated regression, t-test, and ANOVA. This is done using *a priori* power testing.

[299] These limitations can be overcome, but as above, there are relatively few statistical packages that are capable of performing the JN technique which are not, at present, disparate and outdated.

In an updated draft of their 2019 paper, Peer et al. (2020) provide some details on their own power analysis which can be used to estimate a sample size for the moderated regression analysis planned here. While they do not report the R-squared associated with their moderated regressions, they do report an estimated effect size[300] f^2 of 0.1.[301] Using G*Power v3.1.9.4 power calculation software for multiple regression (which follows the methodology of Cohen (1988) and Cohen et al. (2003)), an 80% power level as used by Peer et al. (2020) as well as an effect size of 0.1, three predictor variables following the SLMM (Hayes, 2018) and a probability level (p-value) of 0.05, the estimated sample size required for a moderated regression is 124. Note, this is similar to the rule-of-thump value given by Cronbach and Snow (1977). Given a hypothetical moderated regression sample of 124 would consist of observations from the control group and a nudge-treatment group, the total sample size estimate for the primer group is 310.[302]

G*Power calculation software for two-tailed t-test is used to calculate the estimated sample size for a two-tailed t-test. Cohen (1988) advises that a small effect size has a Cohen's d of 0.2, a medium 0.5, and a large 0.8.[303] However, by reviewing previous literature, a more accurate estimate of effect size, d, may be arrived at. Unfortunately, the necessary details for this calculation are absent from Page, Castleman and Meyer (2020), Schöning, Matt and Hess (2019), Hirsh, Kang and Bodenhausen (2012) and Moon (2002). Peer et al. (2020) do provide the necessary information, but only for a test of their impersonal nudge group versus the personalised nudge group, which they found to have a statistically significant but very small difference. Calculating the effect size using these details gives a value of $d = 0.2359$.[304]

---

300 f^2 is given by $\frac{R^2}{1-R^2}$

301 Soper (2020) notes that a value of 0.02 is considered a small effect, 0.15 is considered a medium effect, and 0.35 is considered a large effect. Given Peer et al. (2019) detect significant moderation effects in 47.5% of their tests (19 out of 40), one may estimate that the effect size to be around the medium value.

302 124 divided by 2 is 62. 62 multiplied by the number of subgroups (5) is 310.

303 Cohen's d for effect size is developed by Cohen (1988).

304 Using Cohen's (1988) equation: $d = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{(\sigma_1^2 + \sigma_2^2)}{2}}}$

Compared to Cohen's (1988) guidance, this would appear to be a relatively small effect, which corroborates the findings of Peer et al. (2019). When 0.2359 is used, the estimated sample size for an 80% power level is 568, or 284 per group. While this may be a valid estimate of the sample size required for the PTG data collection stage,[305] the basis of this calculation is not applicable to the primer group stage. Unfortunately, Peer et al. (2020) do not provide data to calculate an effect size for the primer group.[306] Assuming the effect size to be of a similar size to the estimated effect size used in the moderated regression power analysis (i.e. slightly less than a medium effect size), an effect size of 0.4 is used. At an 80% power level, this produces an estimated sample size of 200 observations, or 100 observations per group.[307] Based on this calculation, a total, minimum sample size of 500 observations is advisable.

Finally, once more using G*Power, an estimated sample size given a one-way ANOVA is examined. Again, none of the literature provide information from which an effect size f^2 can be estimated. Cohen (1988) suggests that a small size takes the value of 0.1, a medium the value of 0.25, and a large the value 0.4. As ANOVA is used in the PTG group, by way of an estimate the calculated effect size of 0.2359 using Peer et al. (2020) is used. For three groups and an 80% power level, this yields a sample size estimate of 177 total, or 59 observations per group. Moon (2002)[308] does report an effect size based on Cohen's (1988) *eta-squared* value of around 0.3873. Using this estimate, for an 80% power level, around 85 total observations would be needed, or 17 observations per group.

Both estimated minimum sample sizes for the primer group (310 and 500) fall within the typical sample range used by RCT experiments (BETA, 2016) and are similar to previously used sample sizes (see Table 1). Furthermore, the estimated minimum sample for the moderated

---

[305] This estimated sample size differs somewhat, also, from the estimated sample size for ANOVA, which is the primary means of analysis in the PTG group. See below.

[306] Peer et al. (2020) also use a WMW-test, presumably because of a non-normal independent variable. Using the same effect size (i.e. $d = 0.2359$), this produces a sample size estimate at the 80% power level of 297, very similar to the two-tailed t-test estimate (yet also larger, congruent with the expectation of lower power associated with non-parametric tests). As is done above, when a larger but still modest effect size of 0.4 is used, the estimated sample size at the 80% level is 104. This is also consistent.

[307] For prosperity, assuming a medium effect size of 0.5 yields an estimated sample size of 53.

[308] Moon (2002) reports multiple eta-squared values which average around 0.15.

regression analysis exceeds Cronbach and Snow's (1977) rule of thumb recommendation, while both minimum estimations are less than the 1,000-observation maximum imposed by CAHOST.

The estimated minimum sample sizes for the PTG data collection based on a two-tailed t-test (284 × 3 = 852), using Peer et al. (2020) as an estimator, is probably infeasible given budget limitations. The use of pilot studies allows the previous minimum sample estimations to be revised given the conditions of this specific project, and so the effect size found in the pilot studies may also be used to re-estimate the minimum sample estimation for the PTG data collection. Alternatively, the minimum sample estimate could follow from the power testing of the one-way ANOVA, which estimates a minimum sample of 177. This sample estimate is much more congruent with previous group estimates (62 and 100) and is within budget limitations. Finally, because a specifically *positive* effect is hypothesised from the PTG group, a one-tailed t-test could also be used. The one-tailed t-test has twice the power of the two-tailed t-test for the same test size, (assuming the direction of the test is correctly specified) so an 80% power level can be maintained using half the estimated minimum sample (i.e. N = 426, n = 142/subgroup). This is also feasible given various limitations.

These minimum sample estimates are summarised in Table 6 below:

Table 6: Minimum Sample Estimates

Test	Effect Size	N/subgroup (Primer)	N/subgroup (PTG)
Two-Tailed T-test	0.2359	284	284
Two-Tailed T-test	0.4000	100	100
One-Tailed T-test	0.2359	142	142
Moderated Regression	0.1000	62	-
One-way ANOVA	0.2359	-	59
One-way ANOVA	0.2500	-	53
One-way ANOVA	0.3873	-	17

A sample congruent with the estimation using the power analysis of the moderated regression is the priority when estimating the size of the sample to collect as moderated regression is the primary means of analysis for the primer group, and matching analysis – which uses the t-test – is only a secondary approach. However, given this would set an initial benchmark sample N of only 310, when – assuming a slightly less than medium effect size – a sample N of 500 would also accommodate the two-tailed t-test at the 80% power level, this higher sample size is set for the primer group sample. Estimates for the PTG vary greatly (from 177 to 852). At present, these estimates are used as a guide, with *post hoc* power analysis being performed on Pilot Study 2 as a means of establishing a firmer estimation.

9.3 – Other Factors

Research is often shaped by the times in which it is conducted. Several extenuating factors are present which could impact – to an extent which likely cannot be known – the research conducted during this experiment. Three factors are of pertinent consideration.

9.3.1 COVID-19

Firstly, all data collection is conducted during a period of global pandemic caused by the novel coronavirus and resulting COVID-19 disease. At the time of writing, the US is one of the worse affected countries, with the highest absolute number of both cases and deaths, and one of the highest per capita figures also (Sullivan, 2020). The US is, as expected, suffering the consequences of such a situation, with unemployment at its highest level in a decade (Badkar, Smith and Politi, 2020). In such an environment, it is reasonable to consider the impact on the current research. For instance, the present levels of risk and uncertainty arising from the pandemic may cause people to exhibit a preference for messages that promise security and certainty. As such, this may skew preferences towards, say, the loss aversion nudge. Methodologically, widespread unemployment coupled with easy access to Mechanical Turk can be expected to increase the user base and thus the number of individual participants who can be drawn upon to participate. Equally, financial strain of unemployment may incentivise less than honest responses in the hope of maximum revenue generation. However, this is likely an issue regardless of economic circumstances. Finally, the pandemic places a great onus, and subsequently scrutiny, on politicians and public officials. Of course, depending on one's political perspective, the response of a governing politician or opposing politician will vary. However, it would be wrong to not acknowledge that the current crisis may cause opinions about government to emerge which would not occur during times of relative normalcy.

9.3.2 Populism

Secondly, political scientists have noted over the past several years a rise in emotive, populist driven politics which stands in stark contrast (both in performance and policy) to the political landscape of the preceding decades (Blyth and Lonergan, 2020). Again, the impact of this changing political landscape is likely a highly subjective one, with some viewing various shifts as positive or negative. Regardless, though, the current political moment is one which is characterised by a sense of uncertainty (Blyth and Lonergan, 2020), which once more might be expected to impact the assessment of the political advertisements in this project.

9.3.3 2020 Presidential and General Election

Finally, the US is due to hold a presidential election in November 2020, several months prior to data collection. With the US Democratic primary race recently concluding (at the time of writing), one can expect that – amongst a not insignificant proportion of the American population – people are looking more critically at political commentary and materials as they consider how they will vote (or act) in the upcoming election. As above, this can be viewed positively in that data collection is conducted during a period where people can be expected to be making similar judgements of political advertisements, adding the realism of this experiment. But it could also be viewed to the detriment of this experiment, with people ascribing onto hypothetical advertisements very real emotions, concerns and complaints. This, to an extent, touches on the previous consideration. It is also – valid or not – one which should be viewed as something to embrace rather than a problem to resolve. If the findings of this thesis work only within a narrow set of circumstances (i.e. an *apolitical* circumstance, for the most part) the impact of this research is minimised.

Indeed, for all of these considerations, little can be done to rectify them. The best that can be done is to acknowledge them as potentially conflating factors, and to draw upon them critically in any subsequent analysis.

Chapter 10 – Summary of Methods

In summary, a review of the methods used in the prominent literature reveals several methodological directions to take when seeking to answer the hypotheses and research questions associated with this thesis, as well as several areas from which to improve and develop the methodology.

Most studies utilise incentivised survey-experiments distributed via Amazon's Mechanical Turk micro-tasking platform. This choice of data collection is subsequently adopted. The choice of nudges – where they can be reasonably called nudges – amongst the previous literature is sparse, and so four common nudges which seem reasonably suitable for the experiment at hand are selected. These are the status quo nudge, the present bias nudge, the loss aversion nudge, and the social norm nudge.

In conjunction with nudge selection, psychometric scales are also selected. Again, the literature is indicative but not definitive. Much of the literature utilises psychometric scales which are lacking specificity such as the Big Five personality scale. Following the use of the GDMS, NFC and CFC scales adopted by a most comparable study – Peer et al. (2019) – this experiment adopts these scales too. Through a review of the literature on psychological traits, nudges and these psychometric scales, the theoretical underpinnings for selecting these psychometric scales is demonstrated.

When it comes to actually nudging, political advertisements which embed the nudge within the campaign slogan are offered. For realism, two different political candidates are utilised, introducing aesthetic differences between the two advertisements beyond the differences introduced with the nudge. To control for this, the survey-experiment design is adapted to use an RCT design. The rationale here is that the aesthetic differences should in roughly half the instances support the nudging effect, and in the other half hinder the nudging effect, and so the net aesthetic effect should be around zero when randomised. To assess the effectiveness of the nudge, in each instance a participant will be presented with a nudge-advertisement and a control advertisement, with the difference in likely voting scores (effectiveness) being attributed to the presence of the nudge.

This stage of the data collection constitutes a primer group stage, where participants complete demographic questions and the psychometric questions before randomly – or *impersonally* – being nudged with one of the four nudges. One in five participants will be assigned to a control-control group which is used to assess the effectiveness of the nudges. The use of two stages of data collection – the impersonal primer group and the personalised treatment group – again follows from Peer et al. (2019).

As does the method of analysis. Key to personalising the delivery of nudges is identifying statistically significant relationships between the nudges and the psychometric variables. Within the literature, the most popular method of doing this has been what is dubbed here the matching approach, a deductive approach whereby data are artificially stratified along often arbitrary lines to determine whether, *in theory*, personalisation would work. The merits of this method have been offered in this Section, as well as the many criticisms which can be made of it. An alternative approach utilised by Peer et al. (2019) is moderation analysis, which uses moderated regression and the Johnson-Neyman technique to identify statistically significant relationships – and the *value-bounds of this significance* – which can then be used to personalise the delivery of nudges. Peer et al. (2019) still present some methodological shortcomings which solutions to have been offered here. However, in contrast with the matching approach, the moderation approach emerges as a superior technique.

Provided statistically significant relationships between the nudges and the psychometric variables have been identified, the personalised treatment data collection can begin. Here, participants are split into 3 subgroups: a choice only personalised subgroup who receive an impersonal nudge which supports a personalised outcome; a delivery only personalised subgroup who receive a personalised nudge which supports an impersonal outcome; and a choice and delivery personalised subgroup who receive a personalised nudge which supports a personalised outcome. Various changes to the functionality of the survey-experiment design are implemented to automate this process.

Delivery personalisation follows the findings of the primer group. Choice personalisation is implemented quite simply by asking respondents to choose from one of four policy areas which is most important to them. The nudge slogan is then adapted to include this policy area, in effect making it the choice personalised nudge. These three subgroups are analysed in conjunction with the impersonal nudging subgroups found in the primer group, as well as the control group also found in the primer group, using various tests of statistical difference. The results of these tests allow the hypotheses established in this thesis to be either rejected, accepted, or adapted.

For all analyses to be performed, appropriate power testing has been utilised to estimate a necessary minimum sample size. This estimate has also been interrogated using various suggestions and examples in the literature. Chapters 11 and 13 present the results of two pilot studies. These studies were conducted for three reasons. Firstly, to evaluate the effectiveness of the nudge-advertisements and the RCT design to determine if any changes need to be made. Secondly, to evaluate the effectiveness of the moderated and matching approaches in identifying statistically significant relationships. Thirdly, to establish further data to be used in power-testing estimates of minimum sample sizes.

Section 3:

Results

Chapter 11 – Pilot Study 1

11.1 – Introduction

In this chapter, data from an initial pilot study (Pilot Study 1) is analysed, and the results discussed. The structure of this chapter is as follows. Firstly, a data summary is offered, before the assumptions of the various proposed statistical tests are investigated. Secondly, the effectiveness of the nudge-advertisements when used impersonally is analysed, as well as the presence of an aesthetic effect owing to aesthetic differences in the advertisements. Following from these results, thirdly moderation analysis is undertaken. Matching analysis is then performed, before finally a conclusion is offered.

The results of Pilot Study 1 are mixed and several areas for development of the experimental approach are revealed. Some of these developments are discussed in this chapter, with the implications of these ideas largely discussed in Chapter 12.

11.2 – Data Summary

A sample of 100 participants from the US were recruited using Amazon's *Mechanical Turk* (MTurk) service and were compensated $0.65 for their participation. After removing responses registered as being completed inappropriately fast (less than 2 minutes) and responses with suspicious results (a standard deviation across *all* psychometric questions of less than 0.2),[309] a sample of N = 95 remained (female = 44%).

Summary statistics are shown in Table 7:

[309] This method of 'cleaning' the data is subsequently determined to be inadequate as it is far too arbitrary and potentially ineffective. The rationale for using a standard deviation measure is that, for disingenuous respondents who choose to answer all questions with the same response, the standard deviation of their responses should be near zero. However, determining how close to zero is so close as to justify removing from the sample is an arbitrary decision. Furthermore, as a method of cleaning the data, there is no reason to believe it is comprehensive. A respondent who alternates their answers between the maximum and minimum scorings could be described as just as disingenuous. However, the resulting standard deviation may be quite similar to a sincere respondent. For these reasons, an attention check is utilised hereinafter.

Table 7: Pilot Study 1 Summary Statistics

Demographic	Frequency	Percentage of N	Average
Education:			3.189
(1) None	0	0	
(2) Highschool	8	8.42%	
(3) Bachelor's Degree	62	65.26%	
(4) Master's Degree	24	25.26%	
(5) PhD	1	1.05%	
Political Identity:			2.895
(1) Left-wing	16	16.84%	
(2) Left-leaning	18	18.95%	
(3) Centre	30	31.58%	
(4) Right-leaning	22	23.16%	
(5) Right-wing	9	9.47%	
Age:			36.789
18-25	16	16.84%	
26-35	37	38.95%	
36-45	10	10.53%	
46-55	26	27.37%	
55<	6	6.32%	

From Table 7, it can be seen that the median respondent is around 37 years of age, is qualified slightly beyond a bachelor's degree, and identifies slightly left of the political centre. A chi-squared test for differences in distribution is used to examine whether the control group and four treatment groups are comparable demographically. No statistically significant difference between the control group and the four treatment groups was found for age (χ^2 (136, N = 95) = 126.0, p = .72), sex (χ^2 (4, N = 95) = 1.5, p = .83), education (χ^2 (12, N = 95) = 9.8, p = .63) or political identity (χ^2 (16, N = 95) = 15.8, p = .47). The control and treatment groups, therefore, appear comparable.

11.3 – Testing of Assumptions

It is first prudent to examine whether the four nudges under consideration appear to be effective or not when used impersonally, which is to say, when participants are randomly assigned to a particular nudge-advertisement. The data are, therefore, first examined to test if there is any violation of the assumptions of a t-test (see Chapter 8).

11.3.1 Normality

The normality of the data is first reviewed.

Figure 8: Pilot Study 1 Histogram Normality Plots

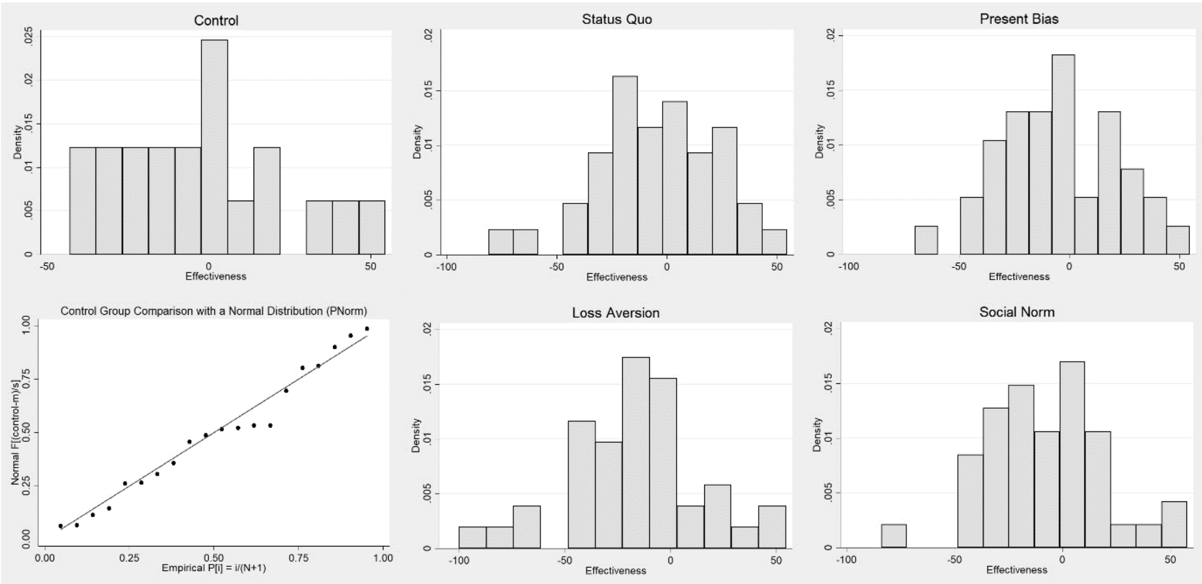


Figure 8 shows histogram plots for each group, with the dependent variable (effectiveness) shown on the x-axis and frequency density shown on the y-axis. All four of the nudge groups have a reasonably normal distribution. The control group diverges from this trend in the histogram, but when plotted against a normal distribution function (bottom-left plot), the data points follow quite closely with the expected normal distribution pattern. Finally, Shapiro-Wilk's test for normality justifies the assumption of normality throughout (control p = 0.6693; status quo p = 0.2411; present bias p = 0.4066; loss aversion p = 0.3448; social norm p = 0.5763).

11.3.2 Homogeneity of Variance

Using Levene's test to investigate the presence of heterogeneity of variance between the control group and each of the four nudge groups reveals no statistically significant evidence of heterogeneous variance (status quo p = 0.6978; present bias p = 0.5787; loss aversion p = 0.5660; social norm p = 0.8756).

The assumptions of a t-test appear to be satisfied and thus no adjustment is made.

11.4 – Impersonal Nudging

The results of these two-tailed t-tests examining the effectiveness of the nudges when used impersonally are presented in Table 8.[310]

Table 8: Pilot Study 1 T-test Results of Impersonal Nudges

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value
Status Quo	-2.170	-8.239	0.6615	0.5125
Present Bias	-2.170	-5.388	0.3584	0.7222
Loss Aversion	-2.170	-31.715	3.2788	0.0022***
Social Norm	-2.170	-13.300	1.2616	0.2148

* p < 10%, ** p < 5%, *** p < 1%

As Table 8 shows, only the loss aversion nudge produces effectiveness scores which differ significantly from the control group. Furthermore, through an examination of the means, the loss aversion nudge is significantly effective compared to the control group, however the sign of the effect is opposite to that which would be expected. In fact, this trend is seen in all the nudges examined.

Three possible explanations emerge. Firstly, there may be a significant aesthetic effect which randomisation has not counteracted. Secondly, there may be moderation effects which, when accounted for, would improve the apparent effectiveness of the nudges. Thirdly, the nudges themselves may simply be ineffective, and/or the control slogan may be having an unexpected influence on participants.

Fortunately, through analysis of the data, the first and second explanation can be examined. Should neither provide an explanation for the apparent ineffectiveness of the nudges, the third explanation seems likely to be correct, and a re-examination of the survey-experiment will be necessary.

11.5 – Testing for the Presence of Aesthetic Effects

[310] Two-tailed tests are used here to avoid the imposition of *a priori* assumptions.

Aesthetic effects may be examined through comparison within the groups, comparing the effectiveness scores for Candidate A to that of Candidate B, where – should no aesthetic effects be present – it would be expected that no statistically significant difference is found. The results of these comparisons are shown in Table 9:

Table 9: Pilot Study 1 Two-tailed T-test Results for Aesthetic Effects

Nudge	Mean (Candidate A)	Mean (Candidate B)	t-Statistic	p-value
Control	56.125	58.295	-0.3017	0.7645
Status Quo	50.000	57.181	-1.0912	0.2839
Present Bias	55.176	60.565	-0.7510	0.4582
Loss Aversion	40.100	71.815	-4.7634	0.0000***
Social Norm	50.140	63.440	-2.0551	0.0468**
* p < 10%, ** p < 5%, *** p < 1%				

The results shown in Table 9 demonstrate a mixed picture. When comparing advertisements in the control group, the status quo group and the present bias group, no statistically significant aesthetic effect is found. However, the loss aversion and social norm groups both suggest a statistically significant aesthetic effect is present, perhaps explaining – at least for the loss aversion nudge – the statistically significant difference found when compared to the control group (see Table 8). Furthermore, even where statistically insignificant, the mean effectiveness associated with the Candidate B advertisement is consistently greater than that associated with the Candidate A advertisement, suggesting – regardless of the nudge – that Candidate B is more popular.

Of course, the criteria for assuming an aesthetic effect of zero under the RCT design is a large sample size and a roughly equal number of observations in each group being compared. As this is a pilot study, the first criterion clearly fails. On the second criterion, however, all groups have an equal number of observations in the compared subgroups. Therefore, a biased random allocation cannot explain any significant difference identified. At least in the case of the loss aversion and social norm nudges, therefore, these results suggest there *may* be an aesthetic effect occurring, though this is inconclusive.

11.6 – Moderation Analysis

As above, it is feasible that the lack of a positive effect arising from the impersonal nudges is due to moderation effects. This is to say, nudges which would be positively effective for an individual who exhibits one set of traits were shown to an individual who exhibits a set of traits which render the nudge ineffective or negatively effective. If evidence of this can be found through moderation analysis, this may provide a route to effective, positive nudging via personalisation.

One conflating factor for moderating regression may be statistically significant differences in psychometric responses across groups. These differences are examined and p-values from associated t-tests and chi-squared tests of differences in distribution are reported in Table 10:

Table 10: Pilot Study 1 Differences in Psychometric Scores

Nudge	Rational	Avoidant	Intuitive	Dependent	Spontaneous	NFC	CFC
Status Quo	0.9052	0.5408	0.0452**	0.4550	0.3673	0.4460	0.8742
	(0.378)	(0.422)	(0.452)	(0.252)	(0.375)	(0.567)	(0.372)
Present Bias	0.8740	0.8927	0.3966	0.9729	0.5936	0.3526	0.4927
	(0.377)	(0.452)	(0.451)	(0.357)	(0.328)	(0.469)	(0.324)
Loss Aversion	0.1510	0.2073	0.2583	0.4250	0.6234	0.5918	0.9022
	(0.375)	(0.261)	(0.378)	(0.338)	(0.338)	(0.516)	(0.325)
Social Norm	0.0900*	0.7664	0.2953	0.2841	0.6059	0.4513	0.4690
	(0.292)	(0.420)	(0.417)	(0.378)	(0.402)	(0.516)	(0.334)

χ^2 p-value shown in brackets, N = 95
* p < 10%, ** p < 5%, *** p < 1%

Across the sample, only two instances of statistically significant differences between the control group and a nudge group are identified. Firstly, the means of the intuitive psychometric score are statistically significantly different across the control and status quo group. However, this variable is not normally distributed, with a peak at around 5 (the high end of the variable). As a result, a t-test may be unsuitable for evaluating difference. Given this, an examination of the differences in distributions is undertaken. Here, the chi-squared value is not significant.

Secondly, the means of the rational psychometric score are statistically significantly different across the control and social norm group, but only at the 10% level. In this instance, the rational variable is normally distributed, and so a t-test appears suitable. However, as with the difference associated with the intuitive variable, no statistically significant difference in distribution is found.

Given there are very few statistically significant differences, and where these exist, they are inconsistent across tests, no adjustment is made in the analysis.

To examine moderation effects, each of the seven psychometric scales discussed previously are averaged, such that each respondent is associated with a single value for, say, NFC or intuition. The validity of this process is assured by measuring Cronbach's alpha for each psychometric scale. These results are shown in Table 11:

Table 11: Pilot Study 1 Cronbach's Alpha Results for Psychometric Variables

Scale	Cronbach's Alpha
GDMS: Avoidant	0.9425
GDMS: Dependent	0.8791
GDMS: Intuitive	0.8454
GDMS: Rational	0.8920
GDMS: Spontaneous	0.9195
NFC	0.8940
CFC	0.8319

As each scale demonstrates a Cronbach's alpha value which is greater than a typically accepted threshold of around 0.6-0.7, averaging each of these scales appears wholly valid. Summary statistics of these psychometric variables are shown in Table 12:

Table 12: Pilot Study 1 Summary Statistics of Psychometric Variables

Psychometric	Mean	Std. Dev.	Min	Max	Median
Rational	3.792	0.726	2.100	5.000	3.800
Avoidant	2.706	1.093	1.000	5.000	2.980
Intuitive	3.361	0.749	1.020	5.000	3.560
Dependent	3.117	0.861	1.000	4.800	3.220
Spontaneous	2.968	0.955	1.000	4.520	3.140
NFC	3.263	0.653	1.117	5.000	3.039
CFC	3.283	0.457	2.575	4.667	3.108

Using these average figures, moderated regression models taking the form:

Equation 8

$$effectiveness_i = \beta_0 + \beta_1 D_i + \beta_2 Psy_\lambda + \beta_3 D_i Psy_\lambda + \varepsilon_i \tag{8}$$

where $effectiveness_i$ is the effectiveness of nudge i, D_i is a dummy variable taking the value of 1 for nudge i, and 0 for all other values, psy_λ is a continuous variable for psychometric λ, and $D_i Psy_\lambda$ is a moderator term, are estimated. Table 13 through Table 16 present the results of moderated regressions for all possible combinations of nudge and psychometric variable.

Table 13: Pilot Study 1 Moderated Regression Results for the Status Quo Nudge

Variable	Model 1A	Model 2A	Model 3A	Model 4A	Model 5A	Model 6A	Model 7A
Dummy	-31.222 [48.481]	-31.913 [34.938]	-30.864 [39.899]	-14.774 [39.804]	-8.966 [33.494]	-44.741 [52.891]	-35.171 [74.646]
Rational	-0.531 [10.306]						
D × Rat.	6.946 [14.648]						
Avoidant		1.837 [7.231]					
D × Avo.		9.974 [11.761]					
Intuitive			-12.551 [9.003]				
D × Intu.			6.15945 [12.426]				
Dependent				1.072 [8.954]			
D × Dep.				2.588 [12.233]			
Spontaneous					0.176 [7.365]		
D × Spon.					1.087 [10.511]		
NFC						-6.303 [10.973]	
D × NFC						11.666 [17.058]	
CFC							1.651 [17.304]
D × CFC							8.929 [23.753]
Constant	-0.231 [34.973]	-7.362 [22.746]	41.770 [32.566]	-5.458 [30.555]	-2.697 [24.391]	19.218 [34.901]	-7.503 [53.956]
R-squared	0.0295	0.0824	0.0615	0.0175	0.0130	0.0322	0.0241
Multicollinearity	0.9996	0.9895	0.8932	0.9844	0.9774	0.9838	0.9993
N	38	38	38	38	38	38	38

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 14: Pilot Study 1 Moderated Regression Results for the Present Bias Nudge

Variable	Model 1B	Model 2B	Model 3B	Model 4B	Model 5B	Model 6B	Model 7B
Dummy	27.736	-5.527	-30.626	24.138	27.067	-42.549	29.703
	[47.318]	[28.233]	[45.495]	[34.426]	[38.159]	[72.153]	[68.923]
Rational	-0.532						
	[10.322]						
D × Rat.	-8.398						
	[13.845]						
Avoidant		1.837					
		[7.233]					
D × Avo.		0.862					
		[9.107]					
Intuitive			-12.551				
			[9.017]				
D × Intu.			-10.656				
			[13.506]				
Dependent				1.072			
				[8.969]			
D × Dep.				-8.895			
				[10.122]			
Spontaneous					0.176		
					[7.377]		
D × Spon.					-9.597		
					[11.753]		
NFC						-6.303	
						[10.990]	
D × NFC						11.901	
						[22.281]	
CFC							1.651
							[17.331]
D × CFC							-9.906
							[21.246]
Constant	-0.231	-7.362	41.770	-5.458	-2.697	19.218	-7.503
	[35.028]	[22.782]	[32.618]	[30.604]	[24.430]	[34.957]	[54.042]
R-squared	0.0286	0.0114	0.0788	0.0252	0.0438	0.0222	0.0128
Multicollinearity	0.9993	0.9995	0.9794	1.0000	0.9918	0.9753	0.9865
N	37	37	37	37	37	37	37

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 15: Pilot Study 1 Moderated Regression Results for the Loss Aversion Nudge

Variable	Model 1C	Model 2C	Model 3C	Model 4C	Model 5C	Model 6C	Model 7C
Dummy	-49.2554 [55.988]	-52.814** [22.503]	-57.548 [50.448]	-60.2174** [29.070]	-63.493** [27.608]	26.016 [44.154]	-63.239 [63.364]
Rational	-0.532 [10.446]						
D × Rat.	5.031 [14.542]						
Avoidant		1.837 [5.950]					
D × Avo.		10.192 [7.879]					
Intuitive			-12.551 [12.529]				
D × Intu.			7.617 [14.443]				
Dependent				1.072 [7.000]			
D × Dep.				10.975 [9.358]			
Spontaneous					0.176 [6.549]		
D × Spon.					11.976 [8.955]		
NFC						-6.303 [8.519]	
D × NFC						-17.180 [13.021]	
CFC							1.651 [12.804]
D × CFC							10.360 [19.346]
Constant	-0.231 [38.661]	-7.362 [17.885]	41.770 [44.330]	-5.458 [22.353]	-2.697 [20.575]	19.218 [29.530]	-7.503 [41.860]
R-squared	0.2248	0.3242	0.2512	0.2947	0.2977	0.3356	0.2354
Multicollinearity	0.9465	0.9585	0.9665	0.9832	0.9936	0.9924	0.9996
N	40	40	40	40	40	40	40

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 16: Pilot Study 1 Moderated Regression Results for the Social Norm Nudge

Variable	Model 1D	Model 2D	Model 3D	Model 4D	Model 5D	Model 6D	Model 7D
Dummy	-2.704 [48.513]	-36.776 [29.471]	-82.085 [50.602]	-68.857* [35.798]	-19.542 [39.980]	9.576 [41.474]	79.894 [63.936]
Rational	-0.532 [10.275]						
D × Rat.	-2.042 [13.900]						
Avoidant		1.837 [7.210]					
D × Avo.		8.677 [8.754]					
Intuitive			-12.551 [8.976]				
D × Intu.			19.872 [14.717]				
Dependent				1.072 [8.928]			
D × Dep.				17.109 [10.840]			
Spontaneous					0.176 [7.344]		
D × Spon.					2.663 [11.422]		
NFC						-6.303 [10.940]	
D × NFC						-6.826 [13.500]	
CFC							1.651 [17.253]
D × CFC							-27.282 [20.351]
Constant	-0.231 [34.870]	-7.362 [22.680]	41.770 [32.471]	-5.458 [30.465]	-2.697 [24.319]	19.218 [34.799]	-7.503 [52.798]
R-squared	0.0425	0.1415	0.0789	0.1595	0.0442	0.1251	0.1258
Multicollinearity	0.9262	0.9976	0.9712	0.9699	0.9929	0.9850	0.9861
N	40	40	40	40	40	40	40

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

As Table 13 through Table 16 demonstrate, no statistically significant moderation effects are identified, and as such, moderation is not a sufficient explanation for the lack of a positive effect from the impersonal nudges. As with the analysis of aesthetic effects, a larger sample may provide greater insights, but even still, the apparent lack of any moderation effects is compelling.

Furthermore, given this lack of moderation effects, further analysis by way of floodlight analysis is not performed.

11.7 – Matching Analysis

Given the consistent lack of moderation effects identified using moderation analysis, the alternative matching analysis is also utilised. This approach uses dummy variables to categorise respondents as being either ‘high’ or ‘low’ scorers on various psychometric variables, and then evaluates the effectiveness scores of the nudge across these manufactured groups. Where a statistically significant difference in effectiveness scores is identified, this difference can be attributed to a tendency to score either ‘high’ or ‘low’ on the psychometric variable used in the analysis.

Determining, however, what constitutes ‘high’ and ‘low’ remains arbitrary. To alleviate such arbitrary selection, a range of constructions of high and low scorers are used: 1) a mean construction, where a person is considered a high-scorer (D = 1) if they score more than the mean, and a low-scorer if they score less than the mean (D = 0); 2) a median construction, where a person is considered a high-scorer (D = 1) if they score within the top 50% of respondents, and a low-scorer if they score within the bottom 50% of respondents (D = 0); 3) and a midpoint construction, where a person is considered a high-scorer (D = 1) if they score above 3 (the midpoint), and a low-scorer (D = 0) if they score below 3.

Evidence of a relationship between a nudge and a psychometric can be determined using the matching analysis via several steps. Firstly, a statistically significant difference across the control and treatment group under examination informs later interpretations. Secondly, a statistically significant difference across high/low subgroups *within* the treatment group identifies possible relationships. Where a statistically significant difference is found, a relationship may be present. Finally, a statistically significant difference between a subgroup and the control group, with a *prior* statistically significant difference between subgroups, provides good evidence to conclude a relationship exists between the psychometric variable and the nudge.

11.7.1 Status Quo Nudge

The results of a two-tailed t-test examining the effectiveness of the status quo nudge across the control and treatment groups is shown in Table 17 (also see Table 8):

Table 17: Pilot Study 1 T-test results for the Status Quo Nudge vs. the Control Group

Nudge	Mean (Control)	Mean (Treatment)	p-value
Status Quo Nudge	-2.170	-8.239	0.5125
N = 38			
* p < 10%, ** p < 5%, *** p < 1%			

As can be seen, the effectiveness score of the status quo nudge group indicates the behaviour of the nudged group is not statistically significantly different from that of the control group. However, this result does not consider that the susceptibility of some participants to the status quo nudge may be moderated by various psychometric effects. As an initial effort to account for these effects, the effectiveness ratings of those in the treatment group who exhibit high psychometric scores are examined against those in the same group who exhibit low psychometric scores. These results are shown in Table 18:

Table 18: Pilot Study 1 T-test Results for the Status Quo Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	-19.763	0.980	0.1494	-18.656	2.178	0.1448	-16.350	-5.921	0.5553
Avoidant	-11.791	-2.657	0.5446	-15.800	-0.678	0.2979	-11.791	-2.657	0.5446
Intuitive	-3.522	-12.956	0.5207	-3.522	-12.956	0.5207	-3.522	-12.956	0.5207
Dependent	-19.000	0.370	0.1802	-13.167	-3.311	0.5019	-10.783	-6.967	0.8075
Spontaneous	-11.250	-4.475	0.6477	-10.167	-6.311	0.7940	-11.250	-4.475	0.6477
NFC	-6.783	-11.150	0.7804	-6.922	-9.556	0.8586	-10.686	-6.682	0.7915
CFC	-6.600	-10.288	0.8040	-5.422	-11.056	0.7025	-5.480	-9.300	0.8168
N = 18									
* p < 10%, ** p < 5%, *** p < 1%									

Consistent with the lack of statistically significant effects in the moderation analysis, regardless of how a ‘high’ and a ‘low’ scoring respondent is constructed, no statistically significant differences between these subgroups are identified. This is re-affirmed when the high/low treatment subgroups are compared once more with the control group:

Table 19: Pilot Study 1 T-test Results for High (Low) Status Quo Nudge Subgroups vs. Control

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-2.170	0.980	0.7814	-2.170	2.178	0.7161	-2.170	-5.921	0.7177
Low	-2.170	-19.763	0.1087	-2.170	-18.656	0.1108	-2.170	-16.350	0.3149
Avoidant									
High	-2.170	-2.657	0.9686	-2.170	-0.678	0.8952	-2.170	-2.657	0.9686
Low	-2.170	-11.791	0.3631	-2.170	-15.800	0.2232	-2.170	-11.791	0.3631
Intuitive									
High	-2.170	-12.956	0.3776	-2.170	-12.956	0.3776	-2.170	-12.956	0.3776
Low	-2.170	-3.522	0.8958	-2.170	-3.522	0.8958	-2.170	-3.522	0.8958
Dependent									
High	-2.170	0.370	0.8213	-2.170	-3.311	0.9213	-2.170	-6.967	0.6642
Low	-2.170	-19.000	0.1313	-2.170	-13.167	0.3204	-2.170	-10.783	0.4689
Spontaneous									
High	-2.170	-4.475	0.8428	-2.170	-6.311	0.7071	-2.170	-4.475	0.8428
Low	-2.170	-11.250	0.4129	-2.170	-10.167	0.4928	-2.170	-11.250	0.4129
NFC									
High	-2.170	-11.150	0.5268	-2.170	-9.556	0.5372	-2.170	-6.682	0.6812
Low	-2.170	-6.783	0.6318	-2.170	-6.922	0.6567	-2.170	-10.686	0.4738
CFC									
High	-2.170	-10.288	0.5085	-2.170	-11.056	0.4430	-2.170	-9.300	0.5001
Low	-2.170	-6.600	0.6744	-2.170	-5.422	0.7693	-2.170	-5.480	0.8006

* p < 10%, ** p < 5%, *** p < 1%

11.7.2 Present Bias Nudge

The results of a two-tailed t-test examining the effectiveness of the present bias nudge across the control and treatment groups is shown in Table 20 (also see Table 8):

Table 20: Pilot Study 1 T-test Results for the Present Bias Nudge Group vs. the Control Group

Nudge	Mean (Control)	Mean (Treatment)	p-value
Present Bias Nudge	-2.170	-5.388	0.7222

N = 37
* p < 10%, ** p < 5%, *** p < 1%

As can be seen, the effectiveness score of the present bias group indicates the behaviour of the nudged group is not statistically significantly different from that of the control group.

However, this result does not consider that the susceptibility of some participants to the present bias nudge may be moderated by various psychometric effects. Again, as an attempt to account for these effects, the effectiveness ratings of those in the treatment group who exhibit high psychometric scores are examined against those in the same group who exhibit low psychometric scores. These results are shown in Table 21:

Table 21: Pilot Study 1 T-test Results for the Present Bias Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	5.171	-12.780	0.2013	-2.544	-8.588	0.6705	9.625	-10.008	0.2299
Avoidant	-5.350	-5.409	0.9968	-11.433	1.413	0.3603	-14.900	3.067	0.1942
Intuitive	9.000	-15.460	0.0739*	7.300	-19.663	0.0425**	12.417	-15.100	0.0482**
Dependent	-8.550	-2.578	0.6741	-4.411	-6.488	0.8841	6.040	-10.150	0.2902
Spontaneous	5.388	-14.967	0.1379	1.167	-12.763	0.3200	5.150	-11.136	0.2634
NFC	-6.277	-2.500	0.8216	-2.367	-8.788	0.6511	-2.500	-7.410	-0.7335
CFC	-8.158	1.260	0.5436	-3.033	-8.038	0.7248	17.925	-12.562	0.0526*

N = 17
* p < 10%, ** p < 5%, *** p < 1%

Contrary to patterns thus far, some evidence of a statistically significant relationship between intuition and the present bias nudge is found. Regardless of how the dummy variable demarcating 'high' and 'low' scorers is constructed, intuition consistently appears to have a statistically significant, negative relationship with the effectiveness of the present bias nudge. This finding is weakest when the dummy variable is constructed using the mean, being only significant at the 10% level. However, given the absence of statistical significance found using moderation analysis, the relative consistency of this result is interesting.

When the high/low treatment subgroups are compared once more with the control group, however, no statistically significant difference is identified:

Table 22: Pilot Study 1 T-test Results for High (Low) Present Bias Nudge Subgroups vs. Control

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-2.170	-12.780	0.3358	-2.170	-8.588	0.5956	-2.170	-10.008	0.4316
Low	-2.170	5.171	0.5166	-2.170	-2.544	0.9715	-2.170	-9.625	0.4118
Avoidant									
High	-2.170	-5.409	0.7587	-2.170	1.143	0.743	-2.170	3.067	0.6123
Low	-2.170	-5.350	0.7992	-2.170	-11.433	0.4195	-2.170	-14.900	0.2888
Intuitive									
High	-2.170	-15.460	0.2225	-2.170	-19.663	0.1333	-2.170	-15.100	0.2128
Low	-2.170	9.000	0.3204	-2.170	7.300	0.3606	-2.170	12.417	0.2257
Dependent									
High	-2.170	-2.578	0.9692	-2.170	-6.488	0.6920	-2.170	-10.150	0.4313
Low	-2.170	-8.550	0.5949	-2.170	-4.411	0.8467	-2.170	6.040	0.5389
Spontaneous									
High	-2.170	-14.967	0.2505	-2.170	-12.763	0.3642	-2.170	-11.136	0.3896
Low	-2.170	5.388	0.4952	-2.170	1.167	0.7565	-2.170	5.150	0.5562
NFC									
High	-2.170	-2.500	0.9823	-2.170	-8.788	0.5487	-2.170	-7.410	0.6013
Low	-2.170	-6.277	0.6779	-2.170	-2.367	0.9863	-2.170	-2.500	0.9795
CFC									
High	-2.170	1.260	0.7983	-2.170	-8.038	0.5942	-2.170	-12.562	0.2914
Low	-2.170	-8.158	0.5571	-2.170	-3.033	0.9401	-2.170	17.925	0.1549

* p < 10%, ** p < 5%, *** p < 1%

11.7.3 Loss Aversion Nudge

The results of a two-tailed t-test examining the effectiveness of the loss aversion nudge across the control and treatment groups is shown in Table 23 (also see Table 8):

Table 23: Pilot Study 1 T-test Results for the Loss Aversion Nudge Group vs. the Control Group

Nudge	Mean (Control)	Mean (Treatment)	p-value
Loss Aversion Nudge	-2.170	-31.715	0.0022***

N = 40
* p < 10%, ** p < 5%, *** p < 1%

Unlike the previously examined nudges, the average effectiveness score for the loss aversion group indicates the behaviour of the nudged group is statistically significantly different from that of the control. However, the sign of this effect is counter to expectations – where the nudge was expected to lead to a *higher* average effectiveness score for the loss aversion group, compared to the control group, the nudge actually lead to a *lower* average effectiveness score.

As above, considering the role of psychometrics may explain this result. The effectiveness of the nudge amongst those with high psychometric scores is compared to those associated those with low psychometric scores. These results are shown in Table 24:

Table 24: Pilot Study 1 T-test Results for the Loss Aversion Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational[311]	-35.060	-28.370	0.6340	-35.060	-28.370	0.6340	-27.714	-32.747	n/a
Avoidant	-45.125	-11.600	0.0108**	-39.170	-24.260	0.2823	-42.585	-11.529	0.0240**
Intuitive	-34.770	-28.660	0.6639	-34.770	-28.660	0.6639	-25.957	-34.815	0.5469
Dependent	-44.989	-20.855	0.0749*	-42.970	-20.460	0.0969*	-37.245	-24.956	0.3805
Spontaneous	-43.020	-20.410	0.0953*	-43.02	-20.41	0.0953*	-42.718	-18.267	0.0708*
NFC	-27.038	-50.425	0.1730	-22.520	-40.910	0.1809	-12.667	-35.076	0.2473
CFC	-26.742	-39.175	0.3824	-28.810	-34.620	0.6795	-27.714	-33.869	0.6763

N = 20
* p < 10%, ** p < 5%, *** p < 1%

As with the present bias nudge, some evidence emerges from this analysis which suggests several relationships between psychometric variables and the effectiveness of the loss aversion nudge. While statistically significant only at the 10% level, spontaneity demonstrates a statistically significant, positive relationship with effectiveness across all dummy constructions.

It is noteworthy, however, that even amongst those with high spontaneity, the loss aversion nudge is still considered rather less effective than the control advertisement (with the loss aversion advertisement consistently having an effectiveness of around -20 for the high spontaneity treatment respondents, compared to the average effectiveness of the control advertisement of around -2). This observation adds to mounting evidence that – regardless of relationships – the implementation of the nudges within advertisements is insufficient.

[311] There are insufficient observations to perform a t-test for the rational decision-making style under the midpoint construction.

A second potential relationship is a positive, significant relationship between avoidance and the loss aversion nudge. While statistically insignificant when the median construction is tested, there is a statistically significant and positive relationship between avoidance and effectiveness at the 5% level using the mean and midpoint constructions. Finally, though somewhat less compelling still, dependence appears to have a positive relationship with the loss aversion nudge, though it is statistically significant only at the 10% level, and is not statistically significant under the midpoint construction. Again, while positive, neither avoidance nor dependence appear to produce effectiveness scores which would see the loss aversion nudge be *more positively effective* than the control advertisement, supporting the hypothesis that the underlying integration of the nudge into the advertisement is inadequate.

When the high/low treatment subgroups are compared with the control group, many occasions of statistically significant difference are identified. However, given the treatment group is overall statistically significantly different from the control group, and the high/low treatment subgroups are generally not statistically significantly different, these results are not surprising:

Table 25: Pilot Study 1 T-test Results for High (Low) Loss Aversion Nudge Subgroups vs. Control

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-2.170	-28.370	0.0237**	-2.170	-28.370	0.0237**	-2.170	-32.747	0.0020***
Low[312]	-2.170	-35.060	0.0050***	-2.170	-35.060	0.0050***	-2.170	-27.714	n/a
Avoidant									
High	-2.170	-11.600	0.3462	-2.170	-24.260	0.0454**	-2.170	-11.529	0.3823
Low	-2.170	-45.125	0.0003***	-2.170	-39.170	0.0022***	-2.170	-45.585	0.0004***
Intuitive									
High	-2.170	-28.660	0.0208**	-2.170	-28.660	0.0208**	-2.170	-34.815	0.0033***
Low	-2.170	-34.770	0.0059***	-2.170	-34.770	0.0059***	-2.170	-25.957	0.0563*
Dependent									
High	-2.170	-20.855	0.0896*	-2.170	-20.460	0.1126	-2.170	-24.956	0.0541*
Low	-2.170	-44.989	0.0003***	-2.170	-42.970	0.0003***	-2.170	-37.245	0.0021***
Spontaneous									
High	-2.170	-20.410	0.0602*	-2.170	-20.41	0.0602*	-2.170	-18.267	0.1086
Low	-2.170	-43.020	0.0016***	-2.170	-43.020	0.0016***	-2.170	-42.718	0.0011***
NFC									
High	-2.170	-50.425	0.0045***	-2.170	-40.910	0.0009***	-2.170	-35.076	0.0006***
Low	-2.170	-27.038	0.0105**	-2.170	-22.520	0.0741*	-2.170	-12.667	0.5755

[312] There are insufficient observations to perform a t-test for the rational decision-making style under the midpoint construction for the low psychometric subgroup.

CFC									
High	-2.170	-39.175	0.0061***	-2.170	-34.620	0.0075***	-2.170	-33.869	0.0046***
Low	-2.170	-26.742	0.0158**	-2.170	-28.810	0.0168**	-2.170	-27.714	0.0394**

* p < 10%, ** p < 5%, *** p < 1%

In every instance where a statistically significant difference is identified, the average effectiveness of the loss aversion nudge is less than that of the control group. This would suggest that even amongst those high in psychometric traits which should find the loss aversion nudge appealing (or low in psychometric traits which should find the loss aversion nudge unappealing), the nudge itself is negatively effective compared to the control advertisement. This adds to mounting evidence that implementation of the nudge, potentially in conjunction with an aesthetic effect or unintended influence from the control advertisement, is inadequate, even when psychometric effects are accounted for.

11.7.4 Social Norm Nudge

Finally, the results of a two-tailed t-test examining the effectiveness of the social norm nudge across the control and treatment groups is shown in Table 26 (also see Table 8):

Table 26: Pilot Study 1 T-test Results for the Social Norm Treatment Group

Nudge	Mean (Control)	Mean (Treatment)	p-value
Social Norm Nudge	-2.170	-13.300	0.2148

N = 40
* p < 10%, ** p < 5%, *** p < 1%

Consistent with previous results, the effectiveness score of the social norm nudge is not statistically significantly different from the control group, and the mean effectiveness is less than the control group. As above, the difference in effectiveness scores between those high in a psychometric variable and those low in a psychometric variable within the treatment group are examined to further elucidate the effectiveness of the social norm nudge:

Table 27: Pilot Study 1 T-test Results for the Social Norm Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	-15.420	-11.180	0.7544	-15.420	-11.180	0.7544	-16.900	-12.900	0.8596
Avoidant	-25.343	-6.815	0.1816	-18.690	-7.910	0.4227	-25.038	-5.475	0.1455
Intuitive	-14.456	-12.355	0.8775	-10.930	-15.670	0.7264	-11.500	-13.900	0.8782
Dependent	-19.238	-9.342	0.4717	-22.770	-3.830	0.1507	-23.900	-8.757	0.2990
Spontaneous	-13.678	-12.991	0.9598	-11.970	-14.630	0.8445	-12.525	-13.817	0.9257
NFC	-10.164	-20.617	0.4771	-12.340	-14.260	0.8875	-14.363	-12.592	0.8982
CFC	-11.562	-16.529	0.7265	-8.730	-17.870	0.4978	-3.725	-15.698	0.4773

N = 20
* p < 10%, ** p < 5%, *** p < 1%

Consistent with the results of the moderation analysis, no statistically significant differences are found between respondents who experience the social norm nudge and score high on a psychometric variable and those who score low on a psychometric variable.

When the high/low treatment subgroups are compared with the control group, few statistically significant differences are identified, with those identified occurring only at the 10% level:

Table 28: Pilot Study 1 T-test Results for High (Low) Social Norm Nudge Subgroups vs. Control

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-2.170	-11.180	0.4566	-2.170	-11.180	0.4566	-2.170	-12.900	0.2560
Low	-2.170	-15.420	0.1679	-2.170	-15.420	0.1679	-2.170	-16.900	0.4539
Avoidant									
High	-2.170	-6.815	0.5836	-2.170	-7.910	0.5467	-2.170	-5.475	0.7056
Low	-2.170	-25.343	0.1029	-2.170	-18.690	0.1720	-2.170	-25.038	0.0854*
Intuitive									
High	-2.170	-12.355	0.3842	-2.170	-15.670	0.2636	-2.170	-13.900	0.2534
Low	-2.170	-14.456	0.2146	-2.170	-10.930	0.3634	-2.170	-11.500	0.4604
Dependent									
High	-2.170	-9.312	0.5192	-2.170	-3.830	0.8767	-2.170	-8.757	0.5221
Low	-2.170	-19.238	0.1052	-2.170	-22.770	0.0579*	-2.170	-23.900	0.0721*
Spontaneous									
High	-2.170	-12.991	0.2597	-2.170	-14.630	0.2142	-2.170	-13.817	0.2067
Low	-2.170	-13.678	0.3525	-2.170	-11.970	0.4048	-2.170	-12.525	0.4292
NFC									
High	-2.170	-20.617	0.2342	-2.170	-14.260	0.3303	-2.170	-12.592	0.3595
Low	-2.170	-10.164	0.3306	-2.170	-12.340	0.2696	-2.170	-14.363	0.2327
CFC									
High	-2.170	-16.529	0.3155	-2.170	-17.870	0.1916	-2.170	-15.694	0.1730
Low	-2.170	-11.562	0.2750	-2.170	-8.730	0.4965	-2.170	-3.725	0.9115

* p < 10%, ** p < 5%, *** p < 1%

Furthermore, where statistical significance is identified, and thus one can conclude the social norm nudge is statistically significantly effective, compared to the control group, the sign of the effect is opposite to previous expectations.

11.8 – Conclusions and Potential Developments

The first pilot study suggests that the selected nudges have either statistically insignificant impacts, when comparing the nudged group to the control group, or statistically significant impacts, but with the sign of the effect being opposite to prior expectations. For three of the nudges – the status quo nudge, the present bias nudge, and the social norm nudge – effectiveness is not statistically significant, in that the nudged group did not behave significanltly differently from the control group. This is not the case with the loss aversion nudge, where effectiveness is statistically significant; implying the nudged group behaved significantly differently from the control group, but the effect of the nudge is that the nudged group has effectiveness scores which are less than that of the control group. This is the case even when no statistically significant difference is identified.

One possible explanation for these results is the presence of aesthetic effects. If one advertisement is more effective than another, regardless of which is using a nudge, the effectiveness of the nudge is expected to be obscured. Evidence for a statistically significant aesthetic effect is mixed, with statistical significance found in the loss aversion and social norm groups, but not in others. However, for all four nudge groups and the control group, the average effectiveness score associated with Candidate B is greater than that of Candidate A. Even though this difference is often not statistically significant, the consistency of this tendency is interesting.

A second explanation is the presence of moderation effects. In principle, a nudge could be ineffective when used impersonally, but also be statistically significantly moderated by a psychometric variable which – when accounted for – could render the nudge effective. Moderation analysis on this pilot sample data, however, reveals no statistically significant moderation effects, which may suggest that the prior beliefs that moderation may be occurring are wrong (see Chapter 8). Given this, an alternative matching analysis is undertaken. This analysis largely corroborates the lack of statistical significance found using moderation analysis, and thus reconfigures expectations – it may not be that there is a lack of significant moderation occurring, but rather the experimental design is flawed.

This leads to a final explanation, which is the survey-experiment set-up for the pilot study is simply inadequate. The four nudges may not be sufficiently incorporated into the advertisement medium to be effective; the control group may be exerting an unintended nudging effect, accounting for that group's consistently higher effectiveness; and the introduction of aesthetic effects following the experiment design may be obscuring the genuine nudge effect. In Chapter 12, adjustments to the survey-experiment design are offered to rectify these issues. Immediately, however, several additional comments can be made.

Firstly, the sample size of the pilot study is rather small given the planned analysis.[313] It may be expected that as the sample size increases, the benefits of the RCT design should reduce the role of any aesthetic effects, assuming no aesthetic effect is expected across a population (Deaton and Cartwright, 2017). As such, the relatively small sample size may account for the influence of some aesthetic effects. Furthermore, as Cronbach and Snow (1977) argue, moderated regression usually requires a minimum of 100 observations. Therefore, each moderation regression shown in this chapter may not reflect the effects seen in a larger sample analysis.

Secondly, regardless of the limitations born from the sample size, several results re-affirm confidence in some elements of the survey-experiment design. Across a relatively small sample no significant demographic differences have been identified, enabling a fair comparison of groups. Furthermore, all psychometric scales demonstrate a high Cronbach's

[313] This is due to budget limitations as the pilot studies and the main study are both incentivised. It was not possible given current budget limitations to conduct extensive data collection for both the pilot studies and the main study.

alpha, which is consistent with the literature and suggests the seven selected scales are suitable for further use. Finally, the relative normality of the dependent variable suggests the assumptions of normality associated with the use of t-test and OLS – and, by extension, moderated regression – are met.

Chapter 12 – Experimental Adjustments

12.1 – Introduction

This chapter considers the implications of the findings from the previous chapter and offers experimental adjustments which are subsequently examined in a second pilot study, which is the topic of Chapter 13.

The evidence which emerges from the first pilot study suggests that the nudges were inadequately embedded into the advertisement format to allow for an assessment of their effect. Furthermore, Pilot Study 1 produces mixed evidence to support the hypothesis that there is a significant aesthetic effect. Finally, given the relative effectiveness of the control advertisement compared to the nudge advertisement, it is possible that the control advertisement is exerting some unintended nudging effect.

Resolving the problem of aesthetic effects – if, indeed, such a problem exists – requires a relatively simple experimental adjustment. Namely, rather than showing participants two advertisements which differ both in nudging (i.e. nudge vs. control) and aesthetics, participants could be shown the same advertisement twice, differing only in the nudge used in each instance. Such an adjustment would control for any aesthetic differences, and thus any difference in response between the two advertisements could be attributed to the nudge. Yet, given evidence for aesthetic effects is mixed, such an adjustment is not immediately undertaken.

12.2 – Dynamic Choice Architecture

The more immediately relevant adjustment, however, pertains to the use of nudges within the political advertisements. As discussed in Chapter 7, an advertisement may be thought of as the sum of its components:

Advertisement = Imagery + Text + Background + Colour Scheme…

In other words, an advertisement A can be understood as consisting of a set of components X such that $A = (X_1 + X_2 + \cdots + X_i)$. Each component, or X, can be used to nudge a decision-maker (Yeung, 2017; Weinmann, Schneider and vom Brocke, 2016). As such, if a nudge embedded within a *single* component (X) of an advertisement fails to exert a positive influence on a decision-maker, a positive influence may be induced by embedding the nudge into *additional* components of the advertisement. In the language of Chapter 7, to increase the effectiveness of the nudges, more choice architecture (components) can be 'switched on'.

Take, for example, advertisements 1 and 2, used in the first pilot study:

Figure 9: Pilot Study 1 Advertisements 1 and 2

Advertisement 1

Advertisement 2

Advertisement 1 is a treatment advertisement using a status quo nudge slogan, "Let's keep going." Advertisement 2 is the control advertisement, with the slogan "Working for you." While aesthetic differences exist between the advertisements (e.g. the candidate pictures and names, or the background images), these features are designed to be generic and not exert any intended, significant influence on decision-makers. As such, these advertisements are designed with the purpose of the control advertisement containing no nudge in any of its components, and the treatment advertisement containing a nudge in only one of its components, the political slogan.

In online and digital settings – as well as in settings where maximal influence is desired, such as a political campaign – such subtlety in nudging may not be typical, or indeed desirable.

Following this rationale, and in response to previous results, several new political advertisements are offered in Table 29 below.

These advertisements differ from the advertisements previously used in several ways.

12.2.1 Background Imagery

In addition to embedding a nudge within the campaign slogan, each advertisement now embeds a respective nudge within the background image used.

1. The Status Quo nudge – The status quo advertisement contains an image on the US Capitol Building, a prominent symbol of American government, which is designed to associate the candidate with this building and institution, thus appealing to the notion of the status quo and incumbency.
2. The Present Bias nudge – The present bias advertisement contains an image of a check-box sheet with the option “Now” ticked. In the faded background, alternative options of “Yesterday”, “Today” and “Tomorrow” can also be seen, emphasising the focus on the present.[314] The layout, similar to that of an election ballot, is also a desirable choice.
3. The Loss Aversion nudge – The loss aversion advertisement contains an image of an agreeable handshake, a typical moment of agreement and certainty which is expected to appeal to the struggle between gains and losses. The choice of a handshake also has synergy with the adjusted loss aversion slogan (see below).
4. The Social Norm nudge – The social norm advertisement features a row of houses in an American suburb spanning into the distance. The suburban imagery is designed to conjure a notion of community akin with the social norm nudge, while the continuity

[314] One could argue that “Now” and “Today” are essentially the same, and such an argument would not be without merit. The decision to use “Now” rather than “Today” is that “Today” could take on more alternative meanings than “Now.” “Now”, in the opinion of this author, is more associated with the immediacy the slogan is trying to invoke than “Today.” For instance, “Today” could mean now; it could also mean *later* today, an interpretation which would have a deleterious effect. “Now” does not suffer such an interpretation.

towards the vanishing point emphasises the magnitude of pre-existing support for the candidate.

It is worthwhile to acknowledge that any imagery selected has subjective interpretations. Indeed, this is likely why the literature on digital nudging (Yeung, 2017; Weinmann, Schneider and vom Brocke, 2016) so often emphasise the personal, dynamic and automated nature of digital nudging – subjectivity may be mediated by large datasets which can test and determine ideal layouts, mediums and content for each individual.[315] In absence of these resources, any selection made here is liable to suffer from clashes in subjective interpretation.[316] Yet, given the rationale for the various selections is offered and seems – at least from this author's *subjective determination* – sensible, these selections in their current forms are used..

In conjunction with the notion of 'switching' choice architecture 'on' and 'off', while the treatment advertisements have the background imagery 'switched on', the control advertisement now has this imagery 'switched off.'

[315] Or so discussions surrounding social media algorithms report, at the very least (Yeung, 2017; Luckerson, 2015; Tucker and Thaler, 2013).
[316] Indeed, the choice of a visual advertisement, rather than a video or written advertisement, demonstrates a subjective determination when one considers the capacities of social media platforms (Luckerson, 2015).

Table 29: Revised Political Advertisements

	Control	Status Quo Nudge	Present Bias Nudge	Loss Aversion Nudge	Social Norm Nudge
“Candidate” A					
“Candidate” B					
Slogan	“A Candidate For You”	“Let’s Keep Going”	“Delivering For You: Now”	“A Safe Pair of Hands”	“Trusted by America”
Background Image	None	The US Capitol Building	A check-box with “Now” selected	An agreeable handshake	An American suburb

12.2.2 Slogans

Three of the four slogans have also been changed. These are the present bias nudge, the loss aversion nudge, and the social norm nudge.

Before discussing these changes, however, it is important to consider a subtle change to the status quo nudge slogan. In terms of language, this slogan has not changed from Pilot Study 1. However, an oversight in Pilot Study 1 was the lack of emphasis. The status quo slogan – "let's keep going" – contains two words from which one might derive a behavioural hypothesis. These are "keep," and "going." The former is designed to invoke the status quo – to *keep* something, one must *already have that thing*. However, the latter – "going" – may suggest some sort of development. For instance, to *keep going* may suggest continuation, but continuation towards a destination not yet reached, changing the primary subject of interest in the phrase. In other words, the word "going" *could* invoke a behavioural response which is not necessarily characteristic of the status quo nudge.

This is merely a hypothesis, but implies a potential confounding effect which should be minimised where possible. As such, by introducing emphasis, the reader's attention can be drawn to the word "keep" rather than "going", thus emphasising the status quo element of the slogan, rather than any behavioural element which may arise from the word "going."

The change to the present bias nudge is largely for greater consistency, as the previous slogan – "Fighting for you today, not tomorrow" – was noticeably longer than other slogans.[317] The revised slogan, "Delivering for You: Now" is of a much more similar length to the others. Furthermore, the word "fighting" has been changed to "delivering" to emphasise the idea of *receiving* something in the present, which the former may communicate less effectively than the latter. In addition, the temporal framing has been changed from "Today" to "Now" in the hopes of emphasising to an even greater degree the present temporality.

[317] One may also note that promising to not fight for a constituent tomorrow may suggest to a voter the candidate believes they will lose.

The loss aversion nudge was the worst performing in the first pilot study, and this is possibly because the nudge was framed as avoiding losses, rather than guaranteeing gains. Several authors (Utych, 2017; Dowling and Krupnikov, 2016; Gunsch et al., 2000) report a complex relationship between negativity and political decision-making, with appeals to the negative often producing a positive reaction amongst partisans (who support the *attacking* candidate), but can produce negative reactions amongst non-partisans (Ordway and Wihbey, 2016). As candidates A and B are deliberating hypothetical and not clearly aligned to either of the two major US parties, it may be the case that the original, negative framing of the loss aversion nudge failed to appeal to participants.

To resolve this, a new slogan "A Safe Pair of Hands" is used. This slogan clearly appeals to the positive frame,[318] and indicates certainty (i.e., safe) over risk or uncertainty.

The final adjusted nudge slogan is that of the social norm nudge. This adjustment is rather small, changing from, "Trusted by Voters" to "Trusted by America." The reason for this change is that, within the experiment, participants are not *actually* voters. Instead, they simply indicate their willingness to vote for each candidate. Thus, while trying to appeal to a social norm (i.e. that voters approve of this candidate), the participant themselves may fail to identify with this group as they are *not actually voting*. To ensure participants have a group to relate themselves to, the social norm slogan is changed to "Trusted by America."[319]

Finally, the control slogan is changed. As indicated by the first pilot study, the previous control advertisement outperformed all the treatment advertisements (though the difference was not always statistically significant). One explanation for this is that the previous control slogan – "Working for You" – was exerting some unintended influence on participants. Of course, the presence of any slogan (or indeed, any choice architecture) can be expected to shape the behaviour of decision-makers somewhat (Thaler and Sunstein, 2008). But given the control slogan itself may have been producing something akin to a nudging effect, such an influence

[318] As opposed to the negative frame, such as "my opponent is not a safe pair of hands."
[319] All participants report to be Americans.

may obscure the effect produced by the intended nudges. To resolve this, a new control slogan – "A Candidate for You" – is used. This slogan is selected as it is, at its core, a statement of fact: the candidate in the control advertisement is a candidate which is presented to the participant for their consideration.

Finally, given the choice personalised slogans are adaptions of the generic nudging slogans shown above, these slogans are also altered from Chapter 8. These slogans are now as follows:

Table 30: Changes to Choice Personalisation Slogans

Policy	Status Quo	Present Bias	Loss Aversion	Social Norm
The Economy	"Let's Keep Building a Better Economy"	"Delivering a Better Economy: Now"	"A Safe Choice for a Better Economy"	"Trusted by America to Protect the Economy."
Education	"Let's Keep Building a Better Education System"	"Delivering a Better Education System: Now"	"A Safe Choice for a Better Education System"	"Trusted by America to Protect Education"
Healthcare	"Let's Keep Building a Better Healthcare System"	"Delivering a Better Healthcare System: Now"	"A Safe Choice for a Better Healthcare System"	"Trusted by America to Protect Healthcare"
National Security	"Let's Keep Building a Secure Nation"	"Delivery a Secure Nation: Now"	"A Safe Choice for National Security"	"Trusted by America to Protect National Security"

12.2.3 Additional Changes

Several additional changes have been made to the political advertisements. Firstly, a logo design following from Kehle and Naimi (2019) has been added.

The same logo is featured on both advertisements, which may have the deleterious effect of prompting participants to believe they would be voting for the same party regardless. Equally, however, by having the same logo on each advertisement, participants may be discouraged from voting for which candidate they think *looks* like a member of their preferred political party. Furthermore, given the American primary system, it is not uncommon for elections to be held between members of the same party.[320] Therefore, the benefits of this change are deemed to outweigh the potential negatives.

Emphasis has been added to a keyword in each of the treatment advertisements using (in most cases) underlining. This emphasis is designed to draw attention to a word within each slogan from which the nudge primarily relies, such as "Keep" in the instance of the status quo, "Now" in the instance of the present bias,[321] "Safe" in the instance of loss aversion, or "Trusted" in the instance of the social norm nudge. No emphasis is used in the control slogan, as there is no keyword to draw attention to.

Finally, the font has been changed. In the previous advertisements, a handwritten-style font was used to suggest the slogan were the words of the candidate. However, on reflection, such a font could be distracting and difficult to read for some participants. Following the stylistic choices of Kehle and Naimi (2019), a more standard and legible font is used.[322]

To examine these changes, a second pilot study is conducted. As evidence for a significant aesthetic effect is mixed at present, the second pilot study continues to use an RCT approach with aesthetically different advertisements. The major change between the first pilot study and the second, therefore, is with the advertisement designs themselves. The new designs given in Table 29 are utilised and evaluated to see if embedding the nudge in more aspects of the advertisement produces a significant, positive effect.

[320] This extends to state elections in some US states.

[321] The emphasis for the present bias nudge differs from the other nudges, as the wording was already embedded within the selected background imagery. Insofar as this is a form of emphasis, this slight deviation is considered unsubstantial, but further demands emphasis be utilised in the other treatment groups.

[322] A slightly higher resolution image of Candidate B is also utilised. However, given the display size of the candidate and the advertisement, and the relatively small increase in image quality (around 100 pixels), this is not considered a substantial adjustment.

Chapter 13 – Pilot Study 2

13.1 – Introduction

This chapter presents data and results from a second pilot study, Pilot Study 2. This pilot study implements the experimental changes discussed in the previous chapter, and informs the procedure undertaken in the main experiment, which is the subject of Chapter 14.

13.2 – Data Summary

A sample of 100 participants from the US were recruited using Amazon's *Mechanical Turk* (MTurk) service and were compensated $0.50 for their participation.[323] After removing responses registered as being completed inappropriately fast (less than 2 minutes) and respondents who failed an attention check, a sample of N = 75 remained (female = 32%).[324]

Summary statistics are shown in Table 31: Pilot Study 2 Summary StatisticsTable 31:

[323] Compensation was revised from the first pilot study following completion-rate times.
[324] Another possible reason for the results found in the first pilot study was poor data quality. This prompted the introduction of an attention check, which subsequently 22% of respondents failed. Therefore, the hypothesis regarding the previous results and data quality may be compelling. Interestingly, following results reported by Oppenheimer, Meyvis and Davidenko (2009), a failed attention check of around 22% is still rather good.

Table 31: Pilot Study 2 Summary Statistics

Demographic	Frequency	Percentage of N	Average
Education:			2.947
(1) None	0	0	
(2) Highschool	15	20.00%	
(3) Bachelor's Degree	51	68.00%	
(4) Master's Degree	7	9.33%	
(5) PhD	2	2.67%	
Political Identity:			2.920
(1) Left-wing	18	24.00%	
(2) Left-leaning	12	16.00%	
(3) Centre	18	24.00%	
(4) Right-leaning	12	16.00%	
(5) Right-wing	15	20.00%	
Age:			36.933
18-25	8	10.67%	
26-35	32	42.67%	
36-45	21	28.00%	
46-55	6	8.00%	
55<	8	10.67%	

From Table 31, it can be seen that the median respondent is around 37 years of age, is qualified slightly below a bachelor's degree, and identifies slightly left of the political centre. These statistics are relatively similar to those compiled in the first pilot study. No statistically significant difference between the control group and the four treatment groups was found for age (χ^2 (124, N = 75) = 129.9, p = .34), sex (χ^2 (4, N = 75) = 0.65, p = .96), education (χ^2 (12, N = 75) = 10.2, p = .60) or political identity (χ^2 (16, N = 75) = 10.5, p = .84). The control and treatment groups, therefore, appear comparable.

13.3 – Testing of Assumptions

As previously, it is first prudent to examine whether the four nudges under consideration appear to be positively effective or not when used impersonally, which is to say, when participants are randomly assigned to a particular nudge-advertisement. The data are, therefore, first examined to determine if there is any violation of the assumptions of a t-test (see Chapter 8).

13.3.1 Normality

The normality of the data is first reviewed.

Figure 10: Pilot Study 2 Histogram Normality Plots

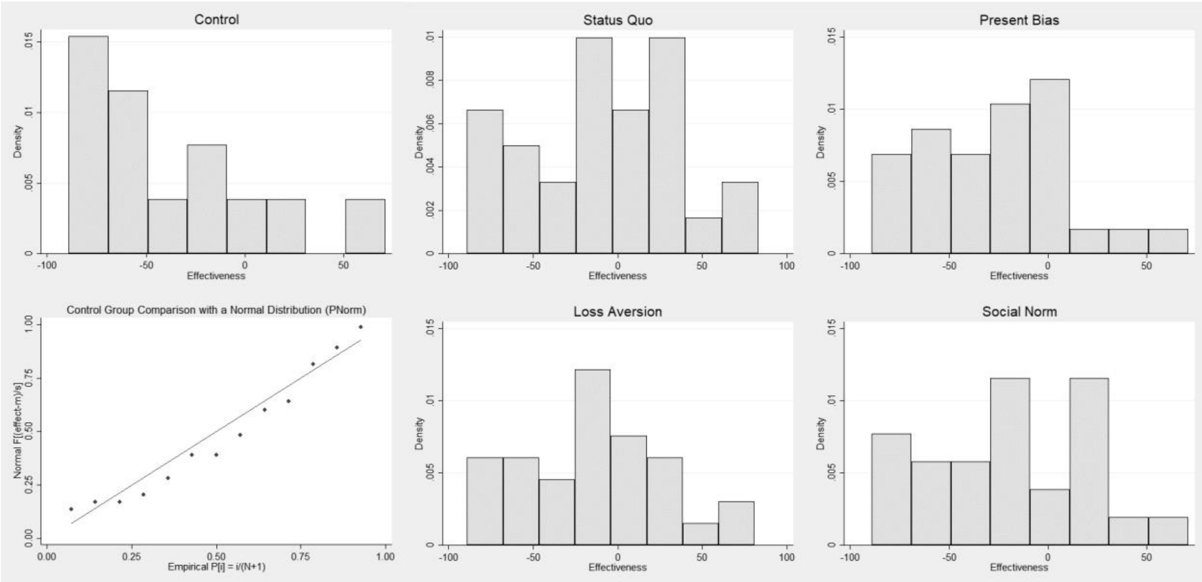


Figure 10 shows histogram plots for each group, with the dependent variable (effectiveness) shown on the x-axis and frequency density shown on the y-axis. There is limited evidence of normality, in particular regarding the control group. However, Shapiro-Wilk's tests for normality suggest these data are normally distributed (control p = 0.1980; status quo p = 0.1715; present bias p = 0.4612; loss aversion p = 0.8057; social norm p = 0.4086). Despite this, the lack of visual confirmation means that the non-parametric WMW-test is also utilised.

13.3.2 Homogeneity of Variance

As with Pilot Study 1, Levene's test is used to investigate the presence of heterogeneity of variance between the control group and each of the four nudge groups. In most instances, no evidence of heterogeneity of variances is found between the control group and the four nudge groups (status quo p = 0.1623; loss aversion p = 0.2251). However, for the present bias nudge at the 5% level (p = 0.0303), and the social norm nudge at the 10% level (p = 0.0779), there is some evidence the variance is heterogeneous. In these instances, a Welch test is used.

13.4 – Impersonal Nudging

These results are presented in Table 32.

Table 32: Pilot Study 2 Two-Tailed T-test Results of Impersonal Nudges with WMW-test Comparisons

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value (T-test)	p-value (WMW-test)
Status Quo	-36.923	10.667	-3.232	0.0033***	0.0032***
Present Bias	-36.923	-14.875	-1.645	0.1115 (0.1410)	0.0653*
Loss Aversion	-36.923	-1.556	-2.689	0.0117**	0.0087***
Social Norm	-36.923	-0.692	-2.457	0.0216** (0.0237**)	0.0128**

Welch's adjusted t-test shown in brackets.
* p < 10%, ** p < 5%, *** p < 1%

As Table 32 shows, the status quo, loss aversion and social norm nudges all have effectiveness scores which are statistically significantly different from scores of the control group, and positively so. This is true for the social norm nudge even allowing for Welch's adjustment. The present bias nudge is also significantly different from the control group at the 10% level when evaluated using the WMW-test, but not using the two-tailed t-test or Welch's t-test.[325] Despite this statistical significance and the mean effectiveness of the nudges consistently being greater than that of the control group mean, only the status quo group produces an absolutely positive effectiveness score. The large negative effectiveness score associated with the control group – which should, in theory, be zero – suggests once more a statistically significant aesthetic effect.

13.5 – Testing for the Presence of Aesthetic Effects

As in the first pilot study, the presence of statistically significant aesthetic effects is examined. These results are shown in Table 33:

[325] As there is an expectation of a positive effect, a one-tailed t-test can also be used. Using the one-tailed t-test, the present bias nudge is statistically significant from the control group at the 10% level (p = 0.0558).

Table 33: Pilot Study 2 Two-Tailed T-test and WMW-test Results for Aesthetic Effects

Nudge	Mean (Candidate A)	Mean (Candidate B)	t-Statistic	p-value (T-test)	p-value (WMW-test)
Control	-68.500	-9.957	-2.812	0.0169**	0.0181**
Status Quo	13.375	7.571	0.361	0.7242	0.3524
Present Bias	-20.600	-5.333	-1.292	0.2172	0.3275
Loss Aversion	-4.875	6.700	-0.735	0.4731	0.2662
Social Norm	-2.375	2.000	-0.298	0.7711	0.7136

* p < 10%, ** p < 5%, *** p < 1%

Generally, these results suggest that aesthetic effects are limited. For each treatment group, no statistically significant difference is found in the effectiveness of the nudge when Candidate A is used compared to Candidate B's nudge-advertisement. However, as with the first pilot study, Candidate B generally produces a higher effectiveness score than Candidate A (the exception being for the status quo nudge).

It is notable that these conclusions are reached under generally suboptimal conditions for the RCT design. The sample size is generally very small (N<20). Furthermore, no group has an equal distribution of observations in their respective subgroups, potentially creating a sampling bias which is amplified by the small sample size. Lack of evidence of an effect must be distinguished from evidence of lack of an effect; it may not be concluded that the general lack of statistically significance indicates no overall aesthetic effect, given the small sample size. However, it seems more prudent to draw no firm conclusions about the presence (or lack thereof) of aesthetic effects given the limitations of the data.

13.6 – Moderation Analysis

As with the first pilot study, the presence of statistically significant moderation effects is examined. Once more, differences in psychometric responses across the treatment groups are examined, with these results shows in Table 34:

Table 34: Pilot Study 2 Differences in Psychometric Scores

Nudge	Rational	Avoidant	Intuitive	Dependent	Spontaneous	NFC	CFC
Status Quo	0.0326**	0.7684	0.9588	0.1321	0.7064	0.9364	0.9814
	(0.096*)	(0.140)	(0.574)	(0.409)	(0.142)	(0.554)	(0.672)
Present Bias	0.6787	0.1384	0.2123	0.4916	0.8212	0.7591	0.2654
	(0.402)	(0.169)	(0.311)	(0.461)	(0.252)	(0.582)	(0.358)
Loss Aversion	0.1400	0.2467	0.8042	0.0989*	0.6572	0.3419	0.4850
	(0.101)	(0.412)	(0.853)	(0.474)	(0.448)	(0.373)	(0.457)
Social Norm	0.0191**	0.8572	0.7410	0.6447	0.5481	0.3789	0.8696
	(0.188)	(0.147)	(0.553)	(0.553)	(0.432)	(0.372)	(0.549)

χ^2 p-value shown in brackets, N = 75
* p < 10%, ** p < 5%, *** p < 1%

As expected, following Pilot Study 1, there are limited instances of statistically significant differences between the control group and a given nudge group. The rational psychometric score in the status quo treatment group is significantly different from the control group, and this significance remains when examined using a chi-squared test for differences in distribution, though only just at the 10% level. This is the only instance of consistent statistical significance between the two groups and suggests it is possible any observed significance of the rational psychometric within the moderated regression for the status quo nudge may be attributed simply to sampling differences between the groups. The rational psychometric score in the social norm treatment group is statistically significantly different from that of the control group, but this is not supported by a chi-squared test. Likewise, the dependent psychometric in the loss aversion treatment group is statistically significantly different from that of the control group, but only just at the 10% level and not when examined using a chi-squared test.

Given these statistically significant differences are few and generally inconsistent in their implications for determining statistical significance of the effects, no adjustment is made to the analysis.

To examine moderation effects, each of the seven psychometric scales discussed previously are averaged once more. The validity of this process is assured by measuring Cronbach's alpha for each psychometric scale. These results are shown in Table 35:

Table 35: Pilot Study 2 Cronbach's Alpha Results for Psychometric Variables

Scale	Cronbach's Alpha
GDMS: Avoidant	0.9316
GDMS: Dependent	0.8965
GDMS: Intuitive	0.9024
GDMS: Rational	0.8425
GDMS: Spontaneous	0.8879
NFC	0.8594
CFC	0.7907

As each scale demonstrates a Cronbach's alpha value which is greater than a typically accepted threshold of around 0.6-0.7, averaging each of these scales appears wholly valid. Summary statistics of these psychometric are shown in Table 36:

Table 36: Pilot Study 2 Summary Statistics of Psychometric Variables

Psychometric	Mean	Std. Dev.	Min	Max	Median
Rational	3.819	0.730	2.200	5.000	4.000
Avoidant	2.952	1.132	1.000	4.800	3.200
Intuitive	3.184	0.950	1.000	5.000	3.400
Dependent	3.331	0.873	1.000	5.000	3.400
Spontaneous	2.965	0.948	1.000	5.000	3.200
NFC	3.136	0.623	1.278	4.778	3.056
CFC	3.339	0.524	2.417	5.000	3.167

Using these average figures, moderated regression models taking the form:

Equation 8Equation 8

$$effectiveness_i = \beta_0 + \beta_1 D_i + \beta_2 Psy_\lambda + \beta_3 D_i Psy_\lambda + \varepsilon_i \qquad (8)$$

where $effectiveness_i$ is the effectiveness of nudge i, D_i is a dummy variable taking the value of 1 for nudge i, and 0 for all other values, psy_λ is a continuous variable for psychometric λ, and $D_i Psy_\lambda$ is a moderator term, are estimated. Table 37 through Table 40 present the results of moderated regressions for all possible combinations of nudge and psychometric variable.

Table 37: Pilot Study 2 Moderated Regression Results for the Status Quo Nudge

Variable	Model 1A	Model 2A	Model 3A	Model 4A	Model 5A	Model 6A	Model 7A
Dummy	159.142**	58.342**	21.491	61.914*	44.085	103.162*	-43.495
	[76.688]	[25.050]	[37.243]	[35.840]	[34.005]	[57.680]	[84.589]
Rational	27.734						
	[16.681]						
D × Rat.	-27.188						
	[18.784]						
Avoidant		12.402*					
		[6.463]					
D × Avo.		-2.945					
		[9.186]					
Intuitive			-0.876				
			[11.318]				
D × Intu.			8.617				
			[13.558]				
Dependent				5.626			
				[6.920]			
D × Dep.				-3.606			
				[10.082]			
Spontaneous					3.149		
					[10.653]		
D × Spon.					1.009		
					[13.406]		
NFC						5.055	
						[11.594]	
D × NFC						-17.620	
						[16.660]	
CFC							-33.510
							[21.057]
D × CFC							27.979
							[22.881]
Constant	-150.418**	-77.181***	-34.256	-57.523**	-45.826*	-52.801	72.200
	[69.328]	[17.484]	[31.769]	[27.837]	[26.723]	[39.040]	[76.754]
R-squared	0.3197	0.3633	0.3030	0.2950	0.2937	0.2975	0.3442
Multicollinearity	0.8361	0.9966	0.9999	0.9149	0.9945	0.9998	1.0000
N	28	28	28	28	28	28	28

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 38: Pilot Study 2 Moderated Regression Results for the Present Bias Nudge

Variable	Model 1B	Model 2B	Model 3B	Model 4B	Model 5B	Model 6B	Model 7B
Dummy	130.266* [75.490]	61.494** [25.050]	24.281 [34.924]	56.466 [36.572]	50.460 [29.984]	93.669** [44.476]	-94.619 [93.771]
Rational	27.734 [16.634]						
D × Rat.	-26.474 [18.315]						
Avoidant		12.402* [6.445]					
D × Avo.		-12.080 [8.478]					
Intuitive			-0.876 [11.285]				
D × Intu.			-0.515 [12.078]				
Dependent				5.626 [6.900]			
D × Dep.				-9.675 [9.442]			
Spontaneous					3.145 [10.622]		
D × Spon.					-9.815 [0.396]		
NFC						5.055 [11.561]	
D × NFC						-23.298* [13.354]	
CFC							-33.510 [20.997]
D × CFC							35.698 [26.744]
Constant	-150.418** [69.129]	-77.181*** [17.434]	-34.256 [31.678]	-57.523** [27.757]	-45.826* [26.646]	-52.801 [39.928]	72.200 [76.534]
R-squared	0.1391	0.1729	0.0922	0.1071	0.1133	0.1670	0.1708
Multicollinearity	0.9935	0.9205	0.9430	0.9823	0.9981	0.9965	0.9543
N	29	29	29	29	29	29	29

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 39: Pilot Study 2 Moderated Regression Results for the Loss Aversion Nudge

Variable	Model 1C	Model 2C	Model 3C	Model 4C	Model 5C	Model 6C	Model 7C
Dummy	148.836* [74.543]	68.191** [28.280]	-12.282 [38.242]	18.530 [39.340]	17.063 [36.805]	112.283* [60.690]	-44.825 [88.505]
Rational	27.734 [16.548]						
D × Rat.	-26.891 [18.231]						
Avoidant		12.402* [6.412]					
D × Avo.		-8.512 [9.573]					
Intuitive			-0.876 [11.228]				
D × Intu.			16.224 [13.074]				
Dependent				5.626 [6.865]			
D × Dep.				7.501 [10.886]			
Spontaneous					3.145 [10.568]		
D × Spon.					6.961 [13.083]		
NFC						5.055 [11.502]	
D × NFC						-22.305 [17.885]	
CFC							-33.510 [20.889]
D × CFC							25.923 [24.980]
Constant	-150.418** [68.775]	-77.181*** [17.344]	-34.256 [31.516]	-57.523** [27.615]	-45.826* [26.510]	-52.801 [38.729]	72.200 [76.142]
R-squared	0.2323	0.2621	0.2681	0.2420	0.2269	0.2266	0.2607
Multicollinearity	0.9264	0.9540	0.9978	0.9089	0.9931	0.9688	0.9830
N	31	31	31	31	31	31	31

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 40: Pilot Study 2 Moderated Regression Results for the Social Norm Nudge

Variable	Model 1D	Model 2D	Model 3D	Model 4D	Model 5D	Model 6D	Model 7D
Dummy	172.509** [78.218]	100.959*** [25.408]	27.644 [52.077]	105.980*** [35.508]	27.168 [42.395]	25.950 [60.281]	-80.109 [97.366]
Rational	27.734 [16.789]						
D × Rat.	-34.201* [19.614]						
Avoidant		12.402* [6.505]					
D × Avo.		-19.765** [8.710]					
Intuitive			-0.876 [11.391]				
D × Intu.			2.752 [16.464]				
Dependent				5.626 [6.965]			
D × Dep.				-19.700** [9.356]			
Spontaneous					3.145 [10.722]		
D × Spon.					2.694 [14.481]		
NFC						5.055 [11.669]	
D × NFC						3.973 [18.409]	
CFC							-33.510 [21.193]
D × CFC							35.705 [28.459]
Constant	-150.418** [69.777]	-77.181*** [17.597]	-34.256 [31.975]	-57.523* [28.017]	-45.826 [26.896]	-52.801 [39.293]	72.200 [77.251]
R-squared	0.2495	0.2844	0.2017	0.2441	0.2112	0.2145	0.2728
Multicollinearity	0.7919	0.9986	0.9954	0.9910	0.9848	0.9676	0.9989
N	26	26	26	26	26	26	26

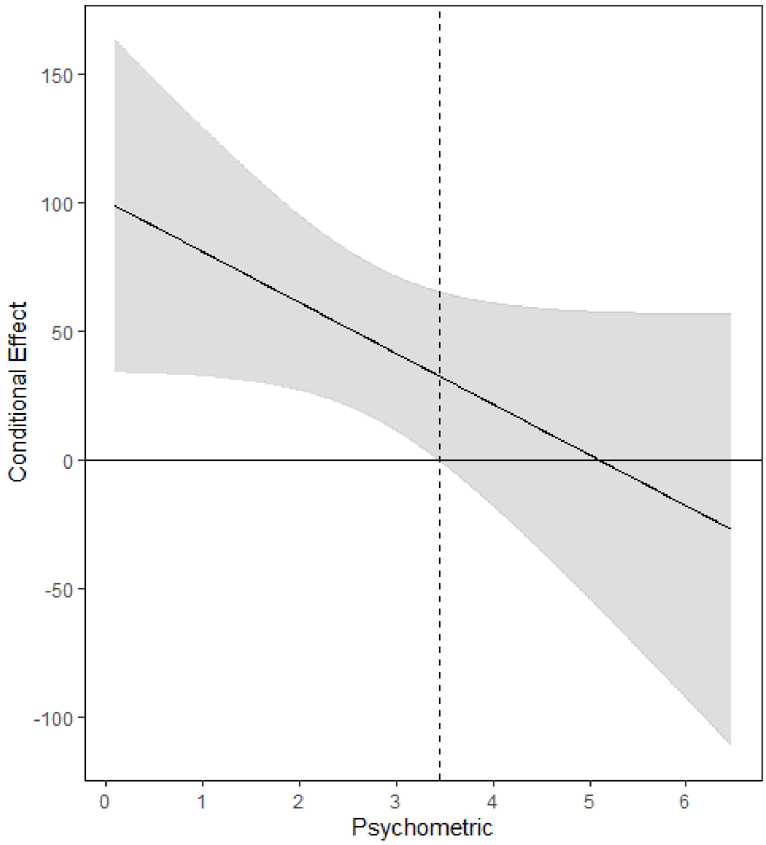
Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Consistent with expectations developed in the first pilot study, the moderated regressions report very limited evidence of statistically significant moderation effects. However, unlike the first pilot study, *some* instances of statistically significant moderation are found, both in the social norm treatment group.

13.6.1 Avoidance and the Social Norm Nudge

The first instance can be found in Model 2D, where the interaction between the dummy variable demarcating the presence of the nudge and the avoidant decision-making style is statistically significant at the 5% level. Using the Johnson-Neyman technique,[326] this result indicates a region of significance lies between the values -26.23 and 3.45 on the avoidance scale. The moderation effect of avoidance is visualised in Figure 11:

Figure 11: Pilot Study 2 Moderation Effect of Avoidance on the Social Norm Nudge



Given a person can only score between 1 and 5 on the avoidance decision-making style, a person who scores less than 3.45 on the avoidance decision-making style will face statistically significant moderating effects when presented with the social norm nudge. However, this

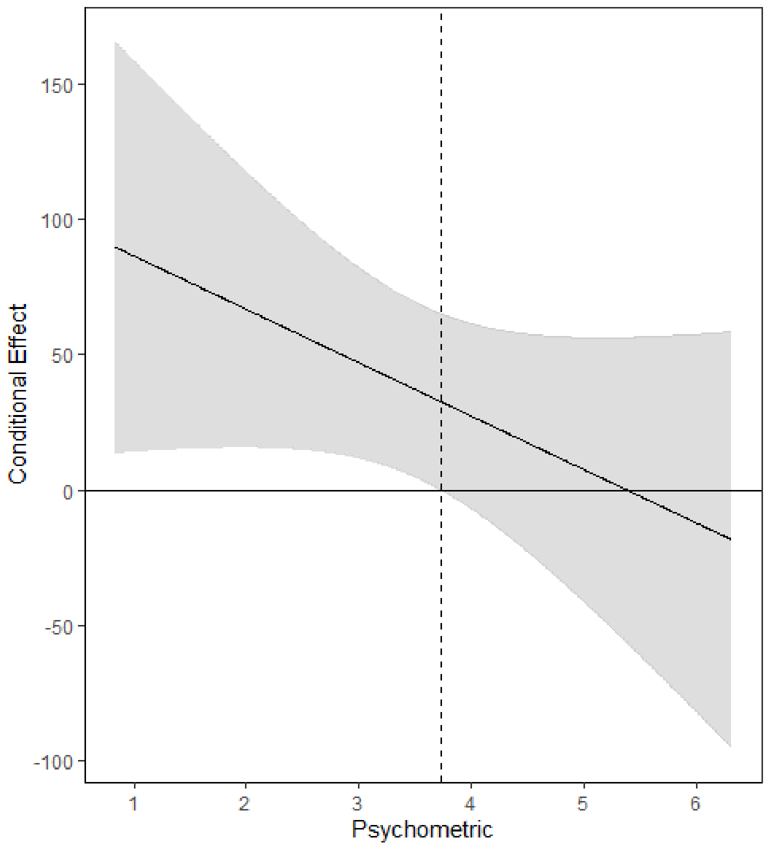
[326] STATA is used for estimating the moderated regression model. When a statistically significant effect is found, data are entered into CAHOST, an Excel program designed to calculate and plot regions of significance. Plots produced by CAHOST are, however, difficult to utilise, and so CAHOST is used much more as a calculator, with the CAHOST output sent to R for plotting purposes.

relationship is negative; a more avoidant person is expected to be susceptible to the social norm nudge than a less avoidant person. As such, a person who scores less than 3.45 on the avoidant scale is expected to be more susceptible to the social norm nudge and this produce a greater effectiveness score than someone who scores above 3.45 .

13.6.2 Dependence and the Social Norm Nudge

The second instance can be found in Model 4D, where the interaction between the dummy variable demarcating the presence of the nudge and the dependent decision-making style is statistically significant at the 5% level. Again, using the Johnson-Neyman technique, this result indicates that a region of significance lies between the values of -2.37 and 3.73 on the dependence scale. The moderation effect of dependence is visualised in Figure 12:

Figure 12: Pilot Study 2 Moderation Effect of Dependence on the Social Norm Nudge



As with the previous JNT result, as a person can only score between 1 and 5 on the dependence scale, the identified region of significance means anyone who scores *below* 3.73

on the dependence scale is expected to face significant moderation effects when presented with the social norm nudge. This relationship is also negative; a more dependent person is expected to be less susceptible to the social norm nudge than a person who is less dependent. As such, a person who scores below 3.73 on the dependent scale is expected to be more susceptible to the social norm nudge, and thus produce a greater effectiveness score than someone who scores above 3.73.

13.6.3 Discussion of Moderation Effects

While these two examples of moderation have been identified, this analysis remains generally disappointing, and warrants further investigation using matching analysis. Of course, these moderation regressions continue to utilise a very small sample which may explain the absence of statistical significance. Nevertheless, the presence of any statistically significant moderation effects marks a distinct improvement from the first pilot study, while the Johnson-Neyman technique demonstrates applicability and utility in this investigation.

The moderation results themselves are interesting insofar as they are counter to prior expectations. In the first instance, there was no hypothesised relationship between avoidance and the social norm nudge. The presence of such a result is therefore intriguing.[327] In hypothesising what relationship may exist, it could be posited that an avoidant person could be susceptible to following the social norm as this allows others to, *implicitly,* make a decision on their behalf. However, the moderation analysis attests to the opposite relationship, namely, a person with less avoidance is expected to be more susceptible to the social norm nudge. An alternative hypothesis, therefore, may be that people who have low avoidance desire to make decisions quickly and so utilise information which is available to them quickly (such as the actions of others) rather than deliberate and search for additional information.

[327] Indeed, even Peer et al. (2019) find no statistically significant moderation effects between avoidance and the social norm nudge.

In the second instance, the hypothesised relationship between dependence and the social norm nudge was a positive one. This is to say those that score high on dependence were expected to be susceptible to the social norm nudge. This was because dependent people can be expected to rely on the judgement of others. The results, however, suggest the opposite relationship is true, with those who demonstrate low dependence expected to be more susceptible to the social norm nudge. At present, no explanation can be offered for such a finding.

13.6.3.1 A Note on Moderation

The above discussion offers potential explanations for these identified moderation effects. In this context it is important to bear in mind, as Hayes (2018) argues, quantitative evidence of a moderation effect may not be sufficient evidence to confidently conclude moderation is occurring; there should also be a *qualitative* explanation or expectation of a moderation effect. Chapter 6 establishes a psychometric map which serves as the basis for the expectations in this thesis. The above discussion is offered primarily to interrogate the qualitative basis of these moderation findings.

13.7 – Matching Analysis

Once more, matching analysis is also utilised to investigate the relationships between psychometric scores and nudge effectiveness. The same criteria as used previously are used to determine ‘high’ and ‘low’ scorers for each psychometric variable. While t-test results for this second sample have broadly aligned with the WMW-test results also utilised, both tests are used throughout this analysis.

13.7.1 Status Quo Nudge

The results of a two-tailed t-test and WMW-test examining the effectiveness of the status quo nudge across the control and treatment groups is shown in Table 41 (also see Table 32):

Table 41: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Status Quo Nudge vs. the Control Group

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value (T-test)	p-value (WMW-test)
Status Quo	-36.923	10.667	-3.232	0.0033***	0.0032***

N = 28
* p < 10%, ** p < 5%, *** p < 1%

As can be seen, without any consideration of psychometric effects, the effect of the advert with the status quo nudge is statistically significantly different from the control advert; a difference which can be attributed to the presence of the nudge. In Table 42, tests for statistically significant differences in effectiveness between high- and low-scorers within the status quo treatment group are reported:

Table 42: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Status Quo Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	7.000	13.875	0.6756 (0.7273)	8.875	12.714	0.8157 (1.0000)	-3.000	14.083	0.3994 (0.3102)
Avoidant	-0.857	20.750	0.1740 (1630)	1.750	20.857	0.2332 (0.2222)	-4.333	20.667	0.1181 (0.1100)
Intuitive	7.714	13.250	0.7365 (0.9075)	11.125	10.143	0.9525 (0.6841)	2.750	13.545	0.5590 (0.5120)
Dependent	7.000	13.875	0.6756 (0.8616)	6.125	15.857	0.5522 (0.7273)	11.500	10.111	0.9341 (0.6786)
Spontaneous	-0.800	16.400	0.3145 (0.2957)	4.250	18.000	0.3975 (0.4855)	-0.800	16.400	0.3145 (0.2957)
NFC	12.111	8.500	0.8295 (0.8128)	9.250	12.286	0.8538 (0.7713)	35.333	4.500	0.1153 (0.1684)
CFC	12.200	7.600	0.7918 (0.9510)	15.500	5.143	0.5266 (0.6419)	4.000	13.091	0.6233 (0.5999)

N = 15
* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

Consistent with the lack of statistically significant effects in the moderation analysis, and regardless of how a ‘high’ and a ‘low’ scoring respondent is constructed, no statistically significant differences between these subgroups are identified. Given the nudge is statistically significantly different from the control group (see Table 41), this difference, therefore, can be likely attributed to the effect of the nudge, and does not appear to be moderated by psychometric effects.

This is re-affirmed when the high/low treatment subgroups are compared with the control group:

Table 43: Pilot Study 2 Two-tailed T-test and WMW-test Results for High (Low) Status Quo vs. Control

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-39.923	13.875	0.0162** (0.0112**)	-39.923	12.714	0.0271** (0.0193**)	-39.923	14.083	0.0039*** (0.0039***)
Low	-39.923	7.000	0.0359** (0.0213**)	-39.923	8.875	0.0206** (0.0123**)	-39.923	-3.000	0.2620 (0.1380)
Avoidant									
High	-39.923	20.750	0.0073*** (0.0059***)	-39.923	20.857	0.0116** (0.0079***)	-39.923	20.667	0.0046*** (0.0045***)
Low	-39.923	-0.857	0.0733* (0.0391**)	-39.923	1.750	0.0427** (0.0269**)	-39.923	-4.333	0.1280 (0.0589*)
Intuitive									
High	-39.923	13.250	0.0182** (0.0101**)	-39.923	10.143	0.0351** (0.0173**)	-39.923	13.545	0.0066*** (0.0045***)
Low	-39.923	7.714	0.0321** (0.0238**)	-39.923	11.125	0.0157** (0.0137**)	-39.923	2.750	0.1313 (0.0888*)
Dependent									
High	-39.923	13.875	0.0152** (0.0112**)	-39.923	15.857	0.0183** (0.0125**)	-39.923	10.111	0.0176** (0.0111**)
Low	-39.923	7.000	0.0381** (0.0214**)	-39.923	6.125	0.0298** (0.0184**)	-39.923	11.500	0.0388** (0.0224**)
Spontaneous									
High	-39.923	16.400	0.0058*** (00043***)	-39.923	18.000	0.0151** (0.0100***)	-39.923	16.400	0.0058*** (0.0043***)
Low	-39.923	-0.800	0.1273 (0.0754*)	-39.923	4.250	0.0349** (0.0224**)	-39.923	-0.800	0.1273 (0.0754*)
NFC									
High	-39.923	8.500	0.0389** (0.0350**)	-39.923	12.286	0.0181** (0.0192**)	-39.923	4.500	0.0107** (0.0083***)
Low	-39.923	12.111	0.0161** (0.0075***)	-39.923	9.250	0.0300** (0.0124**)	-39.923	35.333	0.0324** (0.0368**)
CFC									
High	-39.923	7.600	0.0610* (0.0430**)	-39.923	5.143	0.0394** (0.0238**)	-39.923	13.091	0.0064*** (0.0049***)
Low	-39.923	12.200	0.0117** (0.0069***)	-39.923	15.500	0.0149** (0.0101**)	-39.923	4.000	0.1273 (0.0787*)

* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

As is expected given the status quo treatment group is overall statistically significantly different from the control group, and there is no statistically significant difference between the high and low subgroups, both subgroups when compared with the control are found – in the majority of cases – to be statistically significantly different from the control group. This pattern generally holds across all constructions of the high/low subgroups, and across all psychometrics. Therefore, it seems likely that no statistically significant relationship exists between the psychometric variables and the nudge, and that any difference between the nudge and the control group can be attributed to nudging effects, rather than an interaction of a specific psychometric variable and the status quo nudge.

This is not to say there may not be *signs* of a relationship, only that the data do not suggest any relationship is statistically significant. For instance, the effectiveness scores in the high spontaneity subgroup are not statistically significantly different from those in the low spontaneity subgroup. Yet, under the mean and midpoint constructions, the high subgroup is statistically significantly different from the control group, while the low subgroup is not. An examination of the means shows those with high spontaneity consistently consider the nudged advertisement to be more effective than those in the low spontaneity group. These results suggest a (non-significant) relationship whereby those high in spontaneity are more susceptible to the status quo nudge than those low in spontaneity.

13.7.2 Present Bias Nudge

The results of a two-tailed t-test and WMW-test examining the effectiveness of the present bias nudge across the control and treatment groups is shown in Table 44 (also see Table 32):

Table 44: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Present Bias Nudge vs. Control Group

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value (T-test)	p-value (WMW-test)
Present Bias	-36.923	-14.875	-1.645	0.1115 (0.1410)	0.0653*

N = 29
Welch's adjustment shown in brackets.
* p < 10%, ** p < 5%, *** p < 1%

Without consideration of any psychometric effects, the effectiveness score of the present bias treatment group advert is not statistically significantly different from the control group advert, except when examined using the WMW-test, where there is a statistically significant difference, but only at the 10% level.

As previously, tests for a statistically significant difference in effectiveness scores between high and low groups are performed. These results are presented in Table 45:

Table 45: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Present Bias Nudge Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	-15.000	-14.778	0.9857 (0.8735)	-18.500	-11.250	0.5536 (0.6735)	-4.000	-15.600	n/a (0.7442)
Avoidant	-16.624	-13.125	0.7790 (0.7520)	-16.625	-13.125	0.7760 (0.7520)	-18.889	-9.714	0.4553 (0.4259)
Intuitive	-11.833	-16.700	0.7013 (0.5865)	-20.000	-9.750	0.3992 (0.5274)	-14.200	-15.182	0.9411 (0.8647)
Dependent	-13.600	-17.000	0.7890 (1.0000)	-10.500	-19.250	0.4733 (0.5274)	-18.667	-14.000	0.7671 (0.9462)
Spontaneous	-10.000	-19.750	0.4232 (0.3994)	-10.000	-19.750	0.4232 (0.3994)	-10.000	-19.750	0.4232 (0.3994)
NFC	-4.889	-27.714	0.0485** (0.0798*)	-0.875	-28.875	0.0107** (0.0270**)	1.833	-24.900	0.0207** (0.0296**)
CFC	-14.333	-15.571	0.9206 (0.8735)	-9.625	-20.125	0.3875 (0.3994)	-20.500	-14.071	0.7294 (0.6328)

N = 16
* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

In almost all circumstances, no statistically significant difference is found, suggesting – as with the status quo nudge – that psychometric differences are not influencing effectiveness. This conclusion is also consistent with the lack of moderation effects from in the moderated regression model (see Table 38).

The exception to these conclusions, however, can be seen when statistically significant differences are examined between the high and low subgroups of the NFC psychometric. Here, across all constructions of the dummy variable, a statistically significant difference is identified, with those low in NFC typically finding the present bias advertisement more effective than those high in NFC. This is consistent with the predictions established in Chapter 6. Further evidence to support the possibility of a negative relationship existing between NFC and the present bias nudge can be found in Table 46:

Table 46: Pilot Study 2 Two-tailed T-test and WMW-test Results for High (Low) Present Bias vs. Control

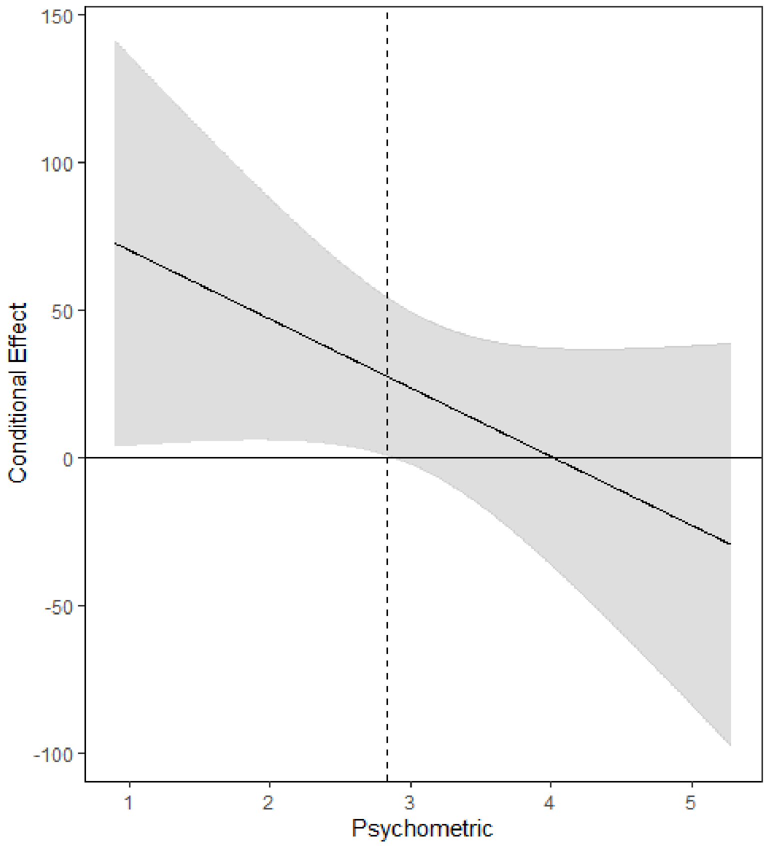
	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-39.923	-14.778	0.2156 (0.1164)	-39.923	-11.250	0.1734 (0.0956*)	-39.923	-15.600	0.1353 (0.0759*)
Low	-39.923	-15.000	0.2627 (0.1422)	-39.923	-18.500	0.3162 (0.1684)	-39.923	-4.000	n/a (0.3842)
Avoidant									
High	-39.923	-13.125	0.1855 (0.0699*)	-39.923	-13.125	0.1855 (0.0699*)	-39.923	-9.714	0.1555 (0.0623*)
Low	-39.923	-16.625	0.2924 (0.2180)	-39.923	-16.625	0.2924 (0.2180)	-39.923	-18.889	0.3211 (0.2164)
Intuitive									
High	-39.923	-16.700	0.2350 (0.1067)	-39.923	-9.750	0.1472 (0.0595*)	-39.923	-15.182	0.1827 (0.0872*)
Low	-39.923	-11.833	0.2335 (0.1600)	-39.923	-20.000	0.3590 (0.2461)	-39.923	-14.200	0.3224 (0.2174)
Dependent									
High	-39.923	-17.000	0.3413 (0.1879)	-39.923	-19.250	0.3288 (0.1472)	-39.923	-14.000	0.1278 (0.0610*)
Low	-39.923	-13.600	0.1727 (0.0938*)	-39.923	-10.500	0.1667 (0.1108)	-39.923	-18.667	0.5367 (0.4587)
Spontaneous									
High	-39.923	-19.750	0.3428 (0.1472)	-39.923	-19.750	0.3428 (0.1472)	-39.923	-19.750	0.3428 (0.1472)
Low	-39.923	-10.000	0.1585 (0.1108)	-39.923	-10.000	0.1585 (0.1108)	-39.923	-10.000	0.1585 (0.1108)
NFC									
High	-39.923	-27.714	0.6349 (0.4511)	-39.923	-28.875	0.6565 (0.4252)	-39.923	-24.900	0.4626 (0.2636)
Low	-39.923	-4.889	0.0688* (0.0299**)	-39.923	-0.875	0.0517* (0.0247**)	-39.923	1.833	0.0699* (0.0351**)
CFC									
High	-39.923	-15.571	0.2934 (0.1907)	-39.923	-20.125	0.3812 (0.2770)	-39.923	-14.071	0.1210 (0.0723*)
Low	-39.923	-14.333	0.1898 (0.0883*)	-39.923	-9.625	0.1292 (0.0503*)	-39.923	-20.500	0.6438 (0.3949)

* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

Here, the difference is only significant at the 10% level (though significant at the 5% level under the WMW-test). Under all constructions of the high/low variable, those who score low in NFC consider the present bias advertisement to be more effective than the control advertisement. Furthermore, those who score high on NFC, under all constructions, find the present bias advert to be (significantly) less effective than their low-scoring contemporaries, and give effectiveness scores which are not statistically significantly different from the control group.

In short, there is compelling evidence that a negative relationship does exist between the NFC psychometric variable and the present bias. Returning to the moderated regression results in Table 38 only further bolsters this evidence; while not investigated because the result was only statistically significant at the 10% level, a statistically significant moderation effect between the NFC and the present bias was previously identified. Given this additional evidence, then, the Johnson-Neyman technique is used to investigate the interaction between the NFC psychometric and the present bias nudge:

Figure 13: Pilot Study 2 Moderation Effect of NFC on the Present Bias Nudge



As visualised in Figure 13, the Johnson-Neyman technique identifies a region of significance between the values of -0.80 and 2.83, which – given the NFC scale can range only from 1 to 5 – means anyone who scores less than 2.83 is expected to be the present bias nudge than someone who scores above 2.83, thus producing a greater effectiveness score. Furthermore, the JNT reconfirms the observed, negative relationship between the NFC psychometric and the present bias nudge found in Table 45 and Table 46.[328]

[328] It is also interesting to note that the mean (3.06), median (3.00) and midpoint (3.00) of the NFC within the present bias treatment group are all close to the region of significance (2.83).

13.7.3 Loss Aversion Nudge

The results of a two-tailed t-test and WMW-test examining the effectiveness of the loss aversion nudge across the control and treatment groups is shown in Table 47 (also see Table 32):

Table 47: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Loss Aversion Nudge vs. the Control Group

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value (T-test)	p-value (WMW-test)
Loss Aversion	-36.923	-1.556	-2.689	0.0117**	0.0087***
N = 31 * p < 10%, ** p < 5%, *** p < 1%					

As with the status quo nudge, the effectiveness of the loss aversion advert is found to be statistically significantly different from that of the control group advert when used impersonally. This result, however, does not consider the effects of psychometric scores. When these effects are evaluated across high/low constructions within the loss aversion treatment group, little evidence of statistically significant difference is identified:

Table 48: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Loss Aversion Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	4.200	1.750	0.7139 (0.9291)	-4.333	7.444	0.4624 (0.5074)	-1.000	1.875	0.9110 (0.8882)
Avoidant	-1.429	3.455	0.7681 (0.9278)	-4.333	7.444	0.4624 (0.5361)	-1.429	3.455	0.7681 (0.9278)
Intuitive	-14.333	9.500	0.1508 (0.1596)	-7.333	10.444	0.2618 (0.2505)	-20.400	10.000	0.0766* (0.0679*)
Dependent	-4.167	4.417	0.6152 (0.7428)	-3.333	6.444	0.5429 (0.7237)	-4.167	4.417	0.6152 (0.7428)
Spontaneous	-4.333	4.500	0.6049 (0.6731)	-10.111	13.222	0.1347 (0.1329)	-4.333	4.500	0.6049 (0.6731)
NFC	1.667	1.333	0.9845 (0.7428)	7.778	-4.667	0.4370 (0.6269)	-5.000	3.429	0.6637 (0.7498)
CFC	2.769	-1.600	0.8085 (1.0000)	4.667	-1.556	0.6996 (1.0000)	-7.000	4.000	0.5695 (0.5236)

N = 31
* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

The only instance of a statistically significant difference is found under the midpoint construction for the intuitive psychometric variable, and even this is only at the 10% level. As with the analysis in Pilot Study 1, these results do not conflict with the lack of moderation effects determined above, and suggest the difference between the effectiveness of the loss aversion advertisement and that of the control advertisement is largely attributable to the effect of the nudge.

Table 49: Pilot Study 2 Two-tailed T-test and WMW-test Results for High (Low) Loss Aversion vs. Control Group

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-39.923	-1.750	0.0799* (0.0645*)	-39.923	7.444	0.0324** (0.0325**)	-39.923	1.875	0.0161** (0.0116**)
Low	-39.923	4.200	0.0313** (0.0130**)	-39.923	-4.333	0.0714* (0.0251**)	-39.923	-1.000	0.3196 (0.1735)
Avoidant									
High	-39.923	3.455	0.0252** (0.0127**)	-39.923	7.444	0.0258** (0.0134**)	-39.923	3.455	0.0252** (0.0127**)
Low	-39.923	-1.429	0.0997* (0.0744*)	-39.923	-4.333	0.0861* (0.0569*)	-39.923	-1.429	0.0997* (0.0744*)
Intuitive									
High	-39.923	9.500	0.0094*** (0.0060***)	-39.923	10.444	0.0196** (0.0111**)	-39.923	10.000	0.0065*** (0.0048***)
Low	-39.923	-14.333	0.2955 (0.1879)	-39.923	-7.333	0.1081 (0.0661*)	-39.923	-20.400	0.4749 (0.3002)
Dependent									
High	-39.923	4.417	0.0171** (0.0097***)	-39.923	6.444	0.0300** (0.0161**)	-39.923	4.417	0.0171** (0.0097***)
Low	-39.923	-4.167	0.1557 (0.1141)	-39.923	-3.333	0.0761* (0.0487**)	-39.923	-4.167	0.1557 (0.1141)
Spontaneous									
High	-39.923	4.500	0.0170** (0.0097***)	-39.923	13.222	0.0109** (0.0075***)	-39.923	4.500	0.0170** (0.0097***)
Low	-39.923	-4.333	0.1575 (0.1141)	-39.923	-10.111	0.1547 (0.0884*)	-39.923	-4.333	0.1575 (0.1141)
NFC									
High	-39.923	1.333	0.1008 (0.0953*)	-39.923	-4.667	0.0919* (0.0661*)	-39.923	3.429	0.0189** (0.0133**)
Low	-39.923	1.667	0.0255** (0.0114**)	-39.923	7.778	0.0236** (0.0111**)	-39.923	-5.000	0.2129 (0.1000*)
CFC									
High	-39.923	-1.600	0.1628 (0.1525)	-39.923	--1.556	0.0702* (0.0487**)	-39.923	4.000	0.0177** (0.0124**)
Low	-39.923	2.769	0.0174** (0.0089***)	-39.923	4.667	0.0325** (0.0161**)	-39.923	-7.000	0.2384 (0.1125)

* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

When the high/low subgroups are compared against the control group, frequent examples of statistically significant differences are found. However, given the effectiveness of the loss aversion nudge was statistically significantly different from that of the control group, and given a lack of evidence to suggest the high subgroups differ from the low, these results are not unexpected.

Several results do appear to be of interest, however. The intuitive, dependent and spontaneous psychometric variables all demonstrate a statistically significant difference from the control group when they are high, but not when they are low. In absence of a statistically significant difference between the high and low subgroups for these psychometric variables (the single occasion with the intuitive psychometric variable being the exception), the conclusion that respective, positive relationships exist between these psychometric variables and the loss aversion nudge cannot be drawn. However, given the relatively low sample size (N = 31), such findings may be indicative of a trend which may be seen using more data.

13.7.4 Social Norm Nudge

The results of a two-tailed t-test and WMW-test examining the effectiveness of the loss aversion nudge across the control and treatment groups is shown in Table 50 (also see Table 32):

Table 50: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Social Norm Nudge vs. the Control Group

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value (T-test)	p-value (WMW-test)
Social Norm	-36.923	-0.692	-2.457	0.0216** (0.0237**)	0.0128**

N = 26
Welch's adjustment shown in brackets.
* p < 10%, ** p < 5%, *** p < 1%

As with the status quo nudge and the loss aversion nudge, when accounting for no psychometric effects, the effectiveness of the social norm advert is found to be statistically significantly different from that of the effectiveness of the control group advert. As two statistically significant interactions were identified between the social norm nudge and the avoidant and dependent psychometrics, respectively, these relationships were expected to be found in the matching analysis:

Table 51: Pilot Study 2 Two-tailed T-test and WMW-test Results for the Social Norm Treatment Group

	Mean			Median			Midpoint		
	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value	Low (D=0)	High (D=1)	p-value
Rational	-4.400	1.625	0.6881 (0.9415)	-6.149	5.667	0.4145 (0.5192)	10.333	-4.000	0.4022 (0.2354)
Avoidant	-3.500	0.556	0.7981 (0.7570)	2.429	-4.333	0.6439 (0.4738)	19.500	-4.364	0.2238 (0.2763)
Intuitive	6.250	-3.778	0.5238 (0.4391)	-5.714	5.167	0.4533 (0.6161)	6.250	-3.778	0.5238 (0.4391)
Dependent	3.143	-5.167	0.5690 (0.5666)	3.143	-5.167	0.5690 (0.5666)	20.000	-4.455	0.2117 (0.1983)
Spontaneous	-5.200	2.125	0.6249 (0.6597)	-4.286	3.500	0.5940 (0.7202)	-3.250	0.444	0.8158 (0.8770)
NFC	-1.875	1.200	0.8381 (0.9415)	-4.143	3.333	0.6089 (0.7745)	-1.875	1.200	0.8381 (0.9415)
CFC	-4.889	8.750	0.3818 (0.3532)	-2.571	1.500	0.7815 (0.7745)	10.000	-3.900	0.4170 (0.3092)

N = 26
* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

It is immediately noteworthy that no statistically significant differences between the high and low subgroups within the social norm treatment group are identified for any psychometric variable, or under any construction. In every case of the dependent psychometric variable, the negative relationship is evidenced via an examination of means, with those scoring low in dependence finding the social norm advertisement more effective than those scoring high in dependence. The same is true in two of the three constructions for the avoidant psychometric variable. However, to reiterate, these differences are not statistically significant. If not for prior expectations arising from the moderation analysis, there would be no reason to focus on these psychometric variables over any others, as other psychometric variables demonstrate – in terms of a means examination – the same consistency of results.[329]

[329] See, for instance, the spontaneity or NFC psychometrics.

Table 52: Pilot Study 2 Two-tailed T-test and WMW-test Results for High (Low) Social Norm vs. Control Group

	Mean			Median			Midpoint		
	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value	Control (D=0)	Treatment (D=1)	p-value
Rational									
High	-39.923	1.625	0.0471** (0.0246**)	-39.923	5.667	0.0576* (0.0351**)	-39.923	-4.000	0.0604* (0.0298**)
Low	-39.923	-4.400	0.1683 (0.0842*)	-39.923	-6.143	0.1223 (0.0520*)	-39.923	10.333	0.1173 (0.0689*)
Avoidant									
High	-39.923	0.556	0.0411** (0.0212**)	-39.923	-4.333	0.1351 (0.0653*)	-39.923	-4.364	0.0519* (0.0256**)
Low	-39.923	-3.500	0.2044 (0.1125)	-39.923	2.429	0.0554* (0.0292**)	-39.923	19.500	0.1247 (0.0889*)
Intuitive									
High	-39.923	-3.778	0.0652* (0.0299**)	-39.923	5.167	0.0536* (0.0282**)	-39.923	-3.778	0.0652 (0.0299**)
Low	-39.923	6.250	0.1093 (0.0697)	-39.923	-5.714	0.1281 (0.0623*)	-39.923	6.250	0.1093 (0.0697*)
Dependent									
High	-39.923	-5.167	0.1390 (0.0653*)	-39.923	-5.167	0.1390 (0.0653*)	-39.923	-4.455	0.0523* (0.0256**)
Low	-39.923	3.143	0.0544* (0.0292**)	-39.923	3.143	0.0544* (0.0292**)	-39.923	20.000	0.1218 (0.0889*)
Spontaneous									
High	-39.923	2.125	0.0450** (0.0247**)	-39.923	3.500	0.0605* (0.0282**)	-39.923	0.444	0.0417** (0.0212**)
Low	-39.923	-5.200	0.1771 (0.0842*)	-39.923	-4.286	0.1164 (0.0623*)	-39.923	-3.250	0.2011 (0.1125)
NFC									
High	-39.923	1.200	0.1018 (0.0431**)	-39.923	3.333	0.0606* (0.0281**)	-39.923	1.200	0.1018 (0.0431**)
Low	-39.923	-1.875	0.0751* (0.0424**)	-39.923	-4.143	0.1162 (0.0623*)	-39.923	-1.875	0.0751* (0.0424**)
CFC									
High	-39.923	8.750	0.0942* (0.0697*)	-39.923	1.500	0.0909* (0.0650*)	-39.923	-3.900	0.0596* (0.0298**)
Low	-39.923	-4.889	0.0709* (0.0299**)	-39.923	-2.571	0.0814* (0.0292**)	-39.923	10.000	0.1200 (0.0689*)

* p < 10%, ** p < 5%, *** p < 1%
WMW p-value shown in brackets

When the high/low subgroups are compared to the control group, as with the status quo and loss aversion nudges, frequent instances of statistically significant differences are found between the nudge and control advertisements. Again, given overall the impact of the social norm nudge is statistically significantly different from the control, and the high-scoring subgroup within the social norm treatment group is not statistically significantly different from the low-scoring subgroup within the social norm treatment group, these results are not unexpected.

What is unexpected is the lack of statistical evidence to corroborate the presence of moderation effects found previously. Neither the avoidant nor dependent psychometric variables demonstrate a statistically significant difference which is consistent with the negative relationships identified by the moderation analysis. This, immediately, prompts an inquiry into why this is so.

Two answers exist in general, though given prior moderation analysis, only one explains the current results. Firstly, it is likely all constructions of the high/low subgroups are obscuring the moderation effect. Recall, the regions of significance associated with the avoidant and dependent psychometric variables are below 3.45 and 3.73 respectively. However, the high/low subgroups are determined at the values of 3.32 (mean), 3.60 (median) and 3 (midpoint) for the avoidant psychometric variable and 3.49 (mean), 3.40 (median) and 3 (midpoint) for the dependent psychometric variable. Unlike in the instance of the NFC psychometric and the present bias nudge, where the region of significance was very near the high/low values (i.e. <0.2 difference), these constructions are generally further away from the region of significance boundaries. This is especially true when considering the midpoint constructions, and the dependent psychometric variable across all constructions. Given the relatively small sample size (N = 26), just a few observations being incorrectly[330] classified high or low may obscure the presence of any relationship.

Secondly, where regions of significance are found at very low (but still observed) values and at very high (but still observed) values, a simple high/low bisection of this sample will result in high/low subgroups that are not statistically significantly different, as both subgroups contain observations that fall within regions of significance and regions of insignificance. Such a bisection – which cannot be avoided using matching analysis – may confound the relationship indicated by moderation analysis. Given previous moderation analyses and uses of the Johnson-Neyman technique, this second answer does not explain the conflicting results found when the matching analysis is performed. This is because, for both the avoidant and dependent psychometric variable, only one region of significance value is observable on the

[330] For the second reason, it is inappropriate to consider the region of significance boundary to be some sort of 'true' value from which high/low subgroups should be determined. However, for these *specific* relationships, it may be appropriate to characterise the region of significance boundary as the 'true' value for determining high/low groups as there is only one value which falls within the observed range of psychometric scores. Where two values fall in this observed range, such a characterisation would be misleading.

psychometric scale. However, this second answer does speak to a fundamental weakness of the matching analysis approach and – as a result – the strength of moderation analysis and the Johnson-Neyman technique. Namely, while matching analysis can identify relationships when the values for constructing high/low subgroups are close to the *single* region of significance value which is observable on the psychometric scale, matching analysis becomes less precise as these values get further away from the single region of significance value, or when *two* region of significance values can be observed on the psychometric scale. For these reasons, matching analysis is not used in Chapter 14.

13.8 – Qualitatively Evaluating Moderated Relationships

Three moderation effects have been identified in this second pilot study. As has been noted above, Hayes (2018) argues, however, the research validity of these effects should not rely solely on quantitative results. Hayes (2018) suggests that while a moderated regression may identify a statistically significant moderation effect, if the effect does not make qualitative sense within the context that the research is conducted, one should be willing to reconsider the significance of the result. Of course, such an approach risks missing details. For instance, rejecting a result simply because no prior explanation can account for it hinders progress. But equally, pursuing a result without *at least a hypothesis explaining it* can also produce conclusions which are lacking formative explanation. It is therefore worthwhile to take a moment to evaluate the three identified moderation effects.

As above, while contrary to expectations, the statistically significant, negative moderated relationship between avoidance and the social norm nudge has a possible explanation: people who have low avoidance may seek to make decisions quickly (decision-impatience), and one way of doing this is to do as others are doing (i.e. follow the social norm).

The significant, negative moderated relationship between dependence and the social norm nudge is wanting an explanation. While the hypothesised psychometric map suggests that those with high dependence have low cognition, and thus rely on social norms more, the reverse relationship might be supposed: those with low dependence have high cognition, allowing them to understand the potential consequences of *not* following the social norm. Yet, this explanation seems rather tenuous, supposing cognition – a trait primarily measured by the NFC – is involved, without NFC statistically significantly moderating the social norm nudge. It also runs counter to intuition – namely, that those who rely more on others (i.e. high dependence) do as others do.

One identified effect which is predicted by the hypothesised psychometric map is the identified significant, negative relationship between NFC and the present bias. Following this result, and the prediction, those with low NFC are expected to have low levels of risk, making them to find appeals to the present more effective than appeals to later periods. This is because, with lower NFC, these individuals may be less likely to consider the longer-term consequences of their decisions. Qualitatively, then, this effect would seem worth pursuing (though, being only significant at the 10% level, one might disregard this result from a *quantitative* perspective).

13.8 – Conclusions and Implications for Chapter 14

The second pilot study demonstrates a marked progression in experimental design from the first pilot study. Most notably, the changes made to the nudge-advertisements appear to be effective, with three of the four nudges now positively, significantly different from the control advert. The exception is the present bias nudge, although this too is significant at the 10% level using the WMW-test and one-tailed t-test.

A second notable development is the presence of significant moderation effects. While the number of moderation effects remains sparse (3 out of 28), the presence of any represents a noteworthy improvement on the results of the first pilot study, in which no statistically significant effects were found.

The presence of moderation has also greatly aided in understanding the advantages and limitations of the matching analysis. In the first pilot study, in the absence of any regions of significance, the arbitrarily constructed high/low subgroups could only infer the presence of relationships with relatively limited capacity to be verified beyond speculation. In the second pilot study, having already established in two instances the existence of regions of significance, the lack of identifying power using matching analysis very much informs the analytical procedure going forward. This is not to say that the matching analysis is without usefulness. For instance, the moderation effect between the NFC psychometric and the present bias was initially found to be significant only at the 10% level, and subsequently overlooked. Only via the matching analysis did this relationship re-emerge as one worthy of investigation.

Nevertheless, moderation analysis had already identified some significance with this relationship. Furthermore, given the lack of arbitrary value selection associated with the Johnson-Neyman technique, moderation analysis clearly emerges as the superior method of analysis. Because of this, matching analysis is not used in Chapter 14.

On the question of aesthetic effects, these results remain mixed once again. The second pilot study finds limited evidence of aesthetic effects but identifies evidence of statistically significant aesthetic effects in the control group. Given the only difference between the control advertisements are candidate names and pictures, such a statistically significant difference can be attributed to these aesthetics. However, no evidence of aesthetic effects is found in the nudge groups. It is once again noted that, in theory, the RCT design should reduce the net effect of aesthetic differences to zero given a sufficiently large sample size. As both the first and the second pilot study find mixed evidence of aesthetic effects using relatively small sample sizes (N = 95 and 75, respectively), no adjustment is made to the main survey-experiment to respond to aesthetic effects.

Some additional comments on the data remain outstanding. Firstly, the sample size of the second pilot study remains very small given the typical criteria for moderation analysis (Cronbach and Snow, 1977). With a larger sample, it may be expected that statistical tests will have increased power, for a given test size, and thus further moderated relationships between psychometric variables and nudges may be statistically significant.

Secondly, a partial reason the small sample size was the use of an attention check, which resulted in some 22% of respondents being removed from the sample (compared to around 4% which were removed using the standard deviation method employed in the first pilot study). The inclusion of the attention check certainly inspires greater confidence in the quality of the data, but in turn inspires a lack of confidence in the quality of the data utilised in the first pilot study. This may explain the generally disappointing results found in the first pilot study. In the main survey-experiment, an attention check is introduced to ensure data quality.

13.9 – Post-hoc Power Testing

By way of further informing the sample selection for the primer group and PTG data collection stages, post-hoc power analysis is performed on the results of the second pilot study.[331] Please see Chapter 9 which contains details on the power testing discussed in this here.

Across the three moderation regression models where a significant moderation effect was identified, all tests appear to be adequately powered. Model 2D produces an f^2 of 0.3229, corresponding to a power of 87.73%. Model 4D produces an f^2 of 0.3974, corresponding to a power of 92.90%. Finally, model 6B produces an f^2 of 0.2005, corresponding to a power of 86.39%. Using the average of these f^2 values (0.3069) in *a priori* power testing for the main study, and accepting a power level of 80%, produces a minimum total sample size estimate of 28.

Considering the use of the two-tailed t-test, adequate power is less consistent. The t-test between the control and status quo groups yields a Cohen's d of 1.2040 and corresponds to a power of 86.39%. However, the comparison between the control and present bias groups yields a Cohen's d of 0.5930 and a power of only 46.22%; between the control and loss aversion groups a Cohen's d of 0.8724 and a power of only 63.95%; and between the control and social norm groups a Cohen's d of 0.9638 and a power of only 72.56%. The relatively

[331] Post-hoc analysis is not performed on the first pilot study owing to a lack of significance.

large effect size (Cohen's d) is assuring, but the low power suggests a larger sample size would be beneficial. Again, using the average of these d values (0.9083) in *a priori* power testing for the main study, and accepting a power level of 80%, produces a minimum sample size estimate peer group of 42.

Chapter 14 – Experiment Implementation and Analysis

14.1 – Introduction

In this chapter, the results of the two-part survey experiment will be presented. The analysis used will follow methods detailed in Section 2 and ideas developed in the pilot studies detailed in Chapters 11 and 13. Where appropriate, comparison between the pilot studies and these results will be made. Given the nature of this project, results are presented in two parts. Firstly, the primer group data are analysed, and the various output produced from this analysis presented. Secondly, the personalised treatment group (PTG) data are analysed, and the hypotheses developed in Chapter 3 considered. A discussion of these results in relation to the wider literature is offered in Chapter 15.

14.2 – The Primer Group

14.2.1 Data Summary

A sample of 762 participants from the US were recruited using Amazon's *Mechanical Turk* (MTurk) service and were paid compensation of $0.50 for their participation. After removing 190 respondents (24.934%) who failed to pass an attention check question,[332] and 8 respondents who completed the survey inappropriately fast (less than 2 minutes), a sample of N = 564 remained (female = 38%).

Summary statistics are shown in Table 53:

[332] Respondents who failed the attention check were not compensated for their participation.

Table 53: Summary Statistics

Demographic	Frequency	Percentage of N	Average
Education:			3.092
(1) None	1	0.18	
(2) Highschool	76	13.48	
(3) Bachelor's Degree	363	64.36	
(4) Master's Degree	118	20.92	
(5) PhD	6	1.06	
Political Identity:			3.005
(1) Left-wing	89	15.78	
(2) Left-leaning	97	17.20	
(3) Centre	172	30.50	
(4) Right-leaning	134	23.76	
(5) Right-wing	72	12.77	
Age:			37.957
18-25	52	9.22	
26-35	236	41.84	
36-45	153	27.13	
46-55	68	12.06	
55<	55	9.75	

From Table 53, the median respondent is around 38 years of age, is qualified slightly beyond a bachelor's degree, and identifies as slightly right of centre. These demographics are very similar to those reported in both previous pilot studies. No statistically significant difference in distribution between the control group and the four treatment groups was found for age (χ^2 (200, N = 564) = 169.0, p = .945), sex (χ^2 (4, N = 564) = 0.28, p = .991), education (χ^2 (16, N = 564) = 18.3, p = .306) or political identity (χ^2 (16, N = 564) = 20.6, p = .194). The control and four treatment groups, therefore, appear comparable.

14.2.2 Testing of Assumptions

14.2.2.1 Normality

The four nudges under consideration are examined to see if they are effective or not when used impersonally, compared to a control group. The normality of the data is first reviewed.

Figure 14: Histogram Normality Plots

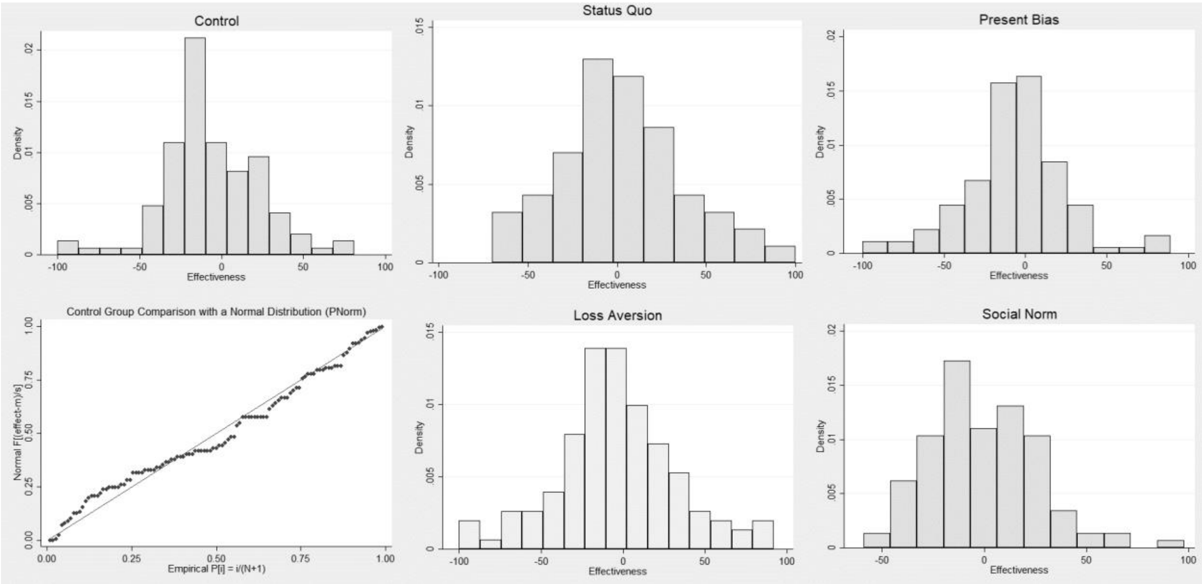
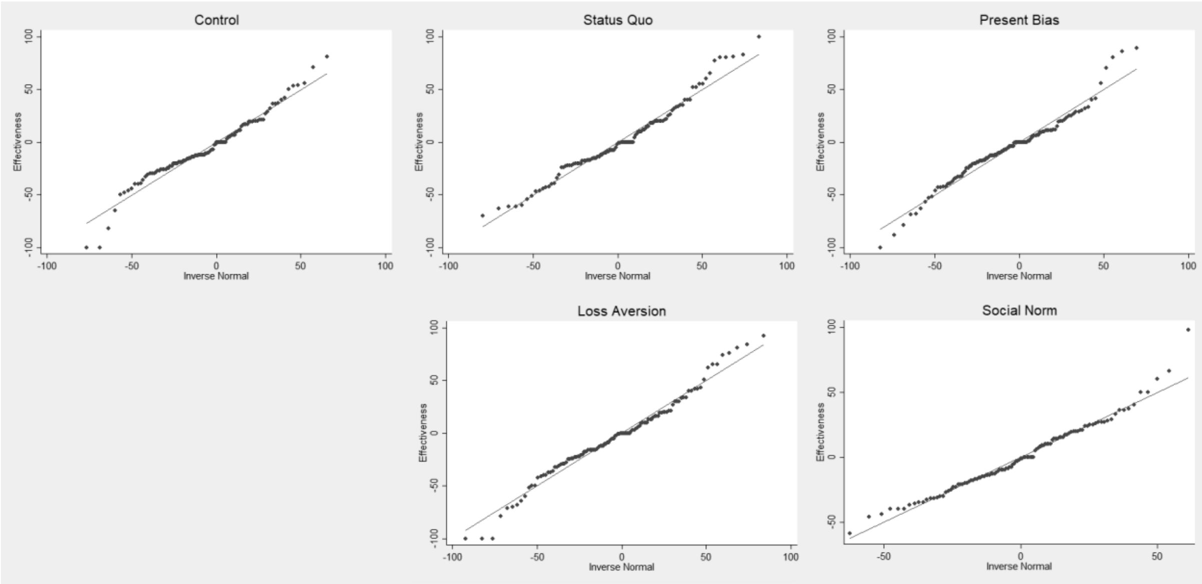


Figure 14 shows histogram plots for each group, with the dependent variable (effectiveness) shown on the x-axis and frequency density shown on the y-axis. As Figure 14 shows, there is good visual evidence of normality in all four nudges and the control group when examining the histogram plots. Plotting the control group against a normal distribution function (bottom-left) also shows a reasonably close alignment. However, despite the clear visual normality, Shapiro-Wilk's tests for all but the loss aversion group suggest the data are non-normal (control p = 0.0074; status quo p = 0.0285; present bias p = 0.0105; loss aversion p = 0.1164; social norm p = 0.0435). Such a situation can arise when data contain outlier values at the upper and lower ends of the distribution. QQplots[333] of the data reveal this to be the case:

[333] Figure 15 shows QQplots for each of the five groups. A QQplot shows the alignment of a dataset – in this instance, the effectiveness dependent variable – against a theoretical, normal distribution. Where the data are perfectly, normally distributed, they will fall on the straight line shown in the QQplot (i.e. $y = x$). Therefore, deviation from this line indicates non-normality.

Figure 15: QQPlots



In each instance, the data align – for the most part – with a normal distribution. Removing outlier values may correct for the apparent lack of normality, however, one should remove outliers only with justified cause. In this instance, there is none; indeed, the nature of this experiment would suggest disparate data should be retained. Therefore, despite the Shapiro-Wilk's results, the data are assumed to be normal given the visual evidence and the explanation for the Shapiro-Wilk p-values.

14.2.2.2 Homogeneity of Variance

Homogeneity of variance is examined between the control group and each of the four nudge groups using Levene's test. There is no statistically significant evidence the variance is heterogeneous in all instances (status quo $p = 0.1237$; present bias $p = 0.7788$; loss aversion $p = 0.0780$; social norm $p = 0.4820$), though Levene's test between the control and loss aversion group does produce a value which is significant at the 10% level.

The assumptions of the t-test are therefore accepted, and this test is used henceforth.

14.2.3 Impersonal Nudging

The results of these two-tailed t-tests are shown in Table 54:

Table 54: Two-tailed T-test Results of Impersonal Nudges vs. Control

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value
Status Quo	-5.902	1.936	-1.758	0.0801*
Present Bias	-5.902	-6.372	0.165	0.8687
Loss Aversion	-5.902	-4.034	-0.374	0.7089
Social Norm	-5.902	-0.468	-1.391	0.1656

* p < 10%, ** p < 5%, *** p < 1%

As Table 54 shows, only the status quo nudge produces an effectiveness score which is statistically significantly different from the advert seen by the control group, though this is only at the 10% level. All other nudges do not produce effectiveness scores which are statistically significantly different from the control group. These results, however, differ from the previous pilot studies in two important ways.

Firstly, unlike the first pilot study, where a similar lack of statistically significant difference was identified, all nudges produce average effectiveness scores which are greater than the average effectiveness of the adverts seen by the control group, except for the present bias nudge. While not statistically significantly different, there is little evidence that the set of nudges under consideration is having a *detrimental* effect, unlike in the first pilot study.

Secondly, the results shown in Table 54 are rather consistent with the results shown in the second pilot study in terms of averages, with both showing the status quo nudge to be the most effective, followed by the social norm nudge, the low aversion nudge and finally the present bias nudge. The major difference between these results and the results of the second pilot study is the average effectiveness for the control group. In Pilot Study 2, the control group has a large, negative effectiveness score (-36.923), while the control group in this experiment – while still negative – has a much smaller score (-5.902). The control group in Pilot Study 2, therefore, appears to anomalous, perhaps as a result of the small sample size used in the second pilot study. This casts doubt on the validity of the statistically significant differences identified between the nudge groups and the control group in Pilot Study 2.

Because there is an *a priori* expectation of the sign of the effect (i.e. that the effect of the nudge will be *positive*), a one-tailed t-test can be used to examine whether there is a statistically significant and *positive* effect. These results are shown in Table 55:

Table 55: One-tailed T-test Results of Impersonal Nudges vs. Control

Nudge	Control Mean	Treatment Mean	t-Statistic	p-value
Status Quo	-5.902	1.936	-1.758	0.0401**
Present Bias	-5.902	-6.372	0.165	0.5656
Loss Aversion	-5.902	-4.034	-0.374	0.3544
Social Norm	-5.902	-0.468	-1.391	0.0828*

* p < 10%, ** p < 5%, *** p < 1%

As Table 55 shows, using a one-tailed t-test, the status quo nudge advert is found to have a positive, statistically significant effect compared to the control. Furthermore, the social norm nudge is found to produce a positive effect which is statistically significant at the 10% level, compared to the control. Thus, there is now *some* evidence impersonal nudging may be effective at influencing political decision-making, but it is limited to only the status quo nudge. These results, therefore, continue to be comparable to the pilot studies, particularly the second pilot study.

Following the second pilot study, one may suspect that once moderation effects are analysed, the statistical insignificance initially identified may be explained. Firstly, as with the pilot studies, the presence of aesthetic effects is analysed.

14.2.4 Testing for the Presence of Aesthetic Effects

The results of an examination of aesthetic effects using a two-tailed t-test[334] are shown in Table 56:

[334] As a note for consistency, a two-tailed t-test is used here, rather than a one-tailed t-test, as there is no *a priori* expectation of the sign of the effect. Indeed, the expectation is that there is *no* aesthetic effect.

Table 56: Aesthetic Testing

Nudge	Mean (Candidate A)	Mean (Candidate B)	t-Statistic	p-value
Control	-8.944	-2.712	1.107	0.2707
Status Quo	2.793	0.961	-0.275	0.7841
Present Bias	-11.222	-0.260	1.830	0.0699*
Loss Aversion	-7.797	0.271	1.109	0.2696
Social Norm	-6.582	5.536	2.505	0.0137**

* p < 10%, ** p < 5%, *** p < 1%

Consistent with both pilot studies, there is some evidence to support the postulation that there is an aesthetic effect occurring. This is despite the RCT design and the larger sample size which, assuming there is no *genuine aesthetic* effect, in theory, should eliminate a bias which might wrongly be attributed to an aesthetic effect. Table 56 shows that the present bias treatment subgroup at the 10% level, and the social norm treatment subgroup at the 5% level, do have statistically significant differences in effectiveness scores when Candidate A is used in the nudge-advertisement compared to Candidate B. Furthermore, the control, present bias, loss aversion and social norm subgroups all suggest Candidate B is more effective than Candidate A, with the status quo nudge being the only exception. This is consistent with the results of the second pilot study.

While across the two pilot studies, and this study, statistical evidence for *some* aesthetic effect is apparent, it is not sufficiently strong to support a conclusion. Furthermore, there is little consistency in these results, with the most consistent tendency being the statistically insignificant indication that Candidate B is preferred to Candidate A.

14.2.5 Moderation Analysis

In contrast to the aesthetic effects, moderation effects *are expected* to influence participants. This follows from both theory and the results of Pilot Study 2. As with the pilot studies, differences in psychometric responses across the treatment groups are examined, with these results shown in Table 57:

Table 57: Differences in Psychometric Scores

Nudge	Rational	Avoidant	Intuitive	Dependent	Spontaneous	NFC	CFC
Status Quo	0.2179	0.2505	0.8068	0.2449	0.5369	0.3419	0.2634
	(0.779)	(0.752)	(0.565)	(0.489)	(0.228)	(0.812)	(0.482)
Present Bias	0.7489	0.7651	0.3641	0.9634	0.2907	0.5285	0.3359
	(0.603)	(0.160)	(0.328)	(0.408)	(0.746)	(0.371)	(0.475)
Loss Aversion	0.5164	0.0661*	0.7860	0.3469	0.9716	0.1106	0.6651
	(0.319)	(0.528)	(0.977)	(0.686)	(0.884)	(0.319)	(0.787)
Social Norm	0.9056	0.2183	0.2542	0.5867	0.5071	0.8012	0.7198
	(0.386)	(0.187)	(0.190)	(0.763)	(0.495)	(0.665)	(0.757)

χ^2 p-value shown in brackets, df =, N = 564
* p < 10%, ** p < 5%, *** p < 1%

Across the sample, only one instance of a statistically significant difference between the control group and a nudge group is identified. This is following a two-tailed t-test for difference in the means of the avoidant psychometric variable between the control group and the loss aversion treatment group. This difference is only significant at the 10% level and is not reproduced in the chi-squared test of distributions. Given this single instance of difference is not compelling, no adjustment is made in the analysis.

Cronbach's alpha scores for each of the seven psychometrics are presented in Table 58:

Table 58: Cronbach's Alpha Results for Psychometric Scales

Scale	Cronbach's Alpha
GDMS: Avoidant	0.9237
GDMS: Dependent	0.8729
GDMS: Intuitive	0.8842
GDMS: Rational	0.8357
GDMS: Spontaneous	0.8767
NFC	0.8781
CFC	0.7832

As each scale demonstrates a Cronbach's alpha value which is greater than a typically accepted threshold of around 0.6-0.7, averaging each of these scales appears wholly valid. Summary statistics of these psychometric variables are shown in Table 59:

Table 59: Summary Statistics of Psychometric Variables

Psychometric	Mean	Std. Dev.	Min	Max	Median
Rational	3.801	0.720	1.000	5.000	3.800
Avoidant	2.813	1.110	1.000	5.000	3.000
Intuitive	3.327	0.895	1.000	5.000	3.400
Dependent	3.259	0.884	1.000	5.000	3.400
Spontaneous	3.013	0.970	1.000	5.000	3.200
NFC	3.177	0.651	1.000	5.000	3.056
CFC	3.288	0.523	1.167	5.000	3.083

As can be seen with a comparison to the second pilot study, these summary statistics are very similar to those previously found in the smaller sample investigation.

Using these average figures, moderated regression models take the form:

Equation 8

$$effectiveness_i = \beta_0 + \beta_1 D_i + \beta_2 Psy_\lambda + \beta_3 D_i Psy_\lambda + \varepsilon_i \quad (8)$$

where $effectiveness_i$ is the effectiveness of nudge i, D_i is a dummy variable taking the value of 1 for nudge i, and 0 for all other values, psy_λ is a continuous variable for psychometric λ, and $D_i Psy_\lambda$ is a moderator term, are estimated. Table 60 through Table 63 present the results of moderated regressions for all possible combinations of nudge and psychometric variable.

Table 60: Moderated Regression Results for the Status Quo Nudge

Variable	Model 1A	Model 2A	Model 3A	Model 4A	Model 5A	Model 6A	Model 7A
Dummy	-23.165 [21.594]	7.031 [13.204]	31.389 [19.042]	-6.254 [18.572]	33.034** [13.785]	28.438 [26.720]	-31.728 [31.256]
Rational	-1.019 [4.191]						
D × Rat.	7.963 [5.818]						
Avoidant		-1.309 [2.926]					
D × Avo.		0.134 [4.292]					
Intuitive			2.552 [3.966]				
D × Intu.			-7.230 [5.594]				
Dependent				-1.941 [3.741]			
D × Dep.				4.281 [5.499]			
Spontaneous					2.276 [2.926]		
D × Spon.					-8.686** [4.244]		
NFC						5.446 [5.860]	
D × NFC						-6.583 [8.494]	
CFC							-6.248 [6.213]
D × CFC							11.846 [9.620]
Constant	-1.858 [15.770]	-1.84 [8.378]	-14.027 [13.612]	4.281 [5.499]	-12.491 [9.954]	-22.826 [18.709]	14.822 [19.651]
R-squared	0.0271	0.0156	0.0253	0.0175	0.0380	0.0201	0.0235
Multicollinearity	0.9931	0.9940	0.9997	0.9939	0.9983	0.9959	0.9943
N	222	222	222	222	222	222	222

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 61: Moderated Regression Results for the Present Bias Nudge

Variable	Model 1B	Model 2B	Model 3B	Model 4B	Model 5B	Model 6B	Model 7B
Dummy	-20.017 [23.837]	-3.401 [14.502]	-6.876 [18.886]	-14.314 [21.188]	7.083 [15.385]	60.858** [24.803]	19.598 [35.147]
Rational	-1.019 [4.190]						
D × Rat.	5.108 [6.467]						
Avoidant		-1.309 [2.926]					
D × Avo.		0.917 [4.547]					
Intuitive			2.552 [3.965]				
D × Intu.			1.754 [5.563]				
Dependent				-1.941 [3.741]			
D × Dep.				4.102 [6.024]			
Spontaneous					2.276 [2.925]		
D × Spon.					-2.585 [4.570]		
NFC						5.446 [5.859]	
D × NFC						-19.805** [8.014]	
CFC							-6.248 [6.212]
D × CFC							-6.413 [11.181]
Constant	-1.858 [15.767]	-1.833 [8.376]	-14.027 [13.609]	0.754 [12.688]	-12.491 [9.953]	-22.826 [18.706]	14.822 [19.647]
R-squared	0.0048	0.0012	0.0106	0.0035	0.0027	0.0510	0.0239
Multicollinearity	0.9995	0.9996	0.9963	1.0000	0.9950	0.9982	0.9959
N	226	226	226	226	226	226	226

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 62: Moderated Regression Results for the Loss Aversion Nudge

Variable	Model 1C	Model 2C	Model 3C	Model 4C	Model 5C	Model 6C	Model 7C
Dummy	-1.749 [24.378]	-10.719 [13.154]	-16.147 [19.477]	-39.278** [19.957]	5.175 [17.346]	22.376 [27.531]	13.082 [31.035]
Rational	-1.019 [4.190]						
D × Rat.	0.907 [6.676]						
Avoidant		-1.309 [2.925]					
D × Avo.		4.494 [4.284]					
Intuitive			2.552 [3.964]				
D × Intu.			5.370 [5.731]				
Dependent				-1.941 [3.740]			
D × Dep.				12.710** [5.924]			
Spontaneous					2.276 [2.925]		
D × Spon.					-1.176 [5.117]		
NFC						5.446 [5.858]	
D × NFC						-6.537 [8.611]	
CFC							-6.248 [6.211]
D × CFC							-3.395 [9.767]
Constant	-1.858 [15.764]	-1.833 [8.375]	-14.027 [13.607]	0.754 [12.686]	-12.491 [9.951]	-22.826 [18.702]	14.822 [19.644]
R-squared	0.0008	0.0077	0.0280	0.0423	0.0032	0.0062	0.0165
Multicollinearity	0.9982	0.9853	0.9997	0.9961	1.0000	0.9889	0.9992
N	231	231	231	231	231	231	231

Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Table 63: Moderated Regression Results for the Social Norm Nudge

Variable	Model 1D	Model 2D	Model 3D	Model 4D	Model 5D	Model 6D	Model 7D
Dummy	-6.213 [20.811]	16.794 [11.396]	19.659 [16.499]	22.502 [15.633]	26.288* [13.650]	18.229 [23.920]	-62.504** [25.812]
Rational	-1.019 [4.191]						
D × Rat.	3.047 [5.573]						
Avoidant		-1.309 [2.926]					
D × Avo.		-4.266 [3.832]					
Intuitive			2.552 [3.965]				
D × Intu.			-4.348 [4.797]				
Dependent				-1.941 [3.741]			
D × Dep.				-5.342 [4.584]			
Spontaneous					2.276 [2.925]		
D × Spon.					-6.920* [4.102]		
NFC						5.446 [5.859]	
D × NFC						-4.125 [7.449]	
CFC							-6.248 [6.212]
D × CFC							20.732** [8.144]
Constant	-1.858 [15.769]	-1.833 [8.377]	-14.027 [13.610]	0.754 [12.689]	-12.491 [9.954]	-22.826 [18.708]	14.822 [19.649]
R-squared	0.0104	0.0338	0.0132	0.0345	0.0233	0.0168	0.0295
Multicollinearity	0.9999	0.9932	0.9941	0.9987	0.9980	0.997	0.9994
N	224	224	224	224	224	224	224

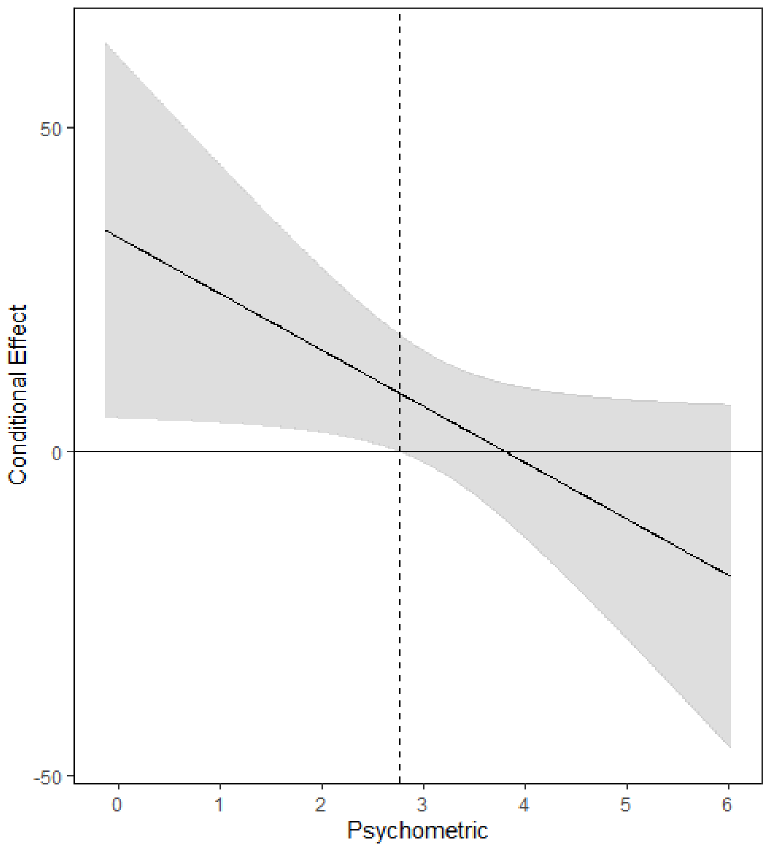
Robust SE shown in brackets
* p < 10%, ** p < 5%, *** p < 1%

Building on the trends identified in the second pilot study, several statistically significant moderation effects are identified across the four nudges and seven psychometrics variables. In total, five moderation effects are statistically significant at the 10% level (17.9% of possible relationships) and four at the 5% level (14.3%). Furthermore, at least one moderation effect is identified for each nudge, with the social norm nudge treatment group producing two moderation effects, though one only is statistically significant at the 10% level. For completeness, all five moderation effects are further analysed using the Johnson-Neyman technique. However, any moderation effects which are significant at the 10% level are not used in the personalisation stage.

14.2.5.1 Spontaneity and the Status Quo Nudge

The first instance of moderation can be found in Model 5A, where the interaction between the dummy variable demarcating the presence of the status quo nudge and the spontaneous decision-making style is statistically significant at the 5% level. Using the Johnson-Neyman technique, this result indicates regions of significance exist for all values less than 2.77 and all values more than 47.84 on the spontaneity scale. The moderation effect of spontaneity is visualised in Figure 16:

Figure 16: Moderation Effect of Spontaneity on the Status Quo Nudge

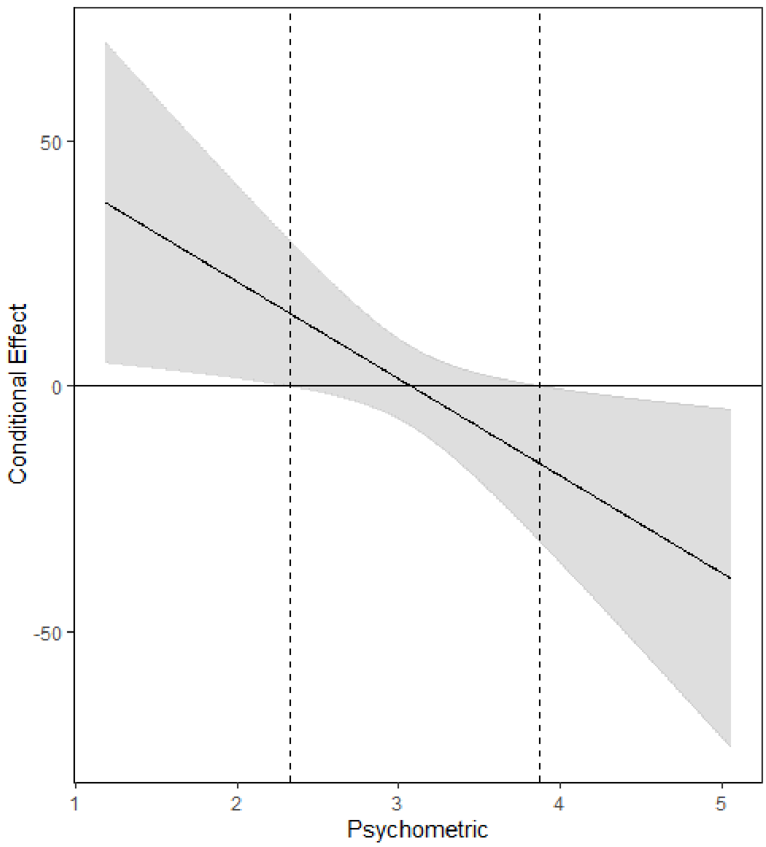


Given a person can only score between 1 and 5 on the spontaneity decision-making style, only the region of significance that exists for all values less than 2.77 is shown. Because the relationship is negative, the status quo nudge is expected to be most effective for those who score below 2.77.

14.2.5.2 NFC and the Present Bias Nudge

The second instance of moderation can be found in Model 6B, where the interaction between the dummy variable demarcating the presence of the present bias nudge and the NFC psychometric variable is statistically significant at the 5% level. This result is noteworthy as it was indicated in the second pilot study, but only at the 10% level. Again, using the Johnson-Neyman technique, this result indicates that regions of significance exist for all values less than 2.33 and for all values greater than 3.87 on the NFC scale. The moderation effect of NFC is visualised in Figure 17:

Figure 17: Moderation Effect of NFC on the Present Bias Nudge

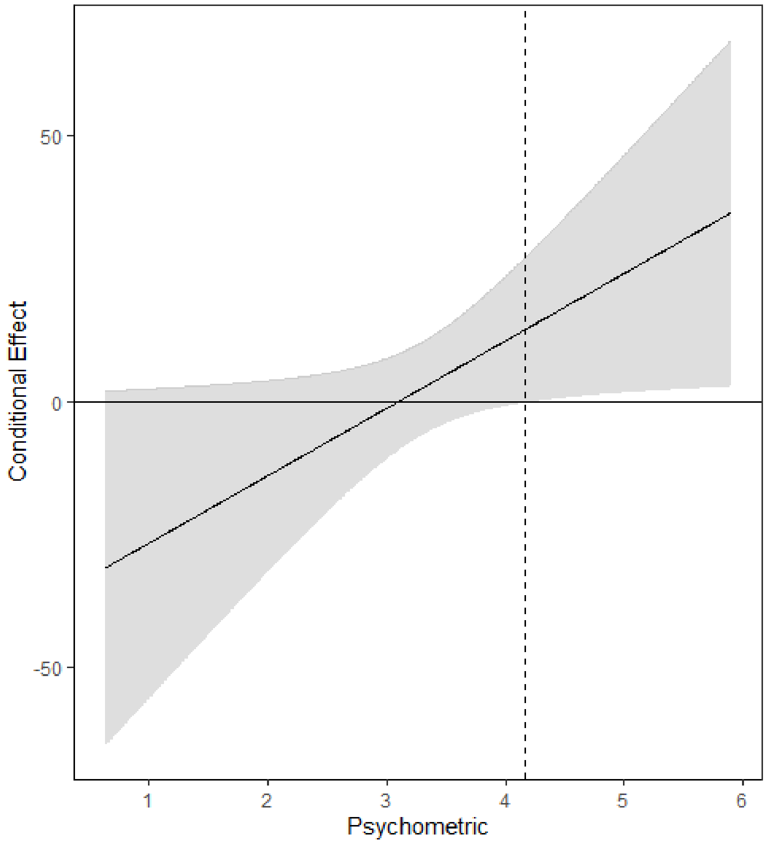


As a person can score both below 2.33 and above 3.87 on the NFC scale, both regions of significance are shown in Figure 17. As the relationship is again negative, the present bias nudge is expected to be most effective with individuals who score below 2.33, and least effective with those that score above 3.87.

14.2.5.3 Dependence and the Loss Aversion Nudge

The third instance of moderation can be found in Model 4C, where the interaction between the dummy demarcating the presence of the loss aversion nudge and the dependent decision-making style is statistically significant at the 5% level. Again, using the Johnson-Neyman technique, this result indicates that regions of significance exist for all values less than -1.59 and for all values greater than 4.16. The moderation effect of dependence is visualised in Figure 18:

Figure 18: Moderation Effect of Dependence on the Loss Aversion Nudge



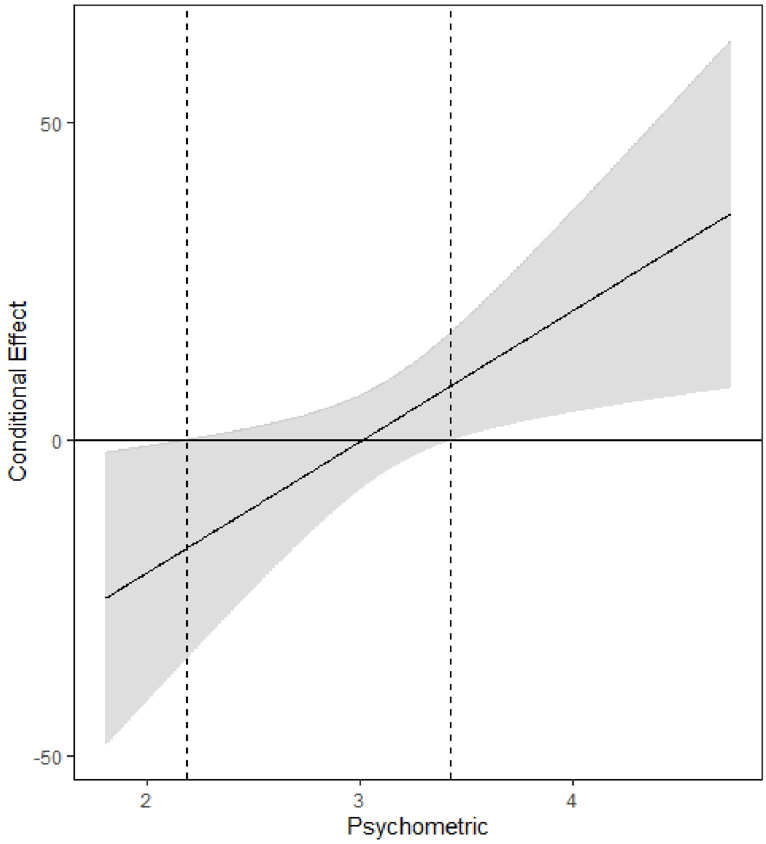
As a person can only score between 1 and 5 on the dependence scale, only the region of significance that exists for all values above 4.16 is shown in Figure 18. Unlike previous moderation effects, this relationship is positive, and so the loss aversion nudge is expected to be most effective for individuals who score above 4.16.

14.2.5.4 CFC and the Social Norm Nudge

The fourth instance of moderation can be found in Model 7D, where the interaction between the dummy variable demarcating the presence of the social norm nudge and the CFC psychometric variable is statistically significant at the 5% level. Interestingly, this relationship was not identified in the second pilot study, and no evidence of moderation between those psychometric variables which were identified (avoidance and dependence) is identified here.[335] Regardless, again using the Johnson-Neyman technique, this result indicates that regions of significance exist for all values less than 2.19 and all values greater than 3.42. The moderation effect of the CFC psychometric is visualised in Figure 19:

[335] Recall, however, that these results had not been previously expected, and in the case of the moderation effect involving the dependence psychometric variable, no adequate qualitative explanation for this interaction was forthcoming.

Figure 19: Moderation Effect of CFC on the Social Norm Nudge

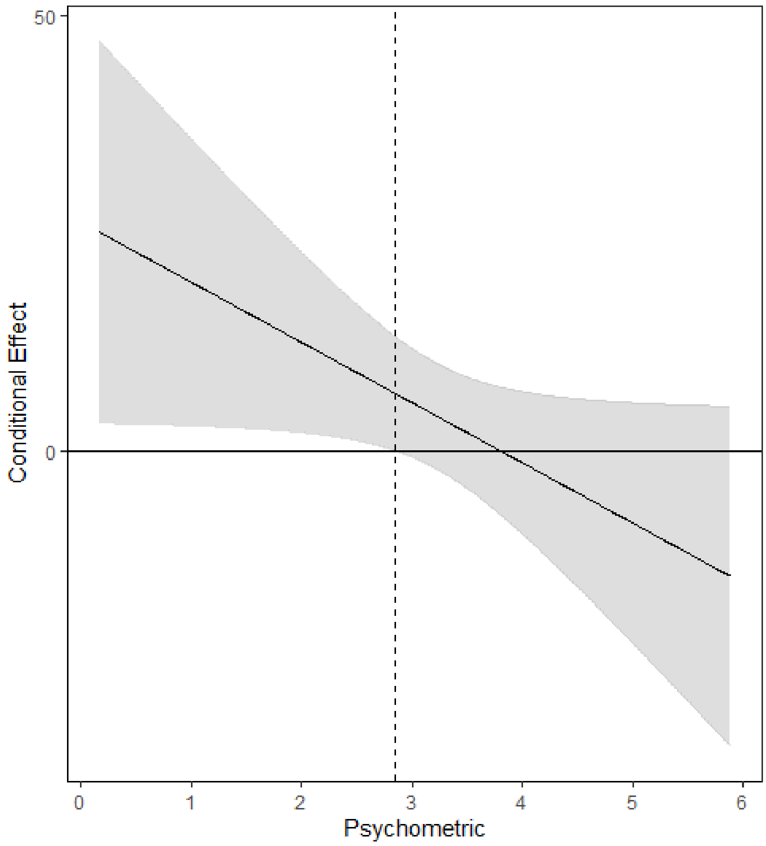


As a person can score below 2.19 and above 3.42 on the CFC scale, both identified regions of significance are shown in Figure 19. As the relationship is positive, the social norm nudge is expected to be most effective for individuals who score above 3.42, and least effective for those who score below 2.19.

14.2.5.5 Spontaneity and the Social Norm Nudge

The final instance of moderation can be found in Model 5D, where the interaction between the dummy variable demarcating the presence of the social norm nudge and the spontaneous decision-making style is statistically significant at the 10% level. Again, using the Johnson-Neyman technique, this result indicates that a region of significance exists between the values of -296.62 and 2.85. The moderation effect of spontaneity is visualised in Figure 20:

Figure 20: Moderation Effect of Spontaneity on the Social Norm Nudge



As a person can only score between 1 and 5 on the spontaneity scale, the identified region of significance, *in practice*, can be said to exist between the values of 1 and 2.85. Because the relationship is negative, the social norm nudge is expected to be most effective for individuals who score below 2.85 on the spontaneity scale.

14.2.5.6 Summary of Moderation Effects

Further discussion of these results will be offered in Chapter 15. At present, these results are summarised in Table 64:

Table 64: Regions of Significance

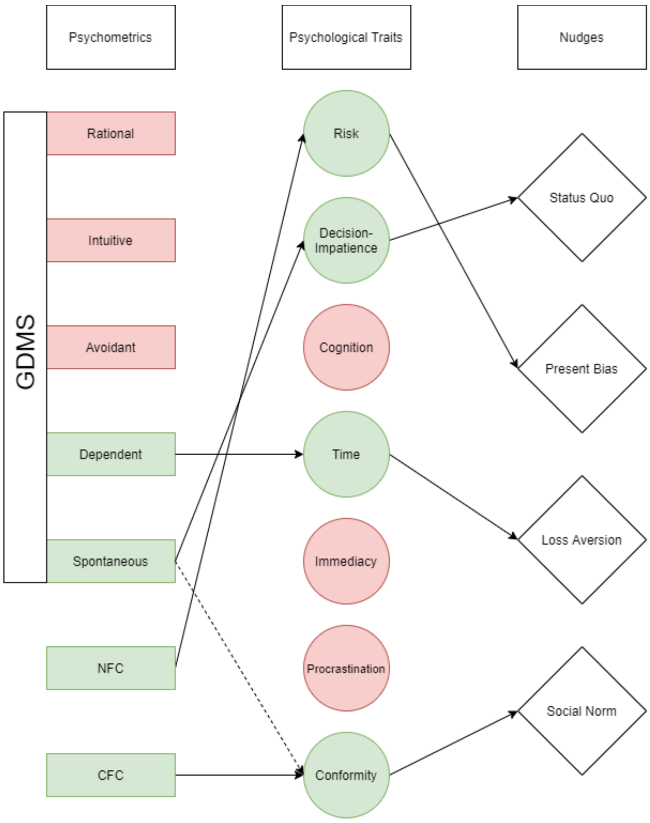
Psychometric	Nudge			
	Status Quo	Present Bias	Loss Aversion	Social Norm
Rational				
Avoidant				
Intuitive				
Dependent			4.16<	
Spontaneous	<2.77			<2.85[a]
NFC		<2.33, 3.87<		
CFC				<2.19, 3.42<

[a] Significant at the 10% level

14.2.6 Empirical Psychometric Map and PTG Survey Design

Using these results, the hypothesised psychometric map given in Chapter 6 is adapted, and in the process greatly simplified, to reflect observation.

Figure 21: Empirical Psychometric Map



As shown in Figure 21, only four of the seven psychometric scales contribute to the map following the moderation analysis. Furthermore, each of the four nudges appears to have only one statistically significant relationship with a psychometric. The exception is the social norm nudge, which has a statistically significant moderated relationship at the 10% level with the spontaneous psychometric, as shown with the dotted arrow in Figure 21.

Interestingly, all relationships identified in this analysis follow only *one* predicted route via a psychological trait. For instance, several traits were predicted to influence the present bias, including risk, but following the moderation analysis, *only risk* can explain the observed statistically significant relationship between NFC and the present bias nudge. Even in the instance of the social norm nudge – which, again, was predicted to be affected by several psychological traits – where multiple relationships between it and the psychometrics variables have been found, *only one psychological trait* – conformity – would explain these observations.

Given the psychometric map would predict *some* relationship between each trait and nudge, these moderation effects motivate the induction of the following qualitative considerations:

1. Low Spontaneity and High Status Quo: A person with low spontaneity is expected to make decisions slowly (Scott and Bruce, 1995). This runs counter to some explanations of the status quo nudge, namely that the status quo is preferred by impatient individuals (Johnson et al., 2012). However, the status quo is also expected to appeal to those who are reluctant to make decisions and is postulated to work via an implicit recommendation mechanism (Tannenbaum and Ditto, 2011; Johnson and Goldstein, 2003; Madrian and Shea, 2001). Patient people may also be understood as those who avoid making decisions by taking a long time to decide. Patient people may also be able to pick up on the implied recommendation of the status quo. In these instances, low spontaneity may explain the appeal of the status quo nudge.
2. Low NFC and High Present Bias: A person with a low NFC is expected not to evaluate the risks associated with their decisions as much as someone who enjoys cognitive tasks (Estelami, 2020; Hadj-Abo et al., 2020; Lin, Yen and Chuang, 2006). The present bias is expected to appeal to those with a low sense of risk who do not evaluate the temporal consequences of their decisions (Benhabib, Bisin and Schotter, 2010;

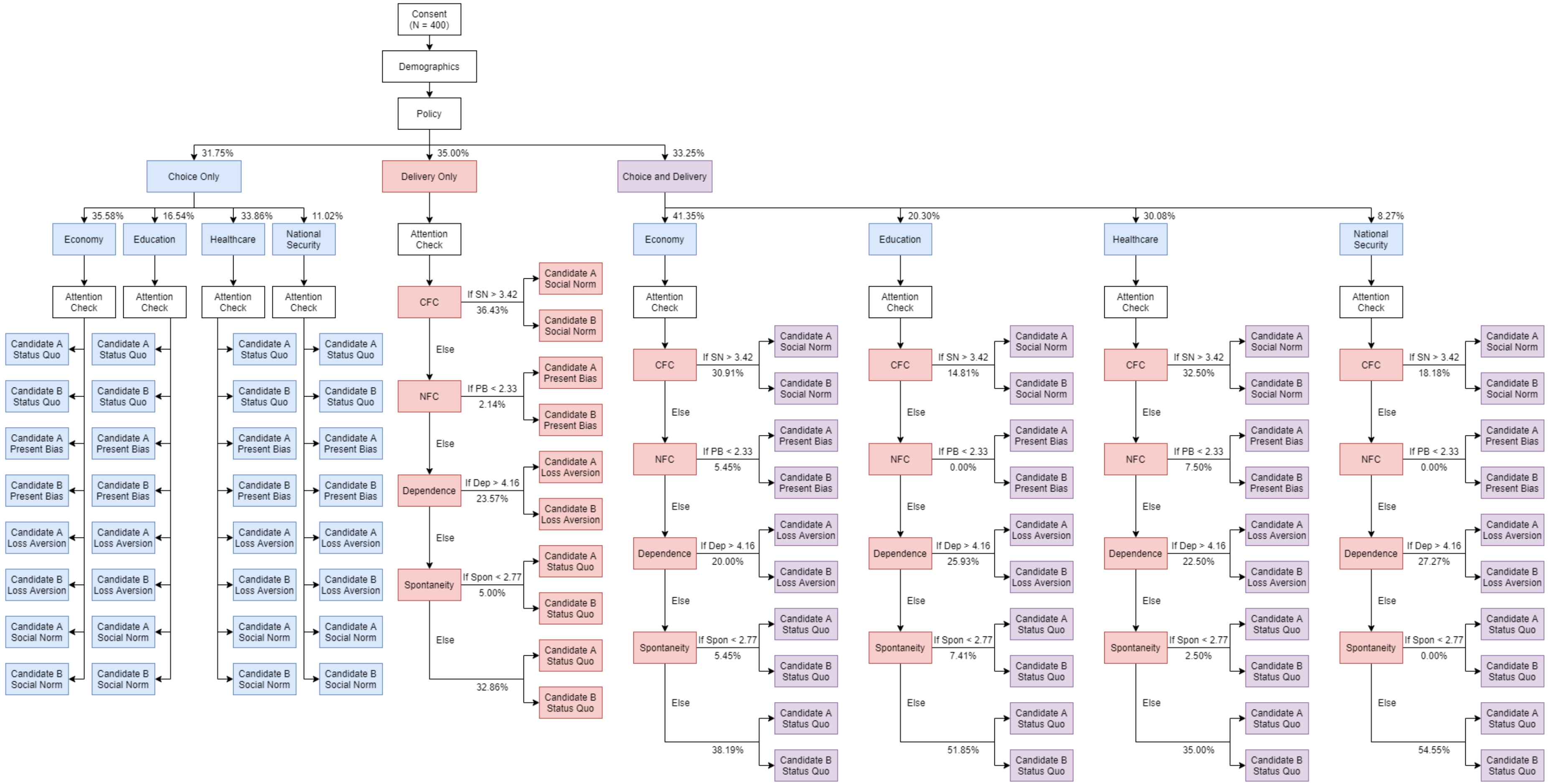
O'Donoghue and Rabin, 1999b). Therefore, low NFC does explain the appeal of the present bias.

3. High Dependence and High Loss Aversion: A person with high dependence is expected to rely on others and regularly search for the opinions of others (Scott and Bruce, 1995). Time is expected to contribute to loss aversion, as people who more regularly evaluate and re-evaluate their decisions are expected to be more loss averse (Benartzi and Thaler, 1995). High dependence may therefore explain the appeal of the loss aversion nudge and highly dependent people regularly search for the opinions of others and thus re-evaluate their own decisions.
4. High CFC and High Social Norm: A person with high CFC is expected to carefully consider the future consequences of their actions and decisions (Strathman et al., 1994). The social norm nudge is expected to appeal to those who conform (Bernheim, 1994), and the risks and consequences of *not* conforming are expected to drive adherence to the social norm (Sunstein, 1996). High CFC, therefore, would explain the appeal of the social norm nudge as people have a greater understanding of the potential consequences of not conforming and thus adopt conforming behaviour (i.e. follow the social norm).
5. Low Spontaneity and Low Social Norm: As above, a person with low spontaneity is expected to make decisions slowly (Scott and Bruce, 1995). This would allow a person time to consider the merits of the social norm, including the potential consequences of not following the social norm (Sunstein, 1996). However, the slower decision-making may also allow a person to evaluate a decision without the influence of the decisions of others, and may lead to low conformity, and thus a low appeal of the social norm nudge. The explanation for this relationship is inconclusive.

In each instance excluding the identified relationship between spontaneity and the social norm nudge, the relationships appear to be qualitatively robust. The single instance where this is not found should not, however, be seen as a concern, as this relationship was significant only at the 10% level and has already been excluded from use in the PTG data collection.

Using the data collected and the results produced in the primer group, the survey-experiment for the PTG data collection can be constructed. Weightings of distributions shown in Figure 22 are as observed in the PTG data sample. See Figure 22 below:

Figure 22: PTG Survey Flow



For the reader's benefit, Figure 22 is coloured coded. Blue items relate to functions and outputs which utilise choice personalisation only. Red items relate to functions and outputs which utilise delivery personalisation only. Finally, purple items relate to functions and outputs which utilise both choice and delivery personalisation.

In practice, the attention check question is embedded within the dependence psychometric scale but is shown as a separate item in Figure 22 for visual ease. Participants in the CO group, therefore, *do* respond to the dependence psychometric scale, but only insofar as it checks they are paying attention. While omitted for clearer visualisation, participants in all groups have an even chance of being shown a nudge-advertisement containing either Candidate A or Candidate B. This survey experiment, therefore, continues to follow an RCT design. Finally, the reported sample size in Figure 22 (N = 400) follows the removal of those who fail the attention check. See below.

The nudge preference used in the survey-experiment follows the ranking of the modulus of moderation term coefficients (β_3) discussed in Chapter 8. Following this procedure, the social norm nudge ($|\beta_3| = 20.732$) proceeds, followed by the present bias nudge ($|\beta_3| = 19.805$), the loss aversion nudge ($|\beta_3| = 12.710$) and the status quo nudge ($|\beta_3| = 8.686$). Again, following Chapter 8, participants who do not score within a region of significance are shown the most effective impersonal nudge, which is identified above as the status quo nudge. Advertisements shown to members of the choice personalisation only (CO) and choice and delivery personalisation (CD) groups in the PTG are shown in Table 65:

Table 65: Nudge-Advertisements with Choice Personalised Slogans

	Status Quo	Present Bias	Loss Aversion	Social Norm
The Economy				
				
Education				
				
Healthcare				
				
National Security				
				

14.3 – The Personalised Treatment Group

14.3.1 Data Summary

A sample of 441 participants were recruited using Amazon's *Mechanical Turk* service and were compensated $0.50 for their participation. After removing 41 participants who failed an attention check (9.297%), 400 participants remained (female = 32%).[336] As the PTG group contained only psychometric questions found to be statistically significant moderators in the primer group, the PTG survey was relatively shorter than the primer survey, and so participants were not removed following rapid completion.

Summary statistics are shown in Table 66:

Table 66: Summary Statistics

Demographic	Frequency	Percentage of N	Average
Education:			3.055
(1) None	1	0.25%	
(2) Highschool	49	12.25%	
(3) Bachelor's Degree	281	70.25%	
(4) Master's Degree	65	16.25%	
(5) PhD	4	1.00%	
Political Identity:			3.123
(1) Left-wing	58	14.50%	
(2) Left-leaning	54	13.50%	
(3) Centre	127	31.75%	
(4) Right-leaning	103	25.75%	
(5) Right-wing	58	14.50%	
Age:			36.94
18-25	62	15.50%	
26-35	176	44.00%	
36-45	80	20.00%	
46-55	44	11.00%	
55<	38	9.50%	

[336] Those respondents who failed the attention check were not compensated. The relatively lower rate of attention check failure (9.297% vs. 24.934%) can be attributed to the relatively shorter survey time for the PTG group.

The demographic breakdown of the PTG sample is very similar to that of previous samples, with the average respondent being around 37 years of age, qualified slightly beyond a bachelor's degree, and identifying slightly right of the political centre. No statistically significant difference between the control group and three treatment groups is found for age (χ^2 (159, *N* = 400) = 153.4, *p* = .61), sex (χ^2 (3, *N* = 400) = 2.2, *p* = .52), education (χ^2 (12, *N* = 400) = 9.8, *p* = .64) or political identity (χ^2 (12, *N* = 400) = 14.4, *p* = .28).

14.3.2 Testing for the Presence of Aesthetic Effects

As with previously, an investigation of the usefulness of the RCT design is carried out using two-tailed t-tests to identify the presence, or lack thereof, of significant differences in nudge effectiveness when Candidate A is in the nudge-advertisement versus Candidate B. These results are presented in Table 67:

Table 67: Aesthetic Effects

Group	Mean (Candidate A)	Mean (Candidate B)	t-Statistic	p-value
Control	-8.944	-2.712	1.107	0.2707
CO	11.475	16.485	-0.782	0.4356
DO	4.353	2.028	0.427	0.6699
CD	10.235	17.815	-1.209	0.2288

* p < 10%, ** p < 5%, *** p < 1%

Results for the control group are repeated from above. As can be seen, no statistically significant difference is found between Candidates A and B for all the groups examined. Consistent with previous results, however, Candidate B does appear to produce higher effectiveness scores than Candidate A. The exception here is in the DO group. The distribution of observations across subgroups are relatively even for all groups.

Based on these results, there does not appear to be any statistically significant aesthetic effects in the PTG sample, and so the RCT design appears to be effective.

14.3.3 Three-way Comparisons

Three-way comparisons of the control and impersonal groups with the choice only (CO), delivery only (DO) and choice and delivery (CD) groups are performed using a one-way ANOVA. Prior to this test, the assumptions of normality and homogeneity of variance are examined.

14.3.3.1 Assumption Testing

Figure 23: Histogram Normality Plots

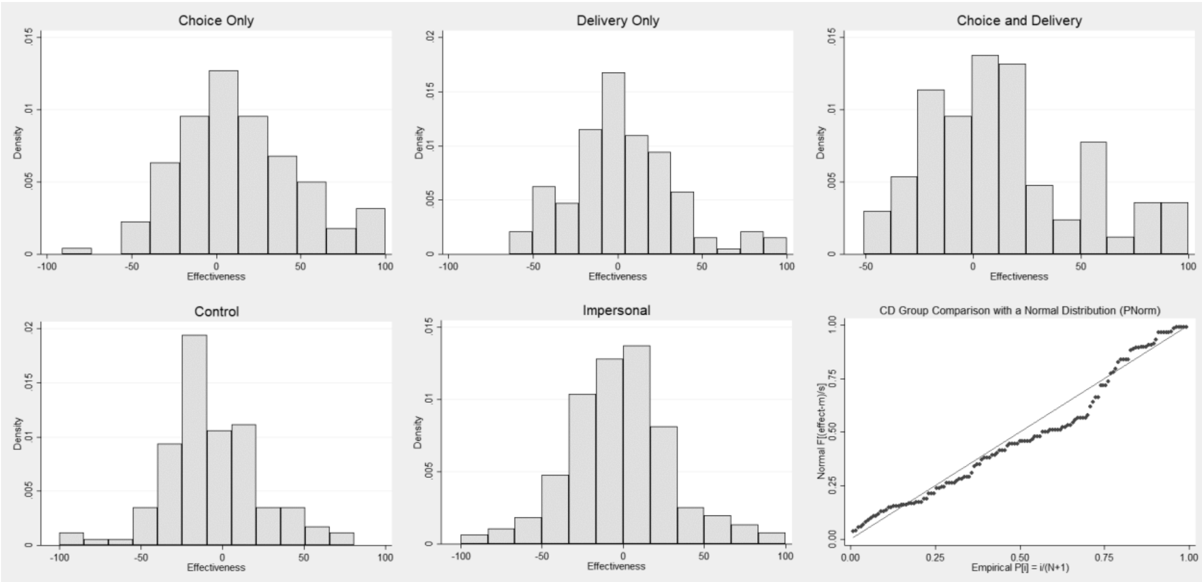


Figure 23 shows histogram plots for the three personalised nudging groups, the impersonal nudging group, and the control group, with the dependent variable (effectiveness) shown on the x-axis and the frequency density shown on the y-axis. All groups appear to be normally distributed based on a histogram plot. The exception may be the CD group, and so a further plot of this group against a normal distribution is produced (bottom-right). This further plot would seem to suggest the CD is reasonably normal also, and so a one-way ANOVA seems suitable.[337] Given the large sample size difference between the impersonal group (N = 451) and the other groups (N of around 130), Levene's test for homogeneity of variance is performed. The finding (p = 0.0849) suggests variance may not be homogeneous at the 10%

[337] Even excepting the non-normality of the CD group, various authors have demonstrated that the normality assumption of a one-way ANOVA can often be relaxed with relatively little impact on the test's reliability. See Glass, Peckham and Sanders (1972), Harwell et al. (1992) and Lix, Keselman and Keselman (1996).

level. In conjunction with an ANOVA, therefore, the Kruskal-Wallis test is utilised given the possibility of non-normality, and Welch's test is used to account for a lack of homogeneity of variances.

An ANOVA model (F (2, 688) = 14.53, p = 0.0000) considering the control group, the impersonal nudge group and the CO group suggests that at least one of these groups is statistically significantly different to another when means are compared, as do ANOVA models for the CD group (F (2, 694) = 14.62, p = 0.0000) and the DO group (F (2, 701) = 2.54, p = 0.0795), though the latter only at the 10% level. Using a Kruskal-Wallis test, a statistically significant difference is found for the CO comparison (p = 0.0001) and the CD comparison (p = 0.0001) at the 5% level and the DO group (p = 0.0960) at the 10% level. Using Welch's test, a statistically significant difference is found for the CO comparison (p = 0.0000) and the CD comparison (p = 0.0000) at the 5% level and the DO group (p = 0.0710) at the 10% level.

14.3.4 Testing Hypothesis 1

Hypothesis 1 states:

Hypothesis 1: *Personalised nudges will be statistically significantly more effective at influencing political decision-making then impersonal nudges, which in turn will be more effective than not nudging.*

Given the evidence of a statistically significant difference between the control group, the impersonal group and the personalised nudging groups, *initial* evidence would seem to support hypothesis 1. To more formally investigate this hypothesis, Tukey's post-hoc test is first utilised. These results are presented in Table 68 through Table 70.

Table 68: Tukey's Test for CO vs. Impersonal vs. Control

Comparison	Contrast	Standard Error	p-value
Impersonal vs. Control	3.391	3.463	0.590
CO vs. Control	19.848	4.257	0.000***
CO vs. Impersonal	16.457	3.307	0.000***
* p < 10%, ** p < 5%, *** p < 1%			

Table 69: Tukey's Test for DO vs. Impersonal vs. Control

Comparison	Contrast	Standard Error	p-value
Impersonal vs. Control	3.391	3.384	0.576
DO vs. Control	8.847	4.069	0.076*
DO vs. Impersonal	5.456	3.113	0.187
* p < 10%, ** p < 5%, *** p < 1%			

Table 70: Tukey's Test for CD vs. Impersonal vs. Control

Comparison	Contrast	Standard Error	p-value
Impersonal vs. Control	3.391	3.471	0.592
CD vs. Control	19.630	4.222	0.000***
CD vs. Impersonal	16.239	3.256	0.000***
* p < 10%, ** p < 5%, *** p < 1%			

From these pairwise comparisons, consistent evidence can be found that suggests the personalised nudging groups are statistically significantly different from *at least* the control group in every case at the 10% level. For the CO and CD groups, this conclusion can be made at the 5% level, and extended to also capture the impersonal nudging group. Two-tailed t-tests expand on these conclusions:

Table 71: Two-tailed T-test of Personalised Groups vs. Control Group

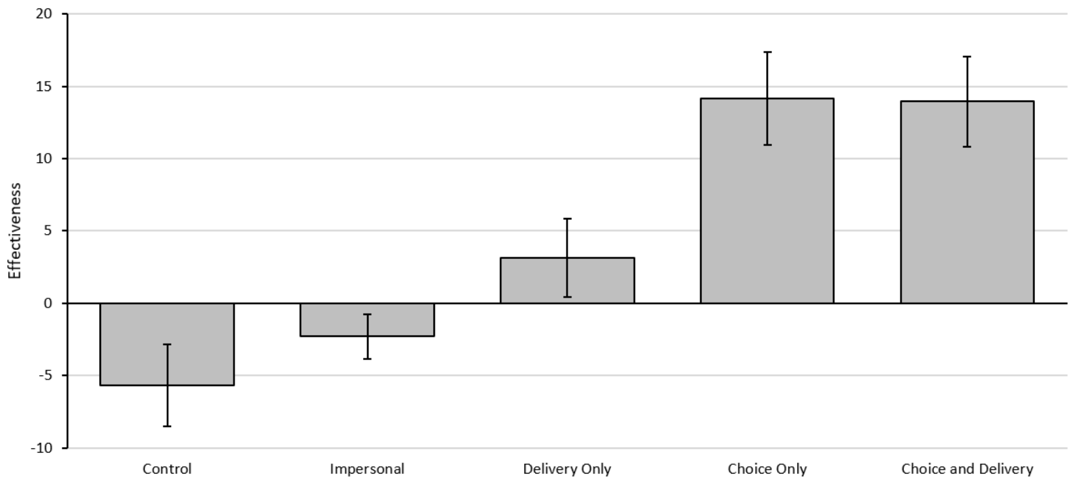
Group	Control Mean	Group Mean	t-Statistic	p-value
CO	-5.690	14.157	-4.615	0.0000***
DO	-5.690	3.157	-2.246	0.0256**
CD	-5.690	13.940	-4.584	0.0000***
* p < 10%, ** p < 5%, *** p < 1%				

Table 72: Two-tailed T-test of Personalised Groups vs. Impersonal Group

Group	Control Mean	Group Mean	t-Statistic	p-value
CO	-2.299	14.157	-4.895	0.0000***
DO	-2.299	3.157	-1.731	0.0840*
CD	-2.299	13.940	-4.905	0.0000***
* p < 10%, ** p < 5%, *** p < 1%				

Two-tailed t-test comparisons find that all personalised nudging groups are statistically significantly different from the control group at the 5% level, and all are statistically significantly different from the impersonal group at the 10% level, with – again – the DO group responsible for this reduced level of confidence. An examination of the means of these groups suggests that this difference is positive, which is to say, the effectiveness of the personalised nudging groups is statistically significantly and positively different from both the control group and the impersonal nudging group. This is visualised in Figure 24:

Figure 24: Comparison of Average Effectiveness



As there is an *a priori* expectation of a *positive effect* from the nudges, however, a one-tailed t-test may be used to investigate the difference in groups. These results are presented in Table 73 and Table 74 below:

Table 73: One-tailed T-test of Personalised Groups vs. Control Group

Group	Control Mean	Group Mean	t-Statistic	p-value
CO	-5.690	14.157	-4.615	0.0000***
DO	-5.690	3.157	-2.246	0.0128**
CD	-5.690	13.940	-4.584	0.0000***

* p < 10%, ** p < 5%, *** p < 1%

Table 74: One-tailed T-test of Personalised Groups vs. Impersonal Group

Group	Control Mean	Group Mean	t-Statistic	p-value
CO	-2.299	14.157	-4.895	0.0000***
DO	-2.299	3.157	-1.731	0.0420**
CD	-2.299	13.940	-4.905	0.0000***
* p < 10%, ** p < 5%, *** p < 1%				

Using the one-tailed t-test, the effect of all personalisation groups is found to be statistically significantly different from both the control and impersonal group effects at the 5% level. There would seem, then, good evidence to support hypothesis 1; albeit only partially. While personalised nudging appears to be more effective than the impersonal and control groups, there is little evidence to suggest the impersonal group is statistically *significantly* more effective than the control group, despite having a *more positive* mean. Tukey's test, for instance, consistently fails to find evidence of statistical significance, and returning to the results of the nudge subgroup comparisons in the primer group, only the status quo nudge was found to be statistically significantly different from the control group, and only at the 10% level.

Table 75: Two-tailed T-test of Impersonal Group vs. Control Group

Group	Control Mean	Group Mean	t-Statistic	p-value
Impersonal	-5.690	-2.299	-1.001	0.3172
* p < 10%, ** p < 5%, *** p < 1%				

This lack of statistically significant difference is confirmed by a two-tailed t-test shown in Table 75 between the impersonal group and the control group.

14.3.5 Testing Hypothesis 2

Hypothesis 2 states:

Hypothesis 2: *Choice and Delivery personalised nudges will be statistically significantly more effective at influencing political decision-making then delivery or choice personalised nudges alone.*

To test hypothesis 2, an ANOVA model comparing the CO, DO and CD groups is estimated. This model ($F(2, 399) = 4.47$, $p = 0.0120$) suggests that there is a statistically significant difference between the means of *at least* 2 of these groups at the 5% level. A Kruskal-Wallis estimation returns the same result ($p = 0.0226$), as well as a Welch test estimation ($p = 0.009$). To identify between which groups this difference lies, Tukey's test is again utilised.

Table 76: Tukey's Test for CO vs. DO vs. CD

Comparison	Contrast	Standard Error	p-value
CO vs. DO	-11.000	4.257	0.010***
CD vs. CO	-0.218	4.310	0.960
CD vs. DO	10.783	4.206	0.011**
* p < 10%, ** p < 5%, *** p < 1%			

As seen in Table 76, the DO group appears to be statistically significantly differently from the CO and CD groups, while these groups do not appear to be statistically significantly different from one another. This conclusion is supported by two-tailed t-tests shown in Table 77 through Table 79:

Table 77: Two-tailed T-test of CO vs. DO

CO Mean	DO Mean	t-Statistic	p-value
14.157	3.157	2.614	0.0087***
* p < 10%, ** p < 5%, *** p < 1%			

Table 78: Two-tailed T-test of CD vs. DO

CD Mean	DO Mean	t-Statistic	p-value
13.940	3.157	-2.607	0.0096***
* p < 10%, ** p < 5%, *** p < 1%			

Table 79: Two-tailed T-test of CO vs. CD

CO Mean	CD Mean	t-Statistic	p-value
14.157	13.940	0.049	0.9613
* p < 10%, ** p < 5%, *** p < 1%			

As can be seen, the effect of the nudges in the DO group is statistically significantly less effective than the CO and CD groups, which are themselves insignificantly different.[338] Hypothesis 2, therefore, must be rejected. While the hypothesis would appear to hold true when only considering delivery personalisation, when choice personalisation is also considered, the effect of the nudges in the CD group is not so different as to be statistically significant and – by the slightest of margins – is actually *less* than choice personalisation alone. Again, these results can be seen visually in Figure 24.

[338] Hypothesis 2 holds that the effect of the nudges in the CD group will be more positive than either the CO or DO groups, and thus, again, a one-tailed t-test can be used. This does not substantially change the conclusions drawn from a two-tailed t-test.

Section 4:

Discussion and Conclusion

Chapter 15 – Discussion

15.1 – Introduction

For the most part, Chapter 14 is structured to present results with little space available for drawing out connections between the findings and the wider literature, or indeed, contextualising the findings within the wider literature. This chapter expands on the preceding analysis.

15.2 – The Implications of Hypothesis 1

Based on the results in Chapter 14, hypothesis 1 is *partially* accepted. The evidence does seem to support the postulate that personalisation – in both the choice and delivery varieties – does produce more positively effective nudges, when compared to both impersonal nudging and not nudging at all. However, the evidence suggests that the adverts using impersonal nudges were not statistically significantly more positvively effective than adverts seen by the control group.

The basis for postulating that the impersonal nudges *would be* more positively effective than not nudging at all stems from the basic claim of nudge theory, namely that changing choice architecture should affect change in human behaviour. Of course, this basic claim is highly generalised and couched in theory. As considered in Chapter 2, the realities of applied nudging are often less definitive. For instance, Thunström, Gilbert and Jones-Ritten (2018) find that a nudge which *overall* appeared to effectively discourage spending was actually harmful when the sample was stratified and re-analysed (which is to say, when heterogeneity was considered).

Equally, nudges are often ineffective when used impersonally, but appear to be effective when heterogeneity is considered (Ruggeri et al., forthcoming). A rather contemporary example of this is in COVID-19 messaging. Kraft-Todd et al. (2020), for instance, have found various nudges embedded within public-safety information messaging in Italy to be ineffective across the population examined, but effective amongst subsamples constructed along age and gender specifications. Staying in the world of health, Ruggeri et al. (forthcoming) argue nudges used in healthcare are often ineffective because the health outcomes being targeted often *impact only part of the population* – not everybody is obese, for example. For these authors, the notion that a nudge might be ineffective when used impersonally, but effective when personalised, is a rather obvious conclusion. This is very much the conclusion to be drawn given the evidence provided in this thesis.

Notably, choice personalisation is found to make the nudges statistically significantly more positively effective than impersonal nudging at the 5% level, with delivery personalisation also producing this result but at the 10% level. These results are based on application of a two-tailed t-test. Given hypothesis 1 holds that personalised nudges will induce a positive effect, a one-tailed t-test can also be used. When this procedure is undertaken, both choice and delivery personalisation are found to make the nudges statistically significantly more positively effective than impersonal nudging at the 5% level. When both choice and delivery are used together, this result is also found at the 5% level. Qualitatively, all personalisation groups produce *absolutely* positive effectiveness scores, compared to the impersonal group for which the score is only positive *relative* to the control group, and is not statistically significantly different from the score of the control group.

The difference between the delivery personalisation group compared to the impersonal group is statistically significant at the 5% level using a one-tailed t-test. As hypothesis 1 expects a *positive* effect resulting from the personalised nudges, a one-tailed t-test is the appropriate and more powerful test to use. Here, the statistical evidence supports the conclusion that delivery personalisation is effective.

Nevertheless, it is notable – even when effective – that delivery personalisation appears to be a statistically significantly less effective personalisation strategy than choice personalisation. Furthermore, the effect of delivery personalisation does not seem to appear in the choice and delivery personalisation group. These observations will be considered shortly.

Immediately, these results allow hypothesis 1 to be partially accepted: inconsistent statistical evidence is found to suggest impersonal nudging is effective – with only the status quo nudge being effective at the 5% level – but consistent statistical evidence is found to suggest both choice and delivery personalisation are effective at the 5% level. This latter result suggests that the problem of heterogeneity may be resolved by personalising the outcomes to which a person is nudged (i.e. choice personalisation) as well as personalising the method by which a person is nudged (i.e. delivery personalisation). Furthermore, the benefits of personalisation are maintained when both choice and delivery personalisation are used in tandem.

15.3 – The Implications of Hypothesis 2

However, when these methods are used in tandem, the resulting effect is not statistically significantly different from using just choice personalisation. On this basis, hypothesis 2 cannot be accepted. Alternatively, combining choice and delivery personalisation did produce a personalisation strategy which is found to be statistically significantly different from using just delivery personalisation, with the effect being positive. On this basis, an argument could be made for accepting hypothesis 2, though given the wording of the hypothesis, this argument is not compelling.

There are two possible explanations for these results, which are not mutually exclusive. The first explanation is that delivery personalisation in this instance is not as effective as one might hypothesise. This notion has been alluded to above and is considered in more detail below. The second explanation is that choice personalisation in this instance is overwhelmingly effective to the point that it is dominating other effects.

The apparent effectiveness of choice personalisation cannot be denied; by simply nudging participants towards candidates who appear to advocate policies which are important to them, the average effectiveness of the nudge-advertisement is found to be statistically significantly different from the control group advertisement, and the adverts seen by the impersonal group *and the DO group*. One explanation for this result is a lack of information.

A frequent idea within the literature on information leakage is the notion of information search (Tannenbaum and Ditto, 2011; McKenzie et al., 2006; Sher and McKenzie, 2006). This idea contends that people search for information when making decisions, even going so far as to *infer* information based on the framing of a prospect (Tannenbaum and Ditto, 2011; Sher and McKenzie, 2006). A lack of meaningful information, for instance, may explain the occasional evidence of an aesthetic effect during the pilot studies and in the primer group. Consider Praino and Stockemer (2018), who find the attractiveness of political candidates can be a significant factor in election outcomes (with the more attractive candidate usually winning), but *only in marginal elections*, where – one might expect – few factors differentiate the candidates.[339]

Further evidence to suggest that the political advertisements which were not choice personalised were informationally equivalent can be found from the survey-experiment respondents. Two respondents communicated,[340] without prompt that:

Respondent A, in the DO group of the PTG data sample: "*I would need some actual information about things like policies and positions before deciding who to vote for*."

Respondent B*,* in the primer group data sample: "*I vote based on issues and policy, not random ads*."

These comments are, of course, subjective. Furthermore, the analytical approach adopted here is not a qualitative one. But equally, these comments come directly from respondents *known to have completed the survey-experiment and interacted with the nudge-advertisements* (albeit a tiny minority). Given this, these comments reinforce the notion that the non-CO advertisements were quite informationally similar.

[339] The role of aesthetic differences is discussed in more detail below.

[340] In MTurk, respondents must enter a completion code generated by the requester to demonstrate they have completed the task. As this is done manually via a text box, respondents can write anything they like. It was via this medium that these comments were communicated.

Accepting this notion supports several implications. Firstly, it suggests information search rather than behavioural bias was driving participant choices, which may explain why the impersonal nudges were not statistically significantly different from the control group. Secondly, extending this explanation offers an explanation for the relatively poor performance of the delivery personalisation group, as that form of personalisation primarily sort to appeal to behavioural bias which was itself not a primary driver of participant decision-making.[341] Thirdly, that when choice personalisation was used, this greatly tipped the balance of informational content towards the choice-personalised advertisement, and, as information search is hypothesised as the main phenomena driving participant response, this led to choice personalisation appearing to be overwhelmingly effective.

Furthermore, in response to hypothesis 2, the notion of information search would explain why the data reject the hypothesis. If information search – which choice personalisation *in this instance* appeals to – is driving most of the participants' behaviour, and behavioural bias – which delivery personalisation *in this instance* appeals to – is driving little of the participants' behaviour, one would expect the effectiveness scores of CD group to very closely resemble the effectiveness scores of CO group, which is observed.

Knowing this, future studies might adopt one or both of two adjustments to the survey-experiment used here. Firstly, the policy might be embedded into the control slogan *also*, such that the CO advertisement and the control advertisement are more informationally equivalent.[342] Secondly, as the advert shown to the CO group uses an impersonal method of nudging, so too might the DO group advert use an impersonal policy within the slogan.[343]

[341] This is only one explanation, but accepting it, one may conclude that the effects of delivery personalisation were actually rather more compelling than they initially appear.

[342] To what extent informational equivalence is desirable, however, is unknown. Nudges frequently involve providing more information to decision-makers (Thaler and Sunstein, 2008), and thus endeavouring to achieve informational equivalency may simply come to stymieing the effect of the nudge.

[343] Again, to what extent this is permissible can be debated. The generic slogan *is also* an impersonal outcome, so simply adding an impersonally assigned policy slogan may just increase information without needing to, perhaps obscuring the effect of delivery personalisation.

These adjustments *may* make the groups' outcomes more comparable, and allow the effects of each method of personalisation to be more accurately measured.

Accepting the results as they are, however, would lead to a clear implication for those seeking to personalise nudges in the future. Namely that, while delivery personalisation (on its own or with choice personalisation) appears somewhat effective, choice personalisation is a significantly *more effective* method, at least *in this instance*. Furthermore, the benefits of using both choice and delivery personalisation are no greater than just using choice personalisation.

15.4 – Delivery Personalisation

While statistical evidence would suggest delivery personalisation is also an effective personalisation strategy, other evidence shows delivery personalisation is statistically significantly less effective than both of the other personalisation strategies. Some space is now given for considering why this may be the case. To an extent, an answer is provided by considering why *choice personalisation* seems so *effective*. But again, an alternative – if not additional – perspective to adopt is one which considers why delivery personalisation seems, relatively, less effective.

15.4.1 Not Enough Personalisation

A primary reason may actually be a *lack of delivery personalisation*. Figure 22 in Chapter 14 details the survey-experiment flow for the PTG sample. Included in this figure are percentage distributions of the sample. For instance, Figure 22 shows that 31.75% of the sample (N = 400) were assigned to the CO group, 35.00% to the DO group and 33.25% to the CD group. Figure 22 also provides more detailed information, showing for instance how many participants within the DO and CD groups experienced the present bias nudge (2.14% and 4.51%, respectively). Furthermore, recall that those participants who did not score within any region which would suggest a significant, positive moderation effect were defaulted into the best impersonal nudge – the status quo nudge.

Figure 22 therefore indicates that anywhere between one-third (32.86%) and one-half (54.55%) of participants in the DO or CD groups did *not* receive a delivery personalised nudge.[344] Of course, these participants did receive the best impersonal nudge – a form of delivery personalisation – but this is a sizeable number of participants receiving a default delivery personalised nudge.

This thesis is not alone in presenting such results, however, *and given this result*, one might come to reinterpret and recontextualise the findings of Peer et al. (2019). They note: "Our simulations estimated that the Crack-Time nudge would be optimal for 85% of the sample, whereas the Meter nudge would be optimal for 15% of the sample" (Peer et al., 2019: 12-13). In Chapter 8, this comment was taken as a weakness of Peer et al. (2019), interpreted as evidence of over-simplification. However, one may speculate that, of those 85% shown the Crack-Time nudge, a sizeable proportion are shown it simply because it is the best impersonal nudge.[345] Certainly, in light the results presented in this thesis, this seems a sensible postulate.

Methodologically, the implications of a large proportion of the sample not receiving delivery personalisation suggest that relevant heterogeneity is not being captured by the four psychometric tests used in the DO and CD groups. Furthermore, given the lack of moderation effects associated with the other psychometric variables, one can conclude relevant heterogeneity is not being captured by the seven-psychometrics used in the primer group either. In other words, there may be heterogeneity in decision-making contained within the one-third to one-half of respondents who receive the best impersonal nudge which is not captured by any of the psychometric scales used here. Therefore, future studies may need to utilise a significantly larger number of psychometric tests, and indeed, the small data approach taken here may speak to the limitations of 'crude' personalisation and the advantages of big

[344] 32.86% represents the percentage of respondents in the DO group who experienced the status quo nudge but did not score within a region of significance on the spontaneity psychometric scale – the scale found to statistically significantly moderate the effect of the status quo nudge. 54.55% is the largest percentage of participants from the CD group who also experienced the status quo nudge as the best impersonal nudge from the four subgroups contained within the CD group. In ascending order, the other subgroup figures are 35.00%, 38.19% and 51.85%, with a CD group average of 44.90%.
[345] Peer et al. (2019) report the Crack-Time nudge was the best impersonal nudge.

data commonly discussed in the personalisation literature (Porat and Strahilevitz, 2014; Sunstein, 2013; Thaler and Tucker, 2013).

An implication of this explanation for why delivery personalisation appears disappointing relative to choice personalisation is that, omitting those who receive delivery personalisation only via the best impersonal nudge, this adjusted DO group should be found to produce more positive effectiveness scores, which are indicative of more effective nudging. With the current data, this hypothesis can be tested. These results shown in Table 80 and Table 81:

Table 80: One- and Two-tailed T-tests of Adjusted DO Group vs. Impersonal Group

Adjusted DO Mean	Impersonal Mean	Test	t-Statistic	p-value
4.404	-2.299	Two-tailed	-1.7767	0.0762*
4.404	-2.299	One-tailed	-1.7767	0.0381**

Levene's test = 0.4544
* p < 10%, ** p < 5%, *** p < 1%

Table 81: One- and Two-tailed T-tests of Adjusted DO Group vs. Best Impersonal Group

Adjusted DO Mean	Impersonal Mean	Test	t-Statistic	p-value
4.404	0.609	Two-tailed	0.6559 (0.7580)	0.5130 (0.4498)
4.404	0.609	One-tailed	0.6559 (0.7580)	0.2565 (0.2249)

Levene's test = 0.0414
Welch's adjustment shown in brackets.
* p < 10%, ** p < 5%, *** p < 1%

Omitting those who receive the best impersonal nudge from the DO sample (Adjusted DO Group) does, as would be expected, increase the DO group mean from 3.157 to 4.404. However, there is no substantial change in the results, with the DO group being statistically significantly different from the impersonal group at the 5% level under a one-tailed t-test. Furthermore, when the DO group is split into those who experience a delivery personalised nudge (i.e. the adjusted DO group) and those who experience the best impersonal nudge, and these subgroups are compared, no statistically significant difference in effect is found.[346]

As such, while the trends in group averages suggest a lack of delivery personalisation owing to missed heterogeneity may be a factor, this explanation alone cannot account for the poor performance of delivery personalisation relative to choice personalisation.

15.4.2 A Mixed Bag

An alternative explanation for the relatively poor performance of delivery personalisation compared to choice personalisation may simply be that delivery personalisation is a less effective strategy than might initially be expected. Peer et al. (2019), for instance, report that delivery personalisation was significantly more effective than impersonal nudging. However, they also note that there was no statistically significant difference between their Meter nudge when it was personalised or delivered impersonally. Similarly, consider Lipman (forthcoming), who also supposes delivery personalisation may be a means of personalising behavioural incentives but fails to find any significant difference between behavioural traits and incentives preferences. By comparison, this thesis presents good statistical evidence which supports the theorised use of delivery personalisation – but delivery personalisation remains a relatively poor strategy compared to choice personalisation.

One should be cautious on two counts. Firstly, the theoretical promise of delivery personalisation (Mills, forthcoming; Ruggeri et al., forthcoming; Peer et al., 2019; Benartzi, 2017) may not exist quite so effectively in practice. Secondly, there is limited research into delivery personalisation,[347] and thus methodological and theoretical considerations may be missing from both Peer et al. (2019) and this thesis. For instance, almost all authors within the personalised nudging literature comment on the emergence of personalisation given the advancement of big data technologies (Mills, forthcoming; Ruggeri et al., forthcoming; Yeung,

[346] Though the best impersonal nudge (the status quo nudge) does appear to perform slightly worse than when impersonally delivered in the primer group (0.609 vs. 1.936).
[347] Peer et al. (2019) attest to be the first, and to this author's knowledge, they are.

2017; Porat and Strahilevitz, 2014; Thaler and Tucker, 2013; Sunstein, 2013) and digital choice environments (Peer et al., 2019; Schöning, Matt and Hess, 2019; Benartzi, 2017; Yeung, 2017).

It may be the case that the principle of delivery personalisation is correct – as some evidence presented here and presented by Peer et al. (2019) would suggest – but that future methods of investigation and application will need to utilise more sophisticated data resources, more detailed measures of heterogeneity, and larger samples to establish statistical significance.[348] For instance, the final formulation of delivery personalisation presented here is one which is algorithmic in nature, taking in an input (i.e. psychometric score), and running this input through an if/else statement to determine which output (i.e. nudge-advertisement) to present. But there is no reason, in principle, why the input in future investigations may not be a *vector* of data, allowing the implicit assumption present in the current model that two people who score the same on a single psychometric measure should necessarily be treated the same to be relaxed, if not abandoned.[349] As Ruggeri et al. (forthcoming) argue, machine learning techniques which can use large amounts of data in vector form may be necessary in complex personalisation tasks.[350]

15.4.3 Complexity

Such complex tasks, within the context of Ruggeri et al. (forthcoming), are medical tasks, but healthcare outcomes are not the only situations which can be expected to benefit from personalisation but are simultaneously complex. Complexity, here, takes a narrow definition with which others may disagree – namely, complexity is taken here to mean any decision where more than one measure of heterogeneity may be *necessary* to effectively personalise

[348] For a recent discussion of these challenges, see Ruggeri et al. (forthcoming). For a discussion of how data challenges may be resolved, see Mills (forthcoming).
[349] This view, for the most part, ignores their scores on all other psychometric measures, as well as innumerable other differences between them which may be relevant.
[350] Moderated regression may also be a valid means of analysis, though models which are more complex than the SLM model used here and by Peer et al. (2019).

interventions.[351] For instance, recall the quote by Benartzi (2017) in their discussion of website personalisation and digital choice architecture: "The logical endpoint is an internet in which the best Web sites and apps customize their appearance based on our demographic background. Are we an educated senior citizen from Poland? Then take away all the colors and give us plenty of text and links. Are we a young Thai man? Then give us lots of bright color and imagery" (Benartzi, 2017: 50). While Benartzi (2017) chooses to emphasise demographic data, he also emphasises *multiple* measures of heterogeneity, from education to age to geography.

Another example of a complex decision, which in turn may entail more complex personalisation, may be political decision-making. By way of demonstration, consider previous studies: Peer et al. (2019) and Guo et al. (2020) examine personalised nudging and password creation; Page, Castleman and Meyer (2020) consider personalised nudging and FAFSA applications; Schöning, Matt and Hess (2019) examine personalised nudging and the disclosure of personal health information; and Hirsh, Kang and Bodenhausen (2012) and Moon (2002) investigate personalisation and advertisements for products.

In each instance, previous studies have investigated fundamentally pro-self decisions – decisions where the costs and the benefits of an outcome are borne, primarily or entirely, by the individual making the decision (Korn et al., 2018; Barton and Grüne-Yanoff, 2015; Hagman et al., 2015). Political decision-making, notably voting, is simultaneously a pro-self decision – as a person votes for the candidate whom they would like to win – and a pro-social decision – as the candidate who wins is based on collective choice.

The consequences of this difference may be substantial. Consider the decision surrounding the creation of a password. The password itself is – by definition – a *private*, personal decision, with a person possibly thinking about how they will remember it and how secure it is (Guo et al., 2020). Voting, by contrast, may invoke personal considerations such as whether a given

[351] This notion applies to both delivery personalisation *and* choice personalisation.

candidate's victory would be in that person's personal interests (Downs, 1957), but may also invoke considerations about whether that candidate's victory would be in the interests of others, or society as a whole. This, in itself, adds complexity to the decision.[352] But this also ignores the interplay between pro-social and pro-self tendencies. For instance, Capraro et al. (2019) argue that pro-social behaviour does not just manifest as acting for the *benefit* of others but can also be a used as an attempt to control the selfishness of others who exhibit pro-self behaviour (Kesberg and Pfattheicher, 2019). Examples of such behaviour in the political context can be seen in the UK General Election in 2019 (GE19), when various websites and online services were created to facilitate tactical voting (Casalicchio, 2019; Sabbagh, 2019) – the act of voting for a political party or person that you would not usually support in order to prevent another party or person from winning (Cambridge Dictionary, 2020). Crucially, such a strategy only succeeds if one adopts a pro-social position of voting against one's *personal* interests and *in conjunction with the actions of others*.

No such interplay can be expected to exist in any of the previous studies of personalisation – an individual purchase or choice of password, typically, is not done with a wider sense of society in mind, let alone a notion of antagonism or coordination with others in society. Accepting that the decision with which personalised nudging is applied here is a more complex decision than the contexts investigated by previous research may explain, partially, the underperformance of delivery personalisation. As above, information is speculated to be playing a larger role than bias in affecting individual behaviour; but so too may several considerations regarding the implication of one's decision on oneself and others.[353] With choice personalisation, the addition of information may work in tandem to make these

[352] Following the definition of complexity given here, one might imagine individuals being modelled by a vector containing two values, one capturing their pro-sociality and another capturing their pro-selfishness.
[353] Very recent research from Kozyreva et al. (2020) examining public attitudes to personalisation algorithms supports this idea. Kozyreva et al. (2020) find personalisation algorithms are more accepted when used to personalise individual experiences – such as recommendations of products – than when used to personalise shared experiences – such as news, comments from friends and politics. They suggest this is because people value exposure to what they consider within the public domain, and that personalisation may operate as a kind of enclosure. See below for a similar discussion of this idea.

considerations easier, while once more, delivery personalisation may be targeting behavioural bias which – in this context – has a relatively small effect on the ultimate decision.

15.5 – Moderation Effects

One result which is consistent throughout both pilot studies and the primer group is the lack of moderation effects, at least in comparison to Peer et al. (2019). As a baseline for comparison, Peer et al. (2019) investigate five nudges and eight psychometric scales, and so have 40 opportunities to identify statistically significant moderation effects, which they do in 19 instances (at the 5% level), or around 47.5% of the time. Recall that in the first pilot study, no statistically significant moderation effects were identified; in the second pilot study, two statistically significant moderation effects were identified out of 28 (7.1%); and in the primer group, four statistically significant moderation effects were identified at the 5% level (14.3%). It is interesting to consider why the disparity in the number of statistically significant moderation effects between this thesis and Peer et al. (2019) has occurred, especially given all the psychometric tests used here are also used by Peer et al. (2019).[354] Several explanations may be induced.

Firstly, as discussed above, the decisional-context investigated by this project (political decision-making) seems to be more complex than the context investigated by Peer et al. (2019) – password creation. Given the simplicity of the latter, the SLM model utilised both here and by Peer et al. (2019) may be better suited to identify simple moderation effects (i.e. a single moderator and a single interaction with a focal variable), while a more complex moderated regression model – or indeed, means of analysis – may be needed to identify more complex interactions.

Secondly, while Peer et al. (2019) identify 19 statistically significant moderation effects, they do not utilise 19 moderation effects, nor do they discuss the qualitative validity of *any* of their

[354] The additional psychometric scale utilised by Peer et al. (2019), the abbreviated numeracy scale, contributes to only 2 of their 19 identified moderation effects.

identified interactions. Addressing the first point, as Peer et al. (2019) only utilise the Crack-Time and Meter nudges in the (delivery) personalisation stage, only 10 of the 19 moderation effects originally identified are actually utilised. As discussed in Chapter 8, the nudge-preference method utilised by Peer et al. (2019) appears to be greatly simplifying, removing nudges from the analysis without any clear explanation as to why.[355] The result is that, despite evidence to suggest significance, nearly half of all the moderation effects identified by Peer et al. (2019) are ignored. Regardless, accepting that only 10 out of the 40 possible interactions are relevant to further investigation, this brings the percentage of possible effects closer to that found in this thesis, though the difference is still sufficient to warrant further consideration (25.0% vs. 14.3%).

Addressing the second point, the lack of behavioural underpinnings in the study undertaken by Peer et al. (2019) mean they are unable to draw on past research to inform whether any of their identified moderation effects make sense from a *qualitative* perspective as Hayes (2018) encourages. Thus, although Peer et al. (2019) identified 19 statistically significant moderation effects, it remains unclear how many of these moderation effects make sense given behavioural theory and past research, and how many may simply emerge as anomalous results or quirks of the sample. Providing such scrutiny may lead to a reduction in the number of identified significant moderation effects on a *qualitative* ground. By contrast, while notably fewer statistically significant moderation effects are identified in this thesis, all which are identified can be grounded in theory and previous behavioural findings.

Thirdly, there may simply be few moderation effects to be identified. Consider Lipman (forthcoming), whose recent work on the *possibilities* of personalised behavioural interventions found no statistically significant heterogeneity effects between those who preferred different behavioural incentives. Lipman (forthcoming): "tailored preferences were not systematically

[355] Again, Peer et al. (2019) only attribute their decisions to an unspecified simulation model. Of course, there may be reasons to remove a nudge if one is attempting to maximise the effectiveness of the nudge. For instance, the moderation effect might suggest that all possible values of the moderator would produce a negative effect. Peer et al. (20190, however, do not state this as a reason for removal.

related to any of the behavioral insights often used to motivate the implementation of behaviorally inspired incentives in practice… Hence, although autonomy is likely increased by allowing individuals full freedom to design their own financial incentives… the results reported here provide no insight into why individuals would prefer particular incentive schemes" (Lipman, forthcoming: 10).

Lipman's (forthcoming) methods can be criticised: they only investigate four psychological (behavioural) variables and four behavioural incentives, producing only 16 instances to identify significant heterogeneity effects; they do not consider the effectiveness of the intervention itself, instead only considering *why* a person choose an intervention; and they use only simple tests of difference and do not engage in methods such as moderated regression or more complex analytical procedures. Yet, Lipman's (forthcoming) findings also serve as an interesting counter to the generally compelling findings of Peer et al. (2019), and in contrasting both studies, the results presented here fall somewhere within a middle ground.[356]

Fourthly, the problem of a high proportion of the sample being shown the best impersonal nudge has been explained with the suggestion that the psychometric scales used were simply not capturing all the relevant heterogeneity. The implication of this explanation is that additional psychometric measures (or additional measures of heterogeneity in general), may be needed to improve from the current study. Relatedly, one would expect that more measures of heterogeneity would identify more statistically significant moderation effects. It cannot, therefore, be overlooked that one explanation for the relatively low number of moderation effects identified here is because of an incomplete set of measures of heterogeneity.

15.6 – Was the RCT Design Successful?

A recurring consideration throughout this thesis has been the presence (or lack thereof) of significant aesthetic effects. As discussed in Chapter 7, the aesthetic differences between Candidates A and B were identified as a potential additional source of variance in the

[356] Also see Schöning, Matt and Hess (2019) and Guo et al. (2020), both of whom find mixed results.

effectiveness variable, potentially obscuring the effect of the nudge. The identified solution to this problem was an RCT design. By assuming no aesthetic effect should persist across the sample, and by randomly assigning participants to one of two aesthetic groups (nudge-advertisements featuring Candidate A versus nudge-advertisements featuring Candidate B), the overall effect of aesthetic differences (aesthetic effect) could be assumed to be insignificantly different from zero.

In several instances, this would appear to be the case. Setting aside the PTG data sample for a moment, in all tests of aesthetic differences, statistically significant aesthetic effects were identified in one or two subgroups per sample. Given the small sample sizes (both generally in the case of the pilot studies and in comparison to some RCTs in the case of the primer group), these results are encouraging, as they suggest the RCT design did work to prevent aesthetic effects confounding the study. This is not to say that apparent *trends* in aesthetic differences cannot be identified. While the effect is statistically insignificant, Candidate B would seem to be more preferable to Candidate A; a trend identified in both pilot studies, the primer group and the PTG group. Furthermore, the presence of *some* statistically significant aesthetic effects in conjunction with this trend suggest that the presence of aesthetic effects should not be discounted entirely. Prior research by Praino and Stockemer (2018) and Lawson et al. (2010) suggest candidate image (i.e. how a candidate looks) can be significant in predicting election outcomes. This expectation would also be congruent with the hypothesis of information search given above – where participants are seeking additional information from which to base their decision, obvious differences – even if “superficial” (Lawson et al., 2010: 561) – may be expected to be drawn upon in the decision making process (Praino, 2018).

Equally, Little et al. (2007) argue that while aesthetic differences (specifically attractiveness) may predict the outcome of elections, there is no robust standard of attractiveness which can be applied across different cultures and constituencies. As such, even where an aesthetic effect might be a realistic hypothesis, the use of an RCT design would be expected to minimise this effect.

Returning to the PTG data sample, no evidence of statistically significant aesthetic effects is identified. Paradoxically, this lack of significance is itself potentially quite significant in the context of personalisation. Where previously, when no personalisation was used, some statistically significant aesthetic effects were identified (and the identification of some statistically significant aesthetic effects was rather consistent), when personalisation is used, no statistically significant aesthetic effects are identified. One conclusion which may be drawn from this, is that the use of personalisation is each group (CO, DO and CD) rendered the nudge effective enough so as to reduce to statistical insignificance any potential aesthetic effects. It is noted, however, that the general – though statistically insignificant – trend of Candidate B being preferred to Candidate A remains.

15.7 – The Role of Other Factors

In Chapter 9, three outstanding factors were identified which, it was suggested, may impact this research. These were the COVID-19 pandemic, the rise of populist politics, and the 2020 US presidential election.

It is likely not possible to evaluate the effect of the current epoch of populist politics, nor the impact of an election year, using this study alone. Indeed, it seems likely the effect of these macro trends – if, indeed, any effect is present at all – may only be understood following a broader analysis of all research into political decision-making during this period. However, hypotheses were made regarding the role of COVID-19 and the nudge preferences, which can be considered in hindsight.

Following Chapter 9, the uncertainty and risk associated with COVID-19 were hypothesised potentially to encourage more loss averse behaviour, and more searching for certainty. The most immediate manifestation of this hypothesis, if correct, would be an apparently more effective loss aversion nudge. However, across both the pilot studies and the primer group, the loss aversion nudge was not found to be especially effective – in fact, only in one instance was the advert utilising the loss aversion nudge even statistically significantly and positively different from the advert seen by the control group.

On the question of certainty, one might look to the status quo nudge. The status quo nudge is expected to convey certainty as it appeals to an outcome which is already known (Kahneman, Knetsch and Thaler, 1991). The search for certainty in uncertainty times, therefore, may spark appeal for the status quo. There may be some evidence for this hypothesis, with the status quo nudge being found to be the best impersonal nudge following an analysis of the primer group. Unfortunately, this is likely the extent to which any conclusions can be drawn. Insofar as all moderation effects identified match concepts previously established in the literature, and insofar as the broad results of this study match expectations, one cannot confidently conclude that COVID-19 is exerting any influence on the data as there is simply a lack of a ‘no-COVID’ control.

15.8 – General Implications of Personalised Nudging as a Strategy and Research Endeavour

There are multiple implications which emerge from this research. The finding that both choice and delivery personalisation may be effective means of nudging – at least when impersonal nudging is ineffective – raises several considerations. Some of these considerations pertain to the implications of personalisation and personalised nudging in general. Furthermore, there is a pertinent methodological consideration which should be addressed, namely, which heterogeneity should be utilised in personalised nudging. Addressing this question requires one to return to the relevancy principle first proposed by Sunstein (2012), but in light of this thesis, a proposed adjustment to the relevancy principle is offered.

15.8.1 Cohesion and Understanding in a Personalised Environment

The first discussion draws from an argument initially developed by Mills (forthcoming). Here, the antagonism between universalism, on the one hand, and personalisation, on the other, is explored. In many domains, a universal approach is desirable, if not absolutely necessary. For instance, criminal justice or civil litigation demands universality – a person who commits a crime should, in principle, be judged solely on the conditions relevant to that criminal act and identically to anyone else accused of and prosecuted for a crime.[357] Equally, in a civil dispute, two identical lawsuits or claims of mistreatment should typically be resolved in the same manner. The notion of universality underpins, in part, Rawls' (1971) theory of justice.

One can imagine, for instance, personalised contract law (Porat and Strahilevitz, 2014). In principle, such measures seek to promote equitable outcomes by accounting for relevant differences between parties before, rather than after, any dispute arises. Porat and Strahilevitz (2014), for instance, advocate the use of personalised default rules to this end. An issue may arise, however, in that the *substance* of any such contract is identical to any other contract,[358] but the *behavioural interpretation* of one person's personalised contract (utilising personalised nudges) may be very different to another's. This is an issue Benartzi (2017) identifies specifically when considering the interpretation of information in digital spaces (which, following Porat and Strahilevitz (2014), law will increasingly come to be provided through), and eloquently summarises: "*function follows form*" (Benartzi, 2017: 52, original emphasis). In personalised contract law, but in principle in other areas such as regulation, the function of these agreements follows the form of these agreements. Where the latter is increasingly personalised, and the behaviour of parties altered accordingly, the former may come to be undermined.

Two ideas related to this discussion are transparency and relevancy (the latter of which more will be said in due course).

Returning to Rawls (1971), they argue that laws and regulations should be sufficiently transparent so as to be easily scrutinised by the public and rejected if necessary. Such an argument is known as the publicity principle (Hansen and Jespersen, 2013; Rawls, 1971), and is commonly invoked in discussions of nudging and transparency (Hansen and Jespersen,

[357] Within the limits of their criminality. This, again, returns to the notion of heterogeneity.
[358] In the same way that a default option does not necessarily reduce the options available to a decision maker.

2013). An advantage of universality, following the publicity principle, is comparability. Universal laws and regulations (and indeed nudges) which afford common experiences to all those who are subject to them can more easily facilitate discussion and thus scrutiny. For instance, a choice architect might wish to increase retirement saving and change the default option for workplace pensions from opt-in to opt-out (Service, 2015). If all employees experience this change, which is to say, if the nudge is introduced impersonally, the effect of the nudge is more obvious because all those subject to the nudge experience the same effect. A by-product of universal or impersonal nudging, therefore, is greater transparency, which should aid scrutiny and thus allow the nudge to be evaluated, accepted or rejected.

By contrast, imagine a nudge to increase workplace pension saving, but the decision to automatically enrol an employee, or to leave them unenrolled, is based on each employee's circumstances (i.e. choice personalisation). In comparing experiences, some employees might remark that they suddenly started paying into a pension plan, while others may remark that their take-home pay has remained the same. This may spark confusion and make it hard for employees to understand why some have been enrolled into the plan and others haven't. The obvious solution, therefore, is transparency, and as such, transparency in how nudges are personalised may be of great importance.[359]

Finally, consider relevancy.[360] Yeung (2017), in their work on the implications of big data and nudging, argues that personalisation often occurs along obvious lines. For instance, an idea popularised by Negroponte (1995) and further deliberated on by Sunstein (2001) is that of the 'daily me,' an information feed built on individual preferences (i.e. personalised). This idea is not unlike that discussed by Thaler and Tucker (2013) or the notion of the echo chamber also discussed – amongst others (Quattrociocchi, Scala and Sunstein, 2016; Massa and Avesani,

[359] This is to apply a Rawlsian principle to nudging, rather than to describe a *Rawlsian* nudge. The latter may be described as something to the effect of changed choice architecture designed to provide the most benefit (or least cost) to those already disadvantaged. This is an interesting idea, but one which is not of relevance to this discussion.

[360] Again, relevancy from a slightly different perspective will be discussed shortly.

2007) – by Sunstein (2001).[361] The issue with personalisation based on obvious, revealed differences is it, “[does] not just reproduce traditional social fault lines but also… exacerbate[s] them” (O’Shea, 2019: 75). This prompts Yeung (2017) to worry that personalisation may lead to ever-more homogenised groups, with personalised nudging – paradoxically – nudging people towards similar experiences, rather than new experiences which allow people to develop a sense of their own identity and autonomy (Verbeek, 2006).

A tendency for this is present in this research. Notably, choice personalisation was facilitated merely by asking participants to reveal that which was important to them and directing them towards a nudge which would support that outcome. Even delivery personalisation as demonstrated here may be guilty of the same sin – merely trying to predict preferences in decision making and construct a model to *accommodate* these decisional preferences, rather than necessarily challenging them. Of course, such a criticism is also contained within context. If one considers, say, the arguments of Ruggeri et al. (forthcoming) and personalisation within a healthcare setting, or indeed Peer et al. (2019) and personalisation in password creation, the dangers of relevancy in this specific sense seem minimal.[362]

15.8.2 Returning to the Relevancy Principle

In a 2012 essay on personalised nudging, Sunstein (2012) argues that, because of the large data resources which could be expected to be associated with personalised nudging, choice architects should ensure that any and all data they are utilising is relevant to the development and administration of the nudge. Sunstein’s (2012) primary concern is personal privacy.

[361] An echo chamber is, according to Quattrociocchi, Scala and Sunstein (2016), the result of “users’ tendency to promote their favored narratives and hence form polarized groups. Confirmation bias helps to account for users’ decisions” (Quattrochiocchi, Scala and Sunstein, 2016: 1).

[362] Indeed, in terms of personalised healthcare, one probably wants a very self-centred experience. This is not to say that these contexts are immune to all the issues discussed. For instance, a lack of transparency in healthcare or cybersecurity could still cause tremendous social difficulties if a patient is aggrieved at their treatment versus another, or if one personalised password tip leaves some more vulnerable to a hack than others.

In Chapter 2, it was suggested the relevancy principle has another implication: because there is seemingly a limitless amount of heterogeneity data which could be collected or imagined, the relevancy principle places restrictions on choice architects who may be tempted to stratify and re-stratify their sample in the search for a significant effect. But this notion is easier stated than realised. One might be tempted, for instance, to argue that relevancy should stem from prediction – where a piece of heterogeneity is predicted to be relevant in the literature, one can claim it is relevant in whatever project is being undertaken. Certainly, the benefits of a firm grounding in the behavioural literature and the creation of predictions via a psychometric map have been shown to be advantageous when qualitatively interrogating the identified statistically significant moderated effects.[363]

However, this leads one into something of an epistemological quandary. Heterogeneity which may need to be addressed via personalisation is, by its very nature, heterogeneity which has not previously be integrated into whatever standard is under scrutiny (be it a law, a regulation, or a nudge). It is not clear how one might be expected to address heterogeneity from a predictive perspective when the presence of heterogeneity itself suggests it has not formally been addressed. As Yeung (2017) argues, this problem can become even worse as the number of data being considered grows; Yeung (2017) goes so far as to suggest in a great many cases, humans simply cannot know what data will ultimately be relevant in personalised nudging.[364]

This issue, furthermore, invokes the problem of transparency highlighted above. While Sunstein (2012) does not consider the role of transparency in their formulation of the relevancy principle, it certainly seems to be a valid component. It seems reasonable to ask: To what end should the relevancy of data in personalised nudging be demonstrative if not to enlighten

[363] Equally, the lack of this grounding seems to be to the detriment of Peer et al. (2019).

[364] By way of a somewhat more lucid argument, Biddle (2018) invokes the science fiction writer Ian Banks' notion of infinite fun space, a place inhabited only by computers engaging in calculations far beyond the understanding of humans, and thus to the exclusion of humans. One can imagine the difficulties that demonstrating relevancy would create when data insights are borne from infinite fun space. See Biddle (2018) for more.

others in an act of transparency? But where choice architects may simply be exploring possibilities, attending to the relevancy principle seems difficult.

Two solutions present themselves. The first is to adopt a marginal approach to personalised nudging. Ruggeri et al. (forthcoming) highlight the importance of considering the margins – where an impersonal nudge is generally effective, the various problems associated with personalised nudging can be reduced by focusing personalisation efforts on those at the margins who exhibit the undesirable behaviour.[365] This would, of course, have notable cost benefits also – something Sunstein (2012) argues should also be a consideration of personalisation.

The second approach may follow the approach adopted here. In this thesis, all personalisation has followed from previous findings in the literature, as well as previous methods established in that same literature. This has allowed the relevancy principle to be satisfied, as there has been clear rationale for the selection of various pieces of heterogeneity data and predictions have been formulated prior to personalisation such that the results may be scrutinised appropriately. Nevertheless, the data suggest anywhere between one-third to one-half of participants in the delivery personalisation groups were still being nudged with the best impersonal nudge.[366] From this result, it is apparent some heterogeneity is being missed, and *on the basis of this result*, one may be able to justify further exploration. This continuous development approach is one advocated by Ruggeri et al. (forthcoming) and Guo et al. (2020).

[365] For lack of a better word. There is a distinct danger in labelling behaviours 'undesirable'.

[366] Furthermore, the 100% allocation rate associated with the choice personalisation groups is only the result of experimental design. If choice personalisation was based on prediction as delivery personalisation was, rather than revealed preference from a finite list, one would not expect 100% allocation in these groups.

Chapter 16 – Conclusion

This thesis' research question, from the outset, was whether personalised nudging could be used to significantly influence political decision-making? The results of this thesis would suggest, on the whole, that the answer to this question is that it can. Of course, there is a degree of nuance with this answer.

The theory of personalised nudging offered here proposes that nudges can be personalised in two ways. Firstly, the method of nudging can be personalised. This is dubbed delivery personalisation. Secondly, the outcome which is being nudged towards can be personalised. This is dubbed choice personalisation.

Using seven psychometric variables contained within three psychometric scales – general decision making style, need for cognition and consideration of future consequences – and four commonly used nudges – the status quo nudge, the present bias nudge, the loss aversion nudge, and the social norm nudge – as well as moderated regression analysis and the Johnson-Neyman technique, four statistically significant moderation effects between each of the nudges and four of the seven psychometric variables were identified. As discussed in Chapter 15, this result is rather less than the number identified (47.5%) or used (25.0%) by Peer et al. (2019) in their analysis of personalised nudging, but remains broadly consistent with other results (Lipman, forthcoming; Guo et al., 2020; Schöning, Matt and Hess, 2019) and robust when interrogated using behavioural theory and literature.

Using these results, nudges were personalised via delivery personalisation. This process produces nudges which are statistically significantly and positively effective in promoting choices which differ from the choices of a control group and an impersonal nudging group. However, the observed preferences of the impersonal nudging group are *not* statistically significantly different from the observed preferences of the control group. The evidence, therefore, would suggest delivery personalisation is effective at influencing political decision-making.

Choice personalisation was also investigated. An investigation of both choice and delivery personalisation together represents a novel contribution of this project. To personalise the outcome towards which a person was nudged (i.e. choice personalisation), participants were asked to choose a political policy most important to them, from a range of four policies. This chosen policy was then integrated into a political advertisement which also used a nudge – in effect, choice personalisation.

Once again, this personalisation strategy is found to be effective. Choice personalised nudges were statistically significantly and positively effective in influencing observed preferences when compared to the preferences of both a control group and an impersonally nudged group.

Comparing both the choice personalisation and delivery personalisation strategies, the statistical evidence suggests that the former is a more effective strategy than the latter, despite both producing a statistically significant and positive effect.

In Chapter 3, it was suggested that the use of choice and delivery personalisation combined may produce nudges which are significantly more effective than either choice personalised or delivery personalised nudges separately. This hypothesis, however, is not supported by the results presented here. While combining choice and delivery personalisation produced personalised nudges which were statistically significantly more effective than impersonal nudges and delivery personalisation nudges, these nudges were not statistically significantly different from simple choice personalised nudges.

Overall, this thesis has demonstrated that personalised nudges can be effective strategies for influencing political decision-making. Furthermore, the results presented here would seem to support the choice/delivery framework for personalised nudging. However, these components do not appear to be equal, and choice personalisation appears to be a significantly more effective strategy than delivery personalisation. Of course, in some situations, one may not wish to personalise outcomes (i.e. choice personalisation). This being so, the problem of heterogeneity which personalisation seeks to resolve can still be tackled, as delivery personalisation still appears to be an effective strategy.

References

Ade, F, Freier, R, Odendahl, C (2014) 'Incumbency effects in government and opposition: Evidence from Germany' *European Journal of Political Economy*, 36(2014), pp. 117-134

Akerlof, G (1991) 'Procrastination and Obediance' *The American Economic Review*, 81(2), pp. 1-19

Akerlof, G, Shiller, R (2017) '*Phishing for Phools'* Princeton University Press: Princeton

Alacreu-Crespo, A, Fuentes, M, Abad-Tortosa, D, Cano-Lopez, I, Gonzalez, E, Serrano, M A (2019) 'Spanish validation of General Decision-Making Style scale: Sex invariance, sex differences and relationships with personality and coping styles' *Judgement and Decision Making*, 14(6), pp. 739-751

Allcott, H (2011) 'Social norms and energy conservation' *Journal of Public Economics*, 95(9), pp. 1082-1095

Allcott, H, Gentzkow, M (2017) 'Social Media and Fake News in the 2016' *Journal of Economic Perspectives,* 31(2), pp. 211-236

Allcott, H, Rogers, T (2014) 'The Short-Run and Long-Run Effects of Behavioral Interentions: Experimental Evidence from Energy Conservation' *The American Economic Review*, 104(10), pp. 3003-3037

Altmann, S, Traxler, C (2014) 'Nudges at the dentist' *European Economic Review*, 72(2014), pp. 19-38

Andreoni, J, Sprenger, C (2012) 'Risk Preferences Are Not Time Preferences' *The American Economic Review*, 102(7), pp. 3357-3376

Appelt, K, Milch, K, Handgraaf, M, Weber, E (2011) 'The Decision Making Individual Differences Inventory and guidelines for the study of individual differences in judgement and decision-making research' *Judgment and Decision Making*, 6(3), pp. 252-262

Ayres, I, Gertner, R (1989) ‘Filling Gaps in Incomplete Contracts: An Economic Theory of Default Rules’ *Yale Law Journal*, 99(1), pp. 87-130

Bachl, M (2015) ‘*Plot of conditional effects and Johnson-Neyman Technique in R’* RPubs. [Online] [Date accessed: 24/06/2020]: https://rpubs.com/bachl/jn-plot

Badkar, M, Smith, C, Politi, J (2020) ‘*US unemployment hits postwar high of 14.7%’* The Financial Times. [Online] [Date accessed: 09/07/2020]: https://www.ft.com/content/2a297508-c8d0-4736-8c2f-9f4d16822f57

Bakir, V, Herring, E, Miller, D, Robinson, P (2019) ‘*Deception and lying in politics*’ *in* Mauber, J (eds.) ‘*Oxford Handbook on Lying*’ (2019). Oxford University Press: Oxford.

Banerjee, A, Chassang, S, Snowberg, E (2016) ‘*Decision theoretic approaches to experiment design and external validity*’ NBER working paper series, no. 22167. [Online] [Date accessed: 20/06/2020]: https://www.nber.org/papers/w22167

Bartke, S, Friedl, A, Gelhaar, F, Reh, L (2017) ‘Social comparison nudges – Guessing the norm increases charitable giving’ *Economics Letters*, 152(2017), pp. 73-75

Barton, A, Grüne-Yanoff, T (2015) ‘From Libertarian Paternalism to Nudging – and Beyond’ *Review of Philosophy and Psychology*, 6(3), pp. 341-359

Bavolar, J, Orosova, O (2015) ‘Decision-making styles and their associations with decision-making competencies and mental health’ *Judgement and Decision Making*, 10(1), pp. 115-122

Beggs, J (2016) ‘*Private-Sector Nudging: The Good, the Bad and the Uncertain’* in Abdukadirov, S (eds.) ‘*Nudge Theory in Action’* (2016), Palgrave Macmillan: London

Benartzi, S (2017) ‘*The Smarter Screen: Surprising Ways to Influence and Improve Online Behavior*’ Profile Books: UK

Benartzi, S, Bhargava, S (2020) '*How Digital Design Drives User Behavior'* Harvard Business Review. [Online] [Date accessed: 18/06/2020]: https://hbr.org/2020/02/how-digital-design-drives-user-behavior

Benartzi, S, Thaler, R (1995) 'Myopic Loss Aversion and the Equity Premium Puzzle' *The Quarterly Journal of Economics*, 110(1), pp. 73-92

Benhabib, J, Bisin, A, Schotter, A (2010) 'Present-bias, quasi-hyperbolic discounting, and fixed costs' *Games and Economic Behavior*, 69(2010), pp. 205-223

Bernheim, B D (1994) 'A Theory of Conformity' *Journal of Political Economy*, 102(5), pp. 841-877

Bernheim, D, Exley, C (2015) '*Understanding Conformity: An Experimental Investigation*' Harvard Business School working paper no. 16-070. [Online] [Date accessed: 25/05/2020]: https://pdfs.semanticscholar.org/e3d5/e680ffbdaa07b77740c348f6c06ed54a950c.pdf

Beshears, J, Choi, J, Laibson, D, Madrian, B, Milkman, K (2015b) 'The Effect of Providing Peer Information on Retirement Savings Decisions' *Journal of Finance*, 70(3), pp. 1161-1201

Beshears, J, Choi, J, Laibson, D, Madrian, B, Wang, S (2015a) '*Who is Easier to Nudge?'* NBER Working Paper. [Online] [Date accessed: 06/02/2020]: http://scholar.harvard.edu/files/laibson/files/who_is_easier_to_nudge_2016.05.27.pdf

Beshears, J, Milkman, K, Dai, H, Benartzi, S (2016) '*Framing the Future: The Risks of Pre-Commitment Nudges and Potential of Fresh-start Messaging'* NBER Working Paper. [Online] [Date accessed: 09/04/2019]: https://static1.squarespace.com/static/5353b838e4b0e68461b517cf/t/583ca5acd2b8571174b28e40/1480369581625/48-Beshears_et_al_2016.pdf

BETA (2016) '*Developing behavioural interventions for randomised controlled trials: Nine guiding questions*' BETA. [Online] [Date accessed: 20/06/2020]:

https://behaviouraleconomics.pmc.gov.au/sites/default/files/files/guide-to-developing-behavioural-interventions-for-randomised-controlled-trials.pdf

Bhargava, S, Benartzi, S (2020) '*How Digital Design Drives User Behavior*' Harvard Business Review. [Online] [Date accessed: 14/07/2020]: https://hbr.org/2020/02/how-digital-design-drives-user-behavior

Bialik, K (2019) '*For the fifth time in a row, the new Congress is the most racially and ethnically diverse ever'* Pew Research Center. [Online] [Date accessed: 18/06/2020]: https://www.pewresearch.org/fact-tank/2019/02/08/for-the-fifth-time-in-a-row-the-new-congress-is-the-most-racially-and-ethnically-diverse-ever/

Blyth, M, Lonergan, E (2020) '*Angrynomics*' Columbia University Press: USA

Boiney, J, Paletz, D (1991) '*In search of the model model: political science versus political advertising perspectives on voter decision making'* in Biocca, F (eds.) '*Television and Political Advertising*' (1991). Lawrence Erlbaum Associates: USA

Borah, A (2019) 'Voting Expressively' *Economic Inquiry,* 57(3), pp. 1617-1635

Bourquin, P, Cribb, J, Emmerson, C (2020) '*Who leaves their pension after being automatically enrolled?'* IFS Briefing Note BN272. [Online] [Date accessed: 19/05/2020]: https://www.ifs.org.uk/uploads/Who-leaves-their-pension-after-being-automatically-enrolled-BN272.pdf

Brambor, T, Clark, W R, Golder, M (2006) 'Understanding Interaction models: Improving Empirical Analyses' *Political Analysis*, 14, pp. 63-82

Brehm, J W (1966) '*A theory of psychological reactance'* New York Academic Press: New York

Brown, C, Krishna, A (2004) 'The Skeptical Shopper: A Metacognitive Accountfor the Effects of Default Options on Choices' *Journal of Consumer Research*, 31(3), pp. 529-539

Busch, C (2017) 'Implementing Personalized Law: Personalized Disclosures in Consumer Law and Data Privacy Law' *The University of Chicago Law Review*, 86(1), pp. 309-331

Butt, A, Donald, M S, Foster, F D, Thorp, S, Warren, G (2018) 'One size fits all? Tailoring retirement plan defaults' *Journal of Economic Behaviour and Organisation*, 145(2018), pp. 546-566

Cacioppo, J, Petty, R (1982) 'The Need for Cognition' *Journal of Personality and Social Psychology*, 42(1), pp. 116-131

Cacioppo, J, Petty, R, Kao, C (1984) 'The Efficient Assessment of Need for Cognition' *Journal of Personality Assessment*, 48(3), pp. 306-307

Cadena, X, Schoar, A (2011) '*Remembering to Pay? Reminders vs. Financial Incentives for Loan Payments*' NBER Working Paper Series no. 17020. [Online] [Date accessed: 03/07/2020]: https://www.nber.org/papers/w17020.pdf

Cadwalladr, C (2019) '*Cambridge Analytica a year on: 'a lesson in institutional failure'* The Guardian. [Online] [Date accessed: 14/07/2020]: https://www.theguardian.com/uk-news/2019/mar/17/cambridge-analytica-year-on-lesson-in-institutional-failure-christopher-wylie

Cadwalladr, C, Graham-Harrison, E (2018) '*Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach'* The Guardian. [Online] [Date accessed: 07/04/2020]: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cambridge Dictionary (2020) '*Tactical Voting'* in *Cambridge Advanced Learner's Dictionary and Thesaurus*. Cambridge University Press: UK

Camerer, C, Loewenstein, G '*Behavioral Economics: Past, Present, Future'* in Camerer, C, Loewenstein, G, Rabin, M (eds.) '*Advances in Behavioral Economics*' (2004). Princeton University Press: Princeton

Cantor, D, O'Hare, B, O'Connor, K (2008) '*The use of monetary incentives to reduce non-response in random digit dial telephone surveys'* in Lepkowski, J, Tucker, C, Brick, M, Leeuw, E, Japec, L, Lavrakas, P, Link, M, Sangster, R (eds.) '*Advances in telephone survey methodology*' (2008) Wiley: New York

Capraro, V, Jagfeld, G, Klein, R, Mul, M, van de Pol, I (2019) 'Increasing altruistic and cooperative behaviour with simple moral nudges' *Scientific Reports*, 9(11880), pp. 1-11

Carden, S, Holtzman, N, Strube, M (2017) 'CAHOST: An excel workbook for facilitating the Johnson-Neyman Technique for two-way interactions in multiple regression' *Frontiers in Psychology*, 8, pp. 1293

Carelli, M, Wiberg, B, Wiberg, M (2011) 'Development and Construct Validation of the Swedish Zimbardo Time Perspective Inventory' *European Journal of Psychological Assessment*, 27(4), pp. 220-227

Carroll, G, Choi, J, Laibson, D, Madrian, B, Metrick, A (2009) 'Optimal Defaults and Active Decisions' *Quarterly Journal of Economics*, 124(4), pp. 1639-1674

Casalicchio, E (2019) '*A third of UK voters will cast ballots tactically'* Politico. [Online] [Date accessed: 12/07/2020]: https://www.politico.eu/article/uk-tactical-voting-statistics/

Cesario, J, Grant, H, Higgins, T (2004) 'Regulatory Fit and Persuasion: Transfer From "Feeling Right"' *Journal of Personality and Social Psychology*, 86(3), pp. 388-404

Cesario, J, Higgins, T, Scholer, A (2008) 'Regulatory Fit and Persuasion: Basic Principles and Remaining Questions' *Social Personality and Psychology Compass,* 2(1), pp. 444-463

Chan, C, Holosko, M (2015) 'An Overview of the Use of Mechanical Turk in Behavioral Science: Implications for Social Work' *Research on Social Work Practice*, 26(4), pp. 441-448

Chang, A (2018) '*The Facebook and Cambridge Analytica scandal, explained with a simple diagram'* Vox. [Online] [Date accessed: 14/07/2020]: https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

Chen, A, Potenza, A (2018) '*Cambridge Analytica's Facebook Data Abuse Shouldn't Get Credit for Trump'* The Verge. [Online] [Date accessed: 07/04/2020]: https://www.theverge.com/2018/3/20/17138854/cambridge-analytica-facebook-data-trump-campaign-psychographic-microtargeting

Cho, E (2016) 'Making Reliability Reliable: A Systematic Approach to Reliability Coefficients' *Organizational Research Methods*, 19(4), pp. 651-682

Cohen, A, Stotland, E, Wolfe, D (1955) 'An experimental investigation of need for cognition' *Journal of Abnormal and Social Psychology*, 51(2), pp. 291-294

Cohen, J (1988) '*Statistical Power Analysis for the Behavioural Sciences*' 2nd edition. Lawrence Earlbaum Associates: New Jersey

Cohen, J, Cohen, P, West, S, Aiken, L (2003) '*Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences*' 3rd edition. Lawrence Earlbaum Associates: New Jersey

Coulter, K (2008) 'The Tri-Mediation Model of persuasion: a case for negative political advertising?' *International Journal of Advertising*, 27(5), pp. 1-23

Cronbach, L (1951) 'Coefficient alpha and the internal structure of tests' *Psychometrika*, 16(3), pp. 297-333

Cronbach, L, Snow, R (1977) '*Aptitudes and instructional methods'* New York: Irvington

Daniels, G (1952) '*The "Average Man"?*' [Online] [Date accessed: 06/02/2020]: https://www.dropbox.com/s/bsedsqvgbohy5wg/The%20%22Average%20Man%22%3F.pdf?dl=0

Das, S, Echambadi, R, McCardle, M, Luckett, M (2003) 'The Effect of Interpersonal Trust, Need for Cognition, and Social Loneliness on Shopping, Information Seeking and Surfing on the Web' *Marketing Letters*, 14(3), pp. 185-202

Deaton, A, Cartwright, N (2017) ‘Understanding and misunderstanding randomized controlled trials’ *Social Science and Medicine*, 210(7), pp. 2-21

del Campo, C, Pauser, S, Steiner, E, Vetschera, R (2016) ‘Decision making styles and the use of heuristics in decision making’ *Journal of Business Economics*, 86(4), pp. 389-412

Desilver, D (2018) ‘*A record number of women will be serving in the new Congress*’ Pew Research Center. [Online] [Date accessed: 18/06/2020]: https://www.pewresearch.org/fact-tank/2018/12/18/record-number-women-in-congress/

Dewberry, C, Juanchich, M, Narendran, S (2013) ‘Decision-making competence in everyday life: The roles of general cognitive styles, decision-making styles and personality’ *Personality and Individual Differences*, 55(1), pp. 783-788

Dinev, T, McConnell, A, Smith, H J (2015) ‘Informing Privacy Research Through Information Systems, Psychology, and Behavioural Economics: Thinking Outside the “APCO” Box’ *Information Systems Research*, 26(4), pp. 639-655

Dolan, P, Galizzi, M (2015) ‘Like ripples on a pond: Behavioral Spillovers and their implications for research and policy’ *Journal of Economic Psychology*, 47(2), pp. 1-16

Dowling, C, Krupnikov, Y (2016) ‘*The Effects of Negative Advertising’* DOI: 10.1093/acrefore/9780190228637.013.51

Downs, A (1957) ‘*An Economic Theory of Democracy*’. Harper & Row Publishers: New York.

Dubois, D, Rucker, D, Galinsky, A (2016) ‘Dynamics of Communicator and Audience Power: The Persuasiveness of Competence versus Warmth’ *Journal of Consumer Research*, 43(1), pp. 68-85

Ebreo, A, Vining, J (2001) ‘How similar are recycling and waste reduction? Future Orientation and Reasons for Reducing Waste as Predictors of Self-Reported Behavior’ *Environment and Behavior*, 33(3), pp. 424-448

Egelman, S, Peer, E (2015) 'The Myth of the Average User' *Proceeding of the 2015 New Security Paradigms Workshop*, *September 2015.* DOI: 10.1145/2841113.2841115

Estelami, H (2020) 'The effects of need for cognition, gender, risk preferences and marketing education on entrepreneurial intentions' *Journal of Research In Marketing and Entrepreneurship*. DOI: 10.1108/JRME-04-2018-0023

Fay, M, Proschan, M (2010) 'Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules' *Statistics Surveys*, 4(2010), pp. 1-39

Felder, F, Soloman, B (1991) '*Index of Learning Styles'*

Fellner, G, Sutter, M (2009) 'Causes, Consequences, and Cures of Myopic Loss Aversion: An Experimental Investigation' *The Economic Journal*, 119(537), pp. 900-916

Fischer, S, Soyez, K, Gurtner, S (2015) 'Adapting Scott and Bruce's General Decision-Making Style Inventory to Patient Decision Making in Provider Choice' *Medical Decision Making*, 35(4), pp. 525-532

Fowler, J, Kam, C (2006) 'Patience as a Political Virtue: Delayed Gratification and Turnout' *Political Behavior*, 28(2), pp. 113-128

Frederick, S, Loewenstein, G, O'Donoghue, T (2002) 'Time Discounting and Time Preference: A Critical Review' *Journal of Economic Literature*, 40(2), pp. 351-401

Friestad, M, Wright, P (1994) 'The Persuasion Knowledge Model: How People Cope with Persuasion Attempts' *Journal of Consumer Research*, 21(1), pp. 1-31

Fryer, R (2011) 'Financial Incentives and student achievement: evidence from randomised trials' *The Quarterly Journal of Economics*, 126(4), pp. 1755-1798

Gambetti, E, Fabbri, M, Bensi, L, Tonetti, L (2008) 'A contribution to the Italian validation of the General Decision-making Style Inventory' *Personality and Individual Differences*, 44(4), pp. 842-852

Gambetti, E, Guisberti, F (2019) 'Personality, decision-making styles and investments' *Journal of Behavioral and Experimental Economics*, 80(6), pp. 14-24

Geisler, M, Allwood, C M (2018) 'Relating Decision-Making Styles to Social Orientation and Time Approach' *Journal of Behavioral Decision Making*, 31, 415-429

Gerber, A, Huber, G, Doherty, D, Dowling, C (2014) 'Why People Vote Estimating the Social Returns to Voting' *British Journal of Political Science*, 46(2), pp. 241-264

Gigerenzer, G (2015) 'On the supposed evidence for libertarian paternalism' *Review of Philosophy and Psychology*, 6, pp. 361-383

Gigerenzer, G (2018) 'The Bias Bias in Behavioral Economics' *Review of Behavioral Economics*, 5, pp. 303-336

Girard, A, Reeve, C, Bonaccio, S (2016) 'Assessing decision-making style in French-speaking populations: Translation and validation of the general decision-making style questionnaire' *Revue européenne de psychologie appliquée*, 66(6), pp. 325-333

Glass, G, Peckham, P, Sanders, J (1972) 'Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance' *Review of Educational Research*, 42(3), pp. 237-288

Glazebrook, K, Larkin, C, Costa, E (2017) '*Improving engagement with pensions decisions: The results from three randomised controlled trials*' BIT. [Online] [Date accessed: 20/06/2020]: http://www.behaviouralinsights.co.uk/wp-content/uploads/2017/10/Pension-wise-trials.pdf

Gomez, B, Wilson, M (2006) 'Cognitive Heterogeneity and Economic Voting: A Comparative Analysis of Four Democratic Electorates' *American Journal of Political Science*, 50(1), pp. 127-145

Griffin, D, O'Cass, A (2010) 'An Exploration of Personality and Speed Limit Compliance' *Journal of Nonprofit and Public Sector Marketing*, 22(4), pp. 336-353

Gunsch, M, Brownlow, S, Haynes, S, Mabe, Z (2000) 'Differential Linguistic Content of Various Forms of Political Advertising' *Journal of Broadcasting and Electronic Media*, 44(1), pp. 27-42

Guo, Y, Zhang, Z, Guo, Y, Guo, X (2020) 'Nudging personalized password policies by understanding users' personality' *Computer Security*. DOI: 10.1016/j.cose.2020.101801

Hadj-Abo, A, Enge, S, Rose, J, Kunte, H, Flieeschhauer, M (2020) 'Individual differences in impulsivity and need for cognition as potential risk or resilience factors of diabetes self-management and glycemic control' *PLoS ONE*, 15(1), pp. 1-13

Haggag, K, Paci, G (2014) 'Default Tips' *American Economic Journal: Applied Economics*, 6(3), pp. 1-19

Hagman, W, Andersson, D, Västfjäll, D, Tinghög, G (2015) 'Public Views on Policies Involving Nudges' *Review of Philosophy and Psychology*, 6, pp. 439-453

Hallsworth, M, Berry, D, Sanders, M, Sallis, A, King, D, Vlaev, I, Darzi, A (2015) 'Stating Appointment Costs in SMS Reminders Reduces Missed Hospital Appointments: Findings from Two Randomised Controlled Trials' *PLoS ONE*, 10(9), pp. 1-14

Hallsworth, M, Chadborn, T, Salles, A, Sanders, M, Berry, D, Greaves, F, Clements, L, Davies, S (2016) 'Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial' *The Lancet*, 387, pp. 1743-1752

Halpern, D (2015) '*Inside the Nudge Unit*' WH Allen: London

Hamlin, A, Jennings, C (2011) 'Expressive Political Behaviour: Foundations, Scope and Implications' *British Journal of Political Science*, 41(3), pp. 645-670

Hansen, P (2016) 'The Definition of Nudge and Libertarian Paternalism: Does the Hand Fit the Glove?' *European Journal of Risk Regulation*, 7(1), pp. 155-174

Hansen, P, Jespersen, A (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*, 1(2013), pp. 3-28

Harwell, M, Rubinstein, E, Hayes, W, Olds, C (1992) 'Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases' *Journal of Educational Statistics*, 17(4), pp. 315-339

Hayes, A (2012) '*PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling*' [Online] [Date accessed: 24/06/2020]: https://www.afhayes.com/public/process2012.pdf

Hayes, A (2018) '*An Introduction to Mediation, Moderation and Conditional Process Analysis: A Regression-Based Approach*' 2nd ed, Guildford Press: UK

Hayes, A, Darlington, R (2017) '*Regression Analysis and Linear Models: Concepts, Applications, and Implementations*' Guildford Press: UK

Hayes, A, Rockwood, N (2017) 'Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation' *Behavior Research and Therapy*, 98(4), pp. 39-57

Haynes, L, Service, O, Goldacre, B, Torgerson, D (2013) '*Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials'* BIT and the Cabinet Office. [Online] [Date accessed: 20/06/2020]: https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials

Hemphill, L, Shapiro, M (2019) 'Appealing to the base or to the moveable middle? Incumbants' partisan messaging before the 2016 U.S. congressional elections' *Journal of Information Technology and Politics*, 16(4), pp. 325-341

Hirsh, J, Kang, S, Bodenhausen, G (2012) 'Personalized Persuasion: Tailoring Persuasive Appeals to Recipients' Personality Traits' *Psychological Science*, 23(6), pp. 578-581

Ho, D, Imai, K (2008) ‘Estimating causal effects of ballot order from a randomized natural experiment: California Alphabet Lottery, 1978-2002’ *Public Opinion Quarterly*, 72(2), pp. 216-240

Hopp, T, Vargo, C (2017) ‘Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data’ *Computers in Human Behavior*, 68(2017), pp. 368-377

Hrynowski, Z (2020) ‘*Several Issues Tie as Most Important in 2020 Election*’ Gallup. [Online] [Date accessed: 25/06/2020]: https://news.gallup.com/poll/276932/several-issues-tie-important-2020-election.aspx

Huitink, M, Poelman, M, van den Eynde, E, Seidell, J, Dijkstra, S C (2020) ‘Social norm nudges in shopping trolleys to promote vegetable purchases: A quasi-experimental study in a supermarket in a deprived urban area in the Netherlands’ *Appetite.* DOI: 101016/j.appet.2020.104655

Itai, H, Inoue, A, Kodama, S (2016) ‘Rethinking Nudge: Libertarian Paternalism and Classical Utilitarianism’ *The Tocqueville Review*, 37(1), p. 81-98

Johnson, E, Goldstein, D (2003) ‘Do Defaults Save Lives?’ *Science*, 302(5649), pp. 1338-1339

Johnson, E, Shu, S, Dellaert, B, Fox, C, Goldstein, D, Häubl, G, Larrick, R, Payne, J, Peters, E, Schkade, D, Wansink, B, Weber, E (2012) ‘Beyond nudges: Tools of a choice architecture’ *Marketing Letters,* 23(2), pp. 487-504

Johnson, P, Neyman, J (1936) ‘Tests of certain linear hypotheses and their application to some educational problems’ *Statistical Research Memoirs*, 1, pp. 57-93

Just, D, Price, J (2013) ‘Using Incentives to Encourage Healthy Eating in Children’ *The Journal of Human Resources*, 48(4), pp. 855-872

Kahneman, D (2003) 'Maps of Bounded Rationality: Psychology for Behavioural Economics' *The American Economic Review*, 93(5), pp. 1449-1475

Kahneman, D (2011) '*Thinking, Fast and Slow*' Penguin Books: UK

Kahneman, D, Knetsch, J, Thaler, R (1991) 'Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias' *Journal of Economic Perspectives*, 5(1), pp. 193-206

Kahneman, D, Tversky, A (1979) 'Prospect Theory: An Analysis of Decision under Risk' *Econometrica*, 47(2), pp. 263-292

Katz, M, Shapiro, C (1986) 'Technology Adoption in the Presence of Network Externalities' *Journal of Political Economy,* 94(4), pp. 822-841

Kehle, J, Naimi, S (2019) 'Every data point counts: Political elections in the age of digital analytics' Honors Thesis. [Online] [Date accessed: 23/09/2019]: https://digitalcommons.lmu.edu/honors-thesis/207

Kesberg, R, Pfattheicher, S (2019) 'Democracy matters: a psychological perspective on the beneficial impact of democratic punishment systems in social dilemmas' *Palgrave Communications*, 4(44), pp. 1-13

Kim, T, Park, J (2019) 'More about the basic assumptions of t-test: normality and sample size' *Korean Journal of Anesthesiology*, 72(4), pp. 331-335

Kirby, J, Moore, P, Schofield, N (1998) 'Verbal and visual learning styles' *Contemporary Educational Psychology*, 13(2), pp. 169-184

Korn, L, Betsch, C, Böhm, R, Meier, N (2018) 'Social nudging: The effect of social feedback interventions on vaccine uptake' *Health Psychology*, 37(11), pp. 1045-1054

Kozhevnikov, M (2007) 'Cognitive Styles in the Context of Modern Psychology: Toward an Integrated Framework of Cognitive Style' *Psychological Bulletin*, 133(3), pp. 464-481

Kozyreva, A, Lorenz-Spreen, P, Hertwig, R, Lewandowsky, S, Herzog, S (2020) '*Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the US'* PsyArXiv. [Online] [Date accessed: 31/08/2020]: https://psyarxiv.com/3q4mg

Laibson, D (1994) '*Hyperbolic Discounting and Consumption*' PhD Dissertation, Massachusetts Institute of Technology

Laibson, D (1997) 'Golden Eggs and Hyperbolic Discounting' *The Quarterly Journal of Economics*, 112(2), pp. 443-478

Langer, T, Wever, M (2008) 'Does commitment or feedback influence myopic loss aversion? An experimental analysis' *Journal of Behavior and Organization*, 67(2008), pp. 810-819

Lapowsky, I (2019) '*How Cambridge Analytica Sparked the Great Privacy Awakening'* Wired. [Online] [Date accessed: 14/07/2020]: https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/

Lavi, M (2017) 'Evil Nudges' *Journal of Entertainment and Technology Law,* 21(1), pp. 1-93

Lawson, C, Lenz, G, Baker, A, Myers, M (2010) 'Looking Like a Winner: Candidate Appearance and Electoral Success in New Democracies' World Politics, 62(4), pp. 561-593

Lee, J (2014) 'Are some people less influenced by others' opinions? The role of internal political self-efficacy and need for cognition in impression formation on social networking sites' *Cyberpsychology, Behavior, and Social Networking*, 17(9), pp. 571-577

Levene, H (1960) '*Robust tests for equality of variances*' in Olkin, I, Hotelling, H (eds.) '*Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*' (1960) Stanford University Press: USA

Lijphart, A (1997) 'Unequal Participation: Democracy's Unresolved Dilemma' *The American Political Science Review*, 91(1), pp. 1-14

Lin, C, Yen, H, Chuang, S (206) 'The effects of emotion and need for cognition on consumer choice involving risk' *Marketing Letters*, 17, pp. 47-60

Lindsay, J, Strathman, A (1997) 'Predictors of Recycling Behavior: An Application of a Modified Health Belief Model' *Journal of Applied Social Psychology*, 27(20), pp. 1799-1823

Lipman, S (forthcoming) 'One size fits all? Designing financial incentives tailored to individual economic preferences' *Behavioural Public Policy*. DOI: 10.1017/bpp.2020.21

Little, A, Burriss, R, Jones, B, Roberts, C (2007) 'Facial appearance affects voting decisions' Evolution and Human Behavior, 28(1), pp. 18-27

Liu, W (2020) '*Abolish Silicon Valley*' Repeater Books: UK

Lix, L, Keselman, J, Keselman, H (1996) 'Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test' *Review of Educational Research*, 66(4), pp. 579-619

Loo, R (2000) 'A psychometric evaluation of the General Decision-Making Style Inventory' *Personality and Individual Differences*, 29(5), pp. 539-561

Luckerson, V (2015) '*Here's How Facebook's News Feed Actually Works'* Time Magazine. [Online] [Date accesed: 12/07/2019]: https://time.com/collection-post/3950525/facebook-news-feedalgorithm/

Madrian, B (2014) '*Applying Insights from Behavioral Economics to Policy Design'* NBER working paper series no. 20318. [Online] [Date accessed: 20/06/2020]: https://www.nber.org/papers/w20318

Madrian, B, Shea, D (2001) 'The Power of Suggestion: Inertia in 401(k) Participation and Savings Behaviour' *Quarterly Journal of Economics*, 116(4), pp. 1149-1187

Mann, H, Whitney, D (1947) 'On a test of whether one or two random variables is stochastically larger than the other' *Annals of Mathematical Statistics*, 18(1), pp. 50-60

Mason, W, Suri, S (2011) 'Conducting behavioural research on Amazon's Mechanical Turk' *Behavior Research Methods*, 44, pp. 1-23

Massa, P, Avesani, P (2007) 'Trust Metrics on Controversial Users: Balancing Between Tyranny of the Majority' *International Journal on Semantic Web and Information Systems*, 3(1), pp. 39-64

Matz, S C, Kosinki, M, Nave, G, Stillwell, D J (2017) 'Psychological targeting as an effective approach to digital mass persuasion' *PNAS*. 114(48), pp. 12714-12719

McKenzie, C, Nelson, J (2003) 'What a Speaker's Choice of Frame Reveals: Reference Points, Frame Selection and Framing Effects' *Psychonomic Bulletin and Review*, 10(3), pp. 596-602

Mill, J S (1859) '*On Liberty'* in Philp, M, Rosen, F (eds.) '*John Stuart Mill: On Liberty, Utilitarianism and Other Essays'* (2015) Oxford University Press: USA

Mills, S (forthcoming) 'Personalized Nudging' *Behavioural Public Policy*. DOI: 10.1017/bpp.2020.7

Mitchel, G (2005) 'Libertarian Paternalism is an Oxymoron' *Northwestern University law review*, 99(3), pp. 1245-1270

Moon, Y (2002) 'Personalization and Personality: Some Effects of Customizing Message Style Based on Consumer Personality' *Journal of Consumer Psychology*, 12(4), pp. 313-325

Moon, Y (2010) '*Different: Escaping the Competitive Herd*' Crown Business: USA

My Lien Rebetez, M, Barsics, C, Rochat, L, D'Argembeau, A, van der Linden, M (2016) 'Procrastination, consideration of future consequences, and episodic future thinking' *Consciousness and Cognition*, 42(2), pp. 286-292

Negroponte, N (1995) '*Being Digital*' Vintage: UK

Nickerson, R (1998) ‘Confirmation Bias: A Ubiquitous Phenomenon in Many Guises’ *Review of General Psychology*, 2(2), pp. 175-220

Nunnally, J, Bernstein, I (1994) ‘*Psychometric Theory’* McGraw-Hill: New York

O’Connor, D, Warttig, S, Conner, M, Lawton, R (2009) ‘Raising awareness of hypertension risk through a web-based framing intervention: Does consideration of future consequences make a difference?’ *Psychology, Health and Medicine*, 14(2), pp. 213-219

O’Donoghue, T, Rabin, M (1999a) ‘Doing It Now or Later’ *The American Economic Review*, 89(1), pp. 103-124

O’Donoghue, T, Rabin, M (1999b) ‘Incentives for Procrastinators’ *The Quarterly Journal of Economics*, 114(3), pp. 769-816

O’Donoghue, T, Rabin, M (2001) ‘Choice and Procrastination’ *The Quarterly Journal of Economics*, 116(1), pp. 121-160

O’Donoghue, T, Rabin, M (2015) ‘Present Bias: Lessons Learned and To Be Learned’ *American Economic Review*, 105(5), pp. 273-279

O’Hara, R, Walter, M, Christopher, A (2009) ‘Need for Cognition and Conscientiousness as Predictors of Political Interest and Voting Strategy’ *Journal of Applied Social Psychology*, 39(6), pp. 1397-1416

O’Shea, L (2019) ‘*Future Histories*’ Verso Books: UK

Oliver, A (forthcoming) ‘Towards a New Political Economy of Behavioural Public Policy’ *Public Policy Review.* DOI: 10.1111/puar.13093

Oppenheimer, D, Meyvis, T, Davidenko, N (2009) ‘Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power’ *Journal of Experimental Social Psychology*, 45(4), pp. 867-872

Orbell, S, Kyriakaki, M (2008) ‘Temporal framing and persuasion to adopt preventive health behavior: Moderating effects of individual differences in consideration of future consequences and sunscreen use’ *Health Psychology*, 27(6), pp. 770-779

Orbell, S, Perugini, M, Rakow, T (2004) ‘Individual differences in sensitivity to health communications: Consideration of future consequences’ *Health Psychology*, 23(4), pp. 388-396

Ordway, D, Wihbey, J (2016) ‘*Negative political ads and their effect on voters: Updated collection of research’* [Online] [Date accessed: 18/06/2019]: https://journalistsresource.org/studies/politics/ads-public-opinion/negative-political-ads-effects-voters-research-roundup/

Page, L, Castleman, B, Meyer, K (2020) ‘Customized Nudging to Improve FAFSA Completion and Income Verification’ *Educational Evaluation and Policy Analysis*, 42(1), pp. 3-21

Peer, E, Brandimarte, L, Samat, S, Acquisti, A (2017) ‘Beyond the Turk: Alternative platforms for crowdsourcing behavioural research’ *Journal of Experimental Social Psychology*, 70(5), pp. 153-163

Peer, E, Egelman, S, Harbach, M, Malkin, N, Mathur, A, Frik, A (2019) ‘*Nudge me right: Personalizing online nudges to people’s decision-making styles’* SSRN. [Online] [Date accessed: 10/07/2019]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3324907

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’* SSRN. [Online] [Date accessed: 26/06/2020]: https://poseidon01.ssrn.com/delivery.php?ID=036104118122119093071064024115065091118034032080036086072001115090126064118102030095039031120001116015022105019064111121006030061050017010060097000098079121123115007088053079114118090103000013124126001077001126109007123118103118090119102093013000009009&EXT=pdf

Pensions Act 2008 (56 Eliz. 2) London: HMSO

Peters, E, Västfjäll, D, Slovic, P, Mertz, C K, Mazzocco, K, Dickert, S (2006) 'Numeracy and Decision Making' *Psychological Science*, 17(5), pp. 407-413

Porat, A, Strahilevitz, L J (2014) 'Personalizing Default Rules and Disclosure with Big Data' *Michigan Law Review,* 112(8), pp. 1417-1478

Praino, R (2018) '*How a candidate's looks may be swinging your vote (without you even realising it)*' The Conversation. [Online] [Date accessed: 08/07/2020]: https://theconversation.com/how-a-candidates-looks-may-be-swinging-your-vote-without-you-even-realising-it-107364

Praino, R, Stockemer, D (2018) 'What Are Good-Looking Candidates, and Can They Sway Election Results?' *Social Science Quarterly*, 100(3), pp. 531-543

Preacher, K, Curran, P, Bauer, D (2006) 'Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis' *Journal of Educational and Behavioral Statistics*, 31(4), pp. 437-448

Prelec, D (2004) 'Decreasing Impatience: A Criterion for Non-stationary Time Preference and "Hyperbolic" Discounting' *Scandinavian Journal of Economics*, 106(3), pp. 511-532

Prelec, D, Loewenstein, G (1991) 'Decision Making Over Time and Under Uncertainty: A Common Approach' *Management Science*, 37(7), pp. 770-786

Prelec, D, Simester, D (2001) 'Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay' *Marketing Letters*, 12(1), pp. 5-12

Quattrociocchi, W, Scala, A, Sunstein, C (2016) '*Echo Chambers on Facebook*' SSRN. [Online] [Date accessed: 13/07/2020]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110

Rawls, J (1971) '*A Theory of Justice*' Harvard University Press: USA

Razali, N M, Wah, Y B (2011) ‘Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests’ *Journal of Statistical Modeling and Analytics*, 2(1), pp. 21-33

Rebonato, R (2014) ‘A Critical Assessment of Libertarian Paternalism’ *Journal of Consumer Policy*, 37(3), pp. 357-396

Redfern, J, Enright, G, Raadsma, S, Allman-Fairnelli, M, Innes-Hughes, C, Khanal, S, Lukeis, S, Rissel, C, Gyani, A (2016) ‘Effectiveness of a behavioural incentive scheme linked to goal achievement: study protocol for a randomized controlled trial’ *Trials,* 17(33), pp. 1-9

Reinecke, K, Gajos, K (2014) ‘Quantifying Visual Preferences Around the World’ *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*.

Resnick, B (2018) ‘*Cambridge Analytica’s “psychographic microtargeting”: what’s bullshit and what’s legit?’* Vox Magazine. [Online] [Date accessed: 07/04/2020]: https://www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what

Rivers, N, Shenstone-Harris, S, Young, N (2017) ‘Using nudges to reduce waste? The case of Toronto’s plastic bag levy’ *Journal of Environmental Management,* 188(1), pp. 153-162

Rizzo, M, Whitman, D (2009) ‘The Knowledge Problem of New Paternalism’ *BYU Law Review*, 2009(4), pp. 905-968

Robinson, P (2019) ‘Expanding the Field of Political Communication: Making the Case for a Fresh Perspective Through “Propaganda Studies”’ *Frontiers In Communication*, 4(26), pp. 1-13

Rokka, J, Airoldi, M (2018) ‘*Cambridge Analytica’s ‘secret’ psychographic tool is a ghost from the past’* The Conversation. [Online] [Date accessed: 07/04/2020]: https://theconversation.com/cambridge-analyticas-secret-psychographic-tool-is-a-ghost-from-the-past-94143

Ruggeri, K, Benzerga, A, Verra, S, Folke, T (forthcoming) 'A behavioral approach to personalizing public health' *Behavioural Public Policy*. DOI: 10.1017/bpp.2020.31

Sabbagh, D (2019) '*Tactical voting across the UK: a region-by-region guide'* The Guardian. [Online] [Date accessed: 12/07/2020]: https://www.theguardian.com/politics/2019/dec/10/tactical-voting-across-the-uk-a-region-by-region-guide

Sætra, H (2019) 'When Nudge comes to Shove: Liberty and Nudging in the Era of Big Data' *Technology in Society*. DOI: 10.1016/j.techsoc.2019.04.006

Samuelson, W, Zeckhauser, R (1988) 'Status Quo Bias in Decision Making' *Journal of Risk and Uncertainty*, 1, pp. 7-59

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), pp. 144-167

Saunders, D (1956) 'Moderator variables in prediction' *Educational and Psychological Measurement*, 16(2), pp. 209-222

Schöning, C, Matt, C, Hess, T (2019) 'Personalised Nudging for More Data Disclosure? On the Adaption of Data Usage Policies Format to Cognitive Styles' *Proceeding of the 52nd Hawaii International Conference on System Sciences,* pp. 4395-4404

Schubert, C (2015) '*On the ethics of public nudging: Autonomy and agency'* Joint Discussion Paper Series in Economics no. 33-2015, Department of Business Administration & Economics, Marburg. [Online] [Date accessed: 13/05/2018]: https://www.econstor.eu/bitstream/10419/125535/1/837886600.pdf

Schubert, C (2017) 'Exploring the (behavioural) political economy of nudging' *Journal of Institutional Economics*, 13(3), pp. 499-522

Schultz, W, Nolan, J, Cialdini, R, Goldstein, N, Griskevicius, V (2007) 'The Constructive, Destructive, and Reconstructive Power of Social Norms' *Psychological Science*, 18(5), pp. 429-434

Scott, S, Bruce, R (1995) 'Decision-making style: The development and assessment of a new measure' *Educational and Psychological Measurement,* 55(5), pp. 818-831

Service, O (2015) '*Automatic Enrolment and Pensions: a behavioural success story*' BIT. [Online] [Date accessed: 09/01/2018]: https://www.bi.team/blogs/automatic-enrolment-and-pensions-a-behavioural-success-story/

Sher, S, McKenzie, C (2006) 'Information Leakage from Logically Equivalent Frames' *Cognition*, 101(3), pp. 467-494

Simon, H (1955) 'A Behavioural Model of Rational Choice' *Quarterly Journal of Economics*, 69(1), pp. 99-118

Singer, E (2002) '*The use of incentives to reduce nonresponse in household surveys'* In Groves, R, Dillman, D, Eltinge, J, Little, R (eds.) '*Survey non-response*' (2002). Wiley: New York

Singer, E, Cooper, M (2008) 'Do Incentives Exert Undue Influence on Survey Participation? Experimental Evidence' *Journal of Empirical Research on Human Research Ethics*, 3(3), pp. 49-56

Singer, E, Kulka, R (2002) '*Paying respondents for survey participation*' in ver Ploeg, M, Moffitt, R, Citro, C (eds.) '*Studies of welfare populations: Data collection and research issues'* (2002) National Academy Press: Washington

Singer, E, Ye, C (2012) 'The Use and Effects of Incentives in Surveys' *The ANNALS of the American Academy of Political and Social Science*, 645(1), pp. 112-141

Sirois, F (2004) 'Procrastination and intentions to perform health behaviors: The role of self-efficacy and the consideration of future consequences' *Personality and Individual Differences*, 37(1), pp. 115-128

Snedecor, G, Cochran, W (1989) '*Statistical Methods'* 8th edition. Iowa University Press: USA

Sohlberg, J (2019) 'Elections Are (Not) Exciting: Need for Cognition and Electoral Behaviour' *Scandinavian Political Studies*, 42(2), pp. 138-150

Soper, D (2020) '*A-priori Sample Size Calculator for Multiple Regression*' [Online] [Date accessed: 26/06/2020]: http://www.danielsoper.com/statcalc

Spicer, D, Sadler-Smith, E (2005) 'An examination of the general decision making style questionnaire in two UK samples' *Journal of Managerial Psychology*, 20(2), pp. 137-149

Spiller, S, Fitzsimmons, G, Lynch, J, McClelland, G (2013) 'Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression' *Journal of Marketing Research*, 50(2), pp. 277-288

Sposini, F M (2019) 'At the borders of the average man: Adolphe Quêtelet on mental, moral and criminal monstrosities' *Journal of Historical Behavioural Science,* 2019(1), pp. 1-17

Srivastava, K, Sharma, N (2012) 'Consumer attitude towards brand-extension incongruity: The moderating role of need for cognition and need for change' *Journal of Marketing Management*, 28(5-6), pp. 652-675

Stone, D (2012) '*Policy Paradox: The Art of Political Decision Making'* W.W. Norton & Co: USA

Strathman, A, Gleicher, F, Boninger, D, Edwards, C S (1994) 'The Consideration of Future Consequences: Weighing Immediate and Distant Outcomes of Behavior' *Journal of Personality and Social Psychology*, 66(4), pp. 742-752

Sullivan, P (2020) '*Why the US has the most reported coronavirus cases in the world'* The Hill. [Online] [Date accessed: 09/07/2020]: https://thehill.com/policy/healthcare/498876-why-the-us-has-the-most-reported-coronavirus-cases-in-the-world

Sunstein, C (1996) 'Social Norms and Social Roles' *Columbia Law Review*, 96(4), pp. 903-968

Sunstein, C (2001) '*Republic.com*' Princeton University Press: USA

Sunstein, C (2012) '*Impersonal Default Rules vs. Active Choices vs. Personalized Default Rules: A Triptych'* SSRN. [Online] [Date accessed: 10/07/2019]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2171343

Sunstein, C (2013a) 'Behavioural Economics and Paternalism' *Yale Law Journal*, 122(7), pp. 1867-1899

Sunstein, C (2013b) '*Why Nudge? The Politics of Libertarian Paternalism'* Yale University Press: USA

Sunstein, C (2017a) '*Misconceptions About Nudges'* SSRN. [Online] [Date accessed: 24/04/2020]: https://ssrn.com/abstract=3033101

Sunstein, C (2017b) 'Nudges that Fail' *Behavioural Public Policy*, 1(1), pp. 4-25

Tannenbaum, D, Ditto, P (2011) '*Information Asymmetries in Default Options*' Unpublished Manuscript.

Thaler, R (2015) '*Misbehaving'* W. W. Norton: USA

Thaler, R, Benartzi, S (2004) 'Save More Tomorrow: Using Behavioural Economics to Increase Employee Saving' *Journal of Political Economy*, 112(1), pp. 164-187

Thaler, R, Sunstein, C (2003) 'Libertarian Paternalism' *American Economic Review*, 93(2), pp. 175-179

Thaler, R, Sunstein, C (2008) '*Nudge'* Penguin Books: London

Thaler, R, Sunstein, C, Balz, J (2014) '*Choice Architecture*' in Shafir, E (eds.) '*The Behavioral Foundations of Public Policy*' (2012). Princeton University Press: USA

Thaler, R, Tucker, W (2013) 'Smarter Information, Smarter Consumers' *Harvard Business Review*, 91(1-2), pp. 44-54

Thunholm, P (2004) 'Decision-making style: habit, style or both?' *Personality and Individual Differences*, 36(4), pp. 931-944

Thunström, L, Gilbert, B, Jones-Ritten, C (2018) 'Nudges that hurt those already hurting – distributional and unintended effects of salience nudges' *Journal of Economic Behaviour and Organisation,* 153(1), pp. 267-282

Tor, A (forthcoming) 'Nudges that should fail?' *Behavioural Public Policy*. DOI: 10.1017/bpp.2019.5

Tversky, A, Kahneman, D (1973) 'Availability: A heuristic for judging frequency and probability' *Cognitive Psychology*, 5(2), pp. 207-232

Tversky, A, Kahneman, D (1992) 'Advances in Prospect Theory: Cumulative Representation of Uncertainty' *Journal of Risk and Uncertainty*, 5, pp. 297-323

Unt, T, Solvak, M, Vassil, K (2017) 'Does Internet voting make elections less social? Group voting patterns in Estonian e-voting log files (2013-2015)' *PLoS ONE*, 12(5), pp. 1-13

Utych, S (2018) 'Negative Affective Language in Politics' *American Politics Research*, 46(1), pp. 77-102

Veblen, T (1899) '*The Theory of the Leisure Class*' (2012) Renaissance Classics: USA

Wade, M (2018) '*Psychographics: the behavioural analysis that helped Cambridge Analytica know voters' minds'* The Conversation. [Online] [Date accessed: 07/04/2020]: https://theconversation.com/psychographics-the-behavioural-analysis-that-helped-cambridge-analytica-know-voters-minds-93675

Weinmann, M, Schneider, C, vom Brocke, J (2016) '*Digital Nudging'* Working Paper, SSRN. [Online] [Date accessed: 06/02/2020]: https://ssrn.com/abstract=2708250

Welch, B (1947) 'The generalization of "Student's" problem when several different population variances are involved' *Biometrika*, 34(1-2), pp. 28-35

Weller, J, Dieckmann, N, Tusler, M, Mertz, C, Burns, W, Peters, E (2012) 'Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach' *Journal of Behavioral Decision Making*, 26(2), pp. 198-212

Wilcoxon, F (1945) 'Individual Comparisons by Ranking Methods' *Biometrics Bulletin*, 1(6), pp. 80-83

Yeung, K (2017) 'Hypernudge: Big Data as a Mode of Regulation by Design' *Information Communication and Society,* 20(1), pp. 118-136

Zuboff, S (2019) '*The Age of Surveillance Capitalism'* Profile Books: UK