

---

**The Effectiveness of  
Commitment Devices:  
Field Experiments on Health  
Behaviour Change**

---

**Manu Manthri Savani**

**Thesis Submitted for the Degree of  
Doctor of Philosophy**

**Department of Political Science  
UCL**

## **Declaration**

I, Manu Manthri Savani, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Manu Savani

## **Abstract**

Behavioural public policy, as popularised by the “nudge” agenda, aims to help people make better choices in the face of their inherent biases (Thaler and Sunstein, 2008), including over diet and weight management (Liu et al, 2014). Present bias can lead to time inconsistency: individuals identify an optimal course of action but when the moment comes to take that action they delay or quit, prioritizing present gains at the expense of longer term benefits (O’ Donoghue and Rabin, 1999).

Time inconsistency is explained in Thaler and Shefrin’s dual-self model (1981) as the result of an internal tussle between a myopic ‘doer’ and a far-sighted ‘planner’. Commitment devices – voluntary strategies to change future behaviours – can help people stay on track with their goals. Emerging empirical evidence from psychology, medicine, and behavioural economics bears out this prediction for health behaviours (Prestwich et al, 2012; Volpp et al, 2008; Giné et al, 2010), but commitment devices remain relatively under-researched (Perry et al, 2015).

The dissertation sets out a fresh analytical framework applying, for the first time, planner-doer theory to health behaviours for weight loss. It also explores how commitment devices might work differently across sub-groups. The empirical strategy, combining quantitative and qualitative methods, centres on two field experiments testing for average and heterogeneous treatment effects of commitment devices on self-monitoring behaviour, participation in a weight loss programme, and weight loss outcomes.

Results indicate commitment devices improve health behaviours, but have mixed effects on weight loss: highlighting the potential for commitment overload, and the importance of choosing the right dose of commitment. Qualitative evidence provides fresh insights for planner-doer theory. Differential impacts on sub-groups imply a need for careful targeting and design of commitment devices. The dissertation concludes there is scope for commitment devices to play an effective role in behaviour change programmes.

# Table of Contents

<b>Abstract</b>	<b>iii</b>
<b>Chapter 1: INTRODUCTION</b>	<b>1</b>
<b>Time Inconsistency, Planning versus Doing, and Commitment Devices for Health</b>	
1. The behavioural puzzle: time inconsistency	2
2. Planner-doer theory and commitment devices	4
3. Time inconsistency and health	7
4. Intended contributions	12
5. Roadmap of the dissertation	15
6. Findings and key results	17
7. Summary	20
<b>Chapter 2: LITERATURE REVIEW</b>	<b>23</b>
<b>A Theory, Taxonomy, and Empirical Review of Commitment Devices</b>	
1. Introduction	24
2. Planner-doer theory of time inconsistency	27
3. Practical applications of commitment devices	38
4. Empirical evidence on commitment device effects	47
5. Literature review conclusions	63
<b>Chapter 3: ANALYTICAL FRAMEWORK</b>	<b>65</b>
<b>Translating Planner-Doer Theory into Testable Propositions</b>	
1. Introduction	66
2. The model	69
3. Three propositions on commitment devices	79
4. Three further propositions on heterogeneous effects	88
5. Conclusions	102
<b>Chapter 4: RESEARCH DESIGN</b>	<b>107</b>
<b>Field Experiments To Isolate Causal Effect and Test Theory</b>	
1. Introduction	108
2. Broad methodological choices	111
3. Forging partnerships for two field experiments	119
4. Field experiment 1: Food Monitor	123
5. Field experiment 2: Camden	137
6. Qualitative analysis plan	150
7. Ensuring internal validity and plausibility	157
8. External validity	177
9. Research design summary	183

<b>Chapter 5: RESULTS AND ANALYSIS (1)</b>	<b>187</b>
<b>Coaches, Refunds, and Overload in the Food Monitor Experiment</b>	
1. Introduction	188
2. Field experiment implementation	190
3. Descriptive statistics	200
4. Outcome data and strategies to address attrition	205
5. Average treatment effects	214
6. Heterogeneous treatment effects	227
7. Discussion	232
8. Conclusions	236
<b>Chapter 6: RESULTS AND ANALYSIS (2)</b>	<b>239</b>
<b>Contracts, thresholds and saturation in the Camden experiment</b>	
1. Introduction	240
2. Field experiment implementation	243
3. Descriptive statistics	250
4. Outcome data	258
5. Average treatment effects	266
6. Heterogeneous treatment effects	280
7. Discussion	288
8. Conclusions	291
<b>Chapter 7: RESULTS AND ANALYSIS (3)</b>	<b>297</b>
<b>Adherence, Planner-Doer Interactions, and Sophistication</b>	
1. Introduction	298
2. Investigating adherence with qualitative data	301
3. The planner-doer tussle	310
4. Sophistication and demand for commitment devices	317
5. Conclusions	329
<b>Chapter 8: CONCLUSIONS</b>	<b>333</b>
1. What did the dissertation set out to achieve, why and how?	334
2. Research findings	341
3. Generalisability and limitations of the study	355
4. Contributions to the scholarly debate	361
5. Implications for policy and future research	370
6. Closing remarks	376
<b>APPENDIX</b>	<b>378</b>
<b>REFERENCES</b>	<b>448</b>

## List of Tables

1	Commitment devices for weight loss	40
2	Overview of literature on commitment devices for weight loss	49
3	Overview of field experiments testing commitment devices on health behaviours	59
4	Capturing myopia through health attitudes	99
5	Factors determining commitment device effectiveness	101
6	Research questions, propositions from the model and hypotheses	105
7	Ex ante sample size calculations (Food Monitor)	127
8	Ex ante sample size calculations (Camden)	141
9	Sub-group analysis	149
10	Overview of qualitative data and intended analysis	151
11	Preliminary coding scheme for interview data	156
12	Key features of the field experiments	184
13	Randomisation balance check (Food Monitor)	197
14	Weight and BMI profile	201
15	Descriptive statistics: myopia	203
16	Descriptive statistics: time preference	204
17	Average weight loss and self-monitoring outcomes	208
18	Attrition patterns across experimental groups	211
19	Can commitment devices boost weight loss?	218
20	Can commitment devices boost self-monitoring behaviour?	224
21	Do commitment devices work differently across sub-groups?	229
22	Recruitment to Camden trial	244
23	Randomisation balance check (Camden)	248
24	Weight and BMI profile	251
25	Descriptive statistics: myopia	253
26	Attrition patterns across experimental groups	259
27	Average participation rates	261
28	Coding scheme for interview data	264
29	Can commitment contracts boost weight loss?	270
30	Use of commitment contracts by the treatment group	274
31	Participation rates across experimental groups	275
32	Can commitment contracts boost attendance and completion rates?	277
33	Do commitment devices work differently across sub-groups?	283
34	Commitment contracts substitute for other commitment elements	286
35	Participant characteristics by adherence groups (Food Monitor)	304
36	Coding high- and low-adherence in the Camden trial	306
37	Participant characteristics by adherence groups (Camden)	308
38	Evidence of planner-doer sub-selves	311
39	Coding sophistication in the Camden trial	319
40	Research questions, selected propositions from the model and hypotheses (reproduced from table 5)	340

## List of Figures

1	Spectrum of commitment devices	46
2	Body Mass Index (BMI) calculator	49
3	Costs of consumption and satiety	73
4	Consumption behaviour and weight loss outcomes	74
5	The model	75
6	Planner identifies a need for a commitment device	80
7	A commitment device changes behaviours	84
8	Complex causal nexus of commitment device effects	91
9	Commitment devices by intensity of design	92
10	Causal effects of commitment devices	102
11	Food Monitor dashboard	124
12	Stages of the Food Monitor experiment	129
13	Commitment contract treatment	139
14	Stages of the Camden experiment	143
15	Food Monitor experiment flow chart (CONSORT)	192
16	Recruitment to experimental groups (Food Monitor)	192
17	Monthly fee message (comparison group)	193
18	Refund treatment offer (limited commitment group)	194
19	Coach treatment offer (reputational commitment group)	194
20	Initial BMI distribution (Food Monitor)	202
21	Weight loss over 4 and 12 weeks	208
22	Usage of all self-monitoring tools	209
23	Weight loss at 4 weeks by group	215
24	Weight loss at 12 weeks by group	215
25	Self-monitoring over 4 weeks by group	223
26	Weight loss and present bias	229
27	Camden experiment flow chart (CONSORT)	245
28	Contract treatment (reproduced from figure 13)	245
29	Initial BMI distribution (Camden)	251
30	Weight loss outcomes	260
31	Weight loss with and without the contract	267
32	Attendance rates by experimental group	276
33	Stronger weight loss with contract and GP referral combined	285
34	Commitment device effect sizes on weight loss (Cohen's d)	348
35	A guide to applying commitment devices for health behaviour change	372

## List of Appendices

A1	Cohen's <i>d</i> calculations	379
A2	Healthy Foundations Segmentation Model	380
A3	CONSORT checklists	383
A4	Interview topic list for Camden	389
A5	Sample size calculations: Food Monitor	390
A6	Sample size calculations: Camden	393
A7	Stata randomisation exercise	395
A8	Informed consent documents	396
A9	Baseline survey	399
A10	Data gathered through Food Monitor systems	404
A11	Reputational treatment take-up decision	405
A12	Food Monitor baseline variables	406
A13	Interpolated outcome data	416
A14	Weight loss outliers: Food Monitor	418
A15	What drives attrition in the Food Monitor trial	421
A16	IPW for 12-week weight loss analysis	423
A17	Robustness checks: chapter 5	424
A18	Heterogeneous treatment effected in chapter 5	430
A19	Corrections for multiple-hypothesis testing in chapter 5	431
A20	Tutors and recruitment details for Camden	432
A21	Camden baseline variables	435
A22	Attrition in chapter 6	438
A23	Robustness checks: chapter 6	442
A24	Weight loss outliers: Camden	443
A25	Summary of interviewees in Camden experiment	445
A26	Heterogeneous treatment effected in chapter 6	446
A27	Corrections for multiple-hypothesis testing in chapter 6	447



## **Acknowledgements**

My PhD journey has been one of personal and intellectual discovery, and I have many people to thank for helping me reaching the final stage.

I thank my supervisors, Professor Peter John and Dr Roland Kappe, for their unfailing encouragement and intellectual support. Professor John gave me the opportunity to switch tracks after my early career in international development, for which I am grateful. The thesis has benefitted immeasurably from his attention to detail and vast experience of conducting field experiments. On a personal level, Professor John's reassurance and understanding made my maternity leave and return to the thesis easier than I could have hoped. Dr Kappe gamely joined at the halfway stage, and provided patient and insightful challenge that has undoubtedly strengthened the thesis. Thanks go also to Dr Jan-Emmanuel de Neve for early supervision and support. I thank my examiners, Prof Albert Weale and Prof David Torgerson for their encouragement and interest in my research.

I owe a great debt to the organisations that hosted my fieldwork. I am hugely grateful – particularly to Verena Trend, Ian Reddington and his team – for accommodating my funny ideas about commitment devices and giving me the space to generate this research. Most of all my gratitude goes to the participants who gave their time and input and made this research possible.

UCL's Department of Political Science has been a wonderful home that has given me so much: from the departmental studentship, and funding for fieldwork, training and conferences, to the lively seminars, ready advice, and PhD networks. Special thanks go to Helen Elliot and Nicky Henson. I thank Dr Niheer Dasandi for the frank advice back in Friends coffee shop in 2011. Dr Lisa Vanhala introduced me to a new world of qualitative methods. Dr David Hudson and Dr Nils Metternich gently but firmly raised my ambitions during the Upgrade, and my thesis is surely richer and bolder for it. My PhD colleagues – with special shout outs to Sofia, Paolo, Orly, Heleen, Jean-Paul – have been a tremendous source of wisdom, fun, and motivation. Moira, thank you for your infectious curiosity, your thoughtful advice and suggestions on my research, making the early replication

exercises so much fun; perhaps most of all for the empathy as a fellow experimenter. Beth, you set an incredible example in grace and grit. As I look back on the PhD I count my finding such a friend as you an unexpected and wonderful blessing. Thank you for your listening ear and your kindnesses to my family.

And so to my family, for whom my gratitude is boundless. To Saroj and Arvin Patel, loving and generous grandparents to my sons, ever ready to step in for late evenings and sick days; thank you for making me a little less missed, and your steadfast encouragement to Hiten and I in all our endeavours. To Anu Manthri, my sister, for Harry Potter, for wit and absurdity, for ludicrous belief in me, and for remembering my research question four years later. To Sushma and Sudheer, to Smt Ramanujamma and late Sri Ramdoss, to Smt Shanta Bai and Sri Ratanlal, and “to all of those names I do not know, but whose blood runs through my veins, this is a song for you...”. To my father, Dr Sudheer Manthri, for believing this was a natural path for me to tread, for being a life-long learner, for being a role model in professionalism and academic pursuit, and for all the life lessons summed up in Telugu proverbs. To my mother, Dr Sushma Manthri, where to begin? For the countless hours of listening and encouraging, for the pride unspoken, but most of all, for the example of combining motherhood with an enduring ambition to improve, to succeed, to overcome, and to achieve.

My husband, Hiten Savani, was at my side from the days I began to love PPE, and a quiet flame was lit. Years later I started talking about a PhD and you were unflinching in your encouragement and support. I have asked you to read and to listen, to juggle and to sacrifice and to carry, more than you signed up for. You kept sight of the finish line when I was myopic. I quite simply could not have done this without you, and would not have wanted to. You have enriched this experience as you have my life, Bob, “all of me loves all of you”. I dedicate this thesis to you.

Finally, to Suraj and Khush Savani, my sunshine and my happiness. You make us want to be better people because we are your parents. Thank you for sharing me with this book. Love, always.

---

## **Chapter 1**

### **INTRODUCTION:**

# **Time Inconsistency, Planning Versus Doing, and Commitment Devices for Health**

---

## **1. THE BEHAVIOURAL PUZZLE: TIME INCONSISTENCY**

At the heart of this thesis is a simple behavioural puzzle: time inconsistency, or the prediction that your future self may not follow the plan you make today (Wilkinson & Klaes 2012). Individuals often identify an optimal course of action but when the moment comes to put it into practice, they delay or quit (Strotz 1955). Anecdotal examples of such behaviour are easily observed in day-to-day situations, such as postponing a flu jab, delaying transfers to a savings account, or swapping exam revision for television. Though these scenarios may appear trivial, going off track in these ways can have a serious impact on health, finances and educational achievement: the postponed vaccination may lead to falling sick, lower savings increases vulnerability to future economic shocks, and insufficient preparation can lead to lower exam scores.

Time inconsistency comes to the fore when making inter-temporal choices, decisions that require a trade-off between benefits now and benefits later. For example, to take out insurance against a natural disaster requires paying a premium. The cost is incurred today, to bring about future benefits in the form of insurance repayments; but those benefits may or may not accrue at a later time. Humans have long been observed to discount a distant and unknowable future relative to the present. In the classic text ‘Capital and Interest’, Bohm-Bawerk asserts that “in circumstances otherwise equal, we prefer a present enjoyment to a future” (1890, p.284); even in the absence of uncertainty, Pigou claims in ‘The Economics of Welfare’ that “everybody prefers present pleasures or satisfactions of given magnitude to future pleasures or satisfactions of equal magnitude, even when the latter are perfectly certain to occur” (1932, I.II.3). But over-emphasising the present payoffs at the expense of those that can only be realised later has been linked to a range of undesirable outcomes including weaker academic performance (Alan & Ertac 2015, p.113), resistance to climate change adaptation

measures (Kunreuther & Weber 2014, p.403), insufficient savings for retirement (Thaler & Benartzi 2004, p.S168).

What might explain such inconsistent and apparently irrational behaviour? Intuitively, it is clear that individuals are not always equipped with full information, perfect foresight or steely willpower; they are not the perfectly rational economic agents of textbook wisdom. The behavioural economics discipline has shown through a robust body of literature that, under such conditions, decision-making can be subject to inherent biases that distort the process of weighing up pros and cons, such as overconfidence, loss aversion, and present bias (Kahneman 2003; Thaler 2016). It is these present-biased preferences that would mean a person has little appetite to forego short-run enjoyment for a long run aspiration. Recognising the power of such biases, and their scope for influencing sub-optimal choices, this dissertation is located in the scholarly debate on behavioural public policy: using the theoretical insights from economics and psychology to understand human decision-making and design better policies (Shafir, 2013).

## **2. PLANNER-DOER THEORY AND COMMITMENT DEVICES**

### **2.1. *The individual has two sub-selves: planner vs. doer***

In order to better understand the internal machinations that lead to time inconsistency, this thesis follows in the tradition of two-self and two-system models (Schelling 1984; Kahneman 2003; Fudenberg & Levine 2006). Time inconsistency can be explained as the natural outcome of dual, competing instincts from two semi-autonomous selves within the individual. In their seminal behavioural public policy book *Nudge: Improving decisions about health, wealth and happiness*, Richard Thaler and Cass Sunstein playfully characterise these two internal influences as Mr Spock versus Homer Simpson, to convey the idea of a far-sighted ‘planner’ sub-self trying to rein in the short-sighted ‘doer’ (Thaler & Sunstein 2008, p.45).

The planner-doer characterisation originates with Thaler and Shefrin’s ‘economic theory of self control’ (1981), and posits that every individual is made up of these dual sub-selves representing competing roles and desires. On the one hand, there is the planner sub-self who cares about long run wellbeing and takes account of future payoffs; on the other hand, there is the doer sub-self who cares only for the present and prioritises immediate payoffs. The planner can set goals and worthy intentions, but it is the doer who takes action.

The stark contrast between the planner and doer – their divergent time horizons, priorities, and ability to act – creates the conditions for inner conflict and self-control problems. As a result of this internal tussle, an individual may deviate from a plan when the moment to take action approaches, generating an observable failure to follow through (Rogers et al. 2015). The issue is not that the

optimal behaviour is unknown; rather that when the time comes to undertake that behaviour, the individual reneges. So what can be done? When an individual anticipates their failure to follow through, can they adopt “tactics to command one’s own future performance” (Schelling 1984, p.2) and avoid procrastinating or quitting? The planner-doer framework points to commitment devices as a means to do exactly this.

## **2.2. *Commitment strategies to tackle future inconsistency***

A commitment device is any arrangement an individual voluntarily pursues to bind their future choices, adhere to the optimal behaviour, and deliver the desired outcomes (Bryan et al. 2010). Seen through the lens of the planner-doer framework, the commitment device is simply some strategy by which the far-sighted planner can align the short-sighted doer’s actions with the plan that delivers higher wellbeing in the long term. Like Odysseus tied to the mast, such commitment strategies lock down future choices in line with the planner’s preferences (Ashraf et al. 2006), to ensure that the right choices are made in the face of temptation to do otherwise. The need for such strategies is clear, with good intentions (set out by the planner) being far from sufficient to ensure good behaviour (by the doer) and ultimately the desired outcomes.

In reality, a commitment device can take the form of something as simple as a personal rule, such as a New Year’s Resolution (Schelling, 1984); it may involve a bolder statement of intent, such as registering for smoking cessation support from a community pharmacist and committing to weekly consultations and medical tests; and it may use money as well as reputation as leverage over one’s future actions, for example placing a bet on weight loss achieved (Burger & Lynham 2010). Such personal gambles are now facilitated through online commitment contracts, with one such

provider stickK.com citing over \$27 million staked since 2008 on a range of personal goals including losing weight and quitting smoking.<sup>1</sup> Commitment strategies have been formally incorporated into financial programmes. Brune et al (2015) discuss an illiquid savings account for farmers in Malawi, who often under-invest in fertiliser despite its high returns to agricultural production. A savings account was designed to help them lock down part of their harvest season profits until a later, pre-specified withdrawal date, reducing the chances of those savings being spent elsewhere ahead of the next growing season, under the influence of present bias and time inconsistency. The programme reported positive effects from the commitment savings account on future agricultural investment and sales, and on household spending.

---

<sup>1</sup> <http://www.stickk.com>, last accessed 27 October 2016.



### **3. TIME INCONSISTENCY AND HEALTH**

#### **3.1. Preventative health behaviours**

Like the investment and savings behaviours discussed above, preventative health behaviours require some investment today – time, money, effort, leisure or pleasure foregone – in order to reap the benefits many years from now. People make plans accordingly, but when the moment comes to take some preventative action (such as immunisation), or to adopt some preventative behaviour (such as reducing salt intake), people find themselves deviating from their plan. A gap opens up between intentions and actions (Rogers et al 2015), with adverse consequences for long run health and wellbeing. Health behaviours provide fertile ground for time inconsistency because they implicitly rely on intertemporal choices, with individuals being forced to trade-off benefits and costs over different time periods.

In a recent application of a formalised dual-system framework to understand dietary decisions, Ruhm demonstrates that food consumption would exceed the optimal level due to the potential for self-control problems between the ‘deliberative’ and ‘reflective’ systems; compared with a simpler case where the deliberative system alone managed decisions (2012, p.793). Time inconsistency in the planner-doer model is closely bound up with the concept of self-control, reflected in the (in)ability of the planner sub-self to effectively govern the doer sub-self. Self-control has been empirically linked both to obesity and to unhealthy exercise and dietary behaviours (Fan & Jin 2013). The literature reports other associations with preventative health behaviours, for example those with time-inconsistent preferences were less likely to put aside money as health insurance when offered simple savings devices, instead requiring stronger reputational commitment devices to do so (Dupas & Robinson 2013).

It is valid to arrive at the conclusion, then, that health behaviours are theoretically and empirically linked to time inconsistency, and provide a sound opportunity to test for commitment devices as an antidote to time inconsistency. Yet, planner-doer theory has not been explicitly applied to health behaviours to date. Ruhm (2012) comes closest to doing so with a two-system approach to analysing over-consumption of food. However, that research makes no mention of commitment devices as a potential solution, discussing instead the role of conventional policies such as taxation, elimination and information; further, the study does not directly test the theorised duality of the decision making process, but only infers plausibility given observational data on health indicators. Thus, there remains a gap in the literature for testing the predictions of planner-doer for health behaviour change, with a particular aim of isolating any causal effects from commitment devices.

### **3.2. *Application: Obesity***

Amongst the wide range of possible health behaviours that could offer these tests, this thesis focuses on weight loss behaviours and outcomes. The motivation for this choice is twofold, driven by the gaps in the scholarly debate, and by the UK policy context.

Firstly, the literature on commitment devices covers a range of health behaviours including smoking cessation (Giné et al. 2010; Halpern et al. 2015), malaria prevention (Tarozzi et al. 2009), and exercise (Prestwich et al. 2012; Royer et al. 2015). Prior research has also examined weight loss, but has not identified the interventions as commitment strategies necessitated by a dual-self decision making problem, precluding the opportunity to test or extend the planner-doer theoretical framework (Volpp et al. 2008; Nyer & Dellande 2010; John et al. 2011; Prestwich et al. 2012). Empirical findings

from these studies have focused largely on average treatment effects, with minimal examination of how commitment devices can work differently across sub-groups of a target population. In sum, there are important knowledge gaps in understanding how commitment devices can be brought to bear on tackling time inconsistency for healthy weight management.

Secondly, obesity is a top public health priority and a “widespread threat to health and wellbeing” in the UK (Department of Health 2011, p.5). The latest Health Survey for England reports that 62% of adults are now overweight or obese compared to 53% 20 years ago, while the proportion of people who have a normal body mass index (BMI) has fallen from 45% to 36% during that time (HSCIC 2014).

The rise in obesity is a biologically and psychologically complex health issue with many contributing factors. From a physiological perspective, evolution has left humans “geared to protect more strongly against weight loss than against weight gain” (Hill 2006, p.751). On the environmental side, the voracity of food marketing and availability of foods, combined with changes in technology and lifestyle has meant “weight gain is the inevitable – and largely involuntary – consequence of exposure to a modern lifestyle” (Butland et al. 2007, p.5). This is the basis for claims that society itself is growing more “obesogenic” (Costa-Font et al. 2013, p.2); with obesity reported to spread through social networks, possibly through shifting norms around the acceptability of being overweight (Christakis & Fowler 2007).

Although the Government’s 2011 Call to Action set itself a target of achieving a downward trend in the level of excess weight across adults by 2020 (Department of Health 2011), it is plausible this target will not be met despite the range of efforts undertaken, including changes to food retailing, unprecedented access to health

information, and traditional campaigns such as Change4Life complemented by digital health initiatives and wearable fitness devices. With advances in medicine and health education, and an ever-growing range of tools and apps at our fingertips, why is overweight and obesity the normal condition for British adults? Put simply, information is not enough to optimise individual decision-making (Downs et al. 2009).

### **3.3. *A role for behavioural public policy: anti-obesity nudges***

Obesity in the UK is justifiably characterized as a public health crisis, but the current arsenal of policies is insufficient. It is clear that tackling obesity is a highly complex challenge, requiring a holistic policy response to effect significant and sustainable change. While it is beyond the scope of this thesis to consider the full range of possible interventions, the focus is on time inconsistency as one of the behavioural drivers of overconsumption. Not only might behavioural policies offer more feasibility in the absence of political appetite for bans or taxes (Thaler & Sunstein 2003), they represent a unique opportunity to address the fundamental biases responsible for poor health choices. Against this backdrop, nudge theory has carved a valuable role in the academic and policy debate.

A ‘nudge’ is any policy instrument that aims to change “people’s behaviour in a predictable way without forbidding any options or significantly changing their economic incentives” (Thaler & Sunstein 2008, p.6). Commitment devices are one amongst a menu of nudges designed to tackle behavioural biases that contribute to obesity (Oliver & Ubel 2014), and are already widely employed in the weight loss sector. Informal commitment devices include strategies such as signing up to a running club to bind future choices about physical activity, or sharing weight loss goals with friends and family on social media. Commercial weight loss programmes incorporate

such reputational and financial commitment devices in the form of publicising weekly weight readings public to the weight loss group, and offering membership fees back if weight loss targets are met; and these commitment strategies have been piloted in NHS programmes (Relton et al. 2011).

The prevalence of commitment devices for weight management begs the question: do they really work, who would benefit most, and could they make a significant difference to policy efforts? Harnessing, and maximising, the potential benefits of commitment devices would contribute to a top health priority. Against this policy and scholarly backdrop, recognising the empirical and theoretical gaps as well as the increasing importance of commitment devices for health behaviour change, two research questions emerge:

**RQ1: Can commitment devices change health behaviours and promote weight loss?**

**RQ2: Do commitment devices work differently across different people?**

#### **4. INTENDED CONTRIBUTIONS**

In answering these research questions, the thesis sets out to make four broad contributions to the knowledge base on commitment devices, time inconsistency, and health behaviour change: through theory development, fresh empirical findings, innovative research design, and practical insights for policy makers.

##### **4.1. *Contributions to theory: a new Analytical Framework for health behaviour change and evidence on planner-doer modelling assumptions***

The thesis has ambitions to contribute to the theoretical understanding of the planner-doer framework, and the causal mechanisms governing the effects of commitment devices within this framework. An original analytical framework is put forward that builds upon and extends the insights from Thaler and Shefrin's planner-doer model. The model aims to make explicit how this dual-self approach applies to health behaviour change, and draws out six testable propositions. One amongst these is the first articulation in the literature of how commitment devices would be expected to exert heterogeneous effects on health behaviour across individuals. The model will consider three broad pathways: the design of the commitment device, individual characteristics, and individual actions to apply the commitment device as intended. Each pathway will be tested using novel measures and data, to further our knowledge both of what may work and how best to capture its effects.

#### **4.2. *New evidence on how commitment devices work***

The two field experiments are designed with a special focus on reputational commitment devices, which are identified as being relatively under-attended in the literature. The field experiments aim to deliver robust findings that offer a fresh comparison with those tested in the literature to date (Volpp et al. 2008; Nyer & Dellande 2010; John et al. 2011; Prestwich et al. 2012). Adherence to commitment devices also comes under the spotlight for the first time in empirical research, testing the assertion that being able to stay committed to the commitment device is a key challenge to their effectiveness (Fan & Jin 2013). The results will also illuminate for the first time the potential interaction effects between these heterogeneity pathways, shedding light on a complex nexus of causal factors that determine how effective a commitment device will be in practice.

#### **4.3. *Innovative research design***

A third set of contributions is expected from the mixed methods research design of the randomised controlled trials. A key distinction of this thesis compared with the published literature on commitment devices is the active combination of qualitative data and analysis to complement and enhance the quantitative data and statistical results. The field experiments make use of new quantitative and qualitative variables to operationalise the theoretical concept of sophistication; and with this data offer fresh insights on how sophistication interacts with the design of the commitment device to effect behaviour change. For the first time, qualitative evidence will be brought to bear on the question of how plausible the planner-doer characterisation is when thinking about time inconsistency, to probe the very foundations of the planner-doer framework.

#### **4.4. *Fresh insights for policy makers***

The findings will also contribute to enhancing the design and application of real world commitment devices. Field experiment results will highlight what role ‘soft’ commitment devices such as public pledges and contracts can play in existing public health programmes to leverage behaviour change and weight loss. Qualitative analysis is expected to highlight what features are most appealing, so that programmes can improve the use of commitment strategies. Sub-group analysis will shed new light on how health programmes can best target those who would benefit most, based on easily identifiable characteristics and prior experiences of participants, to maximize the effect of commitment devices on their health and behaviours.



## **5. ROADMAP OF THE DISSERTATION**

Chapter 2 critically reviews literature from behavioural economics, psychology and public health, which demonstrates a number of gaps in our understanding of commitment devices. For example, few studies investigate heterogeneity to understand who benefits most from commitment devices; and while a key challenge to understanding time inconsistent preferences is to determine the psychological processes underlying it (Wilkinson & Klaes 2012, p.293), this aspect is largely missing from the empirical work surveyed. Other issues identified by the Literature Review include understanding the welfare implications of commitment devices, and examining the grounds for privileging the planner sub-self over the doer sub-self, which is implicitly what a commitment device does. These wellbeing and normative questions – can a commitment device make someone happier? Should a commitment device be used to bind future choices? – are important, but beyond the scope of this thesis.

To begin addressing the theoretical gaps, Chapter 3 presents an original Analytical Framework, building on the planner-doe model by Thaler and Shefrin by spelling out – for the first time and in greater detail – the predictions for why commitment devices would be needed, how they would bring about health behaviour change, and the pathways along which heterogeneity in these effects would arise. Four heterogeneity pathways are brought into focus: three individual characteristics of sophistication (self-awareness), health motivation, and present bias, and individual adherence to the commitment device after it has been created. The framework also clarifies the propositions and concepts arising from the planner-doe theory – such as sophistication, demand for commitment devices, and adherence to the commitment device – so that they become more easily testable with empirical research in later chapters.

Empirical gaps identified in the literature review are best addressed through a field experiment methodology to isolate both average and heterogeneous causal effects. Chapter 4 details the design of two field experiments. The first field experiment is nested within an online weight management service called Food Monitor, and tests the causal effects of a financial and reputational commitment device on clients aiming to lose weight. The second experiment is located within a weight loss programme provided by Camden Council for local residents, and tests a reputational commitment device in the form of a commitment contract to oneself. The dissertation further aims to incorporate qualitative data and analysis to enrich interpretation of the statistical results, provide new opportunities to operationalise concepts such as adherence and sophistication, and gather fresh evidence on whether the planner-doer theoretical assumptions hold in real world behaviour.

Results and analysis are presented over chapters 5 (Food Monitor trial), 6 (Camden trial) and 7 (deeper qualitative analysis from both trials). Section 6 below summarises findings from the empirical analysis over chapters 5, 6 and 7. Chapter 8 draws together these results, reflects on the fit between empirical findings and the analytical model, and summarises contributions from the thesis.

## **6. FINDINGS AND KEY RESULTS**

This dissertation will generate robust evidence to answer the research questions, with three broad findings emerging: firstly, that the average treatment effects are more significant for specific health behaviours around self-monitoring and participation at group weight management programmes than they are for weight loss outcomes. Secondly, heterogeneity effects are uncovered across a number of pathways, including the pre-specified variables around individual trials adherence to the commitment device, and design features of the commitment device. Thirdly, novel evidence is found to support the assumptions of the planner-doer theory and the analytical framework underpinning the thesis. This section expands briefly on each of these broad findings.

### **6.1. *Average treatment effects on health behaviours and weight outcomes***

Both trials report that the commitment devices do not significantly increase average weight loss. This points to both the complexity of the weight loss process, and the subtlety of commitment device effects. Design of the commitment device, and the extent of the psychological tax exerted by it, are key features that determine how well it brings about behaviour change. However, commitment devices can promote positive health behaviours, with the commitment contract boosting participation in Camden's group weight loss programme. There remains a useful role for them in a public health setting, particularly where the intervention relies on people returning week after week to fully benefit from it. Indeed, stronger adherence to a medium term behaviour change programme is the main advantage arising from the commitment contract.

In answer to research question 1, then, the thesis finds that commitment devices are effective in promoting health behaviour

change, but do not exhibit significant, positive effects on weight loss outcomes. Unexpectedly, the Food Monitor trial finds a negative average treatment effect of the reputational commitment device. The idea that more commitment is unambiguously supportive of more behaviour change is refuted with this result; instead, the results suggest, taking on more commitment will at some point start to generate diminishing marginal or even negative effects. The finding raises the prospect of ‘commitment saturation’ and the existence of thresholds beyond which commitment devices do not work as expected. While unexpected, the Food Monitor’s negative treatment effect corroborates other work that suggests ‘less is more’ in terms of reputational commitment devices (Verhoeven et al. 2013). The potential pitfall of such commitment saturation implies a need to carefully design and target these interventions at those who would benefit most rather than view them as a universally applicable solution.

## **6.2. *Heterogeneous treatment effects of commitment devices***

In answer to research question 2, and of particular importance given the finding that design and targeting matter, commitment devices are found to have markedly heterogeneous effects on behaviour change across sub-groups. The commitment contract increased participation amongst participants who were identified as being more sophisticated, meaning they were more self-aware of their propensity for time inconsistency. The financial commitment device was associated with higher weight loss amongst those exhibiting a stronger degree of present bias compared to those who experienced both financial and reputational commitment. Again relating to the idea of commitment overload, the finding suggests that those with a greater propensity to overweight present gains are least likely to respond positively to ever-stronger degrees of commitment.

As predicted by the analytical framework, participants who adhered more faithfully to the commitment device were more likely to experience behaviour changes. The commitment contract raised participation in the weight loss programme for those who were more likely to have a short-termist outlook to their health, in other words those who may have exhibited stronger present bias and a higher propensity for time inconsistency. Underlying health attitudes were also expected to affect how well a commitment device worked, and this was confirmed in both trials.

### **6.3. *Development of the Planner-Doer theory for health behaviour change***

Finally, the Camden study adds new evidence of the key concepts underpinning the planner-doer framework: sophistication, and demand for personal commitment strategies to address the anticipated failure of self-improvement through willpower alone. While the commitment device designed for the Camden experiment was not suited to all participants, qualitative analysis established a number of ways in which people devised their own commitment strategies, further highlighting the scope for co-creation of commitment devices tailored to the individual for maximum adherence and effectiveness.

## **7. SUMMARY**

This chapter has provided an overview of the dissertation, which seeks to answer two research questions: can commitment devices change behaviours, and do they work differently across people? Speaking to the behavioural public policy literature, the thesis will examine how commitment devices can be brought to bear in tackling the behavioural biases that lead to time inconsistency, with a specific application to promoting health behaviour change and weight loss amongst the overweight and obese.

As set out above, there are four intended contributions to the scholarly debate. Firstly it will contribute to theory, by articulating for the first time how the planner-doer framework can be used to understand health behaviours and how commitment devices may generate heterogeneous impacts across people. Secondly, results from two field experiments are expected to fill gaps in the empirical research, particularly on reputational commitment devices and heterogeneous treatment effects. A third contribution arises from a novel mixed methods design, which seeks to combine the rigour of field experiments for causal inference with a nuanced interpretation of results facilitated by qualitative analysis; qualitative methods will also be deployed to uncover new data and proxy variables for sophistication and adherence to a commitment device. Finally, the thesis also aims to provide new insights for policy makers and public health programmes.

The argument that will be made over the coming chapters is that commitment devices do have a role to play in public health programmes, in complementing and not replacing conventional policies. They should not be seen as a silver bullet, and the impact of reputational commitment devices in particular may be too subtle to make a marked difference to weight loss goals across the population of people with excess weight. Where commitment devices can have

their greatest impact is amongst a sub-population of the overweight and obese who need to externalise their commitment and seek accountability outside of themselves because of their propensity for short-termist health attitudes and present bias. Design of the commitment device also matters, and impact can be maximised through ensuring salience and creating the right level of psychological impetus to stick with a health goal over time. Chapter 2 takes up this argument by reviewing the knowledge base on time inconsistency and commitment devices.

---

BLANK PAGE



---

## **Chapter 2**

# **LITERATURE REVIEW: A Theory, Taxonomy, and Empirical Review of Commitment Devices**

---

## **1. INTRODUCTION**

For many people, the experience of setting out to establish healthy habits is swiftly accompanied by the realisation that their future self will not always follow the plan they make today. Such is the essence of the time inconsistency problem introduced in chapter 1: how to stay on track with a health goal where the costs of behaviour change are both immediate and substantial, while the benefits are delayed to an uncertain future.

Maintaining a healthy regimen of diet and exercise is a form of health investment, and an example of an inter-temporal choice where an individual accepts the upfront costs in pursuit of longer-term health benefits and longevity. For example, an individual may accept the hunger pangs while restricting their calorie intake, which is the hallmark of many prescribed diets, in order to reap the benefits of shedding excess weight. Making this choice on a sustained basis is often not easy; indeed for some people it is best characterised a struggle, where they find themselves surrounded by temptation to deviate from an optimal diet, and no shortage of opportunities to avoid active lifestyle choices in favour of more sedentary ones.

Commitment devices – strategies to influence future choices for the better – are expected to combat time inconsistency and support the achievement of health goals. They belong to wider set of nudge solutions in behavioural public policy, designed to address the biases that lead to poor decision making. But do they work in practice? And who might benefit most? These are the questions underpinning this thesis.

The following chapter lays the foundations for answering these questions through a careful and critical review of the literature, with three objectives: (i) to locate the thesis in the scholarly debate around time inconsistency; (ii) to organise a diverse array of real world

commitment devices into a clear framework for analysis; and (iii) summarise the body of empirical work on commitment devices for weight loss. In each section, the aim is to identify what is known, examine disagreements and gaps, and draw implications for the research design (chapter 4).

Section 2 begins by exploring the important, underlying concepts of present bias and self-control in explaining time inconsistency. It then outlines the seminal “planner-doer” model by Thaler and Shefrin (1981), which gives rise to a key prediction: pre-commitment strategies may serve as a potential solution to time inconsistency. Later works in the dual-self tradition derived the same prediction (Laibson 1997; O’ Donoghue & Rabin 1999; Bénabou & Pycia 2002; Fudenberg & Levine 2006) through different modelling approaches. It is argued that the planner-doer framework is a superior lens to view and analyse health behaviour change because it provides enhanced “psychological texture” (Thaler 2016, p.1592) and insight into the root cause of time inconsistency.

Section 3 draws together a diverse array of real world commitment devices and explains how they aim to effect behaviour change, drawing on scholarly contributions from health psychology and behavioural economics. It proposes a taxonomy of commitment devices in the health sector that support weight loss, highlighting their distinct design features.

Section 4 appraises the available evidence base on commitment devices for behaviour change, as they relate to the research questions. Four arguments emerge. Firstly, there is a weak consensus that commitment devices can promote weight loss. While the number of studies addressing this question has grown in recent years, not all offer causal inference. Those that do suggest mixed evidence on the size and sustainability of the treatment effect. Secondly, much of the literature focuses on average treatment effects,

with far less attention for heterogeneity across sub-groups. Where heterogeneous treatment effects are examined, there is rarely a theoretical basis for sub-group selection. Thirdly, the review finds little evidence that studies examining commitment devices focus on the psychological processes – the precise causal mechanisms – underpinning behaviour change, and very rarely have studies attempted to apply qualitative methods to do so. Finally, the review concludes that the low-intensity commitment devices identified in section 3 are relatively under-researched despite claims that commitment contracts are “a way to health” (Halpern et al. 2012). Section 5 concludes by highlighting how the thesis will make a significant contribution to the literature by addressing these gaps in the knowledge base.

## **2. PLANNER-DOER THEORY OF TIME INCONSISTENCY**

### **2.1. Present bias and self-control**

One explanation for time inconsistency points to the tendency for people to weight the present more strongly than the future when faced with a trade-off between two periods of time: they are “present-biased”. This definition was offered most recently by O’ Donoghue and Rabin (1999, p.103), but reflects a longstanding observation amongst economics scholars that even identical payoffs available now and later are not valued equally by the individual. Pigou observed that “everybody prefers present pleasures or satisfactions...to future pleasures or satisfactions”, and blamed this trait on our “defective telescopic faculties”, which entailed that “we therefore see future pleasures, as it were, on a diminished scale” (Pigou 1932).

As argued by O’ Donoghue and Rabin, if an individual is impatient between a reward available now or later, but relatively patient when it comes to choosing between a reward available later or even later, time inconsistency will arise because time passes and ‘later’ becomes ‘now’. Despite the patient strategy the individual originally set out, at the later time the individual grows more impatient for the reward, and changes strategy. Returning to the intertemporal health behaviour choice, if the patient strategy involves eating bland low-salt food to tackle high blood pressure, while the impatient strategy is to continue eating saltier food that tastes good, the individual may decide that tomorrow the low-salt diet begins (patient later), but once tomorrow arrives the individual reneges (impatient now), accepting the likely health costs that brings.

Faced with the immediate choice of enjoying some payoff now or later, a person will tend to choose immediate gratification, unless through force of will we choose to delay the rewards. In other words, people will exhibit some degree of present bias – unless they exercise

self-control, willpower, or “self-command...by which we restrain our present appetites” (Smith 1790). Borrowing from psychology, self-control is “the ability to override or change one’s inner responses, as well as to interrupt undesired behavioural tendencies (such as impulses) and refrain from acting on them” (Tangney et al. 2004, p.274). Self-control defined in this way is a centrally important concept to this thesis, indeed time inconsistency can be understood as “a manifestation of self-control problems” (Wilkinson & Klaes 2012, p.290).

## **2.2. *Behavioural biases lead to sub-optimal health choices***

The relevance of present bias and self-control for health behaviours is clear. Many health choices are intertemporal, with the benefits of a higher quality of life 10 or 15 years down the line predicated on maintaining preventative behaviours and rationing indulgence at the present time – but doing so is not a trivial task. Limited self-control has been empirically linked to poor health outcomes. Fan and Jin report that “a lack of self control capability is associated with poor eating and exercise behaviours, as well as an increase in obesity risk and BMI” (2013, p.18). Cavaliere et al find that the probability of being “overweight or obese increases when consumers are less future-concerned”, and conversely a health BMI is “associated with a high orientation to the future” (2014, p.135). Maintaining sound health behaviours for long-term health requires a strategic deferral of rewards, which is challenging in the context of present bias and limited self-control, giving rise to time inconsistency. So what is to be done?

### **2.3. *Fixing the inconsistent: pre-commitment as a strategy for self-control***

Strotz provided the first formal treatment of time inconsistency. He considered it a puzzle because it suggested people changed their preferences without experiencing any new information that would prompt such a change. Assuming that agents recognised and regretted their inconsistency, Strotz predicted they would either pre-commit themselves, or develop a different plan that would be more feasible for their future self, termed 'consistent planning' (1955, p.165). In their seminal article, "An Economic Theory of Self-Control", Thaler and Shefrin (1981) arrive at the same conclusion: that a commitment strategy may indeed help address time inconsistency, but pose the issue as a principal-agent problem, and delve deeper into the individual's self-control problems arising from their being both the principal and the agent. The logic of the model is presented here, and a more formal treatment is set out in chapter 3.

### **2.4. *Planner versus Doer explains inconsistent behaviour***

Thaler and Shefrin model individuals as having two types of competing sub-selves, a "planner" and a series of "doers", who are locked in a principal-agent relationship over time. The planner is concerned with lifetime utility. The doer is concerned only with maximising utility in the current time period. The doer can be understood as the 'you' who are making decisions that affect you right now, while the planner is the 'you' who is setting goals for the future.

The contrast between the decisions horizons of the myopic doer and the far-sighted planner sets the scene for an internal tussle on intertemporal decisions, where costs and benefits are experienced at different times. The doer will never seek to experience costs now, if

instead it could experience benefits now. The planner, however, recognises that some costs now could pave the way for higher rewards later. The myopia of the doer exaggerates the notion that humans tend to be present-biased, focused solely on the rewards available now, at the cost of any longer term thinking; there is no utility from delaying gratification for the doer sub-self. The doer represents the most impatient and short-sighted version of ourselves.

The doer has decision-making control over the current period. The planner does not make consumption decisions, but derives utility from the doer's consumption in any period. Because the planner cares for long term utility, in a situation where investment (some costly expenditure from current resources) is required in order to increase future returns, the planner will not be fully satisfied with the doer's preferred consumption patterns. Maximising the planner's utility depends crucially on whether he can control the doer's actions in some way, aligning those actions with the planner's preferences. The planner, therefore, requires a "psychic technology" (Thaler & Shefrin 1981, p.395) to influence the doer's behaviour. These psychic technologies are simply some strategy that binds the doer's actions to the planner's preferred behaviour and could take a number of forms, analogous to principal-agent strategies at the level of the firm.

These are broadly categorised as control measures those that allow the doer some degree of discretion, and those that set firm rules for the doer to comply with. They can also vary in the degree of formality, from rules of thumb (mental notes to oneself) to explicitly altering incentives (with rewards or punishments). This menu of commitment strategies has varying degrees of effort. If the costs of monitoring and persuasion under a more flexible system are high, the individual might instead resort to rules. A binding rule essentially eliminates the intertemporal choice and does not require any deliberation by the doer. It is therefore relatively low on psychic



costs, with the trade-off being freedom for the individual to choose at the current time.

The planner-doer framework provides an elegant representation of time inconsistency and successfully combines the concepts of present bias (implicitly) and self-control (explicitly) to enrich our understanding of why we may set a plan for the future, and when the future arrives we delay, procrastinate, or quit. It enjoys consensus on the underlying intuition. Schelling describes the individual “treating himself as though he were occasionally a servant who might misbehave” (1984, p.5). Where Thaler and Shefrin predict the planner’s efforts to foresee and align the doer’s actions, Schelling posits that “the straight self and the wayward self interact strategically” (1984, p.5). Self-regulation theorists Heatherton and Baumeister refer implicitly to internal disjunctions raising the need for self-regulation, some “effort on the part of an agent to alter its own responses” (2013, p.91). All concur that such internal jostling is a part of people’s decision-making.

Thaler and Shefrin (1981) and Strotz (1956) highlight self-awareness in understanding commitment devices. Only if someone is aware of their time inconsistency will they seek to employ some strategy to bring their actions in to line with their intentions. In other words, these individuals are “sophisticated”, to use the terminology of O’ Donoghue & Rabin (1999, p.108), who point out that most individuals will be somewhere on a spectrum between totally sophisticated and totally naïve. With partial naiveté a person may be aware of a future self-control problem but underestimate its magnitude. People who are sufficiently sophisticated might find it beneficial to use commitment devices. If people are naïve, however, then policies may want to make them more sophisticated, or incentivise commitment devices where they do not themselves recognise the need for them (Frederick et al. 2002).

## **2.5. *Extensions and additions to the scholarly debate: temptation models, game theory, hyperbolic discounting and neuroeconomics***

In the years following the publication of Thaler and Shefrin's planner-doer model, a number of academic contributions have lent further theoretical and empirical support to the basic framework of the dual-self model. These theories are briefly discussed below, to demonstrate two points. Firstly, that the sub-theories are all agreed on the likelihood of a commitment device bringing about a positive effect on desired health behaviour change, by reining in the short termist doer's actions; motivating a reasonable expectation of a positive treatment effect from a commitment device intervention. Secondly, that the intuition behind the dual-self conception of human decision-making can be formalised in game theoretic models and bolstered by physical evidence of dual pathways in the brain. The planner and the doer are not only useful as descriptors, but have a deeper appeal in providing a richer psychological explanation of inconsistent behaviour.

A recent branch of theory analyses the 'temptation' at the core of the intertemporal decision. Loewenstein argues that "visceral factors" can manipulate choices in the short run, causing time inconsistency (1996, p.272). Gul and Pesendorfer consider preference reversals, where individuals choose a larger-later reward, but subsequently switch to the smaller-earlier reward if it is available right now. The smaller-earlier reward is the temptation, and could be assigned in a health context to behaviours that allow for short-term indulgence or procrastination at the expense of longer-term health.

While Gul and Pesendorfer make no reference to Thaler and Shefrin, it is clear that their temptation model has a reasonable fit within the planner-doer framework. Firstly, the distinction between the decision maker's present and later selves accepts that an individual is divisible into sub-selves who may not agree over time.

Secondly, the idea that the planner takes advance action to restrict future options, making it easier for the later self to make the right choice, fits well with Thaler and Shefrin's spectrum of commitment strategies. In the Gul and Pesendorfer model, the commitment strategy is a simple one: eliminate the tempting option. This removes all discretionary power from the current self, and as such is a binding rule. Thirdly, Gul and Pesendorfer explicitly incorporate costs of self-control, which Thaler and Shefrin refer to as psychic costs, and provide a plausible explanation of why a costly commitment device would be employed: because the costs of exerting self-control may be even higher at a later date. This final implication is particularly useful in supporting the planner-doer theory's prediction that a commitment device will be taken up, despite the expected short-run disutility for the doer.

Game theory provides an alternative to the principal-agent framework for modelling strategic interaction between sub-selves. Fudenberg and Levine's model focuses on an individual "whose overall behaviour is determined by the interaction of two subsystems" (2006, p.1450) much like the internal tussle described by Thaler and Shefrin that predicts a demand for commitment devices. Their model is concerned with developing formal axioms to predict the Nash equilibrium between sub-selves, but Fudenberg and Levine's conclusions are consistent with Thaler and Shefrin (1981) and Gul and Pesendorfer (2004), on the costs of self control and the benefits of habit formation. Fudenberg and Levine argue that developing habits can lower the costs of self control, as they do not require as much discretionary decision making or repeated exercise of willpower; akin to Thaler and Shefrin's description of a shift from monitoring behaviour to a rule of thumb, which may become eventually a binding rule if the habit is held very strongly. Fudenberg and Levine further refer to commitment strategies in broad terms, as "different mechanisms through which to change the behaviour of

future short-run selves” (2006, p.1450), including through limiting the options available in the future.

Popularised by Laibson (1994, 1997), hyperbolic discounting (or beta-delta) models express how preferences reverse as the individual approaches the moment of decision-making. Individuals are said to exhibit a declining discount rate between now and the next period, but a constant discount rate thereafter; in other words the individual disproportionately underweights the future. The time inconsistency problem is especially (and possibly only) acute in the time period when a difficult choice has to be made, and a commitment strategy that binds the current self at that moment may help to overcome time inconsistency.

The most recent branch of the literature offers a triangulation of planner-doer theory using neuroscience (Mcclure et al. 2004, p.503; Alós-ferrer & Strack 2014, p.4). Akin to the dual selves in Thaler and Shefrin’s model, many other descriptors have been applied to the apparent dichotomy in brain systems, such as system 1 and system 2 (Kahneman 2003, p.1451), to distinguish between “two kinds of thinking, one that is intuitive and automatic, and another that is reflective and rational” (Thaler & Sunstein 2008, p.21), where the reasoning system monitors the intuitive. Physiological data on pupil dilation, skin conductance, and blood flow in the brain allows us to observe how humans respond to different behavioural stimuli. Mapping of brain activity suggests different parts of the brain are indeed associated with different decision-making processes, depending on whether the processes are controlled and based on deliberation, or if they are automatic and occur with little awareness or effort (Camerer et al. 2005, p.11).

A key implication from such evidence is that it is plausible to consider an individual as being composed of different sub-selves who prioritise different things and respond in different ways to external

cues. Camerer et al refer to the possibility of sub-selves competing with one another for mental resources and attention; but assert that it is an unfair contest and “automatic impressions will influence behaviour much of the time” (2005, p.21), in line with the planner-doer model’s prediction that the doer sub-self will dominate without a commitment strategy in place.

## **2.6. *Dual-self theories: metaphor or reality?***

Two-self theories may have flourished since Thaler and Shefrin’s original planner-doer model. But are dual-self models ultimately just “metaphors” for specific aspects of intertemporal choice (Frederick et al. 2002, p.376)? Neuroeconomics evidence is the strongest available in quantitative terms to support the idea that observed human behaviour is the outcome of some hitherto unobserved internal interactions; and the belief that individuals can plausibly be understood as being composed of planner-doer sub-selves. No qualitative evidence on this question has been discovered to date, and this represents an important gap in the evidence base supporting the planner-doer model.

Regardless of whether the dual-self framework is metaphor or reality, Fudenberg and Levine (2006, p.1449) assert a further advantage: dual-self models can provide “a unified explanation” for time inconsistency, accommodating the various models discussed above that pinpoint aspects of a dual-self inner landscape, such as present bias, hyperbolic discounting, the costliness of commitment strategies, and sophistication. The dual-self framework is useful as a “theoretical scaffolding” on which to develop models using further empirical insights (Alós-ferrer & Strack 2014, p.9).

The idea that such a wide range of time inconsistency models can be legitimately organised under one umbrella is indeed plausible. While these models vary in their emphasis on the precise mechanics

of time inconsistency – the waning of self-control, changing preferences as the time for action draws close – they have in common two features. Firstly, they share a common view of the individual as a complex being: a person may hold more than one set of preferences and decision-making facets, and conflicting desires will at times overtake one another. Secondly, they make a common prediction: in the absence of a commitment strategy, an individual is likely to renege on their initial plan at the moment where short term costs loom large and benefits recede into a distant future. Pre-commitment would therefore be an effective way to align the incentives of early and later selves.

The primary aim of this thesis is to test empirically this second prediction of this wider body of literature as it applies to health behaviours, by asking whether commitment devices can change weight loss outcomes.

## **2.7. *Why the planner-doer framework?***

Given the array of approaches to understand time inconsistency, there are two strong grounds for choosing the planner-doer model as the foundation for this dissertation. Firstly, the original model, while sharing much with the later models, has not been fully operationalized – and yet may have much to offer to the understanding of health behaviours in particular, where the internal tussle between current enjoyment and longer term health benefits is writ large.

Secondly, the planner-doer framework expands on the psychological foundations of time inconsistency and the behaviour it engenders, beyond simply describing it with discounting parameters as the hyperbolic discounting models do (Wilkinson & Klaes 2012, p.301). There have been no studies to my knowledge that aim to uncover evidence of planner-doer interactions within individuals who are attempting health behaviour changes. How this might be operationalized is addressed in the research design chapter, and empirical work presented in chapter 7; *why* this might be a useful exercise has been argued here. A stronger understanding of the psychological processes underpinning intertemporal choices will help combat time inconsistency, and give individuals the tools they need to successfully achieve their goals, and improve their health and wellbeing.

The next section provides an overview of how the theoretical commitment strategies discussed so far translate into practical aids for weight loss.

### **3. PRACTICAL APPLICATIONS OF COMMITMENT DEVICES**

#### **3.1. What is a commitment device?**

A commitment device is a voluntary arrangement that restricts or binds future choices, to “fulfil a plan for future behaviour that would otherwise be difficult owing to intra-personal conflict, stemming from, for example, a lack of self-control” (Bryan et al. 2010, p.671). It may take the form of an actual contract with a third party (Halpern et al, 2012), or it may be a more ad hoc arrangement created by individuals as a “promise to oneself” (Benabou & Tirole 2004, p.849).

A commitment device will change the costs of future choices. Bryan et al (2010) specify two identifying criteria of a commitment device, crucially placing weight on the individual’s underlying reason for employing the commitment device:

- i. The arrangement is primarily about changing the individual’s own behaviour, where they are the main risk to achieving their plan; and
- ii. The arrangement does not have a strategic motive in relation to other agents.

These criteria help distinguish commitment devices from other consumer behaviours that involve paying in advance or bulk buying as a means of locking in future choices, but do not have a behaviour change intention. Criterion (i) precludes arrangements made to reduce transaction costs (online shopping), avoid upward price shocks (investing in gold) or reserve a good in high demand (pre-ordering a bestseller); and straightforward exchanges of goods and services. It also emphasises the centrality of self-control in this discussion of commitment devices, with the *self* being the main source of time inconsistency. Criterion (ii) precludes arrangements that are set up with a strategic motive in relation to other actors,



including a variety of institutional, social, legal and political commitments (such as marriage, voting, campaigning, and lobbying behaviour).

These criteria also help distinguish between financial commitment devices and more conventional financial incentives. This dissertation is interested in arrangements driven by a desire to overcome short-term self-control problems and achieve a personal goal. A financial commitment is about shifting or raising the costs of achieving this goal, not just enjoying the payoff itself. In practice there will be a fine line between the two, which is why examining motivation in taking up a commitment device is important.

In practice, a range of gestures and arrangements fall under the definition of a commitment device. Schelling (1984) illuminated a number of commitment devices which were easily observed in everyday life such as denying options, asking for an external intervention, placing a bet on achieving a particular outcome, or threatening oneself with shame. There are various ways to organise the array of existing commitment devices. Thaler and Shefrin's (1981) typology differentiated between methods to alter incentives and to alter opportunities. They also distinguish between internal and external rules for aligning the planner and doer sub-selves, depending on whether the individual relies on some external assistance or applies personal, self-enforced rules. Bryan et al refer to 'soft' and 'hard' commitment devices, depending on whether they involve primarily a psychological or financial cost respectively (2010, p.672), but note that this distinction is not strictly binary, as financial commitments may also have psychological costs to failure attached.

### 3.2. A taxonomy of commitment devices for weight loss

Building on this literature, I identify four broad types of commitment device that meet the two identifying criteria above, and are at present being applied to weight loss behaviours: personal rules, public pledges, paying a voluntary premium, and deposit contracts. These commitment devices are elaborated below.

Table 1: Commitment devices for weight loss		
Commitment device	What's at stake?	Examples
Reputational	1. Personal rule or contract to oneself	Self-image New Year's Resolution Using a pre-written groceries list Plan of action (implementation intention)
	2. Public pledge	Public image Gym pledge board <u>GoodGym</u> runners club Social media posts
Financial	3. Paying a premium	Money (not retrievable) Slimming World/Weight Watchers Pre-paid gym membership Home delivery of calorie-controlled meals
	4. Deposit contract	Money (retrievable) <u>Weight Wins</u> <sup>2</sup> <u>stickK.com</u> Placing a bet

#### 3.2.1. Personal rule

The softest type of reputational commitment device is a personal rule, which can range from relatively informal rules of thumb and one-off resolutions ('no more chocolate today') to more active practices such as self-monitoring. Self-enforcement relies on there being a cost if the doer reneges on the long-term goal. This cost is theorised to lie in the potential damage to the individual's "self-reputation", so the doer's good behaviour is driven by a fear of setting

<sup>2</sup> An organization responsible for the 'Pounds for Pounds' programme with NHS Eastern and Coastal Kent in 2009-10.

bad precedents or losing faith in oneself (Benabou & Tirole 2004, p.849).

Perhaps the most common example of a personal rule is to create a plan of action. Implementation intentions emerged from the recognition of the gap between intention and action, which implies that framing a goal is often not enough. Gollwitzer (1999) highlights the importance of adding crucial detail by specifying the “when, where and how” of a plan, in the form of “when situation x arises, I will perform y” (1999, p.494). For example, when my host offers me a snack, I will ask for fruit. In its simplest form, then, an implementation intention that specifies goals and actions is itself a kind of commitment contract to oneself. The situational cue is understood to transform an intention into an act by automatizing a response ahead of time, prioritising the planner’s preferred choice over the doer’s; at its most effective, creating an “instant habit” (Gollwitzer 1999, p.499). A personal rule of this nature could be informal (a post-it note reminder on your desk) or formal (a signed agreement to respect the rules of the public library when you join); made out to yourself (a gym workout plan) or with others sharing the same goal (a plan to workout together).

A personal rule can also take other written forms. A pre-written grocery list can serve as a guide to help the doer (walking around the shopping aisles) stay on track with the planner’s dietary regimen (Au et al. 2013). A simple contract signed to oneself formalises a personal rule.<sup>3</sup> Websites such as [beeminder.com](http://beeminder.com) offer templates for commitment contracts to improve one’s health. The key

---

<sup>3</sup> For further clarity, the working definition of a commitment contract in this thesis is any arrangement that relies solely on non-financial elements and has some element of ingrained formality, perhaps by being written down or through some verbal agreement with another person. This is in contrast with Rogers et al (2014) who refer to commitment devices, commitment contracts and deposit contracts interchangeably. Their paper advocates for greater use of commitment strategies in public health more broadly; whereas the aim here is to offer precise distinctions between a range of commitment devices based on what costs are attached to future consumption and choices, to begin unpacking the causal mechanism underlying behaviour change.

issue is whether the source of the reputational commitment lies primarily with oneself, or with some external source. The latter is discussed next.

### **3.2.2. Public pledge**

A second type of reputational commitment device is a public pledge. Social psychologists have defined commitment as the “pledging or binding of the individual to behavioural acts” (Kiesler & Sakumura 1966, p.349). The commitment makes an act less changeable. The magnitude of the commitment is associated with how publicly it is stated, because of an individual’s desire to be consistent with what he has declared to others, and to avoid the personal and social disapproval that accompanies inconsistency. Parrott et al believe this lens helps explain “why the use of written and verbal pledges, promises, and contracts has increased compliance with various health care routines” (1998, p.392), and find that the act of making a public commitment as part of a skin cancer campaign led to more people undertaking prevention and detection behaviours.

Pledges are easily incorporated into the planner-doer model. Making a public pledge serves as a commitment device because it alters the individual’s incentives faced by the doer sub-self to encourage behaviour change. Relative to a personal rule, a public pledge may magnify the reputational costs, as the individual’s behaviour is open to wider scrutiny and disapproval. In the weight loss sector, public pledges are used in various guises. Public weigh-ins at a weight loss group serve to hold the individual to account against their stated target; attendance at an exercise club might be encouraged through a promise to a team; and pledge boards are a common feature at gyms. Further, pledges do not have to be very public to be effective. Recent studies have shown that even brief dialogue or written correspondence with a general practitioner can

encourage greater participation in NHS weight loss programmes, linked to a sense of commitment between patient and doctor (Allen et al. 2015; Aveyard et al. 2016). Making the commitment an external one – even if it is a commitment to just one other person – may inspire a sense of accountability that spurs on behaviour change.

### ***3.2.3. Paying a premium***

When internal rules are not powerful enough, individuals may choose to pay a premium to make an internal rule an external commitment. They may pay for a product or service that is expected to help them reach their goal, such as joining a professional weight loss club or private gym. The activities involved in these clubs are often not based on proprietary technology or a unique method. Yet individuals are willing to incur out of pocket costs for what are perceived as premium products that facilitate weight loss, despite their being free or cheaper alternatives. For example, an individual who pays to use the treadmill at the gym could probably have found cheaper or free alternatives by running outside in a public park. Online tools such as calorie counters and weight trackers are freely available online and through smartphone apps, and most individuals could find an alternative group of people to witness their progress through their personal and social networks. It is possible to devise nutritious recipes and meal plans using free resources rather than paying for healthy home catering or going on a detoxification retreat.

Where the main purpose of paying for products aimed at weight loss is behaviour change, and there is no evidence of strategic intent with relation to others, this commitment device meets both the identifying criteria set out by Bryan et al (2010). These actions can plausibly be interpreted as a form of financial commitment, where the out-of-pocket payment is an investment towards their behaviour change goal. The very nature of incurring an upfront cost is an attempt by the planner to rearrange the structure of benefits and

costs, increase salience of the goal, raise the incentive for the doer sub-self to behave in the desired way, and ensure a greater degree of consistency in working towards a goal<sup>4</sup>. The upfront payment is itself evidence of the planner locking in the future behaviour of the doer.

A potential challenge to this interpretation is that because the money is paid upfront and is not returnable, the money is no longer at stake. However, as the money has been spent, the individual arguably has only one way to ensure that money was not spent in vain: to stay on track with their goal and achieve the desired outcome. It is this motivation that is expected to effect behaviour change.

Premium payments have implicitly been interpreted as commitment devices in other research. DellaVigna and Malmendier (2006) ascribe a behaviour change motive to gym membership, and examine whether the upfront gym membership plans lead to increased gym usage. Tarozzi et al (2009) study the effects of malaria bednet retreatment amongst poor households who choose between different purchasing options. One option is to only purchase a bednet by itself, and the alternative is to also pay in advance for that bednet to be retreated, to ensure stronger anti-malaria protection. The latter is described as a contract that “financially ‘commits’ the person who chooses it to comply with future retreatments” (2009, p.232). These studies examine commitment devices that involve money being staked on an outcome, but do not promise a monetary payoff. Individuals make an upfront financial investment towards some desired behaviour change, that can only be recouped in health and wellbeing (not monetary) dividends if behaviour does change and some desired health outcome materialises – such as losing more weight, or avoiding disease.

---

<sup>4</sup> A limitation of this interpretation is where the product that is paid for has some value in the form of higher quality, customer service or brand differentiation which may motivate the purchase. In the case studies this thesis will examine, this is not considered to be significant enough to undermine the proposed interpretation.

### **3.2.4. Deposit contract**

A deposit contract is a financial commitment device with a cash payoff that only becomes available on achieving a certain goal. This is subtly distinct from a straightforward financial incentive or gamble. Unlike an external cash incentive, it involves the individual's own cash, which could be matched by some external agent. Secondly, it involves setting the money aside, then winning it back, so the net financial gain is zero. This form of commitment device evokes the prediction from Prospect Theory (Kahnemann and Tversky, 1979) that individuals are more averse to losing money than to gaining the same amount. In the planner-doer framework, the deposit contract explicitly redefines the incentive structure the doer faces.

Deposit contracts have been popularised through the website [www.stickK.com](http://www.stickK.com), which requires that volunteers signing up for commitment contracts pay upfront a sum of money which will be returned to them if they meet their goal, and donated to a charitable (or anti-charity) cause. This form of commitment device is being actively applied to health behaviours. Weight Wins offered cash rewards for meeting weight loss targets<sup>5</sup>. Conventional gambling also falls into the category of a deposit contract.<sup>6</sup> The US-based Healthy Wage ("win money for losing weight") challenges clients to double their money on losing 10% of their weight<sup>7</sup>, an approach mirrored by websites [fatbet](#) and [dietbet](#).

---

<sup>5</sup> The company has paid out more than £130,000 to successful clients since 2007. Website accessed 31 July 2013. The scheme was piloted by NHS Eastern and Coastal Kent (Relton et al. 2011).

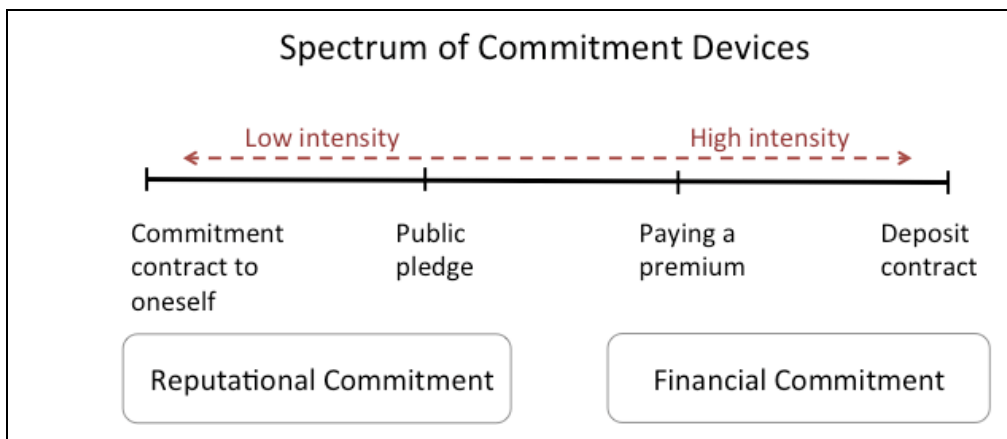
<sup>6</sup> Burger and Lynham reviewed 51 bets placed with William Hill, finding that 20% were successful. Notably, correspondence with participants indicated 70% viewed the bet as a commitment mechanism (2010, p.1163)

<sup>7</sup> Website accessed 31 July 2013.

### 3.3. Summary

Each of the commitment devices discussed here is a mechanism by which the individual's planner sub-self tries to rearrange incentives and increase influence over the doer sub-self's actions. Various underlying mechanisms have been theorised by economics, psychology and public health scholars, and these have been discussed here to further develop the causal mechanisms at work in the planner-doer framework. The analysis suggests a spectrum of commitment devices based on the intensity of perceived costs, as set out in Figure 1. The next section provides a critical review of the empirical evidence on commitment devices changing health behaviours.

Figure 1





#### **4. EMPIRICAL EVIDENCE ON COMMITMENT DEVICE EFFECTS ON HEALTH BEHAVIOURS AND WEIGHT LOSS OUTCOMES**

The literature holds a small but growing evidence base on commitment devices for weight loss and associated health behaviours on diet and exercise. The overarching message from published studies is that commitment devices can boost weight loss, but their influence diminishes over time, and is greater for financial commitment devices than reputational commitment devices. Despite their prevalence in weight management programmes, however, a number of gaps remain in our understanding of how commitment devices work. A critical review of this literature is presented below, focusing on available evidence of average and heterogeneous treatment effects.

##### **4.1. Evidence of causal effects of commitment devices: average treatment effects**

Table 2 compares studies that investigate the causal effect of commitment devices on weight loss health outcomes and related behaviours.<sup>8</sup> The diversity in field setting, timeframes, and treatments is immediately apparent, highlighting the limited scope for drawing general conclusions from the literature. However, it is possible to compare findings broadly using a benchmarking technique such as Cohen's  $d$ , applied below (Thalheimer & Cook 2002).<sup>9</sup> Key messages from this literature are discussed below.

---

<sup>8</sup> This selection is based on a careful review of the literature but is not a systematic review. Studies relying on non-experimental data are not expanded here.

<sup>9</sup> This method, originating in the psychological sciences literature, is increasingly popular as a means to compare average treatment effects across social science studies including meta analyses and systematic reviews (Crutzen 2010). Conventional benchmarking of effect sizes are small ( $d > 0.20$ ), medium ( $d > 0.50$ ) and large ( $d > 0.80$ )

#### ***4.1.1. Financial commitment devices can generate sizeable effects***

The largest effects are reported for financial commitment devices tested in a clinical setting, notably Volpp et al (2008) and John et al (2011). Both trials were based at the same American medical centre. Volpp et al examine the effects of a deposit contract for weight loss on 57 obese participants over 16 weeks. Some participants were allowed to create a deposit contract by staking up to \$3 per day on the final result, but only on those days that their weigh-in reading was progressing in line with their target weight. These stakes were voluntary and matched by the researchers (indicating an additional financial incentive, overlaid on the deposit contract). The control group were invited to a monthly weigh-in only. The average total payoff to the treatment group was \$378.50. This treatment group lost an average 14lb, with 47% achieving the 16lb weight loss goal. The control group in comparison lost an average of 3.9lb with 10.5% achieving the weight loss goal. The weight loss performance for those in the treatment group is significantly greater than the control group, however the design of the treatment (matched financial rewards) may have led to larger weight loss effects than for a deposit contract by itself.

A follow up study by John et al (2011) with 66 obese participants considered the effects of similar deposit contracts on weight loss over 32 weeks, and again found positive and statistically significant impacts: the deposit contract group lost on average 8.7 lbs over 8 months compared to the control group's 1.2lbs. Masked by these averages are a range of experiences in both groups, with many control group members losing weight, and some treatment group members gaining weight over the study – indicating considerable heterogeneity of impacts. As in the Volpp et al study, these effects were found to have dissipated 9 months later, when both groups were

back to their original weight, underscoring the limitations of financial commitment devices in delivering sustained health benefits.

**Table 2: Overview of literature on commitment devices for weight loss**

Study and intervention	N	Study setting	Timeframe	Effect size (Cohen's <i>d</i> )
Volpp et al (2008): Deposit contract for weight loss	51	Medical centre	16 weeks	1.07
Nyer and Dellande (2010): Public pledge for weight loss	211	Leisure centre	8, 16, and 24 weeks	8 wks: 0.52 24 wks: 0.42
John et al (2011): Deposit contracts for weight loss	66	Medical centre	32 weeks	Treat 1: 0.63 Treat 2: 0.49
Prestwich et al (2012): Personal rule for exercise	204	Office	1, 3 and 6 months	Exercise 1 mth: 0.63 Exercise 6 mths: 0.49 Weight 6 mths: 0.50
Royer et al (2015): Deposit contract for exercise	1000	Workplace gym	8 weeks	-
Chapman et al (2015): Personal rule for exercise	254	Office	6 weeks	0.06

*Notes: Volpp et al, John et al and Chapman et al report average outcomes and standard deviations by experimental group and these are used by the author to calculate Cohen's *d*. Nyer and Dellande study reports the short term (3-week) public commitment treatment effects at 8 weeks and 24 weeks. For the long term (6-week) public commitment, effect sizes are 0.36 and 0.44 respectively. Cohen's *d* calculation relies on *F* statistics reported in paper. Prestwich et al effect sizes taken from their reported Cohen's *d* values (p. 491). Data for calculating Cohen's *d* was not readily available for Royer et al. Author's calculations were carried out using the formulae set out in Appendix A2, and triangulated with online effect size calculator at <http://www.uccs.edu/~lbecker/>.*

While these papers offered path-breaking research designs and results, they are not without flaws. In a wider review of the literature on financial incentives for weight loss, Paloyo et al (2014) interpret deposit contracts as a negative financial incentive. They conclude that evidence of their effectiveness is inconclusive, “mostly due to the lack of methodological rigour and the conservative sample sizes”; and where a positive treatment effect is reported the literature has not found convincing explanations for “how it comes about and on what it depends” (2014, p.416). This critique is clearly applicable to Volpp et al (2008) and John et al (2011), where sample sizes were 59 and 66 respectively; and the deposit contract treatment effect could not be isolated from other field experiment design features such as increased contact with medical staff and increased self-monitoring, both of which could have contributed to weight loss and therefore lead to overstated treatment effects of the financial commitment device.

One study that appears to address these concerns tests the effects of a commitment device on exercise at a workplace gym. Royer et al (2015) ask whether a self-funded deposit contract can improve the effectiveness of a health incentive programme on exercise behaviours. All of the 1000 employees taking part in the field experiment were offered a one-month financial incentive to attend the workplace gym. Following this, 346 people were offered the opportunity to stake their own money on a pledge to continue using the gym regularly for a further 2 months. 12% (43 individuals) did so with an average deposit of \$58. Their ensuing gym usage patterns suggest that the commitment option improved the long-run effects of the incentive programme. However, like Volpp et al (2008), Royer et al (2015) are unable to fully account for the psychological processes driving the causal effect as per Paloyo et al’s earlier critique.

#### ***4.1.2. Reputational commitment devices generate modest effects***

Studies with less intense treatment offers – relying on reputational commitment and with no daily contact with researchers – generate smaller effect sizes, for example Nyer and Dellande (2010) and Prestwich et al (2012). In their experiment testing the effect of public pledges on weight loss, Nyer and Dellande recruited 211 women at a fitness centre in India, who were enrolled in a 16-week weight loss programme aiming to lose approximately 15 to 20lb. Two treatment groups displayed their targets on a club noticeboard; one group for 3 weeks (short term public commitment) and the other for 6 weeks (long term public commitment). After 16 weeks, average weight loss in the control group was 89% of the goals specified. In comparison, the short-term public commitment group achieved 97% of the targeted weight loss, and the long-term commitment group achieved 102% (i.e. exceeding the target).

Testing a personal rule with public commitment elements, Prestwich et al recruit 257 working adults and assign 3 different treatments: individual implementation intentions (individual writes and executes exercise plan alone), collaborative implementation intentions (individual and self-selected partner write joint plan for exercise), and partner strategy (individual partners up to fulfil the individual's implementation plan together). Writing and carrying out the implementation plan with a partner was reported to be the most effective strategy: this group exercised more and lost an average of 5.1kg after 6 months, relative to the control group's 0.6kg; and outperformed the other strategies which appeared not to show significant results.

### ***4.1.3. Personal rules may generate low or zero effect***

As intensity of the commitment falls, it is possible that the commitment device exerts no significant effect – but this is reported in only one of the studies in table 2. Chapman et al (2015) recruited 254 people to examine the effect of personal rules on exercise. The authors designed a treatment based on creating an exercise plan and keeping a written copy in a prominent location. This treatment was contrasted with a control group who were given information about the benefits of physical activity, which they too were prompted to print and store somewhere visible. While both groups were found to have increased their exercise behaviours, the difference on average between the groups was not significant. This study highlights the mixed evidence base on personal rules and commitments to oneself, but does not necessarily imply that self-commitment devices do not work overall: both groups registered an improvement, and with a less conservative comparison group (such as no health information), the commitment treatment may have registered a significant average treatment effect.

The Chapman et al study raises important questions about how effective self-reputational commitments and personal rules are, and whether they too have a role to play in public programmes. Wider evidence on the effect of personal action plans (implementation intentions) on health behaviours casts further doubt on whether milder reputational commitment devices can exert positive effects on health outcomes. Prestwich et al (2012) find that three of five studies evaluating implementation intentions bundled the treatment intervention with other measures, so did not allow for precise interpretation; one study showed positive impact but only for a sub-group of overweight and obese participants; and another showed no impact: mixed evidence at best. Given the scope for reputational commitment devices to be a less demanding feature to

incorporate in public health programmes, the question of whether they work and for whom is especially important.<sup>10</sup>

#### ***4.1.4. Short lived positive effects of commitment devices***

The effects of commitment devices appear to diminish over time. The lower effect size from the John et al (2011) study, despite applying a similar research design to Volpp et al (2008), may reflect the fact that outcomes were gathered at 32 weeks rather than the 16 weeks in Volpp et al. In Nyer and Dellande (2010) effect sizes appear to diminish over time for the short-term commitment group, but not so for the long-term commitment group. Royer et al (2015) report that gym usage across all experimental groups tailed off after an initial incentive period, although the commitment group sustained gym attendance for longer.

#### ***4.1.5. The potential for commitment overload is not reported***

None of the studies surveyed above comment on whether adding additional layers of commitment to the treatment affects outcomes, although the idea that stronger commitment devices tend to deliver larger treatment effects would be consistent with the idea that more commitment is more effective. A study by Verhoeven et al, however, rejects this intuition. Their research finds that a single personal rule to reduce unhealthy snacking was more effective than a plan of action containing multiple rules. The authors suggest that information overload (“an interference of information” (2013, p.352) rather than differences in initial motivation or commitment explain

---

<sup>10</sup> Financial commitment devices are arguably more demanding both for the programme (ethical and logistical issues around patients staking money on health outcomes), and demanding for the individual (higher stakes and with it plausibly higher risks to wellbeing). Reputational commitment devices are in principle more feasible, but the case for their effectiveness is weaker.

these results. Although not explicitly framed as an investigation of commitment devices, the study offers an indirect assessment of what happens when layers of personal rules (identified in section 3 as a type of reputational commitment device) are added to a particular strategy, and the counter-intuitive findings suggest this is an important avenue for further research.

#### **4.2. *Exercise as a proxy for health behaviours***

Table 2 highlights a further issue: the literature has tended to focus on exercise to proxy health behaviours, ignoring the equally important self-monitoring of health behaviours as an outcome in itself. Self-monitoring in this context can be defined as the “systematic observation, measurement, and recording of dietary intake, exercise, and weight” (Hutchesson et al. 2016, p.2). Numerous studies agree that self-monitoring is “the cornerstone of behavioural weight-loss treatment” (Peterson et al. 2014, p.1962), and within the planner-doer framework this behaviour can be understood as the means by which the planner sub-self tracks the actions of the doer and identifies the extent of divergence between the long-term optimum and how the individual actually behaves.

A second useful proxy for health behaviours is participation in a weight management programme, which has also been found to be associated with successful weight loss (Stubbs et al. 2015). If people find it difficult to maintain their attendance due to time inconsistency issues – they sign up to a medium-term course in the hope of participating weekly, but find that their interest dwindles over time – then a commitment device may help to bolster attendance rate, and in turn have a meaningful impact on weight loss. In the same vein, staying engaged with digital health tools is also associated with stronger weight loss performance (Johnson & Wardle 2011), and is another example of health behaviours that contribute to the goal of successful weight management. Future



research could usefully shift the discussion beyond exercise to examine the effect of commitment devices on other, relevant, health behaviours.

#### **4.3. *Evidence of causal effects of commitment devices: heterogeneous treatment effects***

Headline results from field experiments have tended to focus on the average effect of an intervention on the target population, but the literature on programme evaluation is increasingly concerned with the possibility that interventions can have different effects on different people, and in different contexts. In other words, far from there being a constant treatment effect that could be captured in a single parameter, research methods should allow for “unlimited heterogeneity” in treatment effects (Imbens & Wooldridge 2009, p.14).

Why might these sub-groups be of interest? From a theoretical perspective, it may help shed light on the causal mechanism, particularly in complex interventions where many different factors can interact to produce an outcome, and causal pathways are non-linear. From a policy perspective, this analysis allows for improved targeting of the intervention to those who may benefit most. In cases where the average treatment effect is close to zero and it appears a service or intervention is not warranted, heterogeneity analysis can help identify those groups for whom the treatment effect is high enough to justify targeted intervention. In other words, heterogeneity analysis provides more granular results and a potentially deeper understanding of cause and effect.

### ***4.3.1. Heterogeneity is under-attended in current research***

Despite the theoretical and practical utility of understanding heterogeneous treatment effects, very little has been established by theorists on whether commitment devices work better for some than others. Thaler and Shefrin (1981) briefly discuss how impatience and time preference may vary with age and, less convincingly, with social class, concluding with the need for further empirical work. Bryan et al (2010) distinguish individuals along a naïvete spectrum, concluding that the question of whether commitment devices only work for the sophisticated and partially naïve has yet to be answered. Table 3 summarises empirical heterogeneity analysis undertaken in the commitment devices literature, highlighting that it is relatively underdeveloped even when the field is broadened to other health behaviours such as smoking.

Giné et al (2010) refer briefly to heterogeneous treatment effects in their field experiment testing commitment devices for smoking cessation, but confuse the issue of heterogeneous treatment effects with a heterogeneous subject pool. The authors cite the 11% take-up rate as evidence of a diverse population with differing views on the treatment and its likely effectiveness, as well as varying appetite for risking their own money on achieving non-smoker status. The low take-up of the deposit contract is taken as evidence of “different consumer types”(Giné et al. 2010, p.229); and while this statement is plausible, it is less convincing that “the 11 percent take-up rate for [the commitment device] implies that our *average* treatment effects mask important heterogeneity” across these consumer types. The heterogeneity of treatment effects is thus assumed rather than explicitly tested by the authors, and the identification of the consumer types is not developed further.<sup>11</sup>

---

<sup>11</sup> Although the paper refers to treatment-on-the-treated estimates, a form of heterogeneity analysis that focuses on a sub-group of people who accepted the

### ***4.3.2. Heterogeneity may exist on demographic, behavioural and social characteristics***

Royer et al (2015) provide a more persuasive discussion of sub-group effects. They report that exercise can be successfully promoted through an incentive programme that is followed up by a commitment device to keep gym members coming back, with “stronger effects [relative to incentive-only treatment]... driven by the availability of the commitment contract” (Royer et al. 2015, p.80). Their motivation for investigating heterogeneous treatment effects is the diversity in take-up of the commitment device: women are significantly more likely to adopt the commitment device offered, as are those who are overweight or obese, and those who already report regular exercise habits.

Testing similar variables for heterogeneity in treatment effects, the authors find that initial levels of exercise, initial gym membership status, gender, and initial weight help explain why the interventions had varying degrees of success in changing exercise patterns amongst participants. While these findings are not pre-specified or derived from a specific theory, they highlight firstly the potential for commitment devices to work differently amongst different sub-groups; and secondly that this variation can be linked to the decision to adopt a commitment device. Compliance with a commitment device treatment is a potentially important indicator of whether and how much it will change health behaviour.

Turning to heterogeneous effects on weight loss outcomes, the evidence base is more mixed. Volpp et al (2008) report that “exploratory subgroup analyses revealed qualitatively similar patterns regardless of age, income or initial BMI” (Volpp et al. 2008, p.2635); although the wide range of outcomes across experimental groups indicates diverse weight loss experiences and unobserved

---

treatment, these effects go unreported with the authors citing an unsatisfactory instrumental variables strategy.

sources of heterogeneity in treatment effects.<sup>12</sup> Taking a different approach, Nyer and Dellande consider the role of a personality trait, ‘susceptibility to normative influence’ (SNI), in determining the effectiveness of a reputational commitment device amongst participants. The authors report that “subjects high in SNI were more likely to be affected by public commitment compared to those lower in SNI” (Nyer & Dellande 2010, p.10), because SNI is linked to the individual’s desire to comply with a publicly stated goal.

---

<sup>12</sup> For example, the deposit contract treatment group recorded a mean weight loss of 6.35 kg (14.0 lbs) with a standard deviation of 4.6 kg (10.2 lbs).

**Table 3: Overview of field experiments testing commitment devices on health behaviours**<sup>13</sup>

<i>Paper</i>	<i>Commitment device</i>	<i>Average effects</i>	<i>Sub-group analysis</i>	<i>Qualitative methods</i>
Volpp et al (2008)	Deposit contract attached to weight loss target.	Positive and significant – deposit contract leads to higher weight loss.	Yes – demographic characteristics. No significant effects.	Not reported.
John et al (2011)	Deposit contract attached to weight loss target.	Positive and significant – deposit contract leads to higher weight loss.	No.	Not reported.
Nyer and Dellande (2010)	Public pledge to lose weight	Positive and reported as significant – public commitment improves weight loss outcomes.	Yes – individuals with higher ‘susceptibility to norms’ were more affected by commitment.	Not reported.
Prestwich et al (2012)	Implementation intention for exercise plan to support weight loss	Positive – particularly for collaborative implementation intentions.	No.	Not reported.
Royer et al (2015)	Deposit contract attached to an exercise target.	Positive – commitment device prolongs the effect of a financial incentive for gym usage.	Yes –baseline level of exercise, gender, and initial weight. Those with lower baseline exercise benefited more from commitment.	Not reported.

<sup>13</sup> This exercise offers an overview of key papers, and is not a systematic review or meta-analysis of the literature.

<i>Paper</i>	<i>Commitment device</i>	<i>Average effects</i>	<i>Sub-group analysis</i>	<i>Qualitative methods</i>
Giné et al (2010)	Deposit contract attached to smoking cessation.	Positive – deposit contract encourages people to quit smoking.	No – speculative discussion of ‘consumer types’, possibly based on sophistication.	3 open-ended follow up interviews with participants.
Halpern et al (2015)	Deposit contract attached to smoking cessation.	Positive – quit rates higher for those with a deposit contract.	No.	Not reported.
Dupas and Robinson (2013)	Health savings product – rotating savings and credit association (ROSCA)	Positive – commitment devices encouraged higher savings for future health needs.	Yes – married women, exhibited present bias, and was a frequent donor to family networks. Some significant effects.	Not reported.
Chapman et al (2015)	Implementation intention for exercise plan	No significant difference relative to information control group, but improvement within group	Yes – based on whether individuals had maintained exercise routine or lapsed	Not reported.

#### **4.4. Summary**

The small but growing evidence base suggests that commitment devices can lead to better health outcomes for some individuals, and some published studies suggest significant positive average treatment effects on weight loss outcomes and improvements in exercise behaviour. These impacts however can be modest relative to the extent of weight management required, are unlikely to be sustained beyond the short term, and weaker for reputational commitment devices than financial commitment devices. Commitment devices then, like many nudges for health behaviours, are not a self-contained solution for tackling excess weight (Loewenstein et al. 2012). However they can be effective, and reputational commitment devices in particular may offer a cheap and easily administered intervention alongside other measures. The issue then is how to optimise the effects of the commitment device, and to answer this question it is important to understand for whom these interventions can work best.

Heterogeneity is an over-looked but promising line of enquiry. Commitment devices should not be expected to deliver uniform results for health behaviour change: individual characteristics around existing habits, motivation, gender, and personality traits can interact with a commitment device to boost the treatment effect, as demonstrated by the studies reviewed above. But the research base is often not explicit about the theoretical basis for sub-group effects.

This thesis will make a contribution to the field by bringing together insights from the behavioural economics and health psychology literatures to highlight a number of theoretically-driven heterogeneity pathways. In Chapter 3 these are broadly grouped into factors relating to the design of the commitment device, such as financial versus reputational commitment device; and factors relating to individual traits and actions, such as motivation for health behaviour change and compliance with the commitment device.

These potential sources of heterogeneous effects are placed within a planner-doer analytical framework and used to derive hypotheses and testable implications, and in Chapter 4 they are operationalized within the broader research design.



## **5. LITERATURE REVIEW CONCLUSIONS**

The planner-doer model put forward by Thaler and Shefrin (1981) offers a broad explanation as to why individuals do not change their behaviours, or if they have initiated that process why they do not carry these goals through to completion, in order to boost their own long run health and wellbeing. The task of delaying the current gain for a distant one is a challenging one and can often thwart the best of intentions. The planner-doer model explains why a person might seek external aids in the form of commitment devices to boost their self-control. The argument has been presented that demand does indeed exist for financial and reputational commitment devices to manage obesity and excess weight in the health sector, bearing out one of the central predictions of the model.

However, a number of important issues are relatively under-researched at present, as argued above. Few studies isolate a causal effect of commitment devices on health-seeking behaviour, and those that do make limited progress in evaluating reputational commitment devices; probing heterogeneous treatment effects; or unpacking and evidencing the internal tussles implied by the planner-doer framework. A critical review of the literature underscores that these features are often missing in the empirical work on commitment devices. The thesis aims to address these gaps, and sets out to make four contributions to the literature.

Firstly, the dissertation will assess two different types of reputational commitment devices: a personal commitment contract and a mild form of public pledge. Secondly, it will offer a more detailed and theoretically grounded exploration of heterogeneous impacts of commitment devices, covering individual traits and behavioural factors and as well as demographic characteristics to investigate for whom commitment devices may be most effective. Thirdly, a key challenge to understanding the anomaly of time

inconsistent preferences is to determine the psychological processes underlying it (Wilkinson & Klaes 2012, p.293), but this aspect is largely missing from the empirical work surveyed. None of the papers explicitly apply a planner-doer theoretical framework to the behaviour being analysed. In contrast, this research project will be grounded more thoroughly in dual-self theory by creating a planner-doer analytical framework as a basis for predicting health behaviour change and improved health outcomes.

The fourth contribution from this thesis is to design a research strategy that combines the rigour of field experimentation to uncover causal effect with nuance and a more granular understanding of context, through qualitative exploration of individuals' behaviour. The studies reviewed here apply experimental research designs with an exclusive focus on quantitative methods. None of the studies discussed above actively incorporate qualitative insights to interpret and contextualise treatment effects, or to characterise the relative influence of planner and doer sub-selves in achieving a plan over a period of time; arguably these are a weakness of the evidence base.

Despite the rich theoretical advances in understanding commitment devices, this chapter has argued that important questions persist. In a first step towards addressing these gaps, the next chapter presents an original analytical framework applying the planner-doer model to health behaviours for weight loss, in order to derive testable hypotheses for the empirical strategy.

---

---

## **Chapter 3**

### **ANALYTICAL FRAMEWORK:**

#### **Translating Planner-Doer Theory into Testable Propositions**

---

## **1. INTRODUCTION**

It was argued in chapter 2 that the Thaler and Shefrin (1981) model offers an elegant and intuitive explanation of intertemporal choice-making, which can be applied to health behaviours. But the theory has limitations. There is little direct evidence of a planner-doer tussle motivating the strategic demand for commitment devices as “anticipatory self-command” (Schelling 1984, p.1) to address self-control problems. Further, the literature review provided ample reason to disagree with Thaler and Shefrin’s assertion that “the most important applications are in the study of individual saving behavior” (1981, p.404). Health behaviours share much in common with savings behaviours due to the nature of the intertemporal choices they demand, and as such provide fertile ground for time inconsistency; and the implication of making poor health decisions is a clear threat to human welfare.

Recent field trials suggest commitment devices can promote health behaviour change, but the knowledge base remains small and results mixed across different commitment devices. Questions around when, and for whom, commitment devices work best are still open. To a large extent these are empirical questions; but to address them the theoretical framework needs to provide a clear set of propositions and implications that can be tested.

In a review of the literature examining time discounting and preferences, Frederick et al assert that few dual-self models “have been used to derive testable implications that go much beyond the intuitions that inspired them” (2002, p.376). The aim of this chapter is to close that gap, by advancing a new, formalised, interpretation of the original Thaler and Shefrin model, which generates six testable propositions on health behaviours. Three of these propositions relate to average and heterogeneous treatment effects that underpin the two central research questions. The model serves two important

purposes: it both illuminates existing hypotheses around the need for and effect of commitment devices on intertemporal choices; and offers new hypotheses to test on heterogeneity of effects across individuals.

Previous endeavours to formalise a two-self model include Benabou and Pycia (2002) and Fudenberg and Levine (2006). Both posit that decision problems can be understood as “a game between a sequence of short-run impulsive selves and a long-run patient self” (Fudenberg & Levine 2006, p.1449), and accordingly employ game theory to formalise the strategic interaction between dual sub-selves. In their paper, Benabou and Pycia draw Gul and Pesendorfer’s (2002) game theoretic model into a planner-doer framework, which serves to highlight the common ground amongst the different dual-self models. Fudenberg and Levine incorporate constructs such as sophistication and cognitive load, and the costs of self-control, into a game theory model.

More recently, Ruhm (2012) creates a dual decision framework to explain a broader health economics problem: the modern propensity for overeating. While it is not explicitly based on TS, it argues that “conflicts between the affective and deliberative systems may be particularly salient for eating decisions” (Ruhm 2012, p.783), and considers strategic behaviour to influence one’s own consumption decisions in the context of weak self-control. Ruhm’s work touches most closely on the aim of this dissertation, with the explicit focus on consumption of food and excess weight problems arising; but it does not incorporate commitment devices as a strategy employed by the planner to leverage power over the doer and achieve a health goal.

In common with the scholars mentioned here, I assume that individuals are made up of dual sub-selves engaged in an internal tussle over consumption choices. At the heart of this tussle is the

divergence between the doer's myopic horizon and the planner's far-sighted horizon. In contrast and addition to the existing literature, my framework explicitly models commitment devices as an instrument to change behaviours, and provides a tractable framework for analysing behaviour change relating to weight loss, that generates predictions around average and heterogeneous treatment effects of commitment devices.

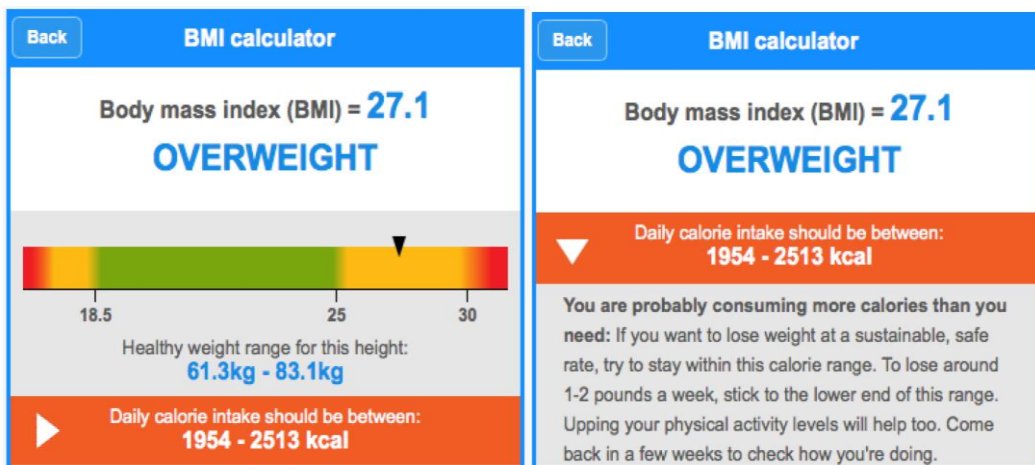
This chapter proceeds as follows. Section 2 lays out the basic tenets of my model, applying the planner-doer framework to weight loss and related health behaviours. It formalises the planner and doer utility functions and clarifies the need for a commitment device to help an individual make healthy choices. Section 3 presents six propositions arising from the model, and section 4 elaborates on the last of these, examining the variables that may determine how effective commitment devices are. Section 5 translates the model's predictions into hypotheses, laying the foundations for the research design in the next chapter.

## 2. THE MODEL

### 2.1. A familiar problem: struggling to lose weight

Consider the following scenario. Paul is 38 years old and works in an office. He readily admits he is fairly inactive, with only 2.5 hours of exercise a week in the form of gentle walking. He weighs 90kg, and at his height of 182cm his body mass index is 27.1. This means he is 'overweight'. At a routine health check he receives advice from his doctor to lose around 10kg at a safe pace, to return to a 'normal' BMI. When he gets home, he browses the NHS weight loss app.<sup>14</sup> It confirms that he weighs more than is good for his long-term health.

Figure 2: Body mass index (BMI) calculator



Paul recognises the benefits of eating a more nutritious diet with sensible portions, and taking more exercise. He makes a mental note to try and lose a few pounds over coming months. But day-to-day, he maintains the old habits, forgetting or resisting small opportunities to be more active or choose lower calorie alternatives to the usual meals and snacks. At his next check up, his weight has increased slightly. With the increase in his BMI, Paul is slowly but surely drifting towards being obese, and his nurse advises him that

<sup>14</sup> [NHS website](#) accessed October 27 2015.

without a change in his diet and exercise patterns he is at risk of early onset of diabetes and high blood pressure.

This hypothetical example easily fits the definition of time-inconsistent behaviour, and readily lends itself to analysis in a planner-doer framework. Paul understands what his health goal is (to lose about 10kg of weight), and the long-term benefits of doing so (preventing chronic ill health, and ensuring his wellbeing into his 40s and beyond). However, he finds himself carrying on in his normal lifestyle and his day-to-day efforts have not given rise to any improvement in his overall weight.

To understand the nature of the problem more precisely, the model below presents planner and doer sub-selves as separate entities, and their behaviour and decisions are observed over discrete time periods to time T.

## **2.2. The doer sub-self**

### **2.2.1. The doer's utility function**

The doer's utility function  $U_D$  is simply based on the level of consumption  $C_t$  in the same time period. Equation 1 formalises the idea that the doer sub-self has no time horizon, and no thought to future actions: only the current choice matters.

$$[1] \quad U_D = f(C_t)$$

Consumption here is measured as the net intake of food energy measured in calories (kcal). Net consumption is based on not only the intake of food, but also activity levels, which determine how many calories are exerted. Net food consumption as a function determined by two concepts: satiety ( $S$ ) and some costs to



consumption ( $K$ ), which can be seen as approximating the benefits and costs respectively of the doer's indulgence at time  $t$ .

Satiety here is a compound of factors such as taste, appetite, and feeling satiated. This is a highly simplified perspective of food consumption decisions, but sufficient for the purposes of this model and grounded in public health literature. Butland et al attempt to map out a complex system to explain obesity and refer to physiological factors, including the "level of primary appetite control in the brain" as a key variable (2007, p.87). Of the multiple determinants of food choices, Glanz et al (1998) find that taste and cost dominate, and these relate broadly to the  $S$  and  $K$  curves modelled here.

$$[2] \quad C_t = f(S_t, K_t),$$

$$[2a] \quad \text{such that } \frac{\partial S_t}{\partial C_t} > 0 \text{ and } \frac{\partial^2 S_t}{\partial C_t^2} < 0$$

$$[2b] \quad \text{and } K = mC \text{ so that } \frac{\partial K_t}{\partial C_t} = m, \text{ with } m > 0$$

### ***2.2.2. The doer's preferred consumption***

Equation 2a implies that increasing consumption is associated with increasing satiation, with diminishing marginal satiety as consumption rises. Eventually the curve flattens as the individual reaches a physical limit to further satiation, beyond which it is not possible to consume more. A high-activity individual will face the same basic experience of food consumption and the same shape of the  $S$  curve as a low-activity individual, so for simplicity exercise is not modelled explicitly but is incorporated implicitly in the measure of net consumption.

In the absence of a budget constraint (for example if food is available cheaply enough that prices do not act as a curb), the doer will choose to consume up to the satiation point, after which further consumption produces physical discomfort and the individual's self-limiting appetite will discourage further consumption.

More likely, the individual does take account of some costs of consumption. These costs have two elements. The first is of a traditional nature referring to the price of food. Assume it is possible to identify a total cost  $K$  for a consumption bundle implied at each level of  $C$ . To nuance this cost curve further, consider a second type of cost: a reputational element in the form of social norms, public image, or self-image.

The reputational cost element is the foundation of a reputational commitment device (Bryan et al, 2010), with the individual experiencing a cost to their reputation either to others or themselves if they do not try to behave in a manner that is consistent with their public statements (Benabou & Tirole 2004). Group-based weight loss programmes are real world examples that rely on this principle, where weekly progress is shared with a tutor and peers, and poor performance can lead to both private psychological costs and a public reputational cost. The field experiment by Nyer and Dellande (2010) which asked Indian gym-goers to make their weight loss target public to the gym community also relied on this reputational element to generate behavior change.

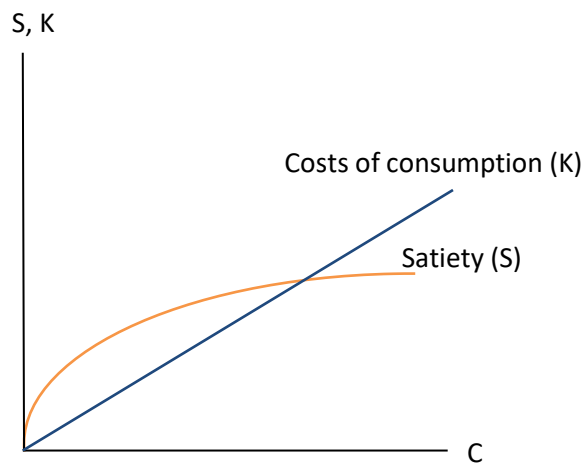
Equation 2b should be understood as containing both financial and reputational costs discussed above. They can be modelled in a linear relationship for simplicity, as set out in equation 2b.<sup>15</sup> As the individual consumes more, he faces a corresponding rise

---

<sup>15</sup> The cost curve could plausibly take a non-linear form even if monetary costs increased at a constant rate; for example if reputational costs increased exponentially beyond a threshold level, when social norms or self-disappointment may accelerate with marginal consumption. The basic implications of the model do

in costs. Equations 2a and 2b are illustrated in Figure 3 below in the form of curves S and K.

Figure 3: Costs of consumption and satiety curves



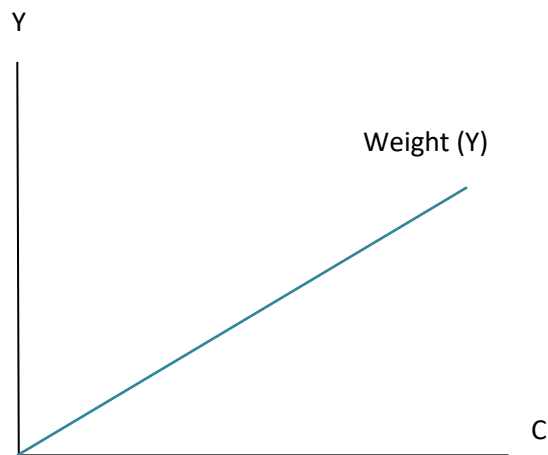
---

not change in this scenario, and for simplicity a linear form is assumed in this exposition.

### 2.2.3. Consumption choices and weight outcomes

Consumption bundles also correspond to health outcome  $Y$ , which is measured as the individual's weight. Again, a simplified linear relationship is assumed: as  $C$  rises, the individual experiences a corresponding gain in weight, represented by an increase in  $Y$ <sup>16</sup>. It can be easily seen that to reduce weight, consumption will need to be constrained; in line with the advice Paul received from the NHS weight loss app. Notably, this implies some short run disutility for the doer.

Figure 4: Consumption behaviour and weight loss outcomes

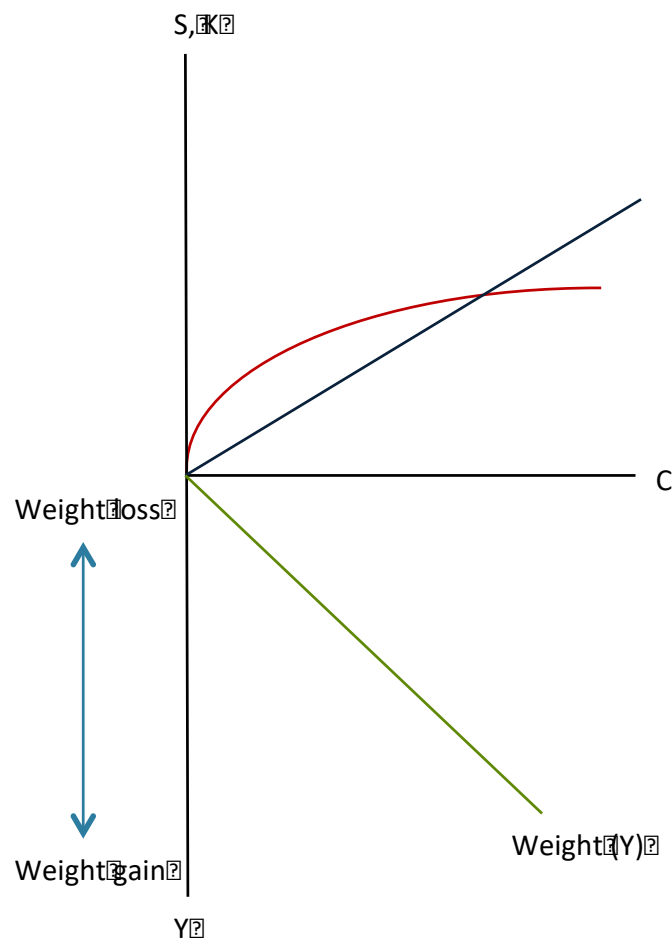


---

<sup>16</sup> In reality this curve is unlikely to be linear, since it is easier to lose weight at higher levels of initial weight, and earlier in the course of an attempt. A straight line used here for simplicity and does not alter the implications of the model.

Combining both graphs along the common axis C allows for a neat visual representation of how the costs and benefits of consumption (S and K) relate to weight loss outcomes (Y) via consumption (C). In figure 5, the weight axis has been inverted so that the further is Y from the origin, the larger its value.

Figure 5: The basic model



### **2.3. The planner sub-self**

#### **2.3.1. The planner's utility**

The planner's utility is based on a series of utility streams generated by successive doer sub-selves, and a function  $V$  that takes into account long-term health implications of the doer's consumption path. This formalises the notion that the planner wants to reach and maintain a healthy weight.

$$[3] \quad U_P = U_D + V$$

$$[4] \quad V = \sum_{n=t}^T \frac{(C_P^* - C_D^*)}{(1+\square)^n}$$

The long-term component of the planner's utility  $V$  is determined by how close actual consumption comes to the healthy benchmark represented by  $C_P^*$ . The numerator expresses the level of consumption that is in excess of planner's preferred consumption. This is discounted over time at rate  $r$ , because although the planner is far-sighted, he does not weight the distant future as much as the immediate future for the standard reasons (such as risk and uncertainty).

### **2.3.2. Excess consumption and planner disutility**

Under excess consumption,  $V$  takes a negative sign. The larger is excess consumption, the more negative is  $V$ , and the larger the downward pressure on the planner's overall utility. When  $V$  is consistently negative, this too will exert a stronger negative effect on the planner's utility.<sup>17</sup>

### **2.4. Health behaviours and health outcomes**

The model takes weight loss as the health outcome, and consumption as the main behavioural outcome from the doer's actions. This has the benefit of a robust evidence base linking net calorie intake to weight management.<sup>18</sup> However, it should be noted that net consumption is a portmanteau for various dietary and physical activity behaviours, going beyond how much is eaten and how much energy is expended; indeed a diverse set of underlying actions exist that can be used to shift net calorie intake, for example self-monitoring of diet and exercise (Boutelle et al. 1999; Johnson & Wardle 2011) and attendance at weight management programmes (Stubbs et al. 2015). These behaviours – attendance and self-monitoring in particular – will be used to test the impact of commitment devices in this thesis, as discussed in chapter 4. The significant point for this chapter is that such behaviours fit easily into the analytical framework presented so far: they help determine net

---

<sup>17</sup> This construction ignores the potential case of under-eating, where  $V$  would be positive even though long-term health effects could be negative, and focuses on cases where an overweight or obese individual to shed a safe amount of weight to reach a normal BMI. Note also there are 2 circumstances where  $V = 0$ . If the individual inherently has a high degree of self-control and does not tend to consume more than this optimal level, the planner and doer are in harmony; or, if the planner had no information regarding  $V$  to suggest any long term disadvantages from consumption level  $C_D$ , there would be divergence between  $U_P$  and  $U_D$ . These are not the scenarios being examined here; the former is a case of a time-consistent individual, and the latter can be thought of as Paul prior to his health check, where  $C_p^*$  is unknown.

<sup>18</sup> See literature on weight management as 'energy balance' (Spiegelman et al. 2001; Hill 2006; Hill et al. 2012).

consumption, which in turn generates the health outcome of weight loss.

## **2.5. Summary**

These are the basic mechanics of the planner-doer model. This section formalised both the divergence between the utility of the planner and the doer, and the reliance of the planner on the doer's actions for long-term utility maximisation. The framework is easily translated into an empirical strategy:  $C_P^*$  and  $Y^*$  can be observed as targets set by the individual, and  $C_D$  and  $Y$  can be observed through data collection on food intake, exercise and weight. The next section highlights the predictions of the model for demand, selection and effectiveness of commitment devices.



### **3. THREE PROPOSITIONS ABOUT COMMITMENT DEVICES: NEED, EFFECT, AND TAKE-UP**

In this section, the model is used to re-state, in more formal terms, the arguments put forward by Thaler and Shefrin regarding the need and effect of commitment devices on time-inconsistent individuals; with an application to health behaviours. Propositions 1 and 2 articulate expected health behaviours and outcomes in the absence and presence of a commitment device respectively. I further show in proposition 3 the model's implication that commitment devices will not be universally attractive, even to self-aware, time-inconsistent individuals, due to the inherent costs of applying control over the doer. For each proposition I consider the available empirical evidence.

#### **3.1. Proposition 1: The planner will identify a need for a commitment device**

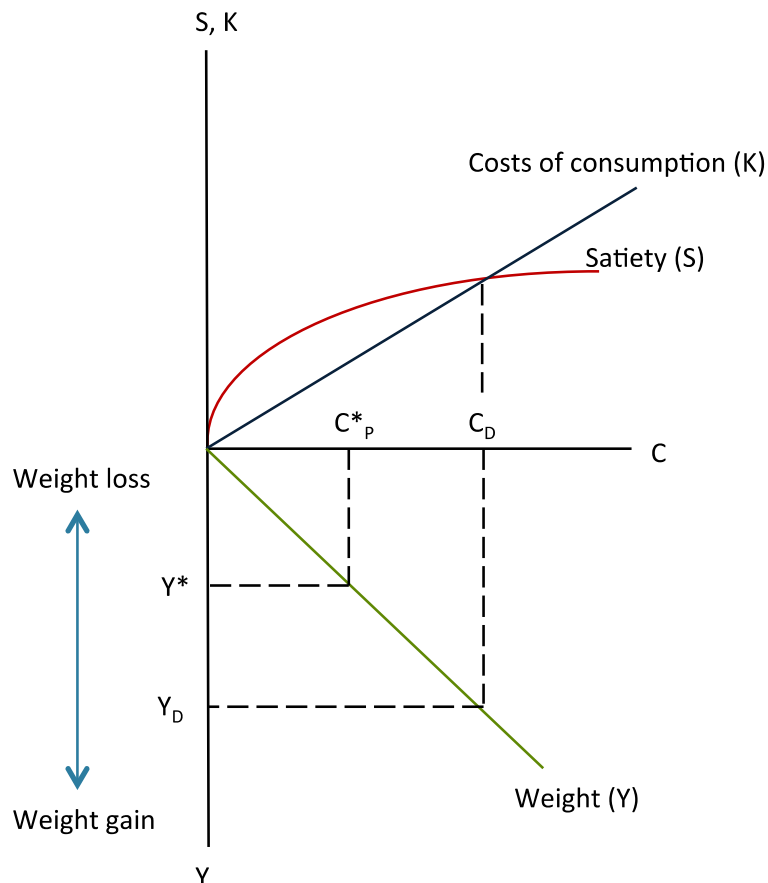
Left to his own devices, the doer sub-self will choose to consume as much as possible within the budget constraint. This myopic outlook considers only present satisfaction, but is a rational outcome for the doer in terms of maximising utility. This is represented in figure 5 below as the doer travelling up curve S until the budget constraint K is reached. At the point where S and K intersect, the doer reaches consumption level  $C_D$  and derives utility  $U_D$ , which corresponds to a weight of  $Y_D$ . This is the outcome that would be observed in the absence of any intervention by the planner sub-self; and can be likened to Paul maintaining diet and exercise behaviours that entrench his overweight status.

The planner seeks to maximise utility, and identifies an optimal level of consumption  $C_P^*$ . For the purposes of the model, the level of optimal consumption is exogenously identified, and may be derived through some external norm such as the recommended daily

net intake of 2000/2500 kcal for women and men respectively; or some customised calculation for the individual based on a goal to achieve a certain weight – in Paul’s case, the advice from the NHS website (Figure 1) was to reduce his daily kcal intake to somewhere in the range of 1950 to 2500 kcal. Having this figure in mind serves as a benchmark both for changing short run behaviours (lower net kcal intake through healthier diet and/or more exercise), and for achieving a target weight ( $Y^*$ ).

If  $C_p^* < C_D$ , this implies  $V < 0$  and therefore at any given time  $U_p^* < U_D$ ; meaning the planner sub-self is unsatisfied with the status quo. The individual has a higher weight than he would ideally like, since  $Y_D > Y^*$ . Empirical evidence supports this implication. Data from the US nutrition and health survey suggest that overweight and obese people are more likely to prefer to weigh less than they actually do (Ruhm 2012, p.789).

Figure 6: Planner identifies a need for a commitment device



This divergence between the doer's actual consumption  $C_D$  and the planner's preferred consumption  $C_P^*$  captures the individual's time inconsistency. In this situation, the planner would want to change the doer's behaviour and achieve an optimal consumption level that maximises long-term utility  $U_P$ . This is what predicts the planner's demand for a commitment device to lock in the doer's actions.

The existence of a market for commitment devices bears out this prediction. As reported in chapter 2, a formal market exists in the form of websites such as stickK.com; a recent browse of the site profiled 8 users who had collectively pledged \$9155 across 8 different commitments for weight loss<sup>19</sup>, with many more users generating contracts with and without financial stakes for weight loss.

### **3.2. *Proposition 2: A commitment device can change behaviour and health outcomes***

Assume that the satiety function is exogenously given, implying that the shape of the  $S$  curve cannot be quickly altered. Arguably, this is a reasonable assumption given the habitual nature of dietary and exercise routines, and the physiological factors underpinning appetite (Butland et al. 2007). The planner must find a way to shift the intersection between  $S$  and  $K$  to a lower level of consumption, aligning  $C_D$  with  $C_P^*$ , in order to achieve weight loss that brings about the desired weight  $Y^*$ . If a planner cannot change the  $S$  curve, he must attempt to alter the  $K$  curve. With exogenously given market value of a food bundle, to change the location of the  $S$ - $K$  intersection in Figure 5, the planner must seek to affect the overall costs of net consumption in other ways, and may turn to a commitment device to do so.

---

<sup>19</sup> From the stickK community journal for weight loss, website accessed 2 November 2015 11:55.

A commitment device generates some influence on the doer's actions, which can be represented by  $\theta$ : a "preference modification parameter" (Thaler & Shefrin 1981, p.395). This parameter is incorporated in Benabou and Pycia (2002) as a "reduced form representation of more concrete incentives (rewards, punishments) or rules put in place by the planner" (p. 422), but there is little beyond these broad definitions to explain how  $\theta$  arises and its psychological underpinnings.

I contend that  $\theta$  represents the commitment device as a 'tax' applied by the planner on the doer. It could be a monetary tax, in the case of a financial commitment device where money has been staked on achieving some health outcome. If the doer sub-self fails to act accordingly, this health outcome will not be secured, and the money is lost. The costs of excess consumption, in other words, have increased. Alternatively, a reputational commitment device applies a "psychological tax" (Miller & Prentice 2013, p.303) that affects self-respect, self-esteem or public image. As with an economic tax, if the individual does not act in a way that is consistent with their goal, the costs of excess consumption increase.

In both cases, the commitment device uses either money or reputation to transform the  $K$  curve and alter the interplay between doer and planner, shifting the consumption outcome towards the planner's desired level.  $\theta$  serves as a tax on the doer, adding penalties to excess consumption and in this way trying to bind future choices. A financial commitment device applies a monetary tax. Any strategy that succeeds in shifting consumption to a lower level would give rise to an immediate and direct cost from the individual feeling less sated, and this will have a negative impact on the doer's utility function.

Now, cost curve  $K'$  can be rewritten with the additive term  $\theta$  as follows:

$$[5] \quad K' = \theta + mC = \theta + K$$

Equation 5 highlights that the intended effect of a commitment device is to shift the cost curve upwards, which serves to alter the structure of costs relative to benefits of the doer's indulgence. This is consistent with, but distinct from, Benabou and Pycia (2002), who frame the problem in terms of internal resources required by the planner and doer in their tussle, with  $\theta$  in their model describing a load on the doer's resources which can nudge the odds of behaviour change in favour of the planner.

The new cost curve  $K'$  is a discontinuous function that begins from the planner's desired consumption level, as shown in Figure 6. Below the point  $C_p^*$ , the planner does not have a need for a commitment device as there is no problem of over-consumption, and the individual would face the original cost curve  $K$ . It is only where the term  $V$  in the planner's utility function is negative, because  $C_D > C_p^*$ , that the commitment device generates penalties.

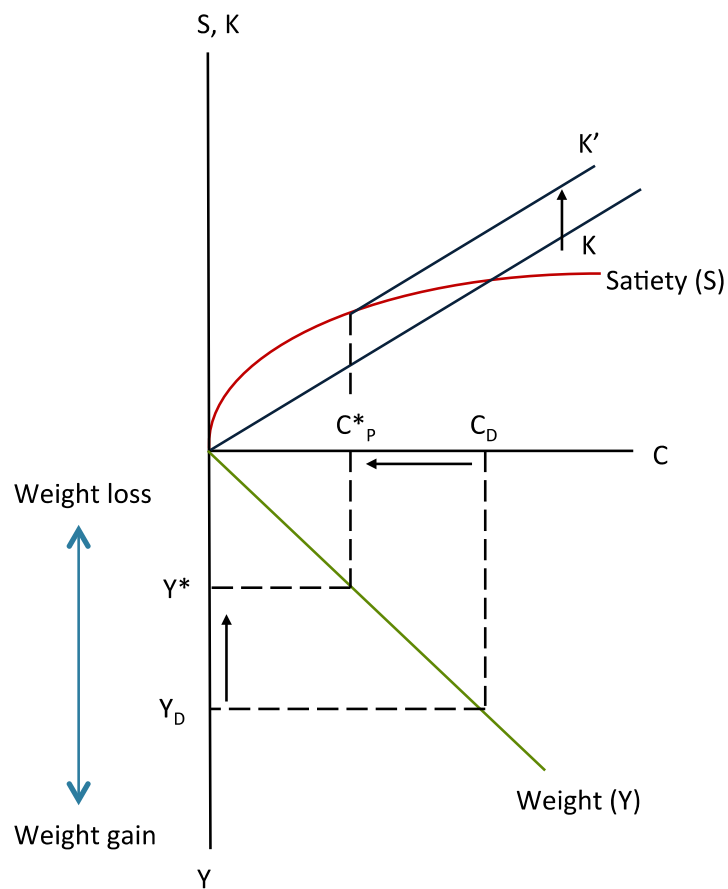
The doer faces the same utility function as in [1], but the consumption function has now changed to:

$$[6] \quad C_t = f(S_t, K'_t)$$

The mechanics of the model laid out so far lead to one clear prediction. Taking a value within sensible bounds,  $\theta$  shifts the individual upwards to the higher cost curve  $K'$ . Now, to enjoy the initial level of consumption  $C_D$ , the doer faces higher costs than before, and more than he would be willing to incur. As before, the doer chooses to consume up to the point where the  $S$  and  $K$  curves intersect, which yields a new intersection is at exactly  $C_p^*$ . Note this

consumption bundle is smaller than  $C_D$ . In Figure 6, this is precisely the behaviour change the planner desires, and corresponds to the desired change in weight over time to optimal weight  $Y^*$ . The model predicts that a commitment device can change behaviours and health outcomes; or, in the language field trials, that a commitment device intervention will yield positive treatment effects.

Figure 7: A commitment device changes behaviours



### **3.3. *Proposition 3: A commitment device will have selective, not universal, appeal***

The previous propositions expanded the two core predictions arising from Thaler and Shefrin (1981). The third proposition explains why commitment devices may not be an attractive option, despite their potential benefits on health behaviours. It sets out for the first time analytical justification for expecting low take-up of commitment devices; an issue which is not addressed explicitly by Thaler and Shefrin (1981), but is discussed implicitly in other dual-self theories (Gul & Pesendorfer 2004) and widely referenced in empirical studies of commitment devices (Giné et al. 2010; Royer et al. 2015).

Given propositions 1 and 2 above, why would any individual shy away from applying a commitment device, when it could solve the time inconsistency problem and improve health outcomes? Assuming the individual is self-aware, any sophisticated individual would be expected to reach for a commitment device to help achieve their health goals. The model explains this puzzle as follows. Any strategy that succeeds in shifting consumption to a lower level would give rise to an immediate and direct cost from the individual feeling less sated, and this will have a negative impact on the doer's utility function, as represented in equation 6 above. The model therefore predicts that a commitment device can change consumption behaviours; but, because the use of a commitment device is likely to lower the doer's welfare, there may be resistance to it.

These inherent costs of taking up a commitment device entail that it is unlikely to have universal appeal. A planner may only be willing to adopt a commitment device that stays above a certain threshold of disutility. It is plausible that initial motivation will need to be high for the planner to opt for one, for example if an individual believes they must act to rein in their excess consumption or they

may face very serious health implications. Further, the willingness to forego future freedoms may be a highly individualistic trait, with only some people ready to tie down their future actions.

These factors collectively predict that the appetite for adopting a commitment device may exist only amongst a rarefied sub-population: people who are aware of a need to change their behaviours, who accept they need an external self control aid, and who are willing to experience the disutility of having their choices bound and face the risk of incurring financial and reputational costs if they fail.

Examining acceptance rates when commitment devices are offered can test this proposition. Empirical evidence from two recent studies testing commitment devices for health behaviours indicate that take-up rates are often low, as predicted by the model. Only 11% of clients accepted a savings account that would return their money if they quit smoking after 6 months (Giné et al. 2010). A baseline variable capturing whether the ‘respondent smells like cigarettes’ was negatively correlated with take-up, indicating that those with a heavy smoking habit were less inclined to restrict their future choices; plausibly because the expected costs of restricting smoking behaviour would create too large a disutility in the short run (for the doer). In a separate study only 12% of participants accepted a deposit contract to exercise more, with evidence suggestive that “demand for commitment is highest among those with partial confidence in their ability to achieve their goals” (Royer et al. 2015, p.75). It is plausible that the planner makes a reasoned assessment of the costs,  $\theta$ , and the likelihood of success in deciding whether to implement a commitment strategy.

In summary, the model allows for commitment devices to have selective appeal to people, even amongst those who are sophisticated and recognise the benefits of binding their future



choices. A number of studies dissect the sources of heterogeneity in take-up decisions, and the proposition will not be tested formally in this dissertation.<sup>20</sup> Rather, the focus now shifts to heterogeneity in treatment effects, which is relatively under-researched and lacks a clear theoretical framework in Thaler and Shefrin (1981) and more recent dual-self models.

---

<sup>20</sup> Giné et al, 2010 and Royer et al, 2015 as discussed above. Also Dupas and Robinson (2013), and in studies of savings behaviour Ashraf et al (2006) and Brune et al (2015).

#### **4. THREE FURTHER PROPOSITIONS ON HETEROGENEOUS EFFECTS OF COMMITMENT DEVICES**

Proposition 2 indicates commitment devices can be expected to generate positive average treatment effects on health behaviours and outcome, but an important question is whether commitment devices will be more effective in some contexts than others.<sup>21</sup> The model as set out above is insufficient to answer this question, and neither is the issue developed in the original planner-doer framework. Thaler and Shefrin (1981) mention briefly the potential role of individual level factors, concluding: “our model stresses the theoretical admissibility of these variables. Only further empirical work can establish their relative explanatory power” (1981, p.404).

Recent empirical work has indeed begun to investigate the sub-groups for whom commitment devices can best bring about behaviour change (as discussed in the literature review), with a focus on individual characteristics such as age, gender, initial behaviours, and personality traits. However, the prior step of theorising why and how commitment devices may have different impacts is still lacking in the literature. The remainder of this section unpacks the issue of heterogeneity in treatment effects, and lays out a series of testable predictions that are developed further in the Research Design (chapter 4).

---

<sup>21</sup> Various studies report that weight loss can take a wide range of values, even after experiencing the same clinical and behavioural interventions, including for commitment device interventions (Volpp et al. 2008; John et al. 2011), indicating heterogeneity of treatment effects.

#### **4.1. *How does the model predict heterogeneity of treatment effects?***

Within the mechanics of the model laid out so far, there are numerous ways in which the effect of a commitment device will vary across individuals. For example, the precise nature of the link between food intake (C) and weight loss (Y) is inherently individualistic, based on factors ranging from genetics, childhood experiences, environmental factors, and lifestyle (Hill 2006). The nature of the satiety curve (S) will vary across people, as will the components of the cost curves (K and K') depending on physiological factors as well as personality and context. The thesis does not aim to identify and analyse an exhaustive set of genetic, social and physiological sources of variation in weight loss performance, but focuses on the potential sub-group effects arising from three specific issues that are argued to affect  $\theta$ : design features of the commitment device, how well the individual adheres to the commitment strategy, and individual traits.<sup>22</sup>

The fundamental reason for expecting heterogeneity lies in the model's assumption that  $\theta$ , the intensity of the commitment device, can take different values. It is this parameter that determines the degree to which the doer's utility is affected in the current period. As  $\theta$  grows, cost curve K is shifted further upwards, meeting the S curve earlier, and consumption falls: in this scenario the commitment device has been highly effective. The upper bound of  $\theta$  is determined by how much cost the individual is willing to experience, both in the take up of the commitment device and the experience of applying it. At lower levels of  $\theta$ , the cost curve shifts little and consumption falls very little: in this scenario, the commitment device is ineffective. In

---

<sup>22</sup> A broader discussion of heterogeneity would also ask whether commitment devices are more effective for some class of behaviours than others, such as savings versus studying; or within the health field for smoking versus exercise. These discussions are beyond the scope of this dissertation, which remains tightly focused on weight loss. Findings relating to other health behaviours are woven in to this chapter where the commitment devices for weight loss literature offers limited insights.

the extreme case where  $\theta = 0$  there is no effect at all (as seen from equation 5), and consumption remains at the original level  $C_D$ . So far this argument explains how heterogeneity can be accommodated within the model. What is lacking from this analysis, however, is an understanding of why  $\theta$  might take different values in different contexts.

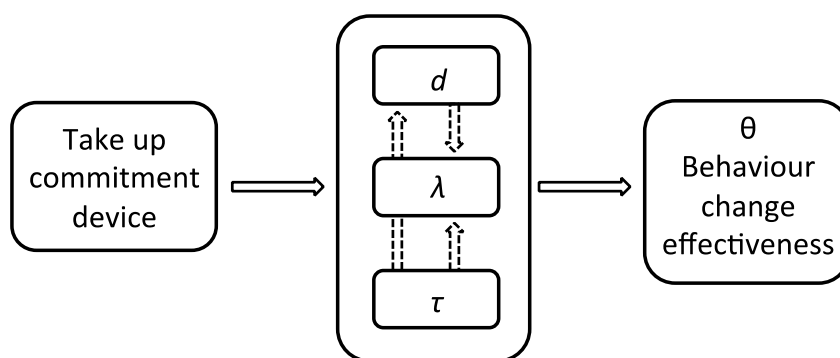
Heterogeneity of commitment device effects is proposed to emerge from three broad sources. Firstly, the design of the commitment device itself, which would be based on issues such as what kind of commitment device it is and how it is arranged. As set out in Figure 1 (chapter 2), different commitment devices can exert different intensity effects on behaviours. For example, is it a deposit contract or personal pledge? Is it a commitment strategy made to and governed by oneself, or is it public? Secondly, the intensity of the commitment device will depend on how faithfully it is embraced by the individual. Few commitment devices will entirely remove discretion for health behaviour decisions; and there will always be some element of choice as to whether or not to adhere to the commitment device. For a given commitment device and health behaviour, not all people will be equally committed to it (Fan & Jin 2013), leading to variation in the actual intensity of the treatment. Thirdly, individual traits also have the potential to affect how commitment devices effect behaviour change.

Taking account of these three proposed factors, Equation 7 decomposes the parameter  $\theta$  accordingly, with variable  $d$  representing the design features of the commitment device,  $\lambda$  the individual's fidelity to the commitment device, and  $\tau$  the individual characteristics of the person applying the commitment device:<sup>23</sup>

$$[7] \quad \theta = f(d, \lambda, \tau)$$

Much of the literature has focused on individual traits  $\tau$ , but this thesis argues that all three factors are critical in determining  $\theta$ . The design features and adherence to the commitment device determine the intensity of the psychological tax experienced by the individual seeking to change their behaviours; and individual traits can work in complex ways to affect  $\theta$  directly, as well as interacting with  $d$  and  $\lambda$ .<sup>24</sup> Figure 8 presents a stylised conception of equation 7 to highlight the likely nexus of interactions amongst the three heterogeneity pathways, which are expanded in the following sections to generate three further propositions.

Figure 8: Complex causal nexus of commitment device effects



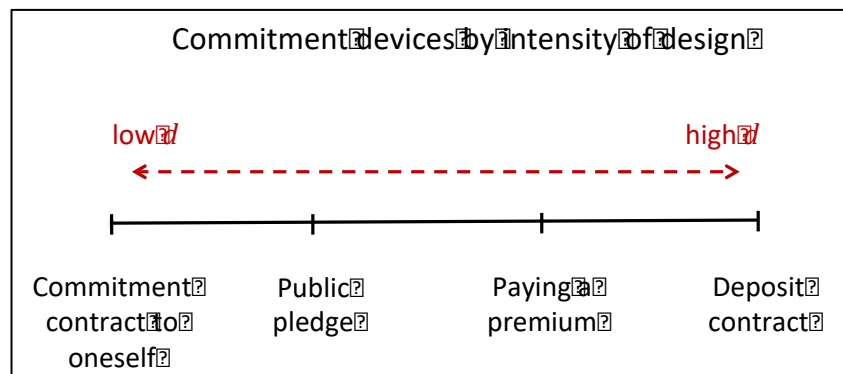
<sup>23</sup> The functional form in equation 7 is deliberately open, reflecting the reality of there being little basis to more tightly theorise the heterogeneity pathways.

<sup>24</sup> The model does not try to incorporate dynamics over time, which arguably could be relevant; this is beyond the scope of this dissertation, and would be better suited to future research once the propositions discussed here have had some empirical tested.

**4.2. Proposition 4: The design of the commitment device will determine its behaviour change effectiveness (d)**

Addressing the design features first, chapter 2 presented a typology of commitment devices in the weight loss sector and argued that they rely on different kinds of costs and mechanisms to effect behaviour change. The financial commitment devices were argued to have a more intense effect than reputational commitment devices; and amongst the latter the wider the public pledge the more intense the commitment. The type of commitment device directly affects  $\theta$  through  $d$ , which can take low or high values to reflect the intensity of the commitment.<sup>25</sup>

Figure 9



The model therefore predicts that for a given individual, such as Paul, a financial commitment device can be expected to have a larger treatment effect than a reputational commitment device; and amongst reputational commitment devices, one that makes a wider public pledge will exert a larger effect than one made to oneself. Empirical evidence provides some support for these predictions. Prestwich et al (2012) compare reputational commitment devices for raising physical activity, and find that exercise plans made with a partner are more effective compared to those made for and by an

<sup>25</sup> Figure 9 reproduces Figure 1 in chapter 2, highlighting the correspondence between variable  $d$  and the intensity of the commitment experienced along the spectrum of commitment device types.

individual. Chapter 4 returns to this prediction and explains how it will be woven into the research design for this dissertation.

**4.3. Proposition 5: Individual adherence to the commitment device will determine its behaviour change effectiveness ( $\lambda$ )**

The concept of fidelity to the commitment device, represented by the term  $\lambda$  in Equation 7, captures how well the individual applies the commitment device. Implicitly, this element assumes that a particular commitment device (holding  $d$  fixed) can be used in different ways by the same individual (holding  $\tau$  fixed) in different situations. It is rare that a commitment device would completely eliminate an option, and so there will always have to be some choice made to comply with the commitment device in order to bring about the behaviour change.

Adherence to the commitment device determines the effectiveness of that commitment device on behaviour change. Specifically, very high adherence would allow for higher levels of  $\theta$ . At the other extreme, where there is no adherence at all,  $\lambda$  would take the value zero; essentially negating the effect of the commitment device altogether because no matter how the commitment device is designed ( $d$ ) or what traits the individual has ( $\tau$ ),  $\theta$  would take the value zero.

The idea of adherence is perhaps open to the critique that  $\lambda$  is simply the residual, the error term that explains why a commitment device did or did not work when all other factors have been exhausted. However, this would be to miss the point about commitment devices requiring ongoing application for them to be

effective. It is rare that a commitment device can rely on a one-off commitment, and still be effective.<sup>26</sup>

The concept of adherence is supported by empirical evidence from a seminal study by DellaVigna and Malmendier, who uncover the puzzling phenomenon of “paying not to go to the gym” (2006, p.694). Members of a leisure centre in the US take up long term contracts and pay upfront, but then fail to utilise the services. Locking themselves in to a gym membership contract is arguably a commitment strategy to encourage attendance at the gym; but it is one that fails for a sizeable majority of clients because they do not adhere to it over time. This is one example of a broader issue with commitment devices: “individuals who understand and use a commitment device still face self-control problems, and often fail to carry out their well-orchestrated plans”, where the plan is the commitment strategy designed to bring about a wider health outcome (Fan & Jin 2013, p.18). Indeed, as pointed out by Fan and Jin, a key implication of commitment devices as a voluntary strategy is they require people to stay committed to the commitment device.

Despite its intuitive appeal, this prediction has not been tested in the literature to date. While visits to the gym in DellaVigna and Malmendier (2006) are an example of how to operationalise  $\lambda$  (the number of visits represent adherence), in that study the variable is not linked to health behaviours (exercise) or health outcomes (lower weight), which are only implicitly assumed to be the motivation behind taking up the gym membership. Chapter 4 sets out how this dissertation will aim to fill the gap by operationalizing  $\lambda$  in the context of the field experiments in this dissertation.

---

<sup>26</sup> An extreme example might be bariatric surgery – even this requires care and counselling to ensure behaviours change.



#### **4.4. Proposition 6: Individual traits will interact with the commitment device to determine its behaviour change effectiveness ( $\tau$ )**

The final variable in equation 7 is individual traits, denoted by the term  $\tau$ . Empirical evidence discussed in chapter 2 suggested that gender and personality traits such as susceptibility to normative influence have been found to interact with the commitment device treatment. While there are many possible traits that could lead to heterogeneous effects, three are selected below on the basis of signposts from theory and empirical evidence reviewed in chapter 2: sophistication, present bias, and health motivation. These are discussed in further detail below, with particular attention to how they may interact with design features and adherence in determining the overall intensity of the commitment device and behaviour change effectiveness.

##### **4.4.1. Self awareness and sophistication**

As discussed in chapter 2, the literature defines sophistication as an individual's ability to "foresee that they will have self-control problems in the future" (O' Donoghue & Rabin 1999, p.104), which can be understood to exist on a spectrum from not self-aware (naïve) to highly self-aware (sophisticated). In reality many individuals would occupy the centre ground, being partially sophisticated. This trait helps predicts whether an individual will demand a commitment device at all (proposition 1), because doing so requires that the individual recognises both the disparity between the planner's optimal consumption and the doer's actual consumption, and the fact that this gap will not be closed without some additional strategy to rein in the doer. It requires a degree of sophistication, in other words, to take up a commitment device. Going beyond this link with take-up of the commitment device, it is plausible that sophistication could be associated with how well the commitment device actually changes

behaviour, but this question remains unanswered in the literature (Bryan et al. 2010, p.694).

Considering the case where an individual is sophisticated enough to recognise the need for a commitment device, their level of sophistication can vary between more and less sophisticated. Two possible relationships can be theorised, both suggesting that sophistication interacts positively with the effectiveness of a commitment device in altering the doer's actions in line with the planner's goal.

Firstly, a more sophisticated individual may be better able to choose an appropriately designed commitment device, which will have enough of a desired effect ( $\theta$ ) to bring about behaviour change. A more naïve individual, on the other hand, may not get the design quite right; either choosing  $d$  such that  $\theta$  is low, and the commitment device appears ineffective, or incorrectly predicting their ability to maintain their fidelity to the commitment device over time. Sophistication, then, could operate through either terms  $d$  or  $\lambda$ .<sup>27</sup>

Regardless of the precise causal mechanism, the prediction from the model is clear: the commitment device is less effective for more naïve individuals. The difficulty, however, lies in testing this. How should sophistication be measured? Few studies have attempted to do so and there is no consensus on operationalizing these concepts. Ashraf et al (2006) highlight the value of understanding sophistication as an individual trait, but are unable to find a good operational measure, resorting instead to related but distinct measures of hyperbolic discounting (2006, p.667). Royer et al (2015) apply “partial proxies for an awareness of a time inconsistency problem” (2015, p.75), but fail to find an association with take-up of a deposit contract, and do not search for an association with outcomes.

---

<sup>27</sup> The research design in chapter 4 explains how qualitative methods will be brought to bear in unpacking some of these theorised interactions.

Chapter 4 sets out proxy measures that will be incorporated in to the field experiments in this dissertation. They aim to make a contribution to the literature by operationalizing and testing the relationship between sophistication and commitment device effectiveness using both quantitative and qualitative methods.

#### ***4.4.2. Short-termism and myopia***

Present bias was defined in chapter 2 as the propensity to weight present gains far more highly than later gains. Theory states that those who are present-biased are more likely to need commitment devices, but does not make a clear-cut prediction on the effect of present bias on how well commitment devices can work. Present bias is not necessarily a binary concept. Amongst those who are time inconsistent, there are degrees of short-termism, how impatient the individual is, which can be incorporated into the model within term  $\tau$ . It is plausible that a highly impatient person, someone who demonstrates a stronger degree of present bias, will not use a commitment device as effectively as someone who is relatively patient; possibly because they find it harder to apply the commitment device in a disciplined manner over time. In other words, as present bias increases, the value of  $\theta$  tends towards 0.

Following the logic of the planner-doer assumptions framed above, the degree of myopia exhibited for health behaviours in particular could be expected to determine how well the commitment device reins in the wayward actions of the doer. The more myopic, the greater the potential good a commitment device could do for an individual; but the less likely the individual is to actually adhere to the commitment strategy and so  $\theta$  is likely to be low. Time-inconsistent individuals who are less myopic, therefore, are argued to benefit more from a commitment device.

The language used to describe this heterogeneity pathway deliberately encompasses a variety of terms – present bias, myopia, and impatience – not to blunt the meaning, but to reflect the nuances of the underlying theoretical insight. The general proposition is that short-termism is a cause of time consistency, and may also determine how well an individual benefits from a commitment device to combat that time inconsistency. Short-termist attitudes, exemplified by the doer sub-self, have not been pinned down in the quantitative research on commitment devices, so how best to operationalise the broad idea of short-termism?

Discount rates can be used to identify degrees of present bias, but there are often methodological challenges to doing so, particularly in a field experiment setting where brief survey questions must be used to elicit discount rates and time preference. Ashraf et al (2006) make the first attempt to understand time preference reversals amongst their participants as a way of modelling both the take-up of a savings commitment device, and its effectiveness. Present-bias is found to have a positive but statistically insignificant interaction with the effectiveness of the savings commitment device tested in the study. But the authors also report that noise drives much of the survey responses on questions aiming to elicit hyperbolic time inconsistency, and where there is evidence of preference reversals these do not predict real behaviour. Chapter 4 discusses further how this concept can be operationalized to test for heterogeneous effects.

Like sophistication and time preference, however, the degree of myopia is a tricky concept to elicit and measure. On these issues, Thaler and Shefrin (1981) and more recent behavioural economics literature is silent, but the health psychology literature offers a measure of health attitudes that provides a proxy variable for short-termism, derived from the Healthy Foundations Segmentation model (HFS) model (Williams et al. 2011). A full explanation of the model's

origins and construction is available in the appendix (section A2). In summary, the model uses a series of questions measuring attitudes to health, and develops five motivational sub-groups as set out in Table 4 below.<sup>28</sup> Live for Today's, Unconfident Fatalists and Hedonistic Immortals are most likely to have short-termist views on their health, which could signal a dominant doer sub-self. These factors imply lower commitment device effectiveness. In contrast, Balanced Compensators and Health Conscious Realists at the higher end of the motivation spectrum are more likely to benefit from commitment devices.

**Table 4: Capturing myopia through health attitudes**

Profile	Beliefs and attitudes
<i>Relatively Myopic</i>	
Live For Today	Short-term view of life. Whatever they do is unlikely to have an impact on their health, so what's the point? Living a healthy lifestyle does not sound like fun; "most likely to be resistant to change" ((Williams et al. 2011, p.60).
Hedonistic Immortal	Feel good about themselves and want to get the most from life. Don't think they will get ill any time soon. Anything enjoyable like smoking and drinking can't be all bad.
Unconfident Fatalist	Negative view of their health, don't feel motivated to act, and think that they are more likely to get ill than others their age.
<i>Relatively Far-Sighted</i>	
Balanced Compensator	People who value their health and have a positive outlook. If they do something unhealthy, they will take steps to make up for it.
Health Conscious Realist	Take a longer-term view of life and prefer not to take risks. Feel good about themselves. "Motivated people who feel in control of their lives and health" (Williams et al. 2011, p.69).

<sup>28</sup> A segmentation approach uses a wider range of indicators to develop consumer sub-groups, in line with Giné et al's suggestion that different "consumer types" respond differently (2010, p.229) to commitment devices.

In summary, the second trait relates to short-termism in health attitudes. It can be operationalised through a time preference measure, attaching a monetary cost to a delayed hypothetical reward. An alternative is to focus on short-termist health attitudes specifically, capturing health myopia through a pre-tested health attitudes segmentation model. Although an individual may evolve from one segment to another over time, the measurement provides a valid snapshot of a person's health motivations and attitudes as a baseline variable. No studies on commitment devices have yet applied such a holistic approach to measuring health motivations, and the HFS instrument promises a richer analysis (relative to discount rates) of how individuals benefit differently from commitment devices based on their initial myopia. Applying two different operational measures of a single theoretical idea also allows for methodological innovation and testing.

#### 4.5. Summary

Design features of the commitment device ( $d$ ), individual characteristics ( $\tau$ ) and adherence ( $\lambda$ ) are proposed to give rise to differences in the extent of behaviour change brought about by commitment devices. Of the full range of heterogeneity pathways that could be identified, these are selected on the basis of theoretical evidence indicating their importance for commitment device effects, as well as their ability to add to the existing knowledge base. Table 5 summarises the individual traits that will be analysed in further detail.

---

**Table 5: Factors determining commitment device effectiveness ( $\theta$ )**

---

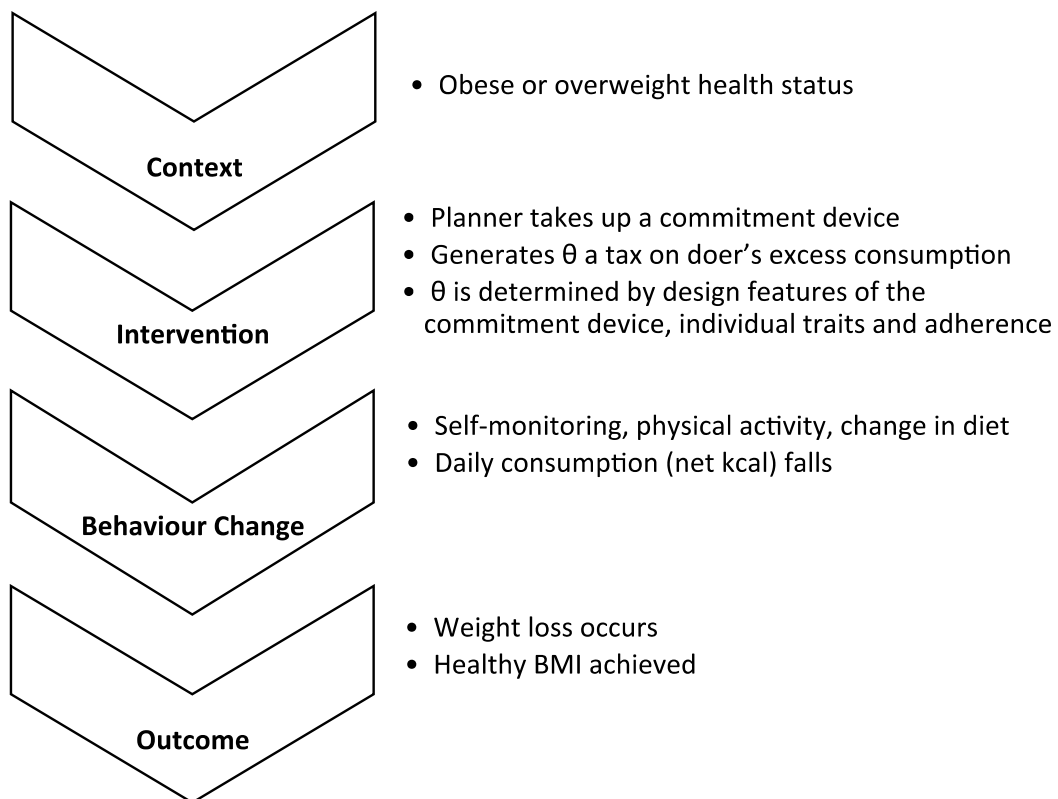
Design features ( $d$ )	The intensity of the commitment device (CD) design, perhaps in terms of money rather than reputational stakes, or scale of publicity for a pledge, will be associated with greater effectiveness.
Sophistication ( $\tau$ )	With greater self-awareness, a person is more likely to choose a suitable CD and recognise the need to adhere to it to keep themselves on track, which will lead to higher effectiveness.
Short-termism and myopia ( $\tau$ )	More impatience signals lower ability to follow through with the CD, and lower effectiveness. Those with far-sighted attitudes to their health will benefit more from a CD.
Adherence ( $\lambda$ )	The more an individual embraces the CD, and sustains their adherence to it through time, the more effective it will be.

---

## 5. CONCLUSIONS

The chapter has presented an original analytical framework to address a health behaviour problem: how can an overweight individual like Paul effectively change his diet and exercise behaviour to ensure he reaches his target weight? The model builds on the Thaler and Shefrin planner-doer model (1981) but makes its own contribution to the dual-self literature, through a careful and methodical theorising of the mechanics of how commitment devices exert an influence over human behaviour. The causal pathway implied by this framework is set out in Figure 8 below.

Figure 10: Causal effects of commitment devices





Scholars have previously proposed models that formalise aspects of the dual-self framework, often applying game theory to model the competition between a doer and a planner to determine whose preferred outcome prevails, or focusing on the descriptive mechanics of discounting parameters; but these models rarely focus on particular types of behaviour, and none consider inter-temporal choices on health in particular. Ruhm (2012) constructs a dual decision model considering health behaviours around diet, and concludes that this duality leads to a propensity to overeat in a modern, obesogenic environment, which is used to explain the growth in obesity rates in high-income countries. While Ruhm comes closest to formalising the intuition of the planner-doer model in a health behaviour setting, this work does not go beyond the prediction that commitment devices may be of value for those struggling with self-control problems, corresponding to proposition one in my model (see point 1 below). In all, six propositions are put forward:

1. In the context of time inconsistency where an individual is unable to rein in his short run consumption patterns despite the longer run benefits of doing so, there will be demand for a commitment device to bind future choices.
2. Commitment devices can bring about behaviour change and weight loss.
3. Taking up and applying a commitment device is costly, and it will have selective rather than universal appeal.
4. Commitment devices can generate more or less change based on their design, which vary the costs associated with failing to achieve the goal.
5. Commitment devices work differently for different people based on their individual characteristics
6. Individual adherence to the commitment device is essential for behaviour change to be effected.

This thesis aims to generate new and rigorous evidence to test propositions 2, 4, 5, and 6, and so answer the two Research Questions. The translation of these propositions into research hypotheses is set out in Table 5. The next chapter outlines the research design of two field experiments to investigate these hypotheses, with strategies to operationalise the variables identified as determinants of heterogeneous effects: sophistication, short-termism and myopia, and adherence to a commitment device.

**Table 6: Research questions, selected propositions from the model, and hypotheses**

<i>Research Question</i>	<i>Model's Prediction</i>	<i>Hypothesis</i>
RQ1: Can commitment devices change behaviour to promote desired health policy outcomes?	A commitment device can change health behaviours and deliver desired weight loss. ( <i>Proposition 2</i> )	1. A commitment device will generate positive average treatment effects on weight loss and health behaviours.
	A commitment device that generates more costs acts as a more severe tax on the doer's consumption and brings about greater effects. ( <i>Proposition 4</i> )	2. A more intense commitment device design will generate larger average treatment effects on weight loss and health behaviours.
RQ2: How does the effect of a commitment vary across people?	Effectiveness will depend on the individual's traits and adherence ( <i>Propositions 5 and 6</i> )	3. A commitment device will work more effectively for more self-aware individuals.
		4. A commitment device will work less effectively for individuals with short-termist and myopic attitudes.
		5. A commitment device will work more effectively for individuals who embrace the commitment device more fully.

BLANK PAGE

---

**Chapter 4**  
**RESEARCH DESIGN:**  
**Field Experiments To Isolate**  
**Causal Effects and Develop**  
**Theory**

---

## **1. INTRODUCTION**

Chapter 1 provided a narrative of the behaviour change problem that concerns this thesis: how to stick to a course of action where the costs loom large in the present and benefits are delayed to an uncertain future? Chapter 2 couched this puzzle in the context of health behaviours, time inconsistency and self-control, critically reviewing the literature on commitment devices and framing two research questions: can commitment devices change behaviours and health outcomes? And, delving more deeply, does it affect different people in different ways? The intuition of Thaler and Shefrin's (1981) planner-doer model was given a fresh analytical structure in Chapter 3, where it was applied to health behaviours for weight loss and produced six testable propositions.

The broad aim of this chapter is to take the hypotheses arising from the formal framework (Chapter 3) and lay out a strategy for testing them (in Chapters 5, 6 and 7). Specifically, this chapter will deliver three objectives. Firstly, it will expand on the chosen research design that underpins the thesis: an enhanced field experimental approach that actively combines quantitative and qualitative data and analysis. Secondly, this chapter will explain in detail the design features of the two field experiments that form the core of this thesis. The first experiment was undertaken with a private company providing an online weight loss service using digital self-monitoring tools, and is titled the Food Monitor trial. The second experiment was nested in a face-to-face weight loss programme run by a local authority, and is referred to as the Camden trial. Finally, the chapter will identify the threats to internal and external validity identified in advance, and explain how specific design features of the Food Monitor and Camden trials aimed to address and mitigate these issues.

The research design aims to maximise the complementarity between quantitative and qualitative methods. The former are used to test the presence and magnitude of a causal relationship between commitment devices and weight loss, and how these effects vary across sub-groups. Meanwhile the qualitative analysis will be used to develop a richer understanding of the average and heterogeneous treatment effects uncovered, and investigate the utility and veracity of the planner-doer model's underlying assumptions. The enhancement of the field experiments with qualitative data is necessary to provide a fulsome answer to both research questions, and represents a contribution both to the scholarly debate on commitment devices and mixed methods field experiments.

Combining qualitative methods within a field experiment is not a new technique (Dunning 2008, p.283; Freedman 2010, p.230). Indeed, "integrating quantitative evidence with qualitative evidence is especially appropriate for field experimental research, which, relative to laboratory experimentation, captures behaviour in complex, real-world settings" (Levy Paluck 2010, p.62). Ignoring these realities and the way in which they can affect statistical results would lead to limited or incorrect understanding of the causal relationship being investigated.

Yet, this combination of techniques is rare in the commitment devices literature as Chapter 2 explained, with much of the existing research focusing on quantitative results and analysis of average treatment effects to identify whether commitment devices can bring about health behaviour change. A key gap in the knowledge base that this thesis aims to address is on heterogeneous effects. To address the question of sub-group effects requires novel data collection on complex concepts such as adherence and sophistication, as discussed in the Analytical Framework. The argument that will be made in this chapter is that a qualitative narrative will strengthen the validity and plausibility of the treatment effects identified, and yield new insights

on causal mechanisms and underspecified variables. Empirical research by health scientists provides instructive examples of how this can be done (discussed in section 2).

The next section explains in greater detail the broad methodological choices of the research design: why are field experiments the chosen empirical strategy? Why and how are qualitative methods woven into them? Section 3 then explains in detail the design of the experiments, following the CONSORT statement's reporting checklist (Schulz et al. 2010; Boutron et al. 2010), with a discussion of how qualitative methods are interwoven in the field experiments through interviews and food journals. Sections 4 and 5 present the analysis plans for quantitative and qualitative data. Sections 6 and 7 discuss threats to internal and external validity and the mitigation strategies put in place to address these challenges. Section 8 concludes the chapter with a summary of the key arguments, and an overview of the two field experiments that form the bedrock of the empirical strategy.



## **2. BROAD METHODOLOGICAL CHOICES**

### **2.1. Why a randomised controlled trial?**

The empirical aim of this thesis is best served by a randomised control trial (RCT) that can isolate the causal effect of the commitment device treatment on the weight loss outcomes and health behaviour changes of interest. A randomised study conducted in a real world setting is a field experiment (Gerber & Green 2012, p.10), and it is this term that best describes the methods chosen to underpin this thesis.<sup>29</sup>

Studies based on observational data struggle to provide conclusive causal inference due to selection bias and endogeneity. For example, the study cited in earlier chapters on “paying not to go to the gym” relies on an observational dataset and is limited to providing a cross-sectional snapshot of what people spend and how that relates to their gym usage. The authors conclude that most consumers are not getting value for money, and would be better off paying per visit than with advance contracts (Della Vigna & Malmendier 2006, p.716). To the extent that the advance contract is a commitment device – a premium payment that aims to lock users into future good exercise behaviour – the results are suggestive that people who take up commitment devices may not always design them optimally, or may overestimate their ability to stay committed to the commitment device (Fan & Jin 2013). However, the study cannot ask or answer the question “did the commitment device work?” because it is unable to frame a counterfactual without the expensive gym membership, to understand whether those same users might have had much less exercise without it. It was also beyond the scope of the paper to consider any health benefits arising, even if expensively obtained, from the commitment device.

---

<sup>29</sup> In the remainder of the chapter, the terms randomised controlled trial and field experiment are used interchangeably.

Similarly, in a study on malaria prevention, Tarozzi et al. (2009) examine whether a financial commitment contract based on advance payment for bednet retreatment did actually increase retreatment rates, and find that 35% of buyers of the standard contract retreated their bednets compared to 79% of those with the commitment contract. Despite the impressive improvement in retreatment rates, the authors cannot conclusively report that the commitment devices caused an uptake of retreatment, because the choice of a standard contract or the retreatment contract was left to buyers. It is possible that some unobserved motivating factor which encouraged the buyers to take up the commitment contract was also the same motivating factor that encouraged them to go through the later effort of having the bednet retreated. The commitment contract is then a signal of willingness to invest in preventative health rather than a cause of health behaviour change. Despite some analysis of observable characteristics that might explain the take up of the contract, it remains the case that “selection on unobservables cannot be ruled out”, and the authors acknowledge the results remain “partly speculative” (Tarozzi et al. 2009, p.235).

Randomised field experiments, on the other hand, are “the methodology that has the best prospect of identifying causal relationships actually at work in the world” (Smith 2002, p.200).<sup>30</sup> They are regarded the most credible research design for uncovering unbiased causal estimates because they “solve the selection problem” (Angrist and Pischke, 2009: 15). With the ‘stable unit treatment value assumption’ (SUTVA), it becomes feasible to apply the potential outcomes framework to derive causal inference from a treatment (Little & Rubin 2000). Random assignment to a treatment creates a counterfactual group whose expected outcomes differ only through their exposure to the treatment (Duflo, Glennerster and Kremer,

---

<sup>30</sup> RCTs are of course not the sole means of drawing causal inference, as the methodological toolkit also includes natural experiments, instrumenting variable strategies that take advantage of exogenous variation, and discontinuity designs.

2007: 8), and a comparison of these outcomes between control and treatment groups can be used to infer a causal effect from the intervention. By designing a treatment around commitment devices, and randomly allocating this to participants in the research project, the field experiment aims to extricate the results from the confounding factor of selection; for example if certain people self-select specific commitment devices.

The limitations of field experiments are also well understood, particularly in the social sciences where it is not always possible or ethical to randomly assign as treatments the interventions and phenomena of greatest interest (Heckman et al. 2000; Deaton & Cartwright 2016). However, commitment devices do lend themselves to being shaped into a discrete intervention that can be feasibly offered to participants in a real world setting, as discussed in chapter 2 (John et al. 2011; Nyer & Dellande 2010; Prestwich et al. 2012). A field experiment, then, is the ideal way to answer the research questions by providing robust estimates of the average treatment effects for the sample as a whole (research question 1) and for particular sub-groups of the population (research question 2).

## ***2.2. Complementing causal inference with qualitative insights***

While RCTs are commonly held as the ‘gold standard’ of evaluation methods, they are often limited to single, narrow questions. An understandable critique is that they apply a black-box approach that tells us little of why a positive or negative (or null) result has come about. A focus on quantitative methods and data within RCTs can serve as a barrier to discovering the complexity of participants’ experiences (Hesse-Biber 2013) and can limit the generaliseability of findings beyond the precise context in which the RCT was implemented. By themselves, it follows, RCTs are insufficient for answering holistic questions about health behaviour

change in a way that would allow findings to be successfully transferred to real world practice.

A more promising research strategy would actively complement the strengths of an RCT with wider methods, to consider both the counterfactual analysis of what would have happened without the intervention, and the factual question of what actually happened (White 2013, p.72). For the latter, qualitative data is key to understand the research process as it took place, and to delve into the experience of participants with the interventions offered. Tarrow talks of “putting qualitative flesh on quantitative bones”, and prescribes that “wherever possible, we should use qualitative data to interpret quantitative findings, to get inside the processes underlying decision outcomes” (Tarrow 2010, p.109). Qualitative methods can illuminate the process by which change came about, but they can also gather information beyond the reach of conventional statistics. A recent trial on weight loss programmes illustrates this well.

Allen et al (2015) report on the qualitative component embedded in an RCT that tested the effects of attending a Weight Watchers programme compared to self-help. The 29 participants selected for qualitative follow up (from a wider sample of 1,269) were asked about the GP referral process, their experience of participating in the commercial programme, and the idea of being overweight as a medical issue. Two themes emerged that are of particular relevance to this dissertation. Firstly, GP referral using a personalised, signed letter was found to play an important role, creating in the users “a sense of moral and financial obligation to the GP” (2015, p.e251). This is arguably a form of reputational commitment the patients formed with their GP, and encouraged greater adherence to the weight loss programme. Secondly, the initiative provided free access to the commercial service, and this “contrasted with the idea that if they had paid for the service themselves this would have given them the right not to attend” (2015, p.e251). These perceptions may not

have been immediately obvious to an outside observer, nor might they have emerged from the statistical analysis.

A mixed methods field experiment would thus offer the advantage of an enriched dataset that can build on the strengths of both approaches: objective outcome measures that isolate a treatment effect, nuanced by more in-depth and discursive data, which ground the overall results in the realities of field experiment conditions and incorporate concepts that may not lend themselves easily to being simply counted.

While the advantages of a mixed methods RCT are clear, in practice several challenges come to the fore, with common pitfalls including poor reporting of sampling and data analysis (Lewin et al. 2009; O’Cathain et al. 2013). Well-established potential pitfalls of selection bias and researcher bias may affect the validity of qualitative conclusions, weakening the argument for incorporating qualitative methods. However, as with quantitative data, applying equivalent high standards to qualitative design, data collection and analysis would deliver robust, new evidence that can be triangulated and subject to robustness checks to ensure credible conclusion (Brady et al. 2010).

### ***2.3. Combining quantitative and qualitative data***

Scholarly debates in the social sciences have articulated various ways in which qualitative and quantitative methods can be successfully combined (Lewin et al. 2009; O’Cathain et al. 2013). Tarrow describes the use of quantitative data as a point of departure for qualitative research; sequencing qualitative and quantitative studies to retest and expand on prior findings; and defines triangulation as the combination of quantitative and qualitative data within a single research project to increase inferential leverage (2010, p.104, table 6.1).

Once both quantitative and qualitative data has been collected, a successful amalgam will ensure that the results of one method are placed in dialogue with another. Successful ‘weaving’ in of perspectives and methods ensures that health and social policy trials explicitly incorporate “the lived experience of those most impacted by the intervention – the study participants” (Hesse-Biber 2013, p.54). They can also illuminate an understanding of the research questions that would have been impossible with quantitative analysis alone, as documented by Starr in a selective review of published social science studies that highlight the value of “using an open-ended method over what would have been possible from a standard close-end approach” (Starr 2014, p.244).

For example, a quasi-experimental study by Valente (2011) uses propensity score matching to identify the treatment effect of land reform in South Africa on beneficiaries. This led to the counterintuitive result that “participants in the land grant scheme were more food-insecure than non-participants with similar socioeconomic, demographic and cultural characteristics”, with variation in this relationship across geographical districts (Valente 2011, p.358). The results prompted further qualitative data collection through interviews and focal groups to understand the apparent failure of the policy. Additional data corroborated the econometric findings, identifying issues such as corruption, and a mismatch between beneficiaries’ skills and the projects for their land grants. These mechanisms were invisible in the original statistical analysis.

Turney et al (2006) provide a second notable example, incorporating interviews into their social policy experiment on the effect of neighbourhoods on employment in Baltimore. The authors found that relocating to better neighbourhoods had no effect on job prospects. This contrarian result was explained through findings from in-depth interviews, which highlighted that job openings were

largely communicated through word-of-mouth along social networks. The move to a different neighbourhood reduced that mode of communication, so the families who relocated were no better able to tap into new opportunities than those who stayed behind in the poorer neighbourhood. Without speaking to the participants of the trial, it would have been impossible to uncover this explanation for a seemingly null result.

#### **2.4. *Summary: a mixed methods research strategy with three objectives***

Field experiments are stronger when designed to offer some explanation of the process of change and insight into the experience of the participants. Despite preconceptions that trials are inherently quantitative, there are many routes to effectively combining qualitative data collection and analysis to generate a richer understanding of the results and uncover new research directions. Placing the quantitative and qualitative insights in conversation with one another allows for a nuanced understanding of statistical analysis, particularly in the context of surprising or null results, and can shed new light on causal mechanisms and pathways that would not come to light through standard metrics alone.

Despite these advantages, the empirical literature on commitment devices, has rarely applied qualitative analysis to understanding average and heterogeneous treatment effects of commitment devices. As reported in chapter 2, a review of leading papers testing commitment devices for health behaviours finds only 1 of 8 published studies refers to any qualitative data collection: Giné et al (2010) report brief summaries of three semi-structured interviews with participants in a smoking cessation experiment.<sup>31</sup>

---

<sup>31</sup> To date I have found no examples of systematic qualitative data collection and analysis to understand the results of field experiments on commitment devices.

The thesis will address this gap by making stronger use of qualitative methods while maintaining the field experiment as the core of the research design. This decision has three specific objectives:

1. To generate new data on how commitment devices are interpreted and applied, in order to understand how differences in adherence generate heterogeneous treatment effects;
2. To contextualise and triangulate the average and heterogeneous treatment effect findings from the statistical modelling; and
3. To investigate for any evidence of the internal strategic interactions implied by the planner-doer theory, to further triangulate the analytical model with the empirical findings.

Objectives 1 and 2 relate directly to the 2 research questions underpinning the overall thesis. Objective 3 allows for some unpacking of the modelling assumptions and theory that generated the six hypotheses laid out in chapter 3. Qualitative methods are uniquely able to examine whether the planner-doer framework is an apt theory for health behaviour change, and whether there is any evidence for the planner-doer theory's assumptions about human behaviour and the internal tussles to pursue short run or long run wellbeing. In this way, qualitative methods can provide a crucial plausibility test of both the analytical model set out in chapter 3, and the statistical analysis that will be presented in chapters 5 and 6. This section aimed to explain the overall research strategy; the next section elaborates the design for two separate field experiments.



### ***3. FORGING PARTNERSHIPS FOR TWO FIELD EXPERIMENTS***

The following section moves the discussion on to the second overarching objective of the chapter: presenting the detailed design of two field experiments that form the core of the thesis. The Food Monitor and Camden trials were designed to test three different commitment devices in two different settings: online and in a weight management group. Both took place in the context of ongoing weight loss programmes, recruiting participants who were already registered in these initiatives, and both have been used as part of NHS efforts to combat obesity in the UK.

#### ***3.1. A mutual interest in the effect of commitment strategies for weight loss***

In both cases, the research partnership came about through a mutual curiosity about the role of commitment devices in encouraging stronger adherence to the weight loss tools available, with a view to improving overall weight loss outcomes. A number of different stakeholders including local NHS staff, councils, private companies, and civil society organisations were contacted during the early stages of the research design to identify partner agencies. Despite challenges in finding appropriate and willing field research partners, two agencies were particularly interested in the research: Food Monitor, a private company who developed a calorie counting tool to support a nutritious diet and weight loss, and Camden Council who provide a group-based weight loss service across the borough.<sup>32</sup> Camden delivered the programme using funding from the NHS, provided on a payment-for-results model, which meant they were

---

<sup>32</sup> The company asked that I sign a non-disclosure agreement, and I have therefore not used their actual brand name in any formal write-ups. For ease I refer to the company as 'Food Monitor', a fictional name. Correspondence with the marketing director of the company and their contact details can be provided. Screenshot graphics are from their website.

looking for innovative ways to improve weight loss and client retention in their programmes.

In-depth discussion in the early phases of the partnerships forged common ground and a strong appetite from the partners to host the trials. For example, I attended the weight loss groups run by Camden to observe first-hand the group dynamics, practical arrangements during the sessions, and the nature and means of existing data collection by Camden as well as scope for additional data collection. During this phase, careful discussion with both partners ensured the field experiments were designed in a way that allowed for testing the Research Questions in a robust and consistent way, and also proved feasible for the partners to participate in the trial alongside their own programme implementation.

For example, Food Monitor was understandably protective of their proprietary client base and this precluded my contacting participants directly. So the research design was drawn up in such a way as to ensure that all data collection took place through the Food Monitor website and administrative apparatus. An advantage of this approach was that data collection was rapid and relatively robust to human error, and allowed for a larger sample in the Food Monitor trial as the online medium reached a large number of prospective participants in a short time. In contrast, Camden were keen that I participate personally in the weight loss groups to register and collect baseline data, because it relieved the administrative burden on their staff; the research design was built accordingly. One advantage here was that I was able to build rapport with participants early on, which proved useful when contacting them for follow-up interviews 3 months later.

### **3.2. Tailoring field experiment design to the policy context**

The field experiments became necessarily distinct to take account of the varying preferences of the partner agencies and their own distinct programmes. While key elements remained uniform across both trials, such as key outcome measures and baseline data collection instruments, the variation in the two trials is arguably a strength of the overall research strategy as it made optimal use of opportunities available while ensuring that the partner agency was fully content, and the trials could feasibly be implemented.

### **3.3. Registration of trials**

The first experiment took place over July 2013 – February 2014 in collaboration with Food Monitor. This focused on testing the effect of a financial and a reputational commitment device delivered through the online tool. The trial was approved by the UCL ethics committee in May 2013 (project ID 4518/002) and has been registered with the American Economic Association’s Social Sciences Registry ([AEARCTR-0000942](#)). The second experiment took place over January 2014 – October 2014 in collaboration with Camden Council’s Active Health team, which tested the effect of a reputational commitment device delivered in person to participants. This trial was approved by the UCL ethics committee in December 2013 (project ID 4518/003) and is also registered with the AEA online registry ([AEARCTR-0000954](#)). The later staging of this experiment allowed for some lesson learning from the Food Monitor experiment. No outside funding was received to run the trials.<sup>33</sup> All data has been collected and stored carefully in line with UCL data protection rules. Quantitative and qualitative data can be made available on request.

---

<sup>33</sup> I was the sole investigator and was supported by a UCL studentship and some departmental resources for fieldwork costs.

The detailed design of these two experiments is described in sections 4 and 5, with a focus on sampling, recruitment, the nature of the programme in which the trial was embedded, the precise design of the treatments, baseline and outcome variables, data collection, and pre-analysis plans.

## **4. FIELD EXPERIMENT 1: FOOD MONITOR**

### **4.1. Target population and health programme context**

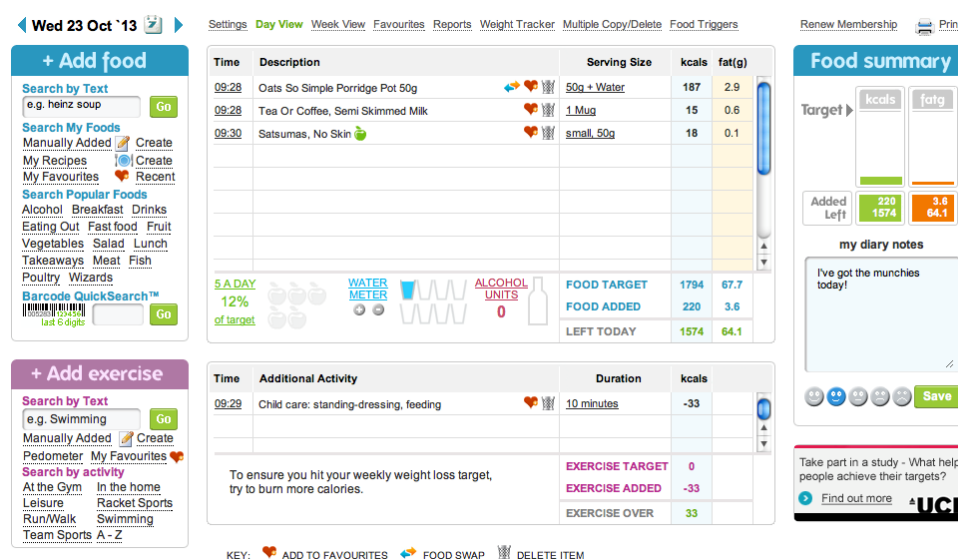
The target population was the Food Monitor client base of 5,650 active users. Food Monitor provides an online calorie counter tool to clients who pay monthly subscriptions of £7-£10. The product itself meets the working definition of a commitment device: it is voluntary; it aims to influence oneself to make good decisions on nutrition and exercise which might otherwise be open to temptation; and it is undertaken with strategic motives in relation to an individual's behaviour change alone. Self-monitoring is a key behaviour it aims to engender, with the wider goal of supporting weight management (see Figure 11 for a sample screenshot). The experiment was designed to run entirely online, in keeping with the digital format of the Food Monitor tool.

Recalling the typology of commitment devices set up in Chapter 2, the subscription to Food Monitor is a financial commitment because individuals pay a monthly fee despite there being free alternatives on the market. Clients could use another, free, calorie-counting tool but instead they choose to pay a subscription, which is interpreted as strategic behaviour by the planner to curb the doer's natural tendencies. By bringing a financial cost into the decision making process, the planner hopes to both increase the salience of the weight loss goal and increase the cost of ignoring it. Both are attempts to align the doer's day-to-day actions with the planner's longer-term goal of achieving a target weight.

To start using the service, clients input their current weight and set a target weight, and the tool determines the recommended daily calorie intake to achieve that goal in a reasonable period of time. Food and exercise entries are logged daily in a calorie calculator, so users quickly grasp how they are performing against

their daily target, and can make adjustments to their choices and behaviour over the day to meet the target. This arrangement easily fits the planner-doer framework: the planner is responsible for setting the weight loss goal and identifying the daily calorie target, and the doer is influenced over the course of the day by the self-monitoring tool, which is designed to help avoid over-consumption of food and encourage exercise to lower net calorie intake

**Figure 11: Food Monitor dashboard**



#### 4.2. Commitment device treatments

The experiment offered a two-pronged test of commitment devices. The financial commitment was the payment of the monthly subscription fee, a premium payment as there are free, alternative calorie-counting tools widely available. A further reputational commitment was introduced in the context of the study, which asked randomly selected participants to name a 'coach' who would check whether they were making progress towards their goal. A mild form of a public pledge, this intervention increased the reputational cost of ignoring the weight loss goal, and by doing so added a further incentive for the doer to make sensible choices.

To assess the effect of the reputational and financial commitment devices on users' weight loss outcomes and behaviours, a separate group of participants was needed who had neither financial nor reputational commitment elements. Individuals in this group were expected to experience far less influence of the planner sub-self on the doer's preferred actions. The collaborating firm agreed to provide refunds on the monthly fee to selected participants. The refund is assumed to weaken the sense of financial commitment amongst these individuals, and as a result decrease the effectiveness of the planner's strategy on the doer's behaviour. As they had signed up with the expectation of paying, it would be fair to assume that the financial commitment element decreases but does not disappear entirely, particularly as payment is likely to resume after the period of the experiment for members on a rolling monthly subscription. Accordingly, this group is assumed to have less (but not zero) financial commitment and no reputational commitment.<sup>34</sup>

### **4.3. *Experimental groups***

In this way, the experiment sets up three participant groups. In increasing order of planner influence over doer actions, these groups experience:

- Limited financial and no reputational commitment (referred to in following discussions as the 'limited commitment' or 'refund' group);
- Financial commitment as usual and no reputational commitment ('financial commitment' or 'monthly fee' group); and
- Financial plus reputational commitment ('reputational commitment' or 'coach' group).

---

<sup>34</sup> Financial commitment is understood as being limited rather than zero because the subscription is often made on a multi-month basis, and a refund on any one month does not mean that the membership is completely withdrawn.

#### **4.4. Sample size calculations**

Ex ante calculations of the sample size conducted suggested the experiment should aim to recruit a sample of 364 participants (see Table 7).<sup>35</sup> The target was taken from a range of estimates that varied the expected mean weight loss and standard deviation, with the baseline scenario assuming a Cohen's *d* treatment effect size of 0.5 and a standard deviation of 2.3 kg (5 lbs). This gave a corresponding difference in weight loss of just under 1 kilogram (2 lbs) for treated individuals over the comparison group; a reasonable assumption when triangulated with commercial weight loss programmes. A recent study found that compared to a control group who exercised only, participants in a Weight Watchers group lost 2.4 kg more, those in the Rosemary Conley group lost 2.2 kg more, and those in the Slimming World group lost 1.6 kg more (Jolly et al. 2011, p.11). The sample size calculation parameter was also in line with the effects reported in the closest available literature at the time on reputational commitment devices in the form of a public pledge, notably Nyer and Dellande (2010). In sum, the assumptions underlying the sample size calculations appeared plausible in this context.

The baseline scenario implied that each experimental group should have 132 participants. From the start it was recognised that the three participant groups would not be of equal size. A quota of 100 was imposed for the 'limited commitment' group by the collaborating firm based on the budget for client refunds. Taking into account this externally imposed cap and the recommended 132 participants in each treatment group, the calculations implied a total sample of 364.

---

<sup>35</sup> Calculations undertaken using lbs, kilogram conversions reported here. Calculations assumed 0.9 power and alpha (two-sided) of 0.05. Appendix section A5 contains a full discussion of assumptions, Stata output, and sensitivity analysis.



Weight loss scenarios	Baseline (0.91 kg difference)	Low (0.45 kg difference)	High (1.8 kg difference)
Baseline SD (2.3 kg)	132	526	33
Lower (1.4 kg)	90	358	23
Lowest (0.45 kg)	69	274	18

It was decided in advance that once the quota was reached, random allocation to the refund group would be closed and additional participants would only be allocated to one of the remaining two groups (for further discussion of randomisation and allocation ratios see below). Recruitment would continue until the overall sample reached 364, with the expectation that this would ensure approximate balance in numbers of 132 participants across the ‘financial commitment’ and ‘reputational commitment’ groups. An overall sample of 364 was observed to be larger than the samples reported in other studies considering commitment devices and weight loss as discussed in Chapter 2: Volpp et al (2008) recruited 59, Nyer and Dellande (2011) recruited 211, and Prestwich et al (2013) recruited 257; all had three participant groups.

The sample size calculations were noted to carry a degree of uncertainty, as they were based on estimates of mean and standard deviation of weight loss, for which actual data was not readily available; they focused on average treatment effects rather than heterogeneous effects; and implicitly assumed minimal and balanced attrition across groups. A larger sample size would have resulted from incorporating the latter two factors, but would likely have pushed the experiment beyond the point of feasibility and acceptability for the partner firm.

#### **4.5. Participant recruitment**

An advertisement for the study was posted to the Food Monitor website, visible only to paying clients (see Figure 11 above). In order to register for the study, interested individuals clicked through and were asked to confirm they were fully paid and registered Food Monitor clients and fit the following eligibility criteria:

- (i) They had already completed Food Monitor's registration process and agreed to the standard terms and conditions of use,
- (ii) They were privately paying clients, not NHS referral clients, and
- (iii) They had at least 4 weeks remaining on their membership from the start of the trial date.

These criteria were designed to ensure that all participants were equally willing to pay for the product, to control for potential variance in motivation. Eligible participants were presented with an information sheet and form for online consent. Those who wished to continue were then asked to complete a baseline survey to gather data on individual weight loss goals, wellbeing, personal and demographic traits, time preference and health attitudes.

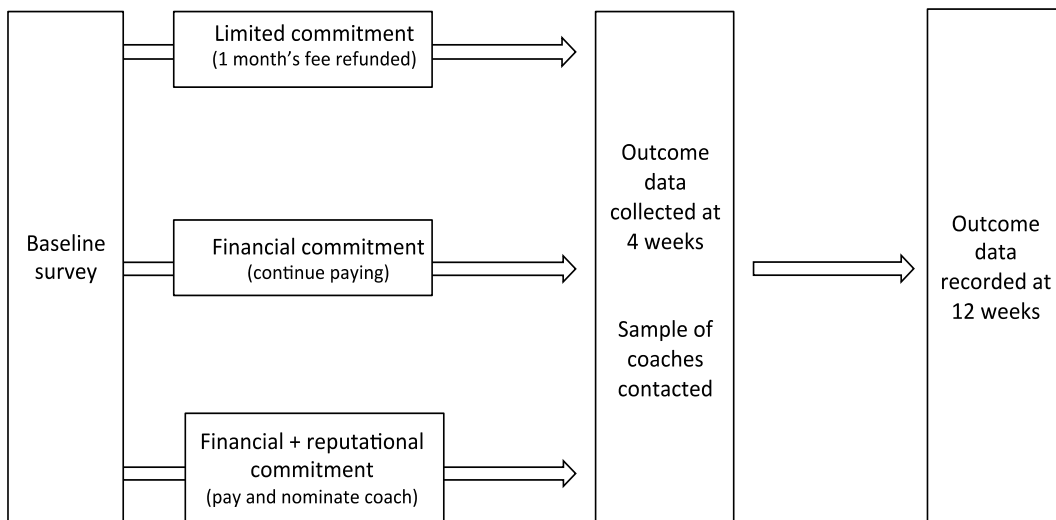
#### **4.6. Stages of the field experiment**

The experiment was staged in four main stages (see Figure 12). The online survey first asked for informed consent and then moved to baseline survey questions. The survey ended with the software's built-in randomisation mechanism assigning participants to one of three experimental groups by displaying different messages (full messages set out in chapter 5).<sup>36</sup>

---

<sup>36</sup> Qualtrics randomization tools are built on the Mersenne Twister algorithm, a pseudo-random number generator.

**Figure 12: Stages of the Food Monitor experiment**



In line with section 4.3 above outlining the experimental groups, the limited commitment group was given the opportunity to receive their next 4 week's subscription for free. The rationale was to diminish the financial commitment that would otherwise have underpinned the use of Food Monitor over the period. The financial commitment group were thanked for completing the survey, asked to use the food diary as they normally would, and continued to pay their subscription. The reputational plus financial commitment group continued to pay their subscription *and* were asked to name a coach, someone encouraging and familiar with the individual's weight loss goals, who might be contacted after 4 weeks by the researcher.

#### **4.7. Baseline survey**

The baseline survey was built on a Qualtrics platform, a popular online survey tool with several applications in research requiring online data collection. The baseline survey gathered data on starting weight and BMI, lifestyle factors (such as diet and exercise), demographic characteristics (including age and gender), and personal traits (to derive operational measures of short-termism). These covariates were selected to control for the variation in individual characteristics across the sample, allowing for more

precise estimation of the average treatment effect (research question 1), as well as to collect data to estimate heterogeneous treatment effects (research question 2)

The online survey platform was designed to be user-friendly and visually straightforward to reduce the risk of respondent fatigue, missing data and low completion rates. Where possible, questions were borrowed from existing surveys – for example, the HFS instrument was replicated using the Health Survey for England (Robinson 2012) – both because this ensured that the formulation of the question had been tried and tested, and the data collected would be comparable to previous survey results and nationally representative samples as a reference point.

The survey was administered ahead of the randomisation process, in order to identify and analyse individual traits amongst compliers and non-compliers. As Gerber and Green advocate, “researchers should take advantage of opportunities to gather background information that may be useful in predicting potential outcomes...it can pay dividends in terms of precision with which the average treatment effect is estimated” (2012: 96).

The survey was pre-tested with staff at Food Monitor and through an interview with an independent weight loss instructor to assess how questions were interpreted, to check that the survey was user-friendly in terms of content, accessibility of language, and length. The aim of pre-testing was to balance robust data gathering with minimal respondent fatigue and dropout, and to ensure that the randomisation and data flow with the Qualtrics platform was working correctly. Participants were to be identified through their baseline surveys using a unique and anonymised client code, which ensured that Food Monitor’s automated reports would accurately gather administrative and outcome data while preserving confidentiality. This step effectively incorporated qualitative insights

before the trial was launched, and supported the process of gathering high quality quantitative data.

All participants were thanked for their time and invited to continue using Food Monitor as they normally would over the following four weeks. The third stage was data collection at the 4-week point, and at this time some coaches were contacted in line with the treatment offered. The final stage was a further period of data collection to 12 weeks in total, after which time the participant would not be monitored further.

#### **4.8. *Randomisation and blinding***

Participants were invited to sign up at a time of their choosing, so it was a recognised limitation that the full sample was not available to randomise in advance, and it was unlikely to be perfectly balanced in numbers across the experimental groups. Given the imposition of a quota on the refund group, it was also necessary to change the treatment allocation ratio from two-in-three to one-in-two during the experiment. Implications from this change are discussed further later in the chapter (see analysis plan below, and the later discussion of threats to validity in section 7).

Participants were not blinded to their own treatment status as the aim of the experiment was to identify whether the change in commitment affected behaviour. Participants were not made aware of the possibility of other treatments. The investigator (myself) had to be aware of treatment status in order to trigger the monthly refund for participants assigned to the limited commitment group, and to initiate follow up with coaches. The partner firm was only made aware of treatment status for those clients who were due a refund, and this refund was delivered through their administrative systems

with a confirmation of the precise amount shared with me afterwards.<sup>37</sup>

#### **4.9. Outcome data collection**

After the survey and treatments were administered, the month-long experimental period began. Data was collected on participants over 12 weeks in total, with the first four weeks interpreted as the treatment period. The primary outcome variable is weight loss at the end of four weeks and 12 weeks, measured as a percentage of initial weight.

Health behaviour change was measured by a secondary outcome variable on how well the individual was self-monitoring, an important aspect of the health behaviour change process often cited as a key ingredient for weight loss success (Butryn et al. 2007; Yu et al. 2015). Public health scholars define self-monitoring as the “systematic observation and recording of target behaviours (Boutelle et al. 1999, p.364), which includes food and exercise diaries (Johnson & Wardle 2011). Self-monitoring was operationalised as usage of the digital tools, measured by the number of logins through the website or Food Monitor app over the 4-week experimental period to use the calorie counter, log a weight reading, or create a journal entry. As discussed in chapter 2, opting for a measure of health behaviour beyond exercise is relatively novel in the commitment device literature, and the Food Monitor trial made this a feasible choice.

In addition to the baseline survey, customised reports from Food Monitor were designed to gather data on the exact amount of monthly refund provided to those in the limited commitment group. Further data collection was planned after the four-week experimental period, to understand how the coach treatment was interpreted and

---

<sup>37</sup> Refunds varied because clients were on different monthly subscriptions, which was largely down to what the going price was when they signed up, and if they moved from an introductory offer to the ‘normal’ fee.

applied by those in the reputational plus financial commitment group. A selection of coaches were to be contacted, and email responses reviewed for qualitative insights into challenges faced for those in the reputational commitment group, and the extent to which the coach was involved in monitoring the weight loss goal. The exercise was designed to identify levels of adherence to the treatment, which was specified in chapter 3 as an important heterogeneity pathway, and is discussed further in the analysis plan below.

#### ***4.10. Quantitative analysis plan***

A quantitative pre-analysis plan was prepared during the research design phase to set out how the data collected would be used to test the hypotheses arising from Chapter 3.<sup>38</sup>

##### ***4.10.1. Statistical model for average treatment effects***

Research question 1 asks whether commitment devices can change behaviour. To take account of the change in treatment allocation rule once the refunds quota was reached, two separate analyses were undertaken to recover the average treatment effect (intent-to-treat estimate) of the refund and the coach on weight loss and on self-monitoring behaviour. The statistical models below are presented with and without covariates. Equations 8a and 8b recover the ATE from the limited commitment group, and are contained to phase one of the trial when all three treatments were available. Equations 9a and 9b recover the ATE from the reputational commitment group and are applied to the full span of the experiment, but include only the coach and comparison groups.

---

<sup>38</sup> Deposited with the AEA trials repository on 11 November 2015, but prior to that had been sent to thesis supervisors in June 2013 as part of an assessed PhD Methodology Paper.

$$[8a] \quad Y_i = \alpha + \beta^R \cdot R + \varepsilon_i$$

$$[8b] \quad Y_i = \alpha + \beta^R \cdot R + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

$$[9a] \quad Y_i = \alpha + \beta^C \cdot C + \varepsilon_i$$

$$[9b] \quad Y_i = \alpha + \beta^C \cdot C + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

Where:

- $Y$  is the outcome variable for weight or self-monitoring.
- Treatment status is captured by dummy variables  $R$  and  $C$ . In both models, the comparison group comprises participants who continued to pay their monthly fee.
- The OLS estimators for  $\beta^R$  and  $\beta^C$  provide the average treatment effects for the 2 treatments R (refund) and C (coach) respectively.
- $W$  is a series of individual covariates  $J$ , with coefficients  $\gamma_j$ . These coefficients offer statistical association with outcome variable  $Y$ , and cannot be used to infer causality.
- $S$  is a series of temporal variables capturing seasonal effects (using month of registration).

The benefit of including covariates is to reduce variance and provide more precise coefficient estimates (Gerber & Green 2012, p.95), and it may also be necessary to account for imbalances between the experimental groups that may occur despite randomisation, particularly in smaller samples (Torgerson & Torgerson 2008, p.61). All statistical analysis was undertaken using Stata v12.

The analytical model (chapter 3) predicts that the presence of a commitment device will bring about behaviour change and weight loss. In the Food Monitor trial, the ‘financial commitment’ group and the ‘reputational plus financial commitment’ group are therefore expected to experience more behaviour change and weight loss than



the group offered a refund who are understood to have ‘limited commitment’. With the comparison group being the ‘financial commitment group’, quantitative analysis is expected to uncover positive and statistically significant coefficients for  $\beta^C$  and negative coefficients for  $\beta^R$  against both weight loss and self-monitoring behaviour, thereby providing a test of hypotheses 1 and 2 (Table 5 page 99).

#### ***4.10.2. Data for sub-group analysis***

Research question 2 moves beyond the simple average across the sample to investigate treatment effects amongst specific sub-groups. Chapter 3 identified two key individual traits – sophistication and short-termism – as important heterogeneity pathways, and noted the challenge of pinning down these concepts with quantitative data. The Food Monitor trial collected data on short-termism using two separate measures. Short-termism in relation to health attitudes is operationalized through the Healthy Foundations Segmentations model, introduced in chapter 3 (see also appendix section A.3).

The theoretical underpinnings for this thesis, unlike those of hyperbolic discounting models, do not require the discount rate parameter itself be calculated, and this means that research design was free to find other workable measures for time preference. Instead of the time-consuming and laborious process of uncovering precise discount rates using a series of choices, the baseline survey instead employed a single question designed to measure the cost of waiting for a modest cash sum (£10) for an additional 1 month and an additional 6 months relative to receiving that cash today. The additional amount required to delay receiving the cash sum is interpreted as the individual’s cost of waiting. The spectrum of values generated is a proxy for patience: the higher the amount entered implying a higher degree of impatience. The range of discount rates elicited in field experiment 1 was to be used to generate two groups:

low and high discounters. The latter are expected to benefit less from commitment devices, in line with the theory set out in chapters 2 and 3 (see appendix section A.12 for more detail on time preference measures).

#### ***4.10.3. Statistical model for heterogeneous treatment effects***

The analysis will employ a statistical model built on treatment-covariate interactions to assess how well commitment devices work for particular sub-groups of people, identified based on some observable trait (Gerber & Green 2012, p.290). The models are set out in Equations 10a and 10b below, pertaining to the refund and coach treatments respectively. The specific interaction terms to be used are summarised in Table 5. The treatment effect for a sub-group can be found by summing the treatment and trait coefficients, to generate a conditional average treatment effect (CATE) (Gerber & Green 2012, p.296).

$$[10a] \quad Y_i = \alpha + \beta^R \cdot R + \beta^{tr} \cdot R \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

$$[10b] \quad Y_i = \alpha + \beta^C \cdot C + \beta^{tr} \cdot C \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

Where, in addition to the model set out in equation 9:

- *Trait* captures the baseline covariate being tested as a heterogeneity pathway
- The linear combination of estimators for  $\beta^{tr} + \beta^R$  and  $\beta^{tr} + \beta^C$  provide the conditional average treatment effects for the refund and coach treatments respectively.
- Covariates *W* are the same as those used in the ATE model.

Short-termism is incorporated through two quantitative measures. A third heterogeneity pathway relates to how well the individual embraced the commitment device (see Equation 7 in

Chapter 3). Further discussion of how adherence will be operationalised is picked up in section 6 below as part of the qualitative analysis plan. Section 4 concludes here, and full details of the CONSORT reporting statement are recapped in the appendix (section A3). The next section provides a similar overview of the second field experiment.

## **5. FIELD EXPERIMENT 2: CAMDEN**

The second trial provided a useful contrast by being embedded in a face-to-face, group-based weight loss programme, which takes as its main treatment a reputational commitment device in the form of a signed contract to oneself.

### **5.1. Target population and health programme context**

Obese and overweight individuals in Camden can access free weight loss services through the Apples and Pears initiative. One of these services is Shape Up, managed by Camden Council's Active Health Team. Shape Up is a 12-week, group-based programme that sets a 5% weight loss target for all clients. The group meets every week to weigh-in with the group tutor, and discuss a different aspect of weight loss and lifestyle as part of the Shape Up programme designed by the non-profit organisation Weight Concern. The tutors measure and record weight data, facilitate group discussion, and coach individuals on their food journal and self-reflection.

The target population is the Shape Up client base. These individuals were screened for eligibility based on their BMI and home postcode, and might have been put forward for the programme either through a GP or self-referral. These same eligibility criteria apply to participation in the field experiment, and did not change during the trial.

Clients chose Shape Up over other similar service providers, and know from their first meeting that if they successfully complete the programme and achieve their goals they will be awarded with free gym sessions at local leisure centres. The experiment remains focused, however, on testing the effects of a commitment strategy and not the role of incentives in encouraging people to lose weight. Under a normal Shape Up programme, the population have made no financial commitment, but arguably may be conscious of a mild reputational commitment to the group tutor. Tutors would often call participants who had missed a class to offer encouragement and support to return the following week.

## **5.2. *Commitment device treatment***

The Camden trial aimed to test a self-reputational commitment device: a commitment contract signed to oneself. The contract was designed to have a certain degree of visual gravitas and formality, on card of A5 size that could be carried around in a handbag or satchel, or stuck on a fridge or wall.

Figure 13: Commitment contract treatment



### **5.3. Experimental groups**

The trial was simply designed with two groups: those who were offered the commitment contract, and those who were not. Treatment offer took place after participants had provided informed consent and completed a baseline survey. The control group were thanked for their time and given a copy of the consent form and information sheet in an A4 brown envelope to take away with them. The only difference for the treatment group was the offer of the commitment contract that they were invited to personalise by writing in their names, signature, and date. They were then advised to keep it somewhere they would see it on most days, and to discuss it with friends and family if they wished but not with the other group members. If they accepted it, they signed the contract then and there, and it was placed in the envelope to take away. Each participant therefore returned to the class with an anonymous-looking brown envelope, whether they belonged to the control or treatment group, and this was a deliberate design feature to minimise the risk of

contamination across groups and associated problems (such as control group members feeling resentful they did not receive a contract, the so called John Henry effect discussed in further detail in section 7 of this chapter).

#### **5.4. Sample size calculations**

Sample size calculations are reported in Table 7 for an individual randomisation design.<sup>39</sup> Sensitivity analysis considered two other scenarios beyond the baseline scenario, allowing for weight loss outcomes and standard deviation to vary. As with the Food Monitor trial, weight loss in kilograms was the outcome variable of interest. The Shape Up programme was assumed to deliver weight loss of 0.23 kg per week for the control group, yielding net weight loss of 2.5 kg over the 11-week course. The treatment group were assumed to secure slightly higher weight loss at the final weigh-in of 2.95 kg. These values were chosen as conservative parameters, but it is worth noting that they imply a moderately strong Cohen's *d* treatment effect size of 0.5. This was consistent with the effect sizes reported in the closest available literature at the time the calculations were undertaken (Nyer & Dellande 2010; Prestwich et al. 2012). In the baseline scenario, standard deviation in both groups was set at 0.9 kg.

As highlighted in Table 8, these calculations implied a total sample size of 170 in the baseline scenario, based on two equal groups of 85 participants.

---

<sup>39</sup> A cluster randomisation design was briefly considered, but it would have required more groups than were available in early discussions with Camden in order to be sufficiently powered, and it was determined early on that an individual randomisation design would be preferable alongside specific measures to avoid contamination across participants who were assigned to different experimental groups while in the same Shape Up Group.

Weight loss scenarios	Baseline (0.45 kg difference)	Moderate (0.68 kg difference)	High (0.9 kg difference)
Low SD (0.45 kg)	22	10	6
Moderate (0.91 kg)	85	38	22
High (1.36 kg)	190	85	48

Other estimates in the table allow for greater weight loss differentials (assuming the treatment group ended with 7 and 7.5 lbs weight loss respectively), and for alternative standard deviations of 1lb and 2 lbs in each group. In line with the sample size calculations reported for the earlier trial, a wide range of estimates emerged from the sensitivity analysis. The calculations were recognised to carry a degree of uncertainty, as they required estimates of the mean and standard deviation of weight loss for which data was not readily available; and they assumed a full sample and no problematic attrition (discussed further in section 7 later).

### **5.5. Participant recruitment**

The aim was to recruit participants early on in the Shape Up programme, ideally in week 1. The introductory week 0 session was a time for tutors to get to know their class and make introductions, and it was felt that with the existing burden of initial paperwork the introduction of this trial would be better deferred to week 1 or 2. Tutors were given a script to read out, which ensured I was introduced to the group in the same way each time (see appendix section A.19).

During the weekly Shape Up classes each participant would provide a weight reading using the tutor’s digital scales. These were often arranged in a quiet corner of the room, and the weigh-ins were essentially private between the tutor and the person. It was agreed that in the sessions where recruitment was taking place, I would go

<sup>40</sup> Calculations undertaken using lbs, kilogram conversions reported here. Calculations assumed 0.9 power and alpha (two-sided) of 0.05. Full details of sample size calculations using Stata are set out in the appendix section A6.

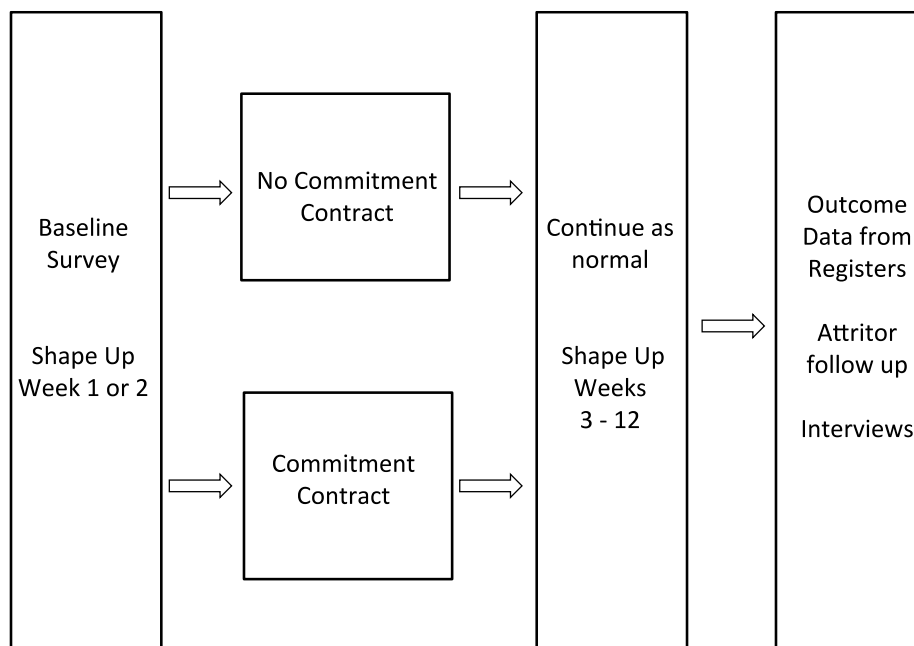
along to the class and manage the weigh-ins for the tutor. The role involved taking the readings from the digital weighing scales and recording them on the class register. It provided an opportunity to introduce myself personally, build rapport with clients, and invite them to take part in my research project. For those who were keen to take part, the registration process involved a brief explanation of the project using the Information Sheet, signing the Informed Consent forms, and filling in the baseline survey.

### ***5.6. Stages of the field experiment***

The field experiment was designed to run in three distinct stages (see Figure 14). Firstly, participants received a short brief about the trial and invited to register, which began with informed consent. They then completed a short baseline survey, and given an information sheet to take away. If they were in the treatment group they would also receive a contract. Secondly, the participants would continue with the Shape Up programme weekly meetings, guided by the tutors through the course syllabus week by week. At the end of the course, data collection would take place, involving both quantitative and qualitative data, and finally analysis.



Figure 14: Stages of the Camden field experiment



### **5.7. Baseline survey**

The Information Sheet and Consent Form mirrored those used in the Food Monitor experiment. In keeping with the first field experiment, the baseline survey asked broadly the same questions on lifestyle, health attitudes, and demographics. Pre-testing with a group tutor suggested the time preference question would not be well understood by participants facing language barriers, and Camden asked that the overall length of the survey be shortened to minimise the amount of time the client was away from the group discussion. Demographic questions on educational background and job status were therefore dropped.

## **5.8. Randomisation and blinding**

Participants were randomised in advance into two groups: control and treatment. Client lists were put together in the weeks running up to the launch of a new class or season of classes, so several randomisation exercises took place using Stata (v12), with the same do-file code. The random number sequence was generated using a seed set to the eight-digit date of the exercise (DD-MM-YYYY) and a sample do-file can be found in the appendix (section A.7).

In the event that participants were not randomised in advance but began attending the Shape Up classes (for example due to administrative errors in the lists provided from Camden), a simple numerical rule was created to allocate the person to an experimental group on the day: to take their eight-digit date of birth, sum all the numbers together, and if the answer was odd they would be assigned to the treatment group. The rule had the advantage of maintaining the 50/50 allocation rule, while also ensuring unpredictability of treatment assignment, thus removing potential subjective bias from entering the randomisation process; and meant that eligible participants were not turned away (as this may have undermined the ability for the trial to reach the desired sample size). Further discussion on the use of this numerical rule for treatment assignment is set out later in the chapter on threats to validity.

Group tutors were blind to the treatment status of the participants. As the person responsible for administering the treatment, I was not blinded; given the nature of the treatment, participants in the treatment group were also not blinded. The implications for bias and validity are discussed in further detail later in section 7.

### **5.9. Outcome data collection**

In line with the first field experiment, the main outcome variable of interest is change in weight at the end of the Shape Up programme, which is measured both in kilograms, as a percentage of initial weight, and a binary variable for whether the 5% weight loss target was achieved or not. A secondary outcome variable, again in line with the Food Monitor trial, captured attendance and completion rates. This variable measures self-monitoring since attendance at a weekly meeting implies a weight reading is taken and discussed with the tutor. It is also a useful indicator of how well the individual was adhering to the Shape Up programme, and the behaviour of returning week after week was cited as a key ingredient for success in the Shape Up plan.

The quantitative outcome data was routinely collected by the group tutors and collated by Camden for their internal monitoring and performance frameworks. The data was shared with me at the end of the Shape Up programmes. Building on experience from the Food Monitor experiment, the Shape Up trial design gave significant attention to potential attrition. Gathering the outcome data from Camden as soon as possible after the group finished the course gave me an opportunity to identify and follow up with attritors who had shared their contact details, with the aim of taking down self-reported readings from them to fill in data gaps around weight loss.

## **5.10. Quantitative analysis plan**

### **5.10.1. Statistical model for average treatment effects**

The statistical models to estimate the average treatment effect are set out in Equations 11 and 12, which are identical to the Food Monitor model with a simplification to reflect there being only one treatment offered:

$$[11] \quad Y_i = \alpha + \beta^C \cdot C + \varepsilon_i$$

$$[12] \quad Y_i = \alpha + \beta^C \cdot C + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

Where:

- $Y$  is the outcome variable for weight and participation
- Treatment status is captured by dummy variable  $C$ , where  $C=1$  if the participant is in the treatment group
- The OLS estimator for  $\beta$  provides the average treatment effects for the commitment contract.
- $W$  is a series of individual covariates  $J$ , with coefficients  $\gamma_j$ . These coefficients offer statistical association with outcome variable  $Y$ , and cannot be used to infer causality.
- $S$  is a series of administrative, group and temporal variables including wave of the study, group tutor, and starting month to capture seasonal effects.

As with the previous analysis plan for the Food Monitor trial, the results are expected to uncover a positive and statistically significant coefficient  $\beta^C$  in line with hypothesis 1.

### **5.10.2. Data for sub-group analysis**

A key innovation of the Camden field experiment is testing for sub-group effects based on self-awareness. The baseline survey for Camden participants asked whether the participant had experience of any previous weight loss programme. This binary variable (yes/no) is expected to be a useful proxy for sophistication, on the basis that prior experience of weight loss programmes would give the individual some insight into their natural tendencies: how they find the weight loss guidance, how well they are able to exercise self-control, and how they find the task of persevering with their behaviour change goals. As discussed in Chapter 3, sophistication is a challenging concept to operationalize, and the variable capturing previous weight loss programme experience promises to offer new insights for this heterogeneity pathway.

As with the previous experiment, the Camden trial will allow for analysis of short-termism using the HFS survey instrument to identify those with myopic health attitudes. The time preference question was omitted from the baseline survey as a result of pre-testing with a group tutor, precluding analysis of present bias as a heterogeneity pathway amongst Camden participants.

### **5.10.3. Statistical model for heterogeneous treatment effects**

Heterogeneity analysis will be conducted using the statistical model set out in Equation 13, which again closely mirror the treatment-by-covariate interaction approach to sub-group analysis (Gerber & Green 2012).

$$[13] \quad Y_i = \alpha + \beta^C \cdot C + \beta^{tr} \cdot C \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

Where, in addition to the model set out in equation 12:

- *Trait* captures the baseline covariate being tested as a heterogeneity pathway
- The linear combination of estimators for  $\beta^{tr} + \beta^C$  provide the conditional average treatment effects for the commitment contract.
- Covariates *W* are the same as those used in equation 12.

### 5.11. Summary

Chapter 3 closed with a summary of the research questions and hypotheses, including three heterogeneity pathways. Table 9 below recalls these pathways out and clarifies how they will be tested across the two field experiments. Adherence and myopia will be investigated in both trials; data on present bias will be collected in the Food Monitor experiment only, and data on sophistication in the Camden experiment only. Adherence will be investigated through qualitative analysis only, in chapter 7. See Table 12 in the concluding remarks for a fuller comparison of the two trials.

	Expected impact	Food Monitor	Camden	Chapter
Short-termism				
- <i>Myopic health attitudes</i>	-	Yes	Yes	5 and 6
- <i>High cost of waiting</i>	-	Yes	x	5
Sophistication	+	x	Yes	6
Adherence	+	Yes	Yes	7

The discussions over sections 4 and 5 have provided a detailed overview of the two field experiments – their context, their design choices, the quantitative data they will generate, and how that data will be used to answer the two research questions. Section 6 examines briefly the qualitative data that will be collected alongside the quantitative and how it will be used to support and nuance the statistical results.

## **6. QUALITATIVE ANALYSIS PLAN**

The following section provides an overview of qualitative data collection and analysis. To recap their objectives, the qualitative components intend to generate new data for sub-group analysis, draw on the lived experiences of the trial participants to help contextualise the quantitative results, and test the assumptions of the planner-doer theory underpinning the hypotheses and statistical modelling. Table 10 provides an overview of these objectives and the qualitative instruments applied to gather the necessary data. The first objective directly contributes to research question 2, by gathering data on how well participants adhered to the contract; the second and third objectives contribute to a more nuanced answer to both research questions.

The Food Monitor trial used a brief email questionnaire to ask coaches how involved they were in the participant's weight loss efforts, to signal how well the participant adhered to this reputational commitment device. For example, if a coach were to respond that he was not aware of the weight loss target, this individual has not adhered to the reputational element of the commitment device. The qualitative coach responses were designed to create a distinction between low and high adherence participants, allowing for basic comparison of outcomes between the two sub-groups (this informs the analysis in chapter 7).

The remainder of section 6 centres on the semi-structured interviews folded in to the Camden trial, which generated the largest volume of qualitative data. The discussion below considers in turn the merits of using interviews, how the objectives informed the topic list and preliminary coding scheme, and the basis for interpreting responses using qualitative content analysis.



<b>Table 10: Overview of qualitative data and intended analysis</b>		
<i>Objective</i>	<i>Variables of interest</i>	<i>Collection instrument</i>
1. Gather fresh data on adherence to the commitment device	Was the coach aware of the weight loss target? How involved were they in supporting the participant?	Responses from Food Monitor coaches
	How often did the participant see the commitment contract?	Interviews with Camden participants
2. Contextualising the average and heterogeneous treatment effects	Motivation, self-awareness, and examples of behaviour change	Interviews with Camden participants; responses from Food Monitor coaches
	Wider circumstances affecting weight management	
3. Exploring evidence of planner-doer interactions	Examples of individual deviating from target consumption goal and reflecting on this later, perhaps with anger or disappointment; or individual sticking to goal and expressing contentment and satisfaction	Interviews with Camden participants

### **6.1. Advantages of interviews**

Conducting interviews promised valuable opportunities for “an in-depth exploration of an aspect of life about which the interviewee has substantial experience, often combined with considerable insight” (Charmaz 2011, p.3). Specifically, the interviews were designed to explore how the participant felt about their weight loss outcome (relative to the standard 5% weight loss target); what behaviours they identified as having changed, if any; and, if they received a commitment contract, how they used it.

The closest alternative to interviews was conducting focus groups. Interviews were the preferred option because they offered the benefit of being able to probe responses, and for the respondent to open new avenues of discussion or close down those they were not comfortable talking about. The semi-structured nature of the interviews, and the fact that each one could be tailored to the respondent, made this this data collection method more attractive. One-to-one interviews also ensured privacy and confidentiality of responses, to mitigate against social desirability bias or ‘group think’, as well as creating a safe space for sharing personal reflections and information of a more sensitive nature. The disadvantage was the relative time-intensity of research effort on interviews, and this had implications for the number of interviews that could feasibly be undertaken.

## **6.2. *Number of interviews***

The literature offers a notoriously diverse range of prescriptions on what sample size is the correct one. An ex ante target of approximately 20 interviews was set. This had the advantage of being feasible given resourcing constraints, and promising a sufficiently good chance of reaching “theoretical saturation” given the relatively narrow set of objectives to the interviews and relatively homogenous population, in the sense that they all were overweight or obese, lived within the borough of Camden, and had been registered on the Shape Up programme. (Guest & Johnson 2006, p.75).

## **6.3. *Recruitment***

Recruitment was based on convenience sampling. Participants who expressed interest in a follow-up interview during the registration process were contacted either by email or phone (up to two attempts made) after their group programme concluded. Those who responded positively were invited to a face-to-face or phone

interview. No major implications were expected from offering both interview mediums, and offering the choice was judged a useful way of providing inclusive opportunities for Shape Up clients to take part in the interviews.

As would be expected with this sampling method, the group of interviewees was expected to over-represent those who completed the programme as easier to reach; those who were either more motivated or more successful at meeting their weight loss target and therefore inclined to discuss their experiences; and those with more free time, who may have been more able to share their time with the research project. Non-probability sampling of this nature has justifiable trade-offs. Tansey (2007) highlights its value when generalisability is not the key concern; rather the objective is to reconstruct a set of events. The aim of the Camden interviews was not to generalise from the sample of interviewees or to pursue causal inference, but to delve more deeply into the experiences and attitudes of those taking part in the Shape Up programme, including (but not solely) those who received a commitment contract. The disadvantages are the potential for selection bias, and this caveat is noted for chapter 7; however the qualitative analysis would still offer opportunities for triangulation, testing of new variables to operationalise theoretical concepts, and generating exploratory insights for future research. Section 7.2.5 below returns to the issue of sample selection within a wider discussion of the plausibility standards for qualitative analysis.

#### **6.4. *Topic list***

The interview topic list asked treated individuals if they recalled the contract, where they had placed it, how often they had seen it during the trial, and if anyone else was aware of it (full topic list in appendix section A.5). Responses allowed for a distinction to be made between low and high adherence to the commitment contract, which could be used to explore differences in outcomes.

This component of the analysis promised to make a contribution to the literature by exploring whether being committed to the commitment device could be traced through to more effective behaviour change and weight loss outcomes. The selective nature of interviews, however, entails that this analysis would not allow for causal inference, but would provide novel evidence on this heterogeneity pathway theorised in Chapter 3.

### **6.5. *Qualitative content analysis***

The interviews were to be recorded and manually transcribed using NVivo for Mac (version 11). Quantitative content analysis would require that I code these texts in order to produce a numerical dataset, perhaps through counting the frequency of a variable or word, or creating a binary variable for themes found or not found. It may be useful to carry out such a count for specific indicators such as, how many people remembered the commitment contract at the end of the Shape Up programme? However, a wholesale reduction of the data to numbers alone would be to lose richness and depth.

Each interview transcript relies on the participant's own recollection and characterisation of their experience, stated in their own words. Given the open-ended nature of questions in the interview schedule, close reading, comparison across entries, and careful interpretation of language and expression will be required to elicit the target information. Some of this target information may be manifest, others will be latent, and will require a close reading between the lines (Halperin and Heath 2012). Qualitative content analysis is much better suited to the task of "scrutinising the text, reading and rereading it, to identify and confirm themes" (Alaszewski 2006, pg 96).

## **6.6. Coding scheme**

Analysis of the interview texts is “a multi-step ‘sense-making’ endeavour” (DeCuir-Gunby et al. 2011, p.137). The process for data condensation and analysis began with a theory-driven coding scheme, set out in brief in Table 11. This was applied to three interview transcripts (15% of the expected total) to test the usefulness of the coding scheme and assess the need to generate further categories and tags. The approach therefore allowed for open coding on a small, random sample of data. Following this early manual coding exercise, the codebook was refined and tags tailored to the separate field experiments, then applied to the remaining interview transcripts and food journal entries. This revised codebook took the form advocated by Mayring (2000), setting out the name of the code, a definition, example text, and any rules for coding. At this second stage the qualitative analysis software NVivo will be used to apply the coding protocol to the full datasets.

The analysis will rely on narrative and quotations (Halperin and Heath, 2012). The standard of evidence is plausibility rather than probability (see section 7 below for further discussion of threats to plausibility). Any emerging patterns will be taken as indicative of actual underlying patterns, which could still be of considerable value and novelty in triangulating with the quantitative data and also in exploring the likelihood of the planner-doer framework operating as expected in real world behaviour change processes. Much of the qualitative analysis planned will be presented in-depth in Chapter 7, but where insights are immediately and concisely applicable to the statistical results presented in Chapters 5 and 6 they will be woven in accordingly.

**Table 11: Preliminary coding scheme for interview data**

<i>Objective 1: contextualising average and heterogeneous treatment effects</i>		
Meta-Theme	Possible tags	Examples from early interviews
What factors do they use to explain the weight loss outcome?	Learning from Shape Up Lifestyle changes Other health issues Job or family related issues Spurred on by CD	“My portion sizes were too big...I’ve reduced the size of my portion...I weigh food” “I started using a pedometer” “[the contract] was an element of [my] self discipline”
How helpful did they find the Shape Up course	Positive feedback on tutor or content Negative feedback	“I don’t think the class was really useful to be honest, but just having the regular meet-ups and weigh-ins was helpful, it helped keep you on track”
For treated group, how did they use the contract/coach?	Contract was salient Contract was not salient Coach was actively involved Coach was not contacted	“yes I remember [the contract]”...it’s in my desk”. “to be honest with you it wasn’t on the top of my mind, no”
<i>Objective 2: evidence of internal planner-doer interactions</i>		
Meta-theme	Possible tags	Examples
What behaviours do they identify as having changed?	Changes in content, regularity and size of meals Exercise more	“forward planning has been a big part of it for me ... we had a brunch and I actually went online and looked at the menu and decided, in advance, what I would have, which was very helpful”
Do they encourage/scold themselves for the choices made?	Having words to oneself	“think how well you’re doing” “I tell myself: ‘no’”
Do they have their own commitment devices to lock in their future choices?	No examples Examples	“I gave away my ice cream to my neighbour” “it’s just a matter of will really, when I set my mind to it I can be really good”

## **7. ENSURING INTERNAL VALIDITY AND PLAUSIBILITY**

The chapter has thus far completed two of its three objectives: firstly, section 2 justified the field experiment research design, and the enrichment of quantitative data with qualitative to provide a fulsome answer to the research questions and test the underlying theoretical framework. Sections 3 to 6 then provided a detailed overview of two field experiments and their analysis plans. However, any field experiment is subject to various threats to internal validity, since conditions cannot be fully controlled in the real-world settings they take place in. Sections 7 and 8 now address the third objective of this chapter: they review potential sources of bias in statistical and qualitative analysis, outlines the mitigation strategies put in place to address them, and address external validity concerns.

### **7.1. Internal validity in field experiments**

Experiments are often lauded for their ability to produce unbiased causal inference, but this result depends on a number of assumptions (Little & Rubin 2000; Torgerson & Torgerson 2008; Gerber & Green 2012). This section details the design choices and features of the two field experiments that aim to avoid and mitigate a series of threats to internal validity, in order to deliver unbiased treatment effects. Internal validity here is the “extent to which an experimenter can be confident that his or her findings result from experimental manipulations” (McDermott 2011, p.28).

A causal effect can be defined as “the difference between two potential outcomes, one in which a subject receives treatment, and the other in which the subject does not receive treatment” (Gerber and Green, 2012: 44). While this is not directly observable, experiments can provide unbiased estimates of the causal effect if three assumptions hold:

1. Random assignment: all units have a known probability between 0 and 1 of being assigned to treatment, and treatment assignment is statistically independent of potential outcomes.
2. Excludability: potential outcomes respond solely to treatment status, not to any indirect implications of assignment.
3. Non-interference: the potential outcome for observation  $i$  reflects only the treatment or control status of  $i$ , and not of other observations  $j, k, l$ , etc...

Ensuring excludability and non-interference allows the Stable Unit Treatment Value Assumption (SUTVA) to hold. Each of the three assumptions is discussed in turn below, and the section then turns to other sources of bias arising in the implementation of the trial, namely: non-compliance, the effects of pre-treatment, attrition, spillover effects, uncertainty, measurement error, the risk of type I errors in multiple comparisons, and researcher bias.

### ***7.1.1. Randomisation***

Assumption one is ensured through the random assignment of consenting participants to the treatment group. The Food Monitor field experiment employs a randomisation feature in the Qualtrics survey suite to ensure that participants are shown one of three possible survey pages at the end of the baseline survey, which determines which experimental group they belong to. The survey software was designed to randomly allocate participants to one of three experimental groups in the first phase, up to the point that a quota of 100 had been reached in the refund group. Beyond this, in the second phase, the software was instructed to randomise participants to one of two groups. Pre-testing indicated the software was conducting the random allocation suitably unpredictably.

As described earlier in the chapter, an externally imposed quota on the refunds meant the allocation ratio had to change during the trial from a one-in-three to a one-in-two probability of being



allocated to the comparison group. The randomisation phase each participant belonged to was clearly documented, and incorporated into the analysis plan in line with Torgerson and Torgerson's prescription: "if the allocation ratio is changed part way through the trial a variable must be included in a regression analysis to control for this" (Torgerson & Torgerson 2008, p.113). The ATE on the refund treatment is recovered from phase one data, while a separate analysis of the coach treatment is undertaken relative to the comparison group across both phases one and two.

In the Camden experiment, every effort was made to ensure that participants were randomised in advance using a random number seed in Stata. Where participants were not included in the client lists, a back-up strategy was to assign them to an experimental group on the day using a simple numerical rule relating to the participant's eight-digit date of birth. This pragmatic randomisation exercise has precedent in the literature, for example Giné et al (2010). Torgerson and Torgerson (2008) further highlight a case where an exogenously determined time stamp is used to undertake randomisation by a phone operator once a call has been taken, using a numerical rule to produce a digit one, two or three that determined group assignment; they judge this "procedure prevents any manipulation or potential subversion of the random allocation". In the same vein, the arithmetical rule used for randomising Camden participants on the day of a Shape Up class also prevented the risk of researcher or some other systematic bias in the allocation of treatments, and maintained a 50% allocation. The process is judged sufficient to ensure that assumption 1 holds.

### **7.1.2. Excludability**

A key risk to the excludability assumption in the Food Monitor experiment is how clients will perceive the refund of the monthly fee. It aims to dismantle temporarily the financial commitment that paying customers are theorised to experience. However, an alternative interpretation is that the refund is a gift or voucher, aiming to generate brand loyalty. This kind of financial incentive may serve to make the product more salient, motivating clients to work harder towards their target. The overall effect on the data is likely to be a combination of two countervailing influences – one encouraging less behaviour change, one encouraging more – which might manifest itself as a treatment effect close to zero. To mitigate the interpretation of the refund as a reward, the treatment message is written in deliberately brief and non-celebratory language, dampening the association of the refund as a gift.

Excludability in the Shape Up experiment is largely secured through the clients experiencing the same programmes whether they are in the control or treatment group. There are two other factors that may influence this assumption. Firstly, there are eight different group tutors facilitating the weekly sessions, and this could have some effect on overall weight loss performance. Since all tutors have received the same briefing, use the same Shape Up manual, and follow the same week-by-week course programme, there is no strong reason to expect this would undermine assumption 2. However tutor and group data will be captured as a control variable to understand whether there are any associations with weight loss outcomes. In the Food Monitor experiment, all participants have access to the same range of facilities on the Food Monitor website and app, and the only known difference in their experience with the service is treatment status.

Treatment in the Camden experiment is assigned in person by the researcher. There is a modest possibility that this interaction compounds the effect of the written contract, adding a further layer of reputational commitment. While this is unavoidable, the researcher's input is early on in the 11-week programme and limited to just one visit to the group; and interaction is largely scripted in advance with great efforts to ensure consistency across all conversations with participants. On balance, this design feature was not deemed to be a significant risk to excludability.

### ***7.1.3. Non-interference***

Assumption three on non-interference could be threatened if participants openly discussed their experiences in the trial, and the commitment devices offered. In the Shape Up groups, such conversations could take place during the weekly meeting; and the Food Monitor app has a discussion forum for users to post notices and contribute to conversations. Those without a commitment device (or in the Food Monitor case, a refund) may feel disappointed or resentful, and this in turn could affect their weight loss performance and contaminate the data.

Interference through informal communication was judged to be a low-risk issue in the Food Monitor experiment. There are a maximum of 100 clients who might be assigned to the refund treatment group, relative to thousands of Food Monitor users. The trial protocol (information sheet and consent form) asked that participants refrain from discussing their experiences on discussion forums, which is the main communication method for the online community. The low-key phrasing of the treatment letter for treatment group 1 was also designed to help dampen any sense of the refund being a membership bonus.

Since registration and treatment assignment was carried out in person in the Camden experiment, it was possible that clients would note if others were offered the commitment contract, and this could affect their own motivation, biasing the treatment effect. To mitigate this risk, as described above, registration for the study took place in a private space and it was not possible to see who took home a commitment contract. The opportunity to discuss the trial with other group members was also limited: clients gather for one and a half hours a week, and the group session is actively facilitated by the tutor who has a busy schedule of topics to cover.

There remains some risk in both trials that contact offline or outside the group sessions leads to some interference across treatment groups, but this was judged sufficiently low as to be acceptable. Following Torgerson (2001) it is reasonable to accept the risk of contamination and deal with it by ex post analytical techniques such as the complier average causal effect estimator. Interviews with participants after the Shape Up programme concludes allows for some investigation of who, if anyone, the treated group discussed their commitment contracts with; and if there is any evidence of contamination with the control group (discussed further under spillovers below).

#### ***7.1.4. One-sided non-compliance***

Compliance is a description of “whether the actual treatment coincides with the assigned treatment” (Gerber & Green 2012, p.132). Experiments run the risk that “the assigned treatment group is no longer the same as the group that is actually treated”. Failure-to-treat might arise if participants have not received the treatment or registered it, they do not want to continue in the trial due to flagging interest, or due to some aversion to the treatment. One-sided compliance is defined very simply in both field experiments as the take-up of the treatment when offered, soon after the baseline

registration process. With the online experiment, anyone who turns down the offer of the refund or refuses to name a coach was to be identified as non-compliers. In the Shape Up experiment, anyone who refused to sign the commitment contract was identified as a non-complier.

One-sided non-compliance is being mitigated *ex ante* through various design features. The trial is entirely voluntary, and it is assumed that those who sign up will be motivated enough to accept the treatment. In both experiments, taking up the treatment offer has been made as simple as possible. With the Food Monitor experiment, the financial commitment group continue with the service as normal; and the limited commitment group simply tick a box in the survey to accept the treatment and trigger a refund. The reputational commitment group are asked for additional information, namely the contact details and name of their nominated coach. Participants could be averse to this treatment, perhaps worried about data protection, or because of the additional effort involved. Complier sub-group analysis will be undertaken if non-compliance rates are high. With the Shape Up experiment, no additional input to the trial is needed beyond accepting the contract, signing it, and taking it away. The treatment itself is non-intrusive, easy to administer on the spot with no further effort, cost, or personal information required.

Despite these mitigation strategies, there will be some participants who refuse to accept a treatment, and in particular the reputational commitment device treatment is likely to have the highest refusal rate. In other studies of commitment devices, take-up has been as low as 11% for a deposit contract (Giné et al. 2010), and this points to considerable self-selection effects within the experiment that lead to downward bias on the average treatment effect. In such a situation it would be challenging to answer the question ‘how does this treatment affect the outcome?’, but still feasible to tackle the narrower question of ‘how does the randomised

offer of this treatment affect the outcome?'. The rich baseline data for both experiments will allow for in-depth exploration of what determines take-up of the treatments. It will remain possible to uncover causal effects using the intent-to-treat estimator; and contrast this estimate with the treatment-on-treated estimator, also referred to as the complier average causal effect (CACE) or local average treatment effect (LATE) (Angrist & Pischke 2009; Gerber & Green 2012). The CACE technique is expected to be of particular relevance to the Food Monitor experiment where non-compliance with the reputational commitment treatment appears most likely.

The definition of compliance applied here makes the assumption that everyone who takes up the treatment experiences it in broadly the same way. For example if participant A names a coach and participant B names a coach, it is assumed that they might have similar degrees of interaction and engagement with their coach, which leads to a similar degree of reputational commitment experienced. Or, if participant A and B both sign the commitment contract and take it away to their homes, it is assumed that the contract retains a similar degree of salience throughout the experiment. In reality there may be partial compliance that “dilutes the effect size because there is less contrast in exposure to the treatment” (Glennester & Takavarasha 2013, p.292). No particular mitigation strategy was adopted here to address this particular risk. On the contrary, adherence is modelled as an explanatory factor, and is fully expected to vary as detailed in chapter 3. In order to gauge whether there is indeed a diversity of experiences in how the treatments are experienced, the research design relies on qualitative insights from interviews and follow-up emails with the weight loss coaches to assess how strongly or weakly the treatment may have been experienced; in other words, what was the strength of the treatment ‘dosage’ in practice, and how adherence may be related to overall effectiveness of the commitment device in changing outcomes.

### **7.1.5. Attrition**

Attrition is “the failure to collect outcome data from some individuals who were part of the original sample” (Duflo et al. 2007, p.58). Any weight loss study expects to suffer some degree of attrition, particularly when relying on self-reported weights (Little & Rubin 2000, p.135). There was little evidence available to suggest that attrition would be especially problematic with the Food Monitor trial, particularly over a 4-week period, given that participants were paying for the tool and so it could be expected they would use it at least one more time after registering for the study. The advertisement for the trial as likely only to be seen by those actively using their Food Monitor account, and the nature of the registration process was such that it was more likely to be concluded by those expecting to use it at least one more time over the trial.

It was therefore expected that attrition will be low and benign, while noting that if attrition was related to treatment status, this would threaten the validity of causal estimates by unwinding the effects of randomisation and introducing selection bias. One *ex ante* mitigation strategy was to collect the weight outcome data through the firm, on the basis that since all treatment group members have access to the online services for the remainder of the trial it is unlikely they would disengage entirely, and even those in the limited commitment group who were theorised to be less committed to their health behaviour change goal were plausibly going to return to the tool to provide one self-reported weight entry over the course of 12 weeks (since their refund, and hence their lower theorised commitment, only applied to the first four weeks). A second mitigation strategy was to supplement the health outcome data with self-monitoring behaviour data. Limited usage and log-ins would still provide useful information about how engaged individuals were with the Food Monitor tool, allowing for comparison of health behaviours

and robust analysis across experimental groups even in the context of participants failing to return to the website.

In addition to the mitigation strategies outlined above, ex post strategies were also considered. The weight loss literature commonly uses techniques such as ‘last observation carried forward’ or ‘baseline observation carried forward’ in order to address missing outcome data; based on the assumptions that weight loss did not change since the last self-report, or that it did not change at all over the course of the trial, respectively (Jolly et al. 2011; Elobeid et al. 2009). Inverse probability weighting is advocated when data is missing independent of potential outcomes (Gerber & Green 2012, p.222).

Based on previous Shape Up programmes, a moderate degree of attrition is expected, which may not be benign. Dropout from the programme is fairly common, with only 50% of clients completing the full 12-week programme in some cases. If attrition is related to treatment status, this will threaten the validity of causal estimates by unwinding the effects of randomisation and introducing selection bias. Early diagnostics will report attrition levels by group, comparing baseline data of attritors with non-attritors (Duflo et al, 2007: 59). Best practice advocates “regular monitoring of missing data and enhanced participant contact” during the trial itself, but this is beyond the scope of the trial (Dziura et al. 2013, p.356). However, prompt identification of missing outcome data and speedy follow-up with those who have dropped may yield a self-reported weight update to mitigate attrition. As discussed above, once the full extent and nature of attrition is analysed, appropriate statistical methods will be applied to uncover a valid treatment effect estimate.



### **7.1.6. Spillovers**

Spillovers can be identified “when the effect on those receiving treatment produces a secondary effect on those who are not treated” (Glennerster & Takavarasha 2013, p.354), but are distinct to the contamination issues discussed above (under the excludability precondition for SUTVA to hold). A spillover is far more likely in the Shape Up experiment than Food Monitor. For example, if the treated participants begin to show good progress towards their weight loss goal, this may through indirect means create a stronger push for their non-treated group peers to accelerate their weight loss efforts; either through a sense of competition, or because of a collegiate sense of group progress and being in it together. Such an upward spiral of weight loss progress would be good news to the participants and group tutor, but if it means that the difference in performance between treated and non-treated participants is diluted, then this could give generate a downward bias on the average treatment effect estimate. Ignoring potential spillover effects implies that the intervention’s effect on the treated is underestimated and the effect on the untreated is not measured at all (Angelucci & Di Maro 2015, p.3).

Although the research questions do not set out to test the presence and magnitude of spillovers, it is nonetheless possible to apply a ‘randomised saturation’ analysis as a means of checking whether spillovers are present, and whether they pose a challenge to the internal validity of the average treatment effects estimated, based on “the extent to which program effects are driven by the percentage of individuals treated” across groups (Baird et al. 2014, p.2). A similar method is applied by Hassan and Lucchino (2015) in a study examining the effect of solar lamp provision on educational outcomes. They report positive spillovers at the classroom level, with the provision of lamps improving the grades of control group students as well as the treated students. Such spillovers are useful to identify both because they help to estimate treatment effects with

greater accuracy, and because they indicate additional benefits from the intervention from a programme delivery perspective that would be valuable for policy makers. The presence of spillover effects will be investigated in the Camden trial applying these methods.

#### ***7.1.7. Uncertainty***

As Gerber and Green point out, “while experiments are unbiased, they are not necessarily very precise” (Gerber & Green 2012, p.55). It will be important to identify and analyse the degree of uncertainty around the coefficient estimates. To mitigate large standard errors, I aim to dampen variance around the dependent variable by measuring outcomes as accurately as possible and controlling for observable differences between treatment and control groups. Some baseline variables allow for greater measurement precision than others – for example, age is measured in number of years in field experiment 2, because it is derived from the registered date of birth in Camden’s administrative records. The baseline survey used to elicit age in field experiment 1, however, offers an age grouping. This was designed to make the survey more user-friendly with drop-down menus, while also addressing the possibility of respondents not wanting to set down their specific age or date of birth. It does however bring a lower degree of accuracy in measuring age as a covariate.

#### ***7.1.8. Measurement error***

Precision and accuracy of causal inference relies on the model actually measuring what it intends to. The primary outcome variable for the Food Monitor experiment is self-reported weight change, and in order to calculate body mass index self-reported height is also elicited. Such self-reported measures could introduce a source of error. The Health Survey for England 2011 compared self-reported estimates with interviewer measurements for height and weight,

finding that, “mean height estimates were consistently higher, and mean weight estimates consistently lower than interviewer-measured estimates”, leading to under-estimation of BMI (Sutton 2012, p.1). The 2012 HSE confirmed this pattern, finding that the average self-reported height was higher (particularly for men), and the average self-reported weight lower (particularly for women), than the measurements taken by the interviewer (Moody 2013).

These findings raise the need for caution in interpreting self-reported measures. However there are grounds for proceeding with self-reported data. Firstly, if optimism and social desirability bias affect all self-reported measures consistently, then the difference in weight recorded over the experimental period remains a valid outcome measure and remains comparable across experimental groups. Secondly, self-reported weight is a voluntary measure, and it is plausible to assume that anyone who uses the online service to keep track of their weight will do so on the basis of a reasonably accurate weight reading they have taken. Thirdly, the degree of error is relatively small – the 2012 HSE reports at most 3.4% for women and 1.9% for men estimating their weight – and unlikely to radically alter the findings of the statistical analysis.

The issue of inaccurate readings is less of a challenge in the Camden field experiment. Shape Up tutors take weight readings from clients using the same digital weighing scales, at the same time each week, and clients are encouraged to remove shoes and heavy clothing to maintain consistency as far as possible. Height readings are taken at the start of the programme. This goes a considerable way towards eliminating the kind of measurement errors that might arise from self-reported weight data. Where clients fail to attend the final sessions and a self-reported weight reading is taken, the risk of optimism bias remains. However this is seen as a necessary trade-off to avoid missing outcome data, which would generate more challenging problems for causal inference.

Secondary outcome measures in both trials are deemed to be robust: in Camden, attendance and completion rates are documented by group tutors and Food Monitor's automated reports gather data on self-monitoring. Amongst the baseline variables, self-reported diet and exercise patterns also rely on respondents for their veracity, but they are deemed fit for purpose on the basis that severe self-deception is not likely amongst participants who have voluntarily signed up for a weight loss programme; and the relative anonymity of the surveys used to provide this information should mitigate against social desirability bias, where participants may want to embellish their responses because they know they are being observed (either by myself or group tutors).

#### ***7.1.9. Pre-treatment***

Gaines and Kuklinski highlight the potential issue of pre-treatment of participants, which would imply that a randomised control trial is capturing “not the discrete effect of treatment, but the average marginal effect of additional treatment conditional on an unmeasured level of real-world pre-treatment” (Gaines & Kuklinski 2011, p.446). The potential for unmeasured and unobserved pre-treatment seems especially pertinent in the weight loss sector, where most individuals have likely been the target of public health campaigns about losing weight by staying active, and improving diets (such as the Change4Life ‘sugar swaps’ campaign). The treatment this thesis is concerned with is commitment devices, and it would be easy to argue that most people have applied some informal commitment strategies, such as a personal rule, to change their weight, improve their diet, or do more exercise. The true extent of such pre-treatment is difficult to measure in the context of any field experiment; but baseline survey questions may help mitigate this information gap, and qualitative follow up with Camden participants

will also play a valuable role. Results will be interpreted in light of the point Gaines and Kuklisnki make.

#### ***7.1.10. Multiple hypothesis testing and false discovery risk***

The research design has set up investigation into two different heterogeneity pathways in each experiment to answer Research Question 2. The statistical analysis may be susceptible to false discovery of treatment effects due to multiple hypothesis testing (Gerber & Green 2012, p.300). Fink et al. (2014) report that it is increasingly common for field experiments to involve sub-group analysis and modelling of interaction effects between the treatment and baseline covariates, with 34 experiments identified in a survey of economic journals over 2005-10. Some articles estimated more than 10 heterogeneity pathways, but none were found to correct these estimates for multiple hypothesis testing, potentially undermining statistical inference.

The Bonferroni correction is a popular technique used to minimise the risk of erroneously accepting a hypothesis that is actually false, by raising the threshold of the p-value needed to signal statistical significance. Specifically, the conventional value for alpha, 0.05, is divided by the number of hypotheses to be tested. P-values reported for coefficient estimates now need to be lower than this new value, alpha dash, to imply statistical significance. The benefits of this method in terms of reducing type I errors can be counteracted by the higher probability of type II errors, or falsely rejecting a hypothesis due to reduced power (Duflo & Kremer 2008). While simple and easy to apply, the Bonferroni correction may be overly conservative. This disadvantage cannot be overlooked in the Camden and Food Monitor experiments with their relatively small samples. A related technique put forward by Benjamini and Hochberg (1995) offers a less conservative approach, where “critical thresholds for a

given false discovery rate are scaled down by a constant factor proportional to the number of hypotheses tested” (Fink et al. 2014, p.50). This means that additional hypotheses are subject to increasingly higher standards testing, rather than applying a blanket rule across all hypotheses as a set (Coppock 2015). This procedure may lead to a higher false discovery rate than the Bonferroni correction but is less taxing on power, and is the preferred approach in chapters 5 and 6.

### ***7.1.11. Researcher bias***

A number of issues fall under this category, for example, Torgerson and Torgerson highlight the possibility that “investigators consciously or unconsciously bias a trial by reporting events more conscientiously in one treatment arm compared with the other arm” (2008, p.57), leading to reporting or detection bias. Boutron et al highlight the importance of blinding researchers of treatment status to ensure subjectivity does not creep in to outcome measurement, but accept that blinding participants is not always “practicable nor possible or even desirable”, as is the case with the field experiments described in this chapter (2010, p.117).

Reporting to CONSORT standards aims to ensure transparency on trial processes even if all sources of bias cannot be completely eliminated. To some extent blinding was possible in that Shape Up tutors did not know which of their class members had received a contract, or who were even part of the trial. Given the small administrative resources available to run the trial, it was not possible to arrange a concealment mechanism or blind the investigator fully from treatment status. However, outcome data were measured using objective and externally verified sources (for example the class registers from Camden, self-reported data from Food Monitor’s systems) and beyond the recruitment stage, Camden participants had no further contact with the investigator until the

Shape Up course had concluded. These measures sufficiently address the potential for researcher bias.

## **7.2. *Plausibility of qualitative analysis***

The standard of evidence for qualitative analysis is plausibility over probability: “the results of a study are valid and reliable to the degree that they are plausible to others” (Halperin & Heath 2012, p.328). Key principles that underpin the qualitative data collection and analysis are: transparency, coder stability, and minimising bias arising from the researcher, the respondent, and sample selection. These issues are discussed in turn below.

### **7.2.1. *Transparency***

Transparency and meticulous record-keeping will be pursued. Following McLellan et al (2003)’s recommendations on data preparation and transcription, each interview transcription will have a coversheet identifying the location, date, time and interviewer, participant background details, linked audio file, and unique identifiers for the interviewee which allow them to be identified in the dataset while maintaining anonymity in line with ethics requirements. Anonymised interview transcripts were carefully recorded and can be made available on request.

### **7.2.2. *Coder stability***

In order to evaluate the reproducibility of the qualitative content analysis, a second coding exercise was planned three months after the initial exercise on a random sample of qualitative data, to assess how similarly the data is coded after a period of time. It was not feasible to employ a second coder, however the second coding exercise should highlight how stable the coding results are, and where there may be any differences of opinion. The exercise aims to

allow for a self-evaluation of whether the coding scheme needs further refinement. The preliminary coding scheme is set out in Table 10 above, and the refined coding scheme is presented in Chapter 7 along with a reflexive discussion of how stable this was found to be over time.

### **7.2.3. Interviewer bias**

No researcher bias, beyond the John Henry or Hawthorne effects highlighted above, were expected from the qualitative analysis in the Food Monitor experiment. However, the interviews in the Camden experiment required planning to avoid researcher bias creeping in to the interviews and affecting the content of the discussion. A topic list was prepared in advance to provide consistency in questions for all interviewees, and ensured that leading questions were avoided. To some extent the interview was designed to follow the narratives provided by the respondents, but any follow-up or probing questions were to be linked back to their own responses and framed carefully to avoid giving the impression to interviewees that there was a 'right' answer that I was searching for.

### **7.2.4. Reliability of self-reported information**

Self-reported information, as discussed above under measurement error, can sometimes be subject to error and inaccuracy, particularly where it is based on recalling events and experiences from some time in the past, or if the respondent edits their narrative due to social desirability bias. For example, if they are embarrassed about some behaviours these may be under-reported, while other more positive behaviours may be emphasised. With the anonymous food journal entries in the Food Monitor experiment, this was deemed to be a minor concern, since the participants were essentially writing notes to themselves and there was no interaction with any external actor. In the Camden trial, however, it was clear



that a one-to-one discussion on health behaviours might create “incentives to misrepresent information about themselves to researchers” (Starr 2014, p.256), since the respondents are aware of the Shape Up guidance they were supposed to be following, and honest responses would often require their being able to admit they deviated from this guidance on occasion.

To address possible error arising from poor recall, the interviews with Camden participants were designed to take place as soon as possible after the final Shape Up class, while the programme and final reflections were uppermost in their minds. To tackle social desirability bias, participants were to be offered either phone or face-to-face meetings, and apart from logistical preferences this allowed for participants who might be naturally shy to avoid having a more intense conversation about themselves in person. Further, it was to be explained at the start of the interview that there were no right or wrong answers, that my primary interest was my own research and not reporting back to Camden or the tutors, and that the purpose of the interview was to understand in their own words how they found the Shape Up experience.

In these ways, the interview was framed to instil confidence in respondents that they were not being tested. Rather, it was conducted in a generally sympathetic manner to prevent any participants from feeling judged, following Halperin and Heath’s guidance to put the interviewee at ease, “never [show] any disapproval of the information received during the interview” and probe sensitively (2012, p.268). Further, the interviews were held in a location where the individuals were able to speak freely about themselves, which was an advantage of offering telephone as well as face-to-face interviews, and transcripts were anonymised to ensure confidentiality.

### **7.2.5. *Sample selection bias***

Given the convenience sampling of interviewees, a degree of self-selection was expected; for example, those who were doing well against their weight loss targets may have been more inclined to take part in an interview. It is unlikely that the sample of respondents in both qualitative datasets is representative of the full sample – a challenge of external as much as internal validity, and the strategy for dealing with this involves: careful interpretation of the results, avoiding generalising where it is not warranted, and looking carefully at the characteristics of the respondents to reflect upon the degree of likely sample selection bias. Chapter 7 will apply each of these techniques to the qualitative analysis. For example, insights from one interview will be cross-referenced and triangulated with other responses to avoid being unduly swayed by anecdotes or outliers.

## **8. EXTERNAL VALIDITY**

The previous sections have outlined the design of two field experiments forming the empirical strategy for this thesis, and discussed the potential threats to internal validity as well as mitigation measures to ensure robust causal inference in chapters 5 and 6. The following section turns to the issue of external validity, which is usually understood as a question of how well the results of an experiment can be generalised to the population as a whole.

Field experiments are often cited as offering greater external validity than the artificial setting of laboratory experiments, where many factors can be tightly controlled. As a trade-off for realism, however, field experiments (particularly those that rely on partners for implementation) are likely to be affected by many context-specific and ex ante unpredictable issues. In considering how well the results of any one study can be generalised to broader theoretical and policy questions, two caveats must be borne in mind. Firstly, it does not follow that sample ATE can be extrapolated to a population ATE (Gerber & Green 2012, p.357). Secondly, it would be risky to claim that internal validity implies external validity (Deaton & Cartwright 2016, p.53). The aim of a sound experimental design is to address threats to internal validity, but the separate question of how well the estimated causal effects can apply to other settings and problems relies partly on whether the results are somehow an artefact of the experimental setting and design, or if they truly capture the effect of the intervention that is being tested.

### **8.1. Challenges to achieving external validity with field experiments**

Simply being in a research project, and the awareness that they are being monitored and studied may be enough to make people behave differently than they would with that intervention under normal circumstances; for example, the treatment group working harder than usual (sometimes referred to as the Hawthorne effect), or the comparison group competing with the treatment group (the John Henry effect) (Glennister & Takavarasha 2013, p.317).

Even if participants behave normally, any extra attention given to the implementation of the intervention by investigators or programme staff delivering the services may generate results that do not transfer well when continued beyond the research project.<sup>41</sup>

Thirdly, if the trial recruits participants are in some way systematically different to the wider target population they are supposed to represent, this too may undermine how well the results can be applied in a wider setting beyond the trial. Recruiting a widely representative pool of participants is therefore seen as a way to enhance external validity (McDermott 2011). Finally, if there are context-specific features that drove the field experiment results, this too may mean that the findings do not travel well to other contexts and delivery environments.

### **8.2. Design features of the trials to improve external validity**

The design of both experiments outlined in this chapter has incorporated a number of features to overcome these issues. To minimise the sense of being observed, the online experiment relied

---

<sup>41</sup> Such factors might explain why seemingly successful pilot phases that are used to run field experiments are not matched in their performance when the programme is delivered at scale.

on a single interaction with the participant during the baseline survey and registration process. Beyond this point, the participants were asked to continue with the service as they normally would. Although the information sheet specified that data would be made available for research purposes, there was no follow-up built into the experiment, and outcome data was collected unobtrusively behind the scenes by the partner firm who had full access to the online accounts. In these ways, contact with the trial as a research endeavour was minimised, with one exception: brief follow up with a selection of coaches nominated by those offered the reputational commitment device. The participants in this treatment group were indeed supposed to feel like they were being monitored – by their coach, not by the research project – to test whether this intervention promoted weight loss.

Other measures aimed to minimise the possibility of researcher bias affecting the strength of weight loss outcomes and health behaviour change, for example through more attention or a different service for treated individuals. With the intervention being delivered through the online medium, it was not possible that participants in some experimental groups would have been treated any differently by the service, because there was very little direct interaction with users; and what contact there was with the Food Monitor service was initiated by users. Although the partner firm were aware of which members had been assigned to the limited commitment group, as they managed the refunds for these members, the nature of the health programme context made it unlikely that bias could have affected the trial from this source.

In the Camden field experiment, as the investigator I was visible to the participants at just one class early on in the programme for registration and recruitment, with no further follow up until the end of that particular 11-week programme. Tutors did not refer back to the trial during their classes, which were closely planned to follow the Shape Up syllabus. These steps aimed to minimise the sense of

being observed in their progress, beyond the normal monitoring processes in the Shape Up programme such as the weekly weigh-ins and class discussions.

Further, tutors were blinded to the treatment status of their participants, making it unlikely they would have singled out any class members for special encouragement or support. Their own incentive structure set by the Camden Active Health team was to maximise the number of participants who completed the course and met their weight loss goals, and this would have been consistent across all groups, and unaffected by the parallel presence of the field experiment.

### **8.3. *Generalisability of the sample***

The field experiments do have a selection issue in the sense they are only recruiting from a sampling frame of people who want to lose weight and who largely need to lose weight for health reasons. But since the commitment device intervention aims to target people who are overweight and obese, and aiming to reduce their BMI to a healthier level, the question of generalisability is less about whether the field experiment results can be applied to the general population, and more about whether they can be applied to the sizeable group of people – 62% of the population in England – who are overweight or obese (Health Social Care Information Centre 2015).

A second selection issue is based on whether those who are part of the sampling frame are more motivated than most people who are obese or overweight. This is a valid criticism of the design, as both field experiments draw on a client base made up of people who have voluntarily signed up for weight loss support. They not only understand that they have or are at risk of health issues, they also make the time or monetary investment to take action. From this perspective, the results may not be generalizable to the full set of

people who have a higher than normal BMI, and only to those within this group who also want to make a change in their lifestyles to tackle this.

Evidence suggests that this applies to a majority of people who are overweight or obese. Data from the US National Health and Nutrition Examination Surveys (NHANES99) suggests that “almost all obese men and women consider themselves to be overweight and would prefer to weigh less” (Ruhm 2012, p.788), and as BMI increases so to do weight loss attempts. Among males 39% of overweight and 51% of obese males attempted weight loss, compared to 12% of those with a healthy weight; and among women the proportion rises to 64% of overweight and 69% of obese, relative to 39% of those with a healthy weight (Ruhm 2012, p.789). These figures highlight that even with a sample selection issue, the results of the field experiments are likely to be relevant to a significant number of people.

#### **8.4. *Integration with research findings***

External validity is not only understood as an issue of generalizability. McDermott argues that no single study is ever able to be large enough or broad enough to offer generalizability. Instead, “external validity results primarily from *replication* of particular experiments across diverse populations and different settings, using a variety of methods and measures” (McDermott 2011, p.34). In this sense, the two field experiments conducted as part of this thesis contribute to the external validity of the existing evidence base on commitment devices for weight loss set out in chapter 2. The external validity of the experiments themselves can be enhanced up to a point through careful design, but arguably the larger test is whether replication of the experiment produces consistent results in different settings (Gerber & Green 2012, p.350). The emphasis this thesis places on heterogeneity of treatment effects helps ensure that the external validity of the results will be carefully defined in terms of the

likely beneficiaries of commitment devices in public programmes, and the specific circumstances under which commitment devices can promote weight loss.

To recap, this section has discussed the factors affecting the external validity of findings from any single experiment. A number of mitigation measures have been put into place to avoid well-established sources of bias such as the Hawthorne Effect, the John Henry Effect, and researcher bias. Ultimately, the value of the two field experiments to the wider field will depend on further research and replication, and this will be aided by careful and transparent reporting of trial protocol and statistical analysis. The following section summarises the contribution of this chapter to the thesis.



## **9. RESEARCH DESIGN SUMMARY**

To provide robust answers to the research questions and uncover the causal effects of commitment devices, this chapter has constructed a research design centred on two field experiments. Kinder argues that “all methods are fallible. None can provide a royal road to truth”, and that “dependable knowledge is grounded in no single method, but rather in convergent results across complementary methods” (2011, p.527). In this vein, the field experiments are designed to combine the strengths of quantitative analysis in providing robust causal inference with the nuance of qualitative analysis.

This enriched field experiment design offers the opportunity to isolate robust estimates of causal effects (research questions 1 and 2); generate new data on adherence to commitment devices (for research question 2); enhance interpretation of statistical results through a deeper understanding of how the trials and participants’ weight loss journeys; and test the theoretical assumptions of the planner-doer model (chapter 3) that gave rise to the hypotheses framing this dissertation. Each of these opportunities represents a contribution to the scholarly debate and our understanding of how commitment devices work.

An overview of the two field experiments is set out in Table 12 below, highlighting their diversity as well as their shared features. Notably, both trials are nested within weight loss programmes, and have a focus on weight loss as a primary outcome variable. Both trials also allow for investigation into health behaviour change, going beyond the conventional focus on exercise (see chapter 2) to consider self-monitoring and attendance at weight loss classes – both key behaviours associated with weight loss success, and important outcomes in their own right. Analysis of heterogeneity pathways are shared across both trials, with qualitative analysis of adherence and

quantitative analysis of myopia in both. Beyond these similarities, however, the two trials offer rich variation in their treatments, mode of delivery, and qualitative analysis.

<b>Table 12: Key features of the field experiments</b>		
	Food Monitor	Camden
Target sample size	364	170
Programme format	Online	Group meetings weekly
Location	UK-wide	North London
Length of trial	12 weeks	11 weeks
Treatment(s)	(i) Subscription refund (ii) Coach nomination	Commitment contract to oneself
Outcome variables	Weight loss Self-monitoring	Weight loss Class participation
Sub-group analysis:		
Myopia	Yes	Yes
Sophistication	-	Yes
Adherence	Yes	Yes
Qualitative data	Coach email survey	Interviews
Timeframe	July 2013 – Feb 2014	Jan 2014 – Mar 2016

The Food Monitor experiment will be run online with a larger target sample of 364, reflecting the more ambitious aim to test two types of commitment device. The financial commitment device is the monthly subscription fee paid by clients of the Food Monitor service, while the reputational commitment device is an added invitation to nominate a coach who supports their weight loss goals and would be able to verify progress after four weeks. A relatively smaller component on qualitative analysis relies on email surveys with coaches focusing on generating data on adherence.

Addressing research question 2, the Food Monitor experiment investigates present bias while the Camden experiment investigates sophistication. The latter aims to recruit a somewhat smaller sample of 170, to be allocated equally across two experimental groups. The treatment is a milder form of reputational commitment device in the form of a contract to oneself. A larger qualitative component was planned in the form of approximately 20 interviews after the trial,

with the aim of generating data on adherence, contextualising the statistical results; and a more ambitious objective of testing the theoretical assumptions of the planner-doer model set out in chapter 3.

In summary, this chapter has made three contributions to the overall thesis. Firstly, it has made the argument for an enriched field experiment research design; secondly it has provided a detailed description of the two experimental designs in line with CONSORT standards; and thirdly it has identified threats to both internal and external validity, and various strategies adopted to mitigate these risks. The next chapters present the results and analysis from the field experiments: beginning with the Food Monitor trial in chapter 5 before turning to the Camden trial in chapter 6, and drawing together insights from detailed qualitative analysis in chapter 7.

---

BLANK PAGE

---

**Chapter 5**  
**RESULTS AND ANALYSIS (1):**  
**Coaches, Refunds, and**  
**Commitment Overload in the**  
**Food Monitor Experiment**

---

## **1. INTRODUCTION**

The dissertation has thus far presented two research questions, a theoretical framework, and a research design to test its hypotheses (chapters 2, 3 and 4 respectively). The planner-doer model informed a new framework to analyse health behaviours as intertemporal choices, and its main predictions are that commitment devices can promote health behaviour change and weight loss (research question 1); and these effects will vary across people (research question 2). Two field experiments were designed and implemented to test these predictions, and this chapter presents the results of the first experiment.

Nested within an online weight loss service called Food Monitor, it tests the effect of two distinct commitment devices on weight loss and self-monitoring behaviours: a financial commitment device in the form of a premium payment, and a reputational commitment device in the form of a mild public pledge to one other person. Quantitative data will be used to derive average treatment effects (hypotheses 1 and 2) and heterogeneous treatment effects based on how present-biased the individuals are (hypothesis 4). Qualitative data is gathered to examine whether commitment device effectiveness is related to how well an individual embraces it (hypothesis 5), and detailed analysis is set out in chapter 7.

The chapter first provides an overview of the implementation of the field experiment, including recruitment, randomisation, and balance checks in line with CONSORT reporting standards (see appendix section A.4 for full checklist). Section 3 then presents descriptive statistics of the sample. Section 4 discusses how quantitative outcome data was collected, investigates attrition rates, and outlines statistical strategies to address potential attrition bias. It also briefly summarises how the qualitative data for the reputational commitment device was collected and coded.

Section 5 presents the average treatment effects to inform answers to research question 1. The results challenge theory: the removal of the financial commitment does not shift the underlying commitment to the health goal; and the application of additional reputational commitment causes lower weight loss rather than greater. Section 6 addresses research question 2, and finds some evidence of heterogeneous treatment effects based on the degree of present bias reported by the individual.

Section 7 discusses the unexpected results around the reputational commitment device. The findings raise new questions about the theoretical framework's assumption that more commitment leads to greater behaviour change; rather, the data suggest that commitment overload might occur when a reputational commitment device is overlaid on existing financial commitment devices. This finding that 'less is more' is new to the literature on commitment devices. The trial provides a testing ground for new operational measures of short-termism, and underscores the importance of design features of commitment devices, given the potential for negative interactions between reputational and financial commitment devices, which has not been reported in the literature to date.

Section 8 concludes by highlighting two important contributions to the thesis. The first concerns the contrast in treatment effects with those reported for deposit contracts in the literature, a distinct type of financial commitment device (introduced in chapter 2), and what this may say about the nature of intrinsic commitment to health goals that it is unaffected when the financial commitment is dismantled temporarily. The second contribution from the field experiment is the test of two new measures of short-termism and myopia, a monetary time preference measure and a health attitudes measure, which proxy the doer sub-self in the

decision-making process. Both issues raise new insights for the Analytical Framework and wider scholarly debate.

## **2. FIELD EXPERIMENT IMPLEMENTATION**

### **2.1. Recruitment**

The trial launched on 26 July 2013 with the online survey accessible at the Food Monitor member pages. It was marketed through emails, Facebook and an online post visible to members who had logged in to the website. Taking part in the survey and trial was entirely voluntary, not linked to any aspect of the Food Monitor membership, and no financial incentives offered. The survey was live for four months and during this time, 435 surveys were initiated, of which 63 failed to complete registration and were not included in the sample, indicating a survey dropout rate of 14.5%. While the survey was constructed to block participants from re-taking the survey based on their IP address, eight clients took the survey twice (2%). In these cases, the second response was excluded from the data and an email sent by the firm to clarify that only the first survey applied.<sup>42</sup> On 21 November 2013, the sample size reached the desired 364 participants (see page 118) and the survey was closed. The sample accounts for 6.5% of the eligible client base.<sup>43</sup>

---

<sup>42</sup> Four of the eight individuals were initially assigned to the reputational commitment group and all declined to name a coach; it could be possible that they later changed their mind and wanted to nominate someone. However they would have had access to my email address if they were very keen to share their coach's name and contact details, but this did not happen. While it is possible that these participants had found out about the other treatment message and tried to take the survey again to get a refund, this is not borne out by any other evidence, and no further correspondence was received from them to suggest they were disgruntled at not receiving the refund.

<sup>43</sup> Calculated using correspondence with the partner firm in June 2013. The 6.5% statistic raises questions about external validity, which are discussed in section 7.4.

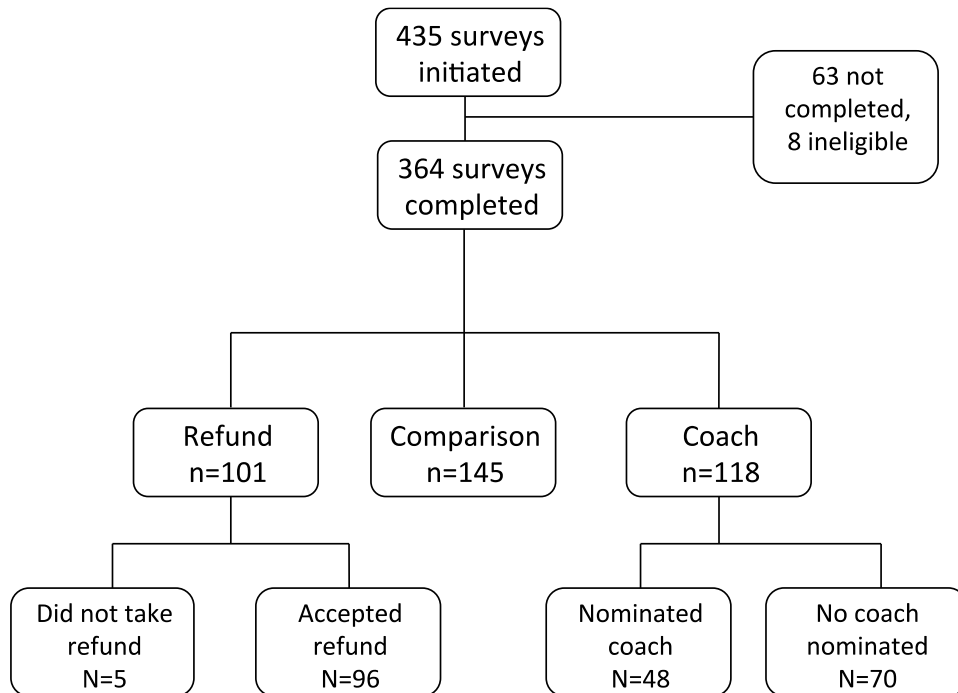


## **2.2. Randomisation**

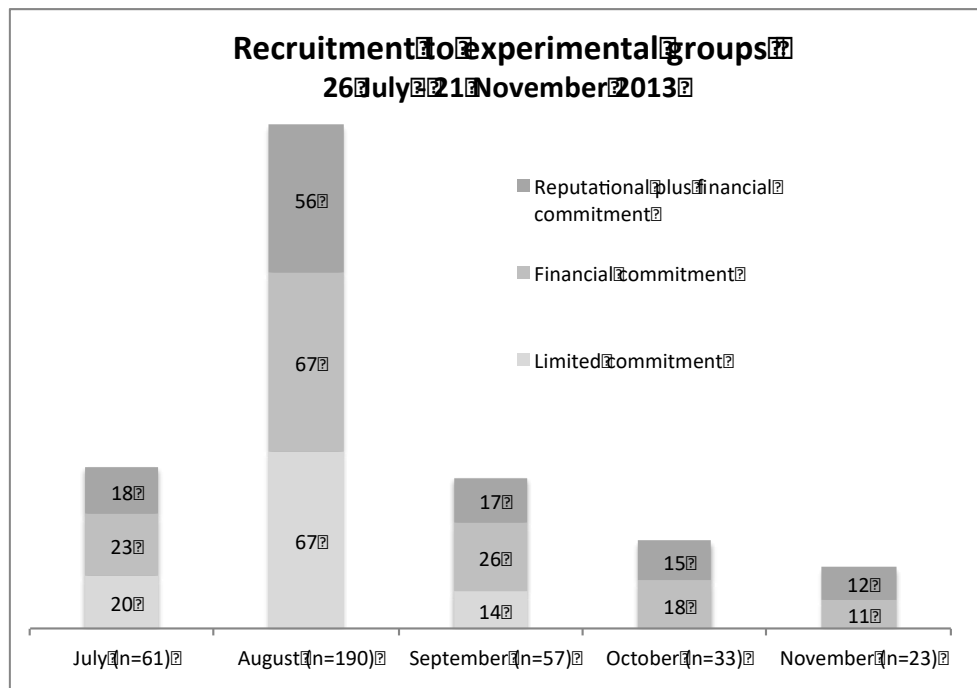
Participants were included as and when they completed the online registration process. The final stage of the baseline survey randomly assigned them to three experimental groups, using three different messages (see Figures 17, 18 and 19 below). The limited commitment group contained 101 participants (28%), the financial commitment group contained 145 (40%), and the reputational commitment group contained 118 (32%), in line with the quota imposed on refunds and ex ante sample size targets (see chapter 4, page 117). The quota was reached in mid-September, and the online randomisation tool thereafter assigned participants to only the comparison and coach groups (see Figure 16). The sample size target of 364 was reached in November 2013 and recruitment ended.

To take account of the change in allocation rule, regression analysis incorporates a dummy variable to capture which randomisation phase the participant entered. Further, robustness checks on the starting month (running from July to November) are reported in the annex, which also serve to highlight any seasonal factors that might affect weight loss performance. The financial and reputational commitment groups were expected to be closer in size. The difference in experimental group size was driven by chance. Robustness checks also apply weighting to take account of the unequal treatment allocation probabilities implied by the final sample sizes (see appendix section A.6), and find no change to the main results discussed in section 5 below.

**Figure 15: Food Monitor experiment flow chart (CONSORT)**



**Figure 16**



### 2.3. Treatment offers

In line with CONSORT reporting requirements, Figure 15 visualises the field experiment recruitment and randomisation processes. The financial commitment group continued to pay their membership, and to all intents and purposes experienced business as usual. In analysis below they form the comparison group. The reputational commitment group were asked to nominate a coach, and 41% complied (n=48). The limited commitment group were offered their monthly subscription fee back and 95% complied (n=96). Non-compliance is discussed further in section 2.5. The refund and coach groups were asked to accept the treatment, and were then shown the same text set out in Figure 17 requesting no cross-talk with other participants about the trial. This aimed to address the potential threat to the validity of the experiment from contamination identified in Chapter 4.

Figure 17: Monthly fee message (comparison group)

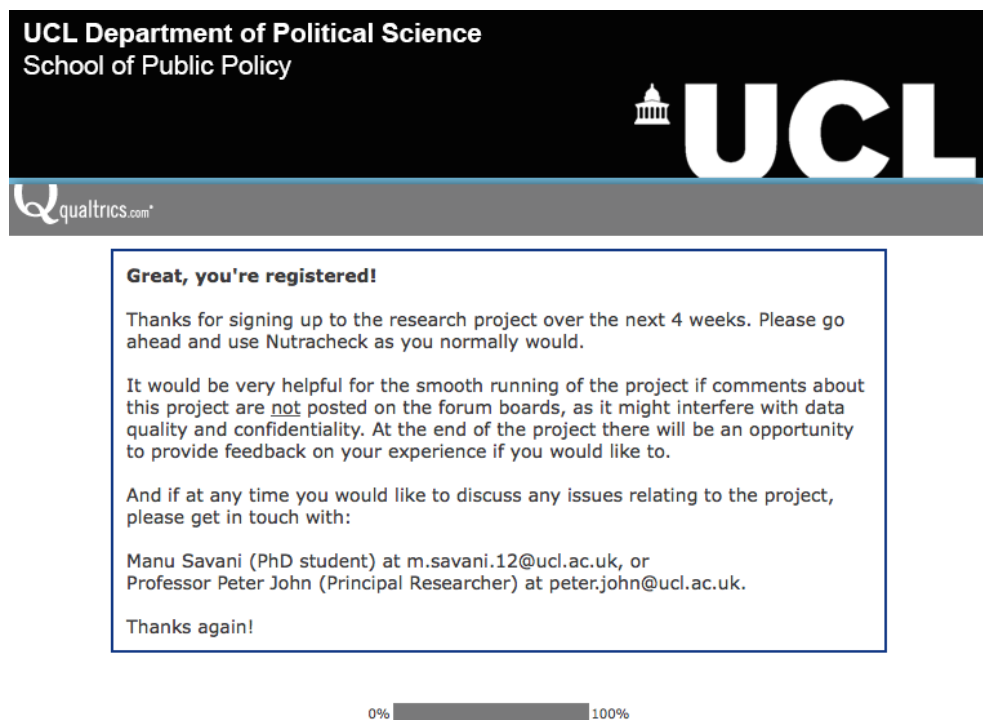


Figure 18: Refund treatment offer (limited commitment group)

School of Public Policy

UCL

qualtrics.com

Thanks, you have completed the survey!

You have been selected to receive the next 4 weeks' subscription for free. Please click below to accept, and a refund will be with you soon.

Do you accept?

Yes

No

<< >>

0% 100%

Figure 19: Coach treatment offer (reputational commitment group)

UCL Department of Political Science  
School of Public Policy

UCL

qualtrics.com

Thanks, you have completed the survey!

How about trying something new? We'd like you to nominate a coach, someone supportive who knows you have a weight goal you are working towards. This could be a good friend or encouraging family member. All they need to do is agree to be contacted either by email or phone in about 4 weeks – so please do check with them.

Please click below to give this strategy a go - it will take you less than 1 minute! Your membership will not be affected either way.

Do you want to nominate a coach?

Yes

No

<< >>

0% 100%

#### **2.4. Randomisation balance check**

Simple randomisation, while easy to administer, may produce statistical imbalances across covariates particularly in smaller samples (Torgerson & Torgerson 2008, p.31). Table 13 reports covariate means in column one for the experimental groups, and modal values for categorical variables.

The groups are generally well balanced, with three exceptions: starting weight, starting BMI, and income groups.<sup>44</sup> Statistically significant associations between covariates and treatment status can occur by chance, and the risk of discovering such associations increases with the number of covariates tested (Glennester & Takavarasha 2013, p.150). It is worth noting that none of these variables remain statistically significant after the Benjamini-Hochberg correction is applied for multiple hypothesis testing. Nevertheless, the variables are investigated further to determine the nature of the imbalances.

Starting weight and BMI appear to be significantly different between the comparison group (financial commitment) and the refund group (limited commitment). A closer look at the BMI categories shows that this difference is driven by a concentration of severely obese individuals in the comparison group (see column 1). Of the 33 individuals with BMI > 40 in the sample, 61% are in the financial commitment group, against 9% in the limited commitment and 30% in the reputational commitment group. A repeat test of equality across groups excluding the severely obese finds that there is no significant difference in initial weight or BMI, confirming that this small subset of high BMI individuals is driving the apparent imbalance (tests now report  $p = 0.676$  between comparison and treatment group 1, and  $p=0.331$  between comparison and treatment group 2). BMI categories will be included as a control variable in all

---

<sup>44</sup> Hypothesis testing for these three variables indicates statistically significant differences in covariate means or modal groups, using ttests, prtests, or rank sum tests.

regressions undertaken in sections 6 and 7 below, to address the imbalance in both starting weight and starting BMI.

Demographic variables are largely similar across the groups with the same modal category throughout. However, there is a notable difference in incomes (see panel B column 2), with the financial commitment group having the largest share of participants reporting less than £19,999 annual household income (19%, relative to 8% in the limited commitment group and 11% in the reputational commitment group). A new binary variable is created to identify participants in the lowest income category. Excluding this income category, the income variable is no longer statistically associated with experimental groups. To address the remaining imbalance at the lower end of the income spectrum, the new 'low income' binary variable is used as a control variable in all regressions reported in this chapter.<sup>45</sup>

---

<sup>45</sup> No harms or unintended effects were brought to my attention during the field experiment.

<b>Table 13: Summary statistics by experimental group</b>			
	Financial commitment (Comparison) N = 145	Limited commitment (Refund) N=101	Reputational commitment (Coach) N=118
	(1)	(2)	(3)
Starting weight (kg)	84.9 (20.6)	79.4 (15.8)	83.4 (18.2)
BMI	31.1 (7.28)	28.9 (5.43)	30.2 (6.30)
Overweight	0.368 (0.484)	0.35 (0.479)	0.291 (0.456)
Obese	0.319 (0.468)	0.34 (0.476)	0.402 (0.492)
Severely obese	0.139 (0.347)	0.03 (0.171)	0.085 (0.281)
Weight loss target (%)	4.13 (2.11)	4.18 (1.53)	4.27 (1.99)
Female (%)	0.903 (0.296)	0.921 (0.271)	0.864 (0.344)
Fruit and veg daily intake	4.13 (2.24)	3.78 (2.07)	3.85 (2.14)
Exercise sessions per week	3.11 (2.72)	3.32 (2.58)	2.95 (3.32)
Experienced major life changes recently	0.290 (0.455)	0.376 (0.487)	0.347 (0.478)
Doing other activities to lose weight	0.883 (0.323)	0.842 (0.367)	0.872 (0.336)
Impatient (%)	0.076 (0.266)	0.069 (0.255)	0.085 (0.280)
Myopic health attitudes (%)	0.572 (0.496)	0.564 (0.498)	0.610 (0.490)
Recruited in phase 1	0.731 (0.445)	1.00 (0.00)	0.720 (0.451)
Age (modal)	40-49 years	40-49 years	50-59 years
Education (modal)	Bachelors	Bachelors	Bachelors
Income (modal)	Up to £19k	£40k - £49k	£50k - £59k
Job status (modal)	Paid employment	Paid employment	Paid employment

*Notes: Mean values reported with standard errors in parentheses for all continuous and binary variables. Modal categories reported for age, education, income and job status categorical variables.*

## 2.5. *One-sided non-compliance*

Participants assigned to both the refund and coach treatments were asked to accept them, and as reported in section 2.3 above not everyone chose to take up the treatment offered (see also Figure 15 above). In the case of the refund, five participants did not receive the refund (either because they declined the refund or did not make it possible for the company to return their monthly fee back). In the case of the coach, 48 participants nominated someone and 70 declined. Reasons given for turning down the coach treatment included not knowing who to name, not wanting to share the coach's contact details, and preferring to pursue the health goals privately.

Non-compliance raises various issues for analysis, as discussed in chapter 4.<sup>46</sup> Firstly, those who accepted treatment have self-selected in, and a comparison of only treated individuals would therefore cause biased inference. Regressions below will therefore rely on intent-to-treat analysis, estimating the causal effect of treatment assignment rather than the actual treatment. Two other investigations are of interest for the coach treatment, which has a much higher non-compliance rate: what determines whether the coach is nominated or not? And might there be distinct causal effects for the group of compliers within the wider treatment group?

To answer the first question, probit regression is used to investigate the take-up decision (see appendix section A.11). It appears that those with more short-termist health attitudes are more likely to take up the coach treatment offer. This might indicate these individuals are more aware of their time inconsistency, hence are keen to seek external commitment aids.<sup>47</sup> To address the second

---

<sup>46</sup> There was no evidence of two-sided non-compliance in the experiment, and the discussion therefore focuses on one-sided non-compliance.

<sup>47</sup> This interpretation is consistent with that reported in Giné et al (2010), who report one-sided non-compliance of 89% for a deposit contract to support smoking cessation. Compliance was correlated with efforts 'to avoid situations that made the participant want to smoke', and the authors interpret this strategic behaviour to control oneself as evidence of sophistication.



question, the complier average causal effect, estimated using instrumental variables, is discussed in further detail in section 5.5.4 below. Section 2 has provided an overview of recruitment, randomisation balance, and compliance in the experiment. The discussion moves on now to consider baseline data and participant characteristics.

### **3. DESCRIPTIVE STATISTICS**

#### **3.1. Participant profile**

A typical participant was female, overweight or obese, aged in their forties or fifties with higher education qualifications, and in paid employment. A fuller discussion of baseline and demographic characteristics is set out in the appendix (section A.12). The remainder of this section focuses on starting weight and BMI, gender, and the two covariates that were pre-specified for sub-group analysis. These variables relate to short-termism and myopia, and are operationalised as a time preference measure and myopia in health attitudes (introduced in chapter 4).

#### **3.2. Starting weight and BMI**

The average weight amongst participants is 83kg, with a large range of values from 48 kg to 157 kg. Figure 20 presents the distribution with a normal density curve drawn for comparison and shows the long tail of individuals with very large starting weight. The top 5% (18 observations) are over 115 kg, while the bottom 5% are under 59 kgs. The larger values are broadly plausible but are associated with extremely high BMI scores including two above 50. They will be included in the baseline models in the statistical analysis below, but excluded as part of the robustness checks.

The average body mass index (BMI) is 30.2, just greater than the level used as the cut off between overweight and obese BMI categories. Table 14 indicates the majority of participants are either overweight or obese, with a small minority (9%) at the severely obese end of the spectrum. Just over 20% of participants had a healthy baseline BMI, but the majority (90%) still aimed to lose rather than maintain weight.

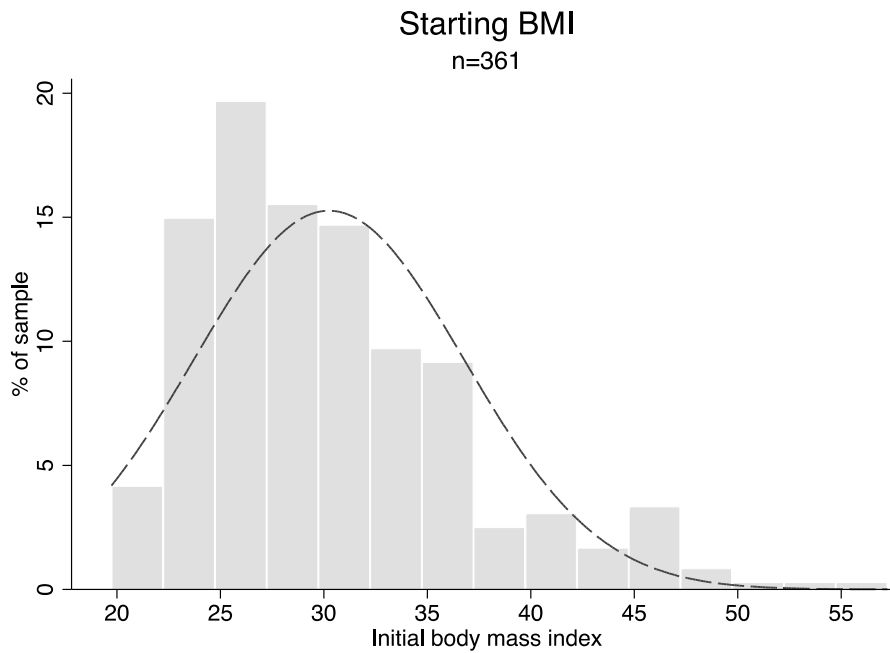
	Mean	Range	SD	N
Weight (kgs)	82.9	47.6 – 156.5	18.6	361
4- week weight loss target (kgs)	3.4	0 – 12.2	1.5	354
Target as % of initial weight	4.2	0 – 16.0	1.9	354
Body Mass Index (BMI)	30.2	19.7 – 55.7	6.5	361
	%	Average	SD	N
Healthy, 18.5 < BMI <24.99	22	23.2	0.15	79
Overweight, 25 < BMI <29.99	34	27.3	0.13	122
Obese, 30 < BMI < 39.99	35	33.6	0.23	127
Severely obese, BMI > 40	9	45.1	0.65	33

The average weight loss goal was an ambitious 3.4 kg over 4 weeks (just under 1 kg a week, or an average of 4% of body weight), and these goals suggest some degree of optimism bias.<sup>48</sup> Not all participants aimed to lose weight over the 4-week trial period: nine individuals (2.5%) reported they wanted to maintain their weight. Of this group, eight started with a healthy BMI, and one was only slightly overweight with a BMI of 25.2; indicating that those who were not explicitly targeting weight loss had reasonable grounds for this decision. Amongst the 97.5% whose aim was to lose weight over the 4-week trial period, 80% were overweight, obese, or severely obese, again indicating reasonable grounds for their decision.<sup>49</sup>

<sup>48</sup> The goals seem ambitious when compared to Camden’s Shape Up programme, which set a target of 5% weight loss over 11 weeks.

<sup>49</sup> Why might the remaining 20% have aimed to lose weight despite having a healthy BMI? Note the average BMI in amongst these people exceeded 23, indicating they were at the top end of the category approaching an overweight BMI. For them, weight loss may have been an important preventative goal. Further, the BMI calculation is a relatively blunt way of diagnosing healthy weight: the measure itself does not take account of an individual’s fat and muscle ratio, and for many individuals a healthy BMI is lower than the standard thresholds based on ethnicity (Ntuk et al. 2014). Finally, social norms (for example around appearance and dress size) may influence weight loss goals more than BMI benchmarks.

**Figure 20**



### **3.3. Gender**

As is typical of weight loss studies, the majority of the sample (90%) are women (Jolly et al. 2011). There are no significant differences in covariates between men and women on the majority of covariates, excepting age and health motivation. While female participants cover the full age range from 18 to 65, male participants are concentrated in the 40s and 50s age range (none are younger than 30). The majority of male participants reported relatively far-sighted health attitudes (58%); women, in contrast, were more likely to be classified as myopic (60%). Health attitudes across genders contrast with those reported in the Health Survey for England 2011 (Robinson 2012), which hints that the experimental sample differs from the wider population (implications for generalisability will be discussed in section 7 of this chapter).

### **3.4. Present bias: myopia in health attitudes**

Chapter 3 hypothesised that the commitment device would be less effective for those with stronger present bias (see Table 6, page

105). The thesis aims to pin down this prediction using two operational measures for time preference. The first relies on an attitudinal measure using the Healthy Foundations Segmentation model. In the Food Monitor dataset, three particular segments are identified as exhibiting short-termist attitudes: Hedonistic Immortals, Live for Today's, and Unconfident Fatalists, suggesting greater influence of the doer. In contrast, Balanced Compensators and Health Conscious Realists are identified as being more far-sighted, suggesting greater control of the planner over intertemporal choices. In this way, the categories are collapsed into a binary variable to proxy whether an individual has short-termist health attitudes (58%) or not (42%).

**Table 15: Descriptive statistics: Myopia**

Short-term health attitudes (n=364)	%	N
Myopic	58%	212
Far-sighted	42%	152

### **3.5. Present bias: cost of waiting**

The second measure of time preference is derived from a more general behavioural economics question in the baseline survey. It asks how much the respondent needs to be compensated in order to delay receiving £10 now, where the delay is 1 month; and repeats the question with a delay of 6 months. The responses generate a 'cost of waiting' variable, where those who are relatively impatient – the present-biased – have a high cost of waiting; conversely the relatively patient have a low cost of waiting. Those with a high cost of waiting disproportionately weight present gains over future gains, and so can be described as myopic. While the measure does not refer to health gains, it offers a useful triangulation with the previous measure using health attitudes. Full details on how the variable was constructed are available in the appendix (section A12).

While the average cost of waiting an extra month for £10 is approximately £5.50, the figure varies across the sample, with a distinct sub-group of impatient individuals reporting a higher cost of waiting. As shown in Table 16, the cost of delaying the £10 hypothetical payoff is over £40 for the impatient, compared to less than £3 for the patient. Both measures of time preference will be used in section 6 below to answer research question 2.<sup>50</sup>

**Table 16: Descriptive statistics: Time preference**

Time preference (n=351, 97% of sample)	Mean	Range	SD
Cost of waiting 1 month among 'impatient'	40.2	15 – 90	22.8
Cost of waiting 1 month among 'patient'	2.6	0 – 10	3.7

*Note: 'impatient' sub-group (n=28) identified as having top 8% of required compensation to delay payoff (see appendix for full details).*

### **3.6. Summary**

This section provided a detailed description of the Food Monitor dataset, with further detail available in the annex. The next section examines covariate balance across the three experimental groups – limited commitment, financial commitment, and reputational commitment – in order to establish broad parity between the three groups as a basis for inferring causal effects from the treatment.

<sup>50</sup> In regressions below the continuous variable 'cost of waiting' is used to capture present bias. Robustness checks in the appendix (section A17) present results when the regression model relies on the binary 'impatient' variable instead. In the randomisation balance check above, the proportion of people who are impatient is used to assess balance of traits across the groups.

## **4. OUTCOME DATA AND STRATEGIES TO ADDRESS ATTRITION**

### **4.1. Gathering data on weight loss**

The primary outcome variable is percentage weight loss over four weeks and 12 weeks, derived from self-reported weight logged by users in their Food Monitor accounts. Self-reported weight entries were included if they fell in a window of seven days before or after the four-week target date, and 14 days before and after the 12-week target date; with the closest date match preferred, and later observations recorded where two equally distant time stamps were recorded. The four-week period fits closely with the trial period and duration of the treatments, for example the refund covered one month's subscription, and the coaches were due to be contacted one month after the participant registered for the trial. With the 12-week period, it is possible to see whether the effects of the commitment device are sustained over time.<sup>51</sup>

In the interests of ensuring the trial participants had as realistic an experience as possible (addressing Hawthorne Effects as discussed in chapter 4), these 'virtual' weigh-ins were not mandatory. Participants were encouraged to use the service as they normally would. The trade-off was a reliance on participants returning to the website to share outcome data, generating the potential for attrition bias (chapter 4). At the design stage, it was expected that Food Monitor users who were motivated enough to complete the baseline survey and sign up to the trial were more likely to be regular users of the tool, including the weigh-in feature. Nearly all participants had a clear weight loss goal, and it seemed reasonable to assume they appreciated the importance of the Food Monitor tool for tracking

---

<sup>51</sup> As the registration process only required that participants have at least 4 weeks remaining on their membership, it is also possible that over the 12-week period some members reached the end of their subscription and did not renew. This would entail attrition at the 12-week stage, and it was not possible to track how many participants this might have applied to.

progress (given they self-selected into the client base). These assumptions proved to be too optimistic, as the following discussion demonstrates.

#### **4.2. Attrition on end weight readings**

The dataset contains 3,224 weight entries, but with uneven coverage: 98 participants never entered an eligible weight reading, while at the other extreme one participant recorded their weight 34 times. For those that did use the weigh-in tool, readings were only taken if they fell within the four-week and 12-week window. The dataset had end weight readings for 187 participants at four weeks and 162 participants at 12 weeks, implying attrition rates of 49% and 55% respectively. It was not feasible to incorporate second round sampling as a mitigating strategy, however there was an opportunity to improve the coverage of outcome data at four weeks through a simple, linear interpolation exercise using outcome data at 12 weeks. Full details of this exercise can be found in the appendix.

Even with the augmented dataset, however, attrition rates at four weeks remain high at 38%. The challenge of attrition is well documented in the field experiments (Gerber & Green 2012, p.211) and health sciences literature (Torgerson & Torgerson 2008, p.51; Little & Rubin 2000, p.135). In a meta-analysis of weight loss RCTs, Elobeid et al report a range of attrition rates from 6% to 46% (2009, p.5, table 2). Even against this range, the Food Monitor attrition rates are high, and potentially threaten the validity of causal inference if they unwind the effect of random assignment.

Further investigation of attrition patterns is essential to determine what statistical techniques are best suited to uncover causal inference, and this issue is taken up in sections 4.6 and 4.7 later in this chapter. First, this discussion turns briefly to consider outliers in the weight loss data series, and presents descriptive



statistics on weight loss and self-monitoring performance for the sample as a whole.

### **4.3. Outliers**

Early investigation into this augmented dataset detected the presence of some outliers, and this prompted a more detailed look of the top 5% and bottom 5% of weight loss performers. In a small number of cases there appeared to have been inputting errors at the baseline or endline weight readings. Triangulating against other weight readings for those participants allowed for these errors to be ‘corrected’. Detail on the steps involved, the original readings and revised readings are set out in the annex.<sup>52</sup> All numbers reported here, and in the summary statistics above, take account of the cleaned and corroborated weight readings. Most of the outliers remained in the sample uncorrected, as there was sufficient evidence to suggest the self-reported readings were not an inputting error. Of the 18 outliers investigated, one was dropped from the 12-week dataset, and four were amended (1% of all observations involved in this data cleaning).

### **4.4. Weight loss outcomes**

A first look at the outcome data suggests that at 4 and 12 weeks average weight loss is positive but modest, with participants losing on average 1.0 kg over 4 weeks and 1.70 kg over 12 weeks (see Table 17 below). A small proportion of individuals at both ends of the distribution have lost considerably more weight than the average (a maximum of 6.8 kg at 4 weeks and 15 kg at 12 weeks), or gained considerably more weight (2.7 kg at 4 weeks and 8 kg at 12 weeks). Excluding those who aimed to maintain rather than lose weight shifts the mean values upwards slightly to weight loss of 1.76 kg at 12 weeks but no meaningful change at 4 weeks. Figure 21 shows the

---

<sup>52</sup> This exercise was undertaken blind to the group allocation.

distribution of values around the median (for those who were not intending to lose weight).

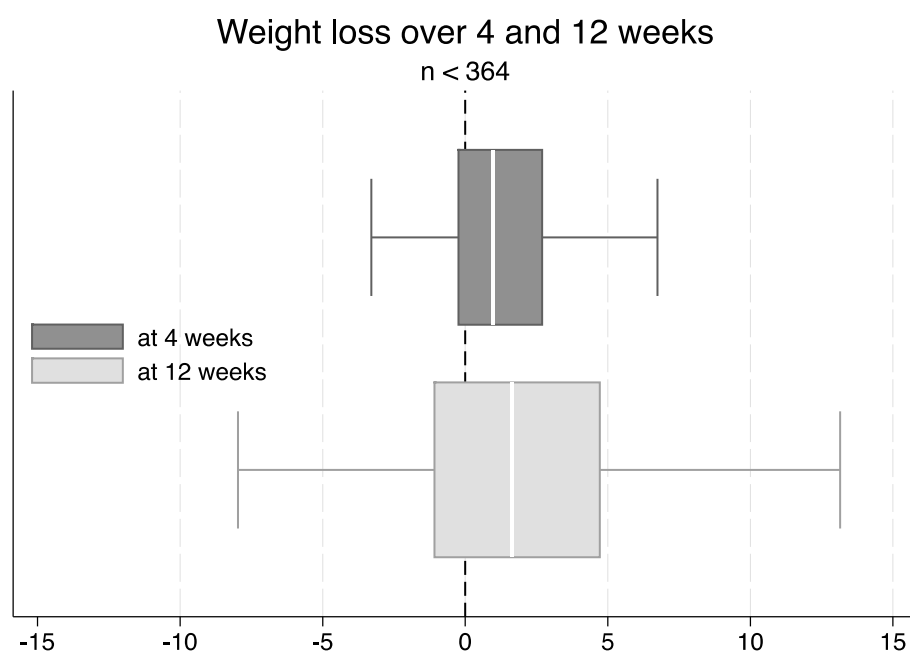
It is somewhat surprising that mean weight loss at 12 weeks does not show a stronger improvement on the 4 week data, but this reflects the larger dispersion of values around the mean at 12 weeks, and the longer tails at both ends of the distribution. After the initial trial period, the potential trajectories of different weight loss journeys became much more dispersed, leading to a much larger standard deviation around the mean at 12 weeks. The overall picture implies a wide range of weight loss experiences, masked by a fairly modest mean value.

**Table 17: Average weight loss and self monitoring outcomes**

	Mean	SD	Range	N
4 weeks weight loss kg	1.03	1.80	[-2.72, 6.81]	218
% of initial weight	1.19	2.09	[-3.30, 6.74]	218
12 weeks weight loss kg	1.76	3.97	[-8.2, 15.0]	157
% of initial weight	1.98	4.55	[-12.2, 13.1]	157
Self-monitoring with all tools	20.5	13.3	[0, 64]	364
With calorie counter	16.1	10.0	[0, 28]	364

*Notes: Weight loss reported for those seeking to lose (not maintain) weight*

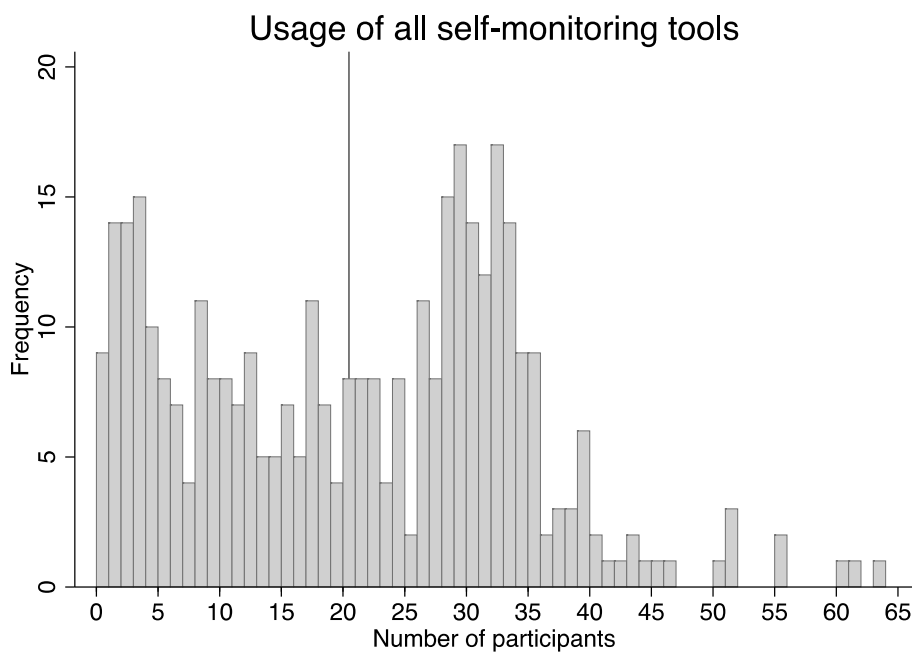
**Figure 21**



#### 4.5. *Self-monitoring outcomes*

A secondary outcome variable is usage of the service over the 4-week trial period, measured by the number of times any of the self-monitoring tools available were used, of which the calorie counter tool was the most popular (79% of all usage), followed by weigh-ins (13%) and food journals (8%). Any one of these tools supports self-monitoring. Figure 22 demonstrates that around the average of 20 uses over four weeks there is much diversity, with some participants rarely using any of the tools available, and others making use of the different tools over 60 times. The 'usage' variable does not suffer from attrition, as non-usage was coded 0.

Figure 22



#### **4.6. *Investigating differential attrition***

The earlier discussion confirmed the high rates of attrition, and this motivates further investigation into attrition patterns. The remainder of this section examines what factors drive attrition, and identifies suitable statistical techniques to address attrition bias in the regression analysis.

As a first step, Table 18 looks more closely at attrition rates across treatment groups. Column 1 shows that attrition across the dataset as a whole increases over time. Column 2 sets out the attrition rates in the comparison group, and columns 3 and 4 present the difference in attrition rates for the refund and coach groups respectively. In essence, weight outcome data is more likely to be available for the limited commitment (refund) group at 4 weeks, but no more likely at 12 weeks. Conversely, attrition is higher in the reputational commitment (coach) group in both periods. None of these associations, however, are statistically significant (p-values for hypothesis tests reported in parentheses), which is an important and reassuring finding. The alternative, differential attrition linked to treatment status, would have reintroduced selection and made it difficult to recover valid estimates of the average treatment effect in later regressions. Probit regression analysis offers further insight into the attrition patterns (see appendix) and confirms there are no statistically significant associations with treatment status at either four or 12 weeks.

% of missing weight observations (N)	All participants (1)	Comparison (Monthly fee) (2)	Limited commitment (Refund) (3)	Reputational commitment (Coach) (4)
Baseline (%) (N)	0.8% (3)	0.6% (1)	+ 0.3% (1) (0.796)	+ 0.2% (1) (0.884)
4 weeks (%) (N)	37.6% (137)	37.9% (55)	- 7.2% (31) (0.242)	+ 5.3% (51) (0.384)
12 weeks (%) (N)	55.5% (202)	54.5% (79)	- 1.0% (54) (0.876)	+ 4.0% (69) (0.518)

*Notes: p-values in parentheses for hypothesis tests of equal attrition rates between comparison group and treatment group, i.e. (2) = (3) and (2) = (4) at each stage of data collection.*

#### **4.7. Statistical strategies to deal with attrition**

While the dataset does not suffer from differential attrition, the loss of outcome data, particularly at the 12-week stage, reduces the effective sample size. The additional 21% observations plugged in through interpolation improve the four-week dataset somewhat. Given “dropouts and missing data are nearly-ubiquitous in obesity randomised controlled trials” (Elobeid et al. 2009, p.1), it is useful to learn lessons from the literature in how to address sizeable attrition.

A popular strategy in the weight loss literature (John et al. 2011; Augurzky et al. 2012) is to assume that those without outcome data experienced no weight loss. Essentially, this is equivalent to carrying forward the baseline weight to the end point. It is a relatively straightforward assumption about what happened in practice to the missing outcome data, and does not require more demanding assumptions about the mechanism driving attrition.<sup>53</sup> The ‘baseline

<sup>53</sup> A variation of this is to use the last available observation, which may be more recent for those individuals who provided weigh-in entries sporadically up to 12 weeks, but who may have been coded as missing because their data did not correspond to the 12 week ‘window’ outlined above. For current purposes the former assumption will be made, i.e. that there was no weight loss amongst those

carried forward' approach is unlikely to overestimate the effects of the commitment device, as it exerts a downward pressure on treatment effects towards zero. Following the intuition of Manski (1989), this method effectively places a restrictive upper bound on the average treatment effect, and is likely to produce very conservative estimates of effect size. It will be applied to both the four- and 12-week data in regression analysis below.

A recent body of literature (Augurzky et al. 2012; Tauchmann 2014) employs the non-parametric technique of placing bounds around the treatment effect. The 'Lee Bounds' estimator offers an interval of values rather than a point estimate, based on a scenario of "worst case sample selection biases" (Lee 2002, p.12). Under this conservative approach, these "extreme value bounds tend to be successful in bracketing the true average treatment effect" (Gerber and Green 2012: 227). The approach is premised on the data having two characteristics: firstly that treatments are as good as randomly assigned, as confirmed by Table 12 above; and secondly, that those who are observed (or missing) would be observed (or missing) regardless of treatment status. The latter assumption is justified through the attrition investigations (Table 17 and appendices A.13 and A.15), implying Lee Bounds can be calculated for both 4-week and 12-week weight loss outcomes.

A third statistical strategy is inverse probability weighting (IPW), which is advocated when data is judged to be missing independent of potential outcomes (MIPO) and conditional on covariates (Gerber & Green 2012, p.222). This strategy will be applied on the 12-week outcomes, given evidence in the appendix suggesting that covariates are significantly associated with attrition, indicating that data are MIPO at 12 weeks. The appendix lays out in detail how weights were derived for this approach.

---

who dropped out, which has the benefit of being more straightforward and less reliant on arbitrary date cut-offs.

Other strategies have been considered but rejected: statistical modelling to fill in missing data, selection models (Heckman, 1976) or multiple imputation (Graham 2009). These approaches would require much stronger assumptions to be made on the nature of the missing data mechanism; assumptions which may be “tenuous and untestable” (Little and Rubin, 2000: 137). In general, the modelling approach is foregone in this analysis, because of the inherent “tension with the agnostic style of experimental investigation” (Gerber and Green 2012: 226).

In summary, three statistical strategies are chosen to address attrition in the Food Monitor dataset: baseline carried forward, inverse probability weighting (for 12-week data only), and non-parametric Lee bounds (for 4-week and 12-week data). Estimates of commitment device effects from these three strategies are compared to the ‘complete case’ analysis. The latter essentially ignores the problem of attrition and offers a benchmark to understand how the different methods diverge from the raw data available. The next section presents the average treatment effect results based on the statistical model presented in chapter 4, with a focus on the fully specified model with covariates.

## **5. AVERAGE TREATMENT EFFECTS**

### **5.1. Weight loss outcomes: comparison of means and graphical evidence**

A comparison of mean weight loss across experimental groups suggests little difference (Figures 10 and 11).<sup>54</sup> The financial commitment group lost on average 1.7% of their body weight over four weeks, compared to 1.3% in the refund group and 1.4% in the coach group. These differences are not statistically significant. At 12 weeks, the comparison group lost 2.5% of body weight, matched exactly by the refund group, while the coach group performed considerably less well with 1.1% weight loss ( $p=0.078$ ).

Figures 23 and 24 below suggest that the reputational commitment device treatments had little effect in the short run, and may have had a negative effect in the medium run, which runs counter to the expected relationship (hypothesis 1). Meanwhile, the limited commitment group were expected to perform worse than the comparison group because their financial commitment for the month was removed; however they show only a minor change relative to the comparison group in the short term, and are back on a par with the financial commitment group by 12 weeks.

---

<sup>54</sup> The weight loss results focus on those participants who intended to lose not maintain weight, which brings the available observations at 4 weeks to 178 from 186, and at 12 weeks to 157 from 162. Robustness checks using the full sample are reported in the annex and show no difference to the overall story.



Figure 23

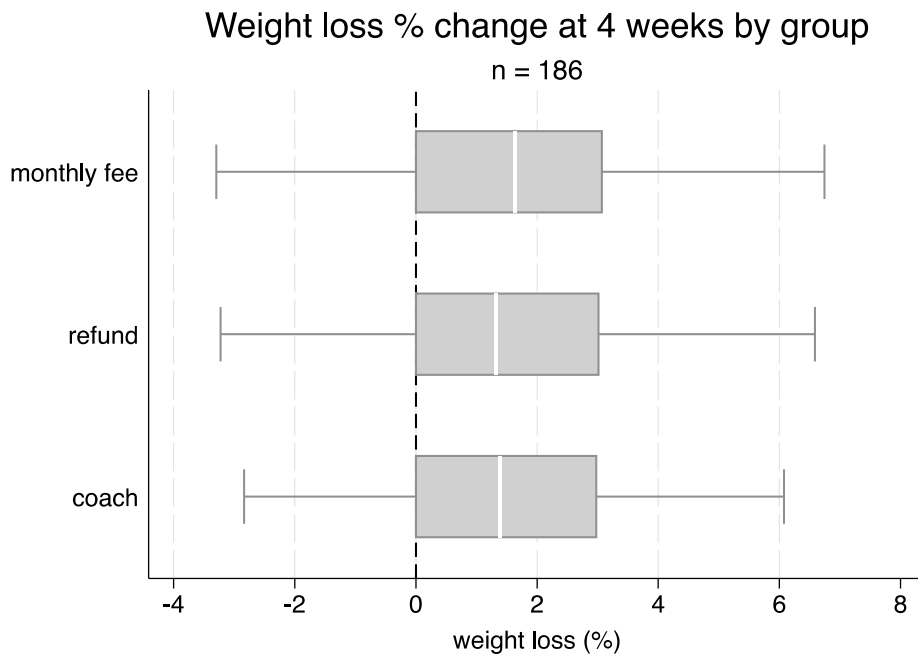
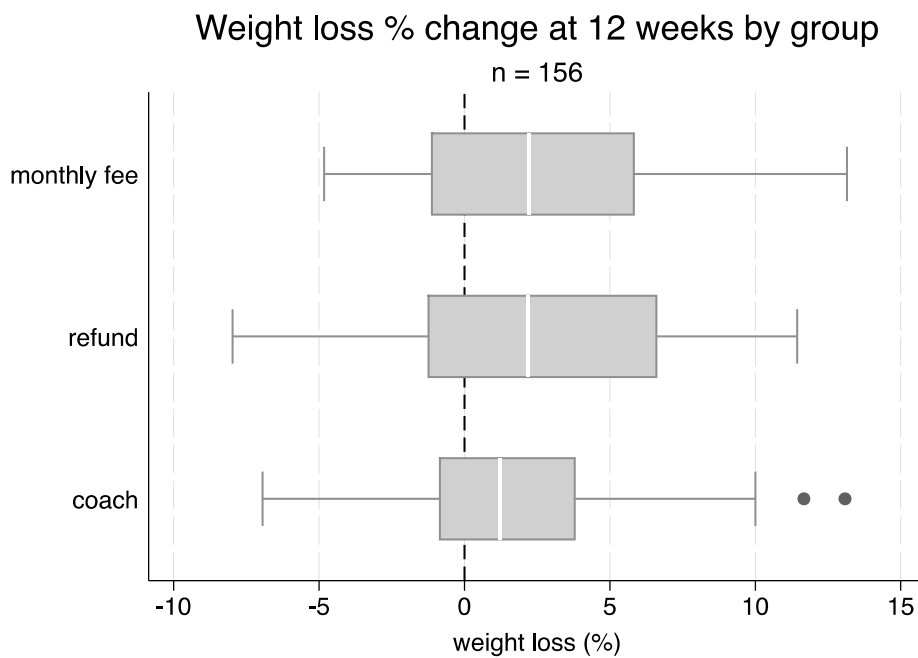


Figure 24



## 5.2. *Regression model*

As described in Chapter 4, OLS regression results aim to uncover the average treatment effect, or intent-to-treat estimate, through the coefficient on the binary treatment variable.<sup>55</sup> Robust standard errors are clustered at the individual level. To address the change in treatment allocation rule during the trial, causal effects are identified through two separate analyses for each treatment. Equation 15a recovers the ATE for the refund, while equation 15b recovers the ATE for the coach:

$$[15a] \quad Y_i = \alpha + \beta^R \cdot R + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \varepsilon_i$$

$$[15b] \quad Y_i = \alpha + \beta^C \cdot C + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \varepsilon_i$$

Equation 15a is run using data from phase 1, when all three treatments were offered, and controls for the coach group to allow for a treatment effect to be estimated against the comparison group in phase 1. Equation 15b compares the participants in the coach experimental group with those in the comparison group across the full span of the trial. In both models,  $Y$  measures end weight outcomes (kg). Treatment status is captured by dummy variables  $R$  (refund) and  $C$  (coach), where  $R=1$  and  $C=1$  if the participant was offered the treatment. The OLS estimators for  $\beta^R$  and  $\beta^C$  uncover the average treatment effects for the commitment contract.  $W$  is a series of baseline covariates  $J$ , with coefficients  $\gamma_j$ . These coefficients offer statistical association with outcome variable  $Y$ , and cannot be used to infer causality. The model incorporates individual traits (including gender, age, and variables to capture myopia and time preference); lifestyle and behavioural variables (exercise and diet, experiencing major life changes and taking part in other activities to pursue weight

---

<sup>55</sup> Chapter 4 also set out an equation without baseline covariates, however the results from this approach had such low explanatory value (very low R-squared) that their contribution to the analysis here is trivial. They are reported in appendix section A.18 for completeness.

loss); and finally, a series of time dummies to capture the effects of different starting months.

### **5.3. Regression results**

Table 19 presents treatment effect estimates on four- and twelve-week weight outcomes, with panel A reporting on the refund offer and panel B reporting on the coach offer. Columns 1 and 3 present complete case analyses (ignoring attrition); columns 2 and 4 use baseline observations carried forward where necessary; and column 5 applies inverse proportionality weights.

The results point to two headline messages. Firstly, in the short term, neither the refund nor the coach treatment significantly affected weight loss. This runs counter to the hypotheses, but is triangulated with the boxplots that show a large dispersion around very similar mean values for weight loss. Lee Bounds estimates on 4-week weight loss data offer further triangulation: the interval for the refund treatment effect being  $[-1.95, 1.42]$ , with the lower bound approaching statistical significance ( $p=0.057$ ). The corresponding result for the reputational commitment treatment effect is  $[-1.47, 1.08]$ .<sup>56</sup> Both imply that the true ATE lies in an interval that includes zero. Arguably, the evidence points to the commitment devices exerting zero treatment effect on weight loss outcomes in the short term.

Secondly, treatment effects are detected at 12 weeks, when the reputational commitment group perform considerably worse than the rest of the sample. Negative treatment effects from the coach offer are reported consistently across the modelling strategies, with the IPW estimate implying that the coach offer reduced weight loss by 1.6 kilograms ( $p=0.034$ ), equivalent to -1.7% of starting body weight on average. While the p-values do not withstand the

---

<sup>56</sup> Lee Bounds estimates on weight loss (percent).

Benjamini-Hochberg correction, the weight of evidence across the three models, triangulated with a Lee Bounds interval of [-1.7, -0.34], clearly argues that the reputational commitment treatment worked in the opposite way to that implied by theory. The effect size in terms of Cohen's *d* is -0.34.<sup>57</sup>

Two immediate questions arise: why was there no effect from either treatment in the short term, and why did the reputational commitment device discourage weight loss in the medium term?

**Table 19: Can commitment devices promote weight loss?**

	4 weeks CC (1)	4 weeks BCF (2)	12 weeks CC (3)	12 weeks BCF (4)	12 weeks IPW (5)
Panel A:					
Refund	-0.303 (0.330)	-0.302 (0.151)	-0.145 (0.865)	-0.183 (0.641)	-0.060 (0.942)
N	171	271	121	270	121
R <sup>2</sup>	0.992	0.995	0.967	0.981	0.967
Panel B:					
Coach	0.189 (0.553)	0.205 (0.296)	1.606* (0.040)	0.664 (0.053)	1.632* (0.034)
N	145	245	106	244	106
R <sup>2</sup>	0.992	0.995	0.969	0.984	0.970

*Notes: OLS regressions on end weight. All models have same covariates, including start weight, which are reported fully in appendix. Columns 1 and 3 report complete case analysis, columns 2 and 4 carry forward the baseline observation, and column 5 applies inverse proportionality weighting. Panel A presents results from equation 15a, panel B presents results from equation 15b. Panel A includes the coach treatment participants as a control dummy variable. Treatment effects are compared to the comparison group, the financial commitment group. P-values in parentheses \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$*

<sup>57</sup> Calculated using weight loss (percent).

#### **5.4. Discussion: why no weight loss effects at 4 weeks?**

##### **5.4.1. Persistence of commitment despite refund**

It was expected that the refund would temporarily dismantle the financial commitment arising from the monthly subscription fee. This change was predicted to reduce the psychological tax on the doer's over consumption ( $\theta$ ), and lead to weaker behaviour change and weight loss. Indeed, the financial commitment group did record the strongest weight loss at four weeks (figure 23), and the sign on the refund treatment coefficient was negative; however these effects were not statistically significant (table 19 panel A). Why did the limited commitment group not register a stronger fall in weight loss performance when the commitment device was removed?

One possibility is that because participants were already signed up to the Food Monitor website, indeed had already made their monthly payment, the temporary dismantling of the financial commitment had no effect on their innate motivation to lose weight during that period. If so, this finding has wider significance for theory, and a fuller discussion is set out in section 7 below. A second explanation for the low impact of the refund is that it had little psychological effect on clients because the amounts were fairly modest to begin with, with an average monthly fee of £5.50, ranging from less than £4 to £8. With a higher premium payment, the effects of being unshackled from the financial commitment may be different. A third possibility is that the study was picking up the negative effects hypothesised from the refund, but was not sufficiently powered to find significant results (and hence  $p > 0.05$ ). This is discussed further in section 7 later in this chapter, and is noted as a limitation of the research design.

### ***5.4.2. Low adherence and compliance with coach treatment***

Chapter 4 hypothesised that by combining commitment elements together – the monthly fee plus the nominated coach – the individual would experience a greater sense of commitment, a stronger psychological tax spurring on behaviour change, and this would be manifested as higher weight loss outcomes. Instead the data shows zero effect in the short term (and weight gain in the medium term – the subject of section 5.6 below). There are three potential explanations for this.

Firstly, nominating a coach may have encouraged a substitution from online tools to offline support; but with the adverse consequences of being less successful in losing weight. The data indicates that those offered a coach were less likely to use the Food Monitor service ( $p=0.045$ ) than the rest of the sample. Self-monitoring behaviour is explored in greater detail in the next section, but this evidence is suggestive of substitution effects away from the Food Monitor tool towards external accountability from a coach. Yet, this cannot be the full explanation, since most participants offered the coach treatment did not nominate one (40% compliance).

A second explanation is low adherence amongst the compliers. A brief survey of coaches at the end of the trial suggested only half were actively involved in the participant's weight loss efforts and recognised their role as a coach. As set out in the Analytical Framework, fidelity to the commitment device is a key aspect of maintaining the psychological tax of a commitment device ( $\lambda$ ). Without this, it cannot be expected to work, hence the lack of effect on weight loss at 4 weeks. This discussion is taken up in greater detail in chapter 7 using qualitative analysis.

A third explanation is that the intent-to-treat estimates are too conservative in the context of non-compliance: perhaps the analysis would detect stronger effects when actual treatment rates are investigated. Complier average causal effects (CACE) are estimated for the sub-group of those nominating a coach (see appendix for details of instrumenting variables approach), and results confirm there is no effect from the treatment on weight loss ( $p > 0.05$ ).

A final explanation is heterogeneity: while individual level treatment effects may in reality be sizeable (and in the case of the coach, positive), if these effects operate in both positive and negative directions they will average out to zero. This is investigated further in section 6 below.

### **5.5. Discussion: why negative effects from the reputational commitment device at 12 weeks?**

The discussion thus far offers explanations for a zero treatment effect at four weeks, but cannot explain why participants in the coach group went on to do worse at 12 weeks; losing less weight than those who received the refund or continued paying their monthly fee as usual. The results refutes hypothesis 2, that a stronger commitment device will generate larger effects on weight loss.

One explanation, as discussed above, lies with non-compliance rates, which may have diluted the average treatment effects. If this were true, we would expect the CACE to be positive and significant for the sub-set of people in the reputational commitment group who were treated, in line with hypotheses 1 and 2. Instead, the findings grow more puzzling: CACE estimates on 12-week data reveals a more negative treatment effect. Those who nominated the coach recorded 5kg higher end weight, equivalent to over 5% less weight loss (complete case and IPW models, both  $p < 0.05$ ); in other words compliers fared worse than those who were offered the treatment but

declined. While the difference was not significant in the short run, it grew more pronounced by 12 weeks (full details in appendix).

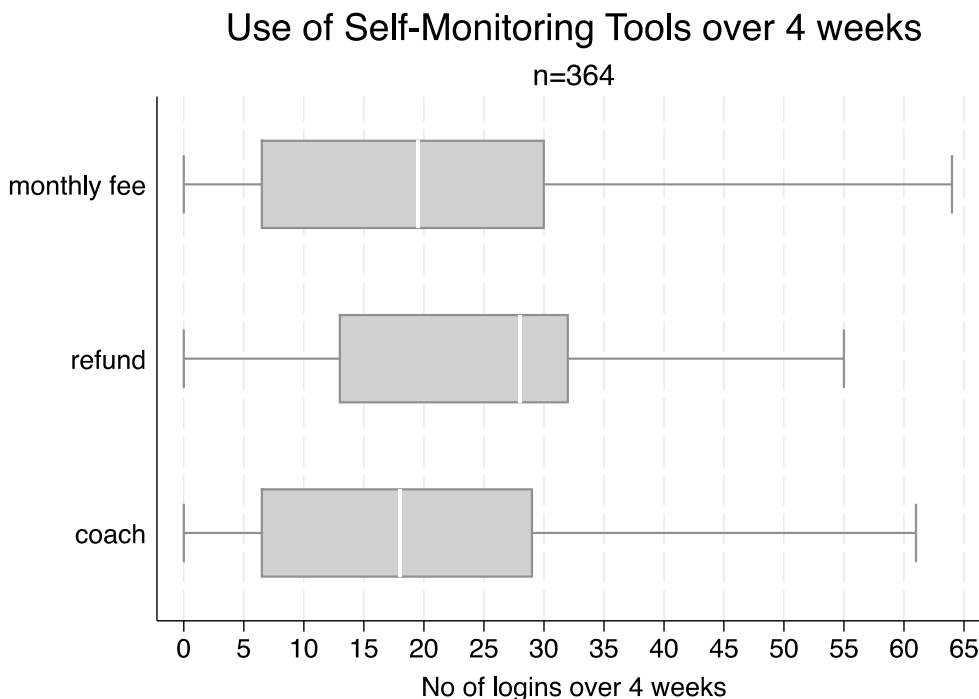
A second possibility, then, is that participants in the coach group experienced some form of commitment overload, or saturation, through the offer of the reputational commitment on top of the monthly fee. The Analytical Framework made an implicit assumption of a linear and monotonic relationship between commitment and health behaviour change. What the model did not allow for is the idea that increasing commitment does not always lead to increasing health behaviour change. The possibility of commitment overload indicates some form of upper threshold, a ceiling beyond which more commitment has adverse effects. It is not possible to delve further in to the potential mechanisms explaining the negative treatment effects, but the findings shed light on new research questions around the optimal level of commitment that motivates action; and how commitment overload generates adverse consequences.



### 5.6. *Self-monitoring outcomes: comparison of means and graphical evidence*

The secondary outcome variable is patterns of usage of the Food Monitor tools for tracking diet, exercise and weight, to assess whether the commitment devices play a role in improving self-monitoring behaviours. Figure 25 presents self-monitoring outcomes across experimental groups, and it is immediately clear that the limited commitment group used the service the most, followed by the financial commitment group and closely after by the reputational commitment group. The patterns are somewhat surprising, since theory would suggest that commitment devices encourage greater self-monitoring, as a form of positive health behaviour change that requires effort in the short term.

Figure 25



## 5.7. Regression analysis

Regression analysis delves deeper into these patterns (Table 20), and corroborates the graphical evidence above: the refund group were significantly more likely to use the self-monitoring tools, with an additional four to five logins over a four-week period ( $p < 0.01$ ), translating to a Cohen's  $d$  effect size of 0.36.

**Table 20: Can commitment devices boost self-monitoring behaviour?**

Panel A: Limited commitment		
	(1)	(2)
Refund	5.119** (0.001)	4.861** (0.009)
Controls	No	Yes
Observations	292	278
R <sup>2</sup>	0.034	0.126
Panel B: Reputational commitment		
Coach	-0.949 (0.563)	-0.304 (0.858)
Controls	No	Yes
Observations	263	250
R <sup>2</sup>	0.001	0.069

*Notes: OLS regressions on usage of Food Monitor tools. Panel A models equation 15a to recover the ATE on the refund treatment; panel B models equation 15b to recover the ATE on the coach treatment. Covariates listed in appendix. Treatment effects are compared to the comparison group, the financial commitment group. Robustness check using Poisson regression corroborates findings in appendix.*

*P-values in parentheses \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$*

### **5.8. Discussion: why does reduced commitment spur self-monitoring?**

The relationship again runs in the opposite direction to what theory would suggest: the refund group (least commitment) experienced positive treatment effects, while the coach group (greatest commitment) had no significant effects.

The refund may have increased the salience of the Food Monitor accounts, by creating a sense of brand loyalty or reciprocity. Both factors may have diluted the intended treatment of dismantling the commitment device. The size of the refund had no association with subsequent self-monitoring ( $p > 0.05$ ); rather, simply having a free month's subscription may have encouraged clients to take advantage of 'a good deal'. The treatment message was designed carefully to avoid any sense of it being a reward or a gift from the firm (see figure 18 above), but it was not possible to double check how the refund was interpreted by the participants due to external constraints on follow up.

Secondly, the lack of positive effect on self-monitoring for the coach treatment group is also surprising. An increase in overall commitment to the weight loss goal (by nominating a coach) was expected to increase self-monitoring so these participants would know if they were on track with their goal, and report progress to their coaches after the four-week period. It may be the case that rather than using their Food Monitor accounts, they relied on offline monitoring tools. The results may also corroborate the earlier discussion on commitment overload, and are consistent with a zero treatment effect on weight loss reported in table 19.

### **5.9. Discussion: why does self-monitoring not lead to more weight loss?**

The results also beg the question, why did the limited commitment group not perform significantly better on weight loss given the improvements in self-monitoring? The results appear to challenge the received wisdom on self-monitoring (Boutelle et al. 1999; Yu et al. 2015). One way to square the circle is if self-monitoring is necessary but not sufficient for weight loss. Those who self-monitor more have better chances of losing weight, but only if other health behaviours also change. The expected benefits of self-monitoring on other health behaviours (for example exercising more and eating less, linked to tracking calories) may not have been realised by people in the refund group. On the other hand, those who do not self-monitor (such as the coach group) are less likely to lose weight or may even gain weight; because if they cannot undertake this relatively simple health behaviour change, they may struggle with more demanding changes around diet and physical activity.

### **5.10. Summary: research question 1**

Section 5 presented average treatments of two commitment devices on weight loss and self-monitoring behaviour, generating surprise findings in response to research question 1. In contrast to the hypothesised relationships, reducing commitment boosts self-monitoring behaviour; and adding reputational commitment elements to an existing financial commitment has the opposite effect in reducing weight loss. These results suggest new dynamics at work – such as commitment overload and the persistence of a financial commitment despite a temporary dismantling of the monetary investment – that can inform the theoretical framework and wider literature, and are taken up for discussion in section 7 later. The analysis now turns to research question 2 and models heterogeneity of treatment effects across sub-groups.

## 6. HETEROGENEOUS TREATMENT EFFECTS

### 6.1. Two heterogeneity pathways

This section investigates whether commitment devices vary in their effectiveness based on individual traits, as predicted by the Analytical Framework. Hypothesis 4 states that a commitment device will work less effectively for those exhibiting present bias and short-termism, and this will be tested below using two operational variables: myopia in health attitudes, and time preference measured as the cost of waiting for a future payoff.

### 6.2. Regression model

The statistical models below estimates the conditional average treatment effect (CATE) from the linear combination of the treatment coefficient and treatment x trait coefficient (as introduced in the Research Design). Two separate equations are estimated in order to address the change in treatment allocation rule, in line with the earlier approach on weight and self-monitoring outcomes. The CATE is expected to be negative for the coach treatment (from equation 16b), and positive for the refund treatment (from equation 16a), for both weight and self-monitoring outcomes.

$$[16a] \quad Y_i = \alpha + \beta^R \cdot R + \beta^{tr} \cdot R \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

$$[16a] \quad Y_i = \alpha + \beta^C \cdot C + \beta^{tr} \cdot C \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

Results are corrected for multiple hypotheses testing using the Benjamini-Hochberg procedure (also discussed in chapter 4) (Coppock 2015). All control variables are identical to those used in the statistical model applied above for average treatment effects. For brevity, only CATE estimates are presented in Table 21 below (see appendix for full set of results).

### **6.3. Regression analysis**

#### **6.3.1. Myopic health attitudes**

In support of hypothesis 4, participants with myopic health attitudes experienced stronger self-monitoring when the commitment device was relaxed (through the refund). Table 21 reports a positive coefficient on the treatment-covariate term, although it is not statistically significant at the 5% level (see column 1 panel A). No other statistically significant results were found, suggesting that this heterogeneity either did not extend to weight loss outcomes; or that the study was not able to detect modest differences across sub-groups.

#### **6.3.2. Time preference (cost of waiting)**

Some support was found for heterogeneity in treatment effects on weight outcomes (at 12 weeks) based on time preference. The subgroup of people with a higher degree of present bias experienced negative effects from the reputational plus financial commitment device as expected (see column 3 panel B), and this result remained statistically significant after correcting for multiple hypothesis testing<sup>58</sup>. Why is there no relationship found for short-term weight outcomes and self-monitoring? The cost of waiting measure may be too blunt an instrument to fully capture the kind of short-termism that would affect health choices in particular; and in this sense is open to a wider criticism of the methods that aim to elicit discount rates and measure present bias (Frederick et al. 2002). Graphical investigation corroborates, however, that there is a relationship between degree of present bias (impatience) and weight loss performance. In general, impatient participants are less successful at losing weight (Figure 13). The effect of present bias may be so entrenched that the commitment devices tested here are simply

---

<sup>58</sup> The corrected threshold is  $p=0.025$  based on two hypotheses being tested within that model on 12-week outcomes.

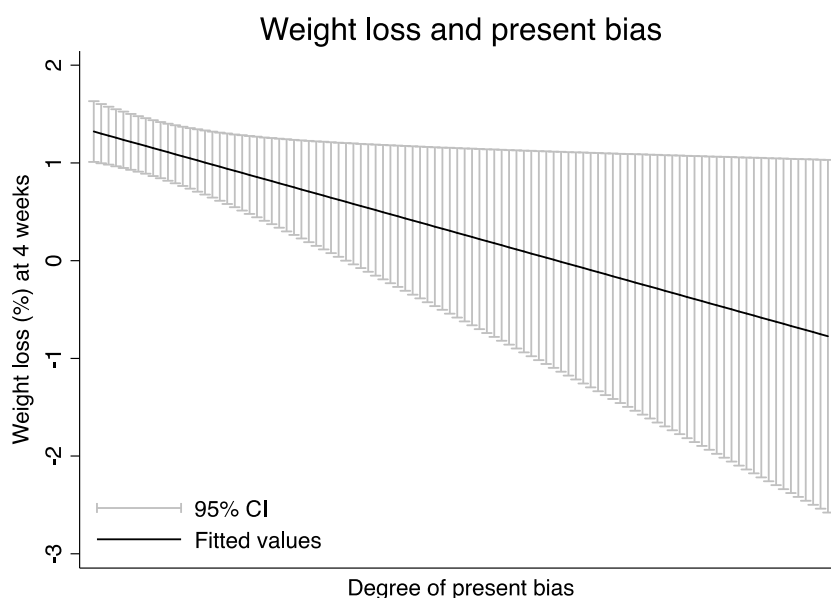
unable to make any difference to that trend in the short run (at 4 weeks); and in the medium run (12 weeks), there was a negative effect of the coach treatment.

**Table 21: Do commitment devices work differently across sub-groups?**

	Self-monitoring (1)	Weight 4 weeks (2)	Weight 12 weeks (3)
<b>Panel A: Limited commitment</b>			
x myopia	4.03 (0.079)	-0.064 (0.881)	-0.884 (0.448)
x present bias	4.11 (0.147)	-0.474 (0.316)	0.761 (0.543)
Observations	271	171	121
R <sup>2</sup>	0.131	0.992	0.968
<b>Panel B: Reputational commitment</b>			
x myopia	0.907 (0.701)	-0.289 (0.528)	0.656 (0.527)
x present bias	-0.027 (0.992)	0.645 (0.177)	2.82* (0.015)
Observations	250	145	106
R <sup>2</sup>	0.075	0.992	0.970

*Notes: OLS regressions on usage of Food Monitor tools in column 1 and weight outcomes in columns 2 and 3. Weight outcomes only for sample aiming to lose weight. Panel A for phase 1 only to recover CATEs on refund treatment. Panel B for phases 1 and 2 to recover CATEs on coach treatment. Column 3 applies inverse proportionality weighting. Original p-values reported here, Benjamini-Hochberg thresholds in appendix.*

**Figure 26**



#### **6.4. Discussion: why weak heterogeneous effects?**

Despite the large range of weight loss outcomes set out in the earlier boxplots (figures 23 and 24), the sub-group analysis conducted here sheds little light on why commitment devices worked differently across the sample. It is possible that the analysis was under-powered, particularly for weight loss outcomes; section 7 returns to this issue to reflect on research design limitations. It is also possible that short-termism was not the key cleavage determining sub-group effects. Regression analysis of average treatment effects in section 5 above raised questions over the role of age and taking part in other activities to lose weight, and future research could consider whether such covariates are better able to predict heterogeneous treatment effects. The second experiment (chapter 6) will consider the role of sophistication, and chapter 7 returns to the idea of adherence to commitment devices as another potential driver of heterogeneity.

The results may also reflect on the precise variables used to operationalise short-termism. Present bias and myopia are fundamental concepts in the planner-doer framework, and more generally the dual-self tradition for explaining time inconsistency (as argued in chapters 2 and 3), and yet is rarely operationalised in the commitment devices literature. Two innovative measures were developed in this trial to pin down these concepts, on the basis that it could predict variation in commitment device effectiveness. The measures were necessarily novel, and in this sense the trial provided a methodological testing ground.

A number of insights emerge. While the two measures relate to the same theoretical idea, the correlation in their findings appears weak: note there are no areas of overlap between panels A and B in Table 21. The cost of waiting measure in particular may be too blunt an instrument to fully capture the kind of short-termism that would



affect health choices in particular; and in this sense is open to a wider criticism of the methods that aim to elicit discount rates and measure present bias (Frederick et al. 2002). The health attitudes measure has the advantage of being better tailored to the health behaviours and outcomes of particular interest to this thesis, and the second field experiment offers a further opportunity to test this measure and triangulate with the Food Monitor results.

### **6.5. Summary: research question 2**

The results provide modest evidence that commitment devices effects vary based on individual traits, with two significant interaction effects detected. Firstly, self-monitoring is especially boosted by the refund for those who are short-termist in their health attitudes. This finding is in line with the Analytical Framework's prediction that those who strongly value immediate gains over deferred gains would benefit less from a commitment device; and indeed they do experience health behaviour improvements when released from a financial commitment. Secondly, the sub-group of more impatient people experienced negative treatment effects from the reputational plus financial commitment device. These findings suggest that the very traits which engender a need for a commitment device might be the ones that make it hardest for it to effectively change behaviours – an insight that helps contextualise the modest effect sizes, but underscores the limitations of applying commitment devices in a health programme. The analysis raises valuable insights into the two new measures used to test short-termism. Wider lessons are taken up for further discussion in the next section.

## **7. DISCUSSION: WIDER IMPLICATIONS**

Sections 5.3 and 5.4 presented a number of interesting findings from the trial. The following section unpacks the puzzles and surprises emerging from that analysis, and draws out the wider implications for the literature on commitment devices. To begin, section 7.1 discusses what the performance of the financial and limited commitment groups implies for how premium payment commitment devices work, and what makes them different to other designs of financial commitment device. In section 7.2 the discussion returns to the idea of commitment overload, as an explanation for the reputational plus financial commitment group experiencing negative treatment effects on weight loss. Section 7.3 argues that design features matter, and section 7.4 reflects on lessons learned for the research design, including external validity issues.

### **7.1. *Diverse mechanisms underpinning financial commitment devices***

The weak treatment effects of the refund on weight loss indicate that removing the monetary aspect of a commitment device does not remove the commitment entirely. To some extent this was expected (hence the group was described in chapter 4 as the limited, not zero, commitment group), but the findings here go further. There are potential insights from this experiment for the question of intrinsic versus extrinsic motivations for health behaviour change, and the diverse effects of financial commitment devices depending on their design.

Recall the typology of commitment devices (chapter 2), which described two different financial commitment devices: a deposit contract, and a premium payment. The Food Monitor membership is a premium payment, unlike the commitment device tested by Volpp et al (2008) where money was lost unless weight loss stayed on track.

The principle of loss aversion that underpins the deposit contract – staking money on an outcome, and losing it unless that outcome is achieved – is different to the willingness to pay extra in order to bind oneself to a desired course of action (in this case, using Food Monitor to frame goals, track progress, and support weight loss). The results from this trial suggest that deposit contracts do not generate a psychological tax ( $\theta$ ) in the same way as a premium payment. The monetary investment that underpins a deposit contract plays a stronger role in changing behaviour, but when removed the psychological tax dwindles rapidly (John et al. 2011). In a premium payment commitment device, the commitment element arises from a willingness to pay, and is ‘sticky’: it persists even when the monetary component is removed. Further analysis finds no relationship between the size of the refund and subsequent weight loss performance at 4 or 12 weeks ( $p > 0.05$ ), supporting the argument that participants valued the Food Monitor membership beyond the precise financial commitment.

## **7.2. Commitment overload**

The Analytical Framework made an implicit assumption of a linear and monotonic relationship between commitment and health behaviour change. What the model did not allow for is the idea that increasing commitment does not always lead to increasing health behaviour change. The possibility of commitment overload indicates some form of upper threshold, a ceiling beyond which more commitment has adverse effects. It is not possible to delve further in to the potential mechanisms explaining the negative treatment effects, but the findings shed light on new research questions around the optimal level of commitment that motivates action; and how commitment overload generates adverse consequences. This finding is new to the literature on commitment devices, but echoes the findings from Verhoeven et al in a wider literature on personal rules that ‘less’ can sometimes be ‘more’ (Verhoeven et al. 2013).

### **7.3. *Design features (d) can interact negatively to reduce $\theta$***

The suggestion to include a public commitment element may have jarred with clients' natural demand for a more private commitment device, which Food Monitor offered, and the overall impact was to undermine motivation. If the negative result from the coach treatment is due to a clash between commitment device design features (online and offline, personal and public commitment), this supports the Analytical Framework's emphasis on design as a predictor of commitment device effectiveness ( $d$ ), but reveals a weakness in the model by not allowing for negative interactions between design features. A wider lesson is the importance of tailoring a reputational commitment design to the target population, considering means of delivery and use (digital or in-person), and preferences for the nature of the commitment (financial or reputational, private or in partnership with others).

### **7.4. *Improving the research design***

There are two prominent ways the research design could have been improved. Firstly, attrition reduced the effective sample size; in hindsight, the sample size calculations could have built in more pessimistic assumptions about drop-out rates. Despite lower than expected effective sample size, statistically significant and valid results are recovered to inform answers to both research questions (including after corrections for multiple hypothesis testing in subgroup analysis) using statistical techniques to mitigate attrition bias.

Secondly, the explanation of the statistical results would have benefitted from qualitative follow-up with participants. For example, interviews would have allowed for testing of how the refund was interpreted, and why the coach treatment led to weight gain. Qualitative analysis could have probed the concept of commitment saturation, exploring how and whether participants experienced a

sense of overload, in order to corroborate or refute the earlier discussions on the unexpected findings.

As argued in the Research Design, all field experiments face the challenge of external validity, and this study is similarly limited in how far results can be generalised. Baseline statistics showed that the sample is over-represented with those who are overweight, obese, and severely obese; those who wish to lose weight; and those who are more short-termist and fatalistic in their health attitudes than the population as a whole. Yet there remains important lessons from the study that are of value to the sizeable sub-population in the UK who similarly are hoping to shed excess weight and tackle their innate health myopia to do so. Another criticism could be levelled at the Food Monitor trial being an online digital health tool, which may appeal to only a limited proportion of that sub-population. However, people are increasingly turning to digital health tools as behaviour change aids (Imison et al. 2016), and the results of this trial are highly relevant to improve the design of such methods.

## **8. CONCLUSIONS**

The chapter presented the results of an experiment testing two commitment devices: one reputational, relying on sharing a weight loss goal with another person to create a sense of external commitment; and one financial, to test the effect of the monthly commitment made by subscribing to the Food Monitor online service. Despite efforts to mitigate attrition and non-compliance, the field experiment experienced both. Nevertheless, the chapter makes three key contributions to the thesis and wider literature: firstly, it informs answers to the research question. Secondly it offers fresh insights for the scholarly debate on commitment devices, both by drawing a clear distinction between the premium payment and deposit contract types of financial commitments, and by flagging the prospect of commitment overload. Thirdly, it informs the theoretical framework applied to this thesis. This concluding section to chapter 5 summarises these contributions.

### **8.1. Contribution to the research questions**

In answer to research question 1, commitment devices can affect weight loss and self-monitoring, but in unexpected ways: the coach treatment exerted negative treatment effects on weight loss at 12 weeks, and the refund had no significant effect relative to those continuing with a financial commitment. In answer to research question 2, the data finds modest evidence of heterogeneity, with significant interactions between short-termism and the effectiveness of commitment devices.

## ***8.2. Fresh empirical evidence on financial and reputational commitment devices***

The results contrast with the literature on deposit contracts, implying that premium payments and deposit contracts have distinct behavioural effects. The refund exerted zero effect on weight loss outcomes in the short or medium term, and it was argued that low cost premium payments (such as the monthly subscription to Food Monitor) have ‘sticky’ commitment elements. Once a payment has been made, sealing a financial commitment, returning the money does not erode the commitment quickly.

The study also showed that design matters, and the behavioural effects of a commitment device will be sensitive to how well the commitment device is tailored to the target population. The reputational commitment treatment encouraged participants to form new external accountability and commitment with another person; but this was perhaps not the optimal design for the kind of individual that signs up to an online self-monitoring service. Participants may have opted in to the Food Monitor service because they value privacy, and prefer to hold themselves to account via the anonymity of digital self-monitoring tools. Having already made a financial commitment in the form of premium payment, perhaps the additional layer of reputational commitment was a step too far; the ensuing commitment overload would explain the low take-up of the coach treatment (40% compliance), low fidelity to the treatment, and the negative effects on weight loss at 12 weeks ( $p < 0.05$ ).

### **8.3. *Lessons for the theoretical framework***

Finally, the trial raises new lessons for theory. Commitment device design and adherence, factors set out in the Analytical Framework, help explain the unexpected zero treatment effect on weight loss and self-monitoring in the short run. However, the model did not take account of the potential effects of commitment overload, because the assumption was that increasing commitment would lead in a linear fashion to increasing behaviour change and weight loss. Rather, the study suggests for the first time that commitment devices may reach thresholds of effectiveness, beyond which they exert negative effects on desired outcomes. The framework's assertion that the three factors jointly affecting the magnitude of  $\theta$  appears borne out; but the model did not account for negative interactions amongst design features that could cause reduce  $\theta$ . The next chapter turns to the second field experiment, testing commitment contracts in a group weight loss scheme.

---



---

# **Chapter 6**

## **RESULTS AND ANALYSIS (2):**

### **Commitment contracts, thresholds and saturation in the Camden experiment**

---

## **1. INTRODUCTION**

Findings from the Food Monitor experiment (chapter 5) indicated that a reputational commitment device in the form of a pledge to a family member or friend did not generate significant weight loss improvements. In contrast, the unshackling of participants from financial and reputational commitment increased their self-monitoring using an online weight loss tool, and boosted their weight loss results. These results run counter both to the hypotheses arising from the planner-doer framework (chapter 3), and a body of published empirical work (chapter 2); they provoke questions about the nature of commitment, and its scope for exerting predictable and desirable influence on complex health behaviour change.

In this second results and analysis chapter, testing continues for causal effects from commitment devices on health behaviour change and weight loss. A randomised controlled trial was carried out in collaboration with Camden Council, nested within an 11-week weight loss group programme called Shape Up. The experiment tests a commitment contract signed to oneself: another test of a reputational commitment device, but one that relies on the principle of self-commitment rather than public commitment as in the Food Monitor experiment.

A contract to oneself can be seen as the weakest sort of commitment (as argued in chapter 2, Table 1). Reneging on the initial intentions bears no monetary costs and relatively limited reputational costs, assuming that psychological costs increase with publicity. Yet it is worth testing, and aims to make a number of contributions to the thesis. Firstly, results will contribute to a relatively sparse literature on commitment contracts, and provide a comparison against different commitment device designs tested in the Food Monitor experiment and wider literature. These results will

speak directly to hypothesis 2 (Table 5, page 103). Secondly, by testing attendance and completion of a publicly provided weight management programme, the results will shed new light on health behaviours beyond exercise, which are rarely discussed in the commitment devices literature (chapter 2). Thirdly, if this form of commitment device works, it is the one that is most easily administered in the context of public health programmes, and may be of value in supporting participation in weight management programmes. Fourthly, the experiment incorporates qualitative methods in the form of semi-structured interviews to delve deeper into how participants approached the commitment device, what influence it had on them and the health outcomes, and to explore wider commitment strategies that are employed in their weight loss journeys – issues that go to the heart of the Analytical Framework, and which promise a useful test of planner-doer theory in describing behaviour change.

The chapter proceeds as follows. Section 2 describes the implementation of the experiment and presents a balance check across the experimental groups. Section 3 summarises baseline data on the 197 participants with a focus on key variables used in the subgroup analysis for research question 2. Section 4 discusses the quantitative and qualitative outcome data and investigates attrition. Section 5 presents average treatment effects on weight loss and participation outcomes. Section 6 presents heterogeneity analysis based on the pre-specific pathways of health motivation and sophistication; with added exploratory analysis on commitment priming, using referral route into the programme, and attendance at an introductory class as proxies for early commitment elements experienced before the contract itself. Section 7 discusses the wider implications of the trial for the scholarly debate on commitment devices and their practical applications.

Section 8 concludes with a summary of key findings from the experiment. Namely, the commitment contract raised attendance and completion rates in the weight management programme, thereby improving self-monitoring and salience of the health goal. It was too mild an intervention to exert significant influence on weight loss itself. However, commitment contracts may yet have a role to play in supporting health behaviour change: both because of the useful improvements to participation in the weight loss programme, and because of its value to many who received it. Participants described it as a useful aid to reinforce their self-discipline, and help keep their health goals from being forgotten in their day-to-day lives.

In answer to research question 2, the contract had differential results for sub-groups as expected. Confirming hypothesis 3 it generated significantly greater participation for sophisticated people. In contrast with hypothesis 5, the contract worked better for those with a myopic and fatalistic outlook to their health. Heterogeneous treatment effects were also examined for two variables that were not pre-specified. The exploratory analysis suggests, firstly, that the contract was particularly effective for participants referred by their GPs rather than self-referred; and secondly, that the contract was especially useful for those people who did not attend an early motivational class. The findings raise interesting questions about the possibility of substitution and saturation effects of commitment devices, and the role of commitment priming in health programmes.

## **2. FIELD EXPERIMENT IMPLEMENTATION**

### **2.1. Recruitment**

Recruitment to the trial commenced in January 2014 and continued over three distinct waves to January 2016 (see Table 22), at which time 197 participants had been recruited. Participants were drawn from Camden's client base spread over 27 Shape Up groups. To be eligible for the Shape Up programme, participants had to demonstrate they were overweight or obese and were local residents. Further detail on the process and timing of the recruitment waves, groups, and tutors are set out in the annex. In total, 208 participants were recruited, with 197 continuing as eligible participants, as mapped out in Figure 27 below (for full CONSORT reporting see Appendix Table A.4).

More participants were approached than strictly required by the sample size target (of 170) as part of lessons learned both from the Food Monitor trial and the first wave of the Camden trial, which registered non-zero drop out rates. Recruiting participants beyond the precise target built in some allowance for later attrition. Recruitment continued up to a natural pause in Shape Up group programmes in spring 2016, imposed by Camden for administrative reasons.

In all, 268 participants approached, of whom 78% agreed to participate. The common reason given by those who declined was difficulty in reading and comprehending the Information Note and Consent Form because English was not their first language.<sup>59</sup> Some participants also cited a lack of time to engage with the registration process, which involved filling out a baseline survey. Of the 208 participants who did register for the experiment, 11 were later excluded due to unforeseeable events that meant they had to leave

---

<sup>59</sup> Developed to comply with UCL ethics standards.

the Shape Up group (of which, 7 were in the treatment group and 4 in the comparison group). These included: serious health deterioration (7), changing health status (2), and administrative grounds for cancelling one of the groups (2). This left 197 participants in the study, of which 49% were offered a commitment contract to sign and take home with them. The target sample size of 170 was met.<sup>60</sup>

**Table 22: Recruitment to Camden trial**

Wave	Duration	Number (% of total)	Groups
1	Jan 2014 – March 2014	57 (29%)	1 – 9
2	April 2014 – Aug 2014	61 (31%)	10 – 19
3	Sept 2015 – March 2016	79 (40%)	20 – 27
	Total participants	197	

Camden employed eight Shape Up tutors over the course of the trial. They are identified by a unique tutor number in the dataset and used as a control variable (see appendix for full list of tutor and group). The intention was to recruit participants as early in the 11-week Shape Up course as possible, but Camden suggested avoiding the first session (termed week zero) when the tutors were focused on introductions and relationship-building. The majority of participants were recruited in week one of the course (51%), but logistical issues sometimes required recruitment took place in week two (34% of participants) or week three (15%). The implication is that some participants will have experienced the commitment contract treatment for a longer period of time over the programme. Data on recruitment week is included as a control variable in later analysis to assess whether recruitment week was associated with attendance or weight loss performance, and suggests no significant effects.

<sup>60</sup> Ex ante sample size calculations presented in chapter 4.

Figure 27: Field experiment flowchart (CONSORT)

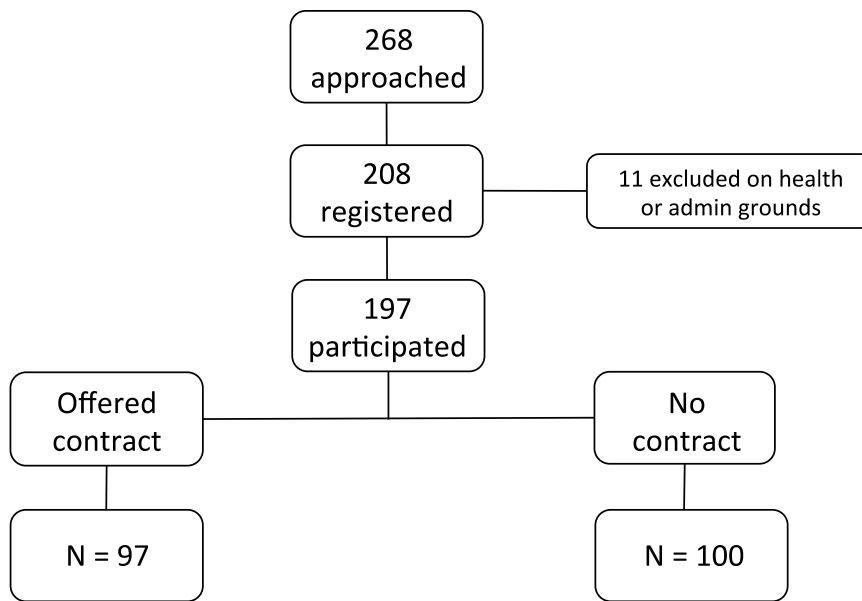


Figure 28: Contract treatment



## **2.2. Randomisation**

The experiment intended to randomise participants in advance using Camden's client lists, which were put together in the days and weeks running up to a new Shape Up group being launched. Participants were randomised using a random number generator in Stata (simple randomisation). Due to administrative inefficiencies, some participants had to be randomised on the day; either because some client lists were not made available in advance, or because some participants to join the programme after it had officially begun (for example in week 2). Their names were not on the group lists shared in advance, and so these participants had to be randomised on the day.<sup>61</sup> In total, 72% of participants were randomised in advance and 28% had to be randomised on the day.

Randomisation on the day was carried out using a simple rule: adding up the digits of their six-figure date of birth, if the sum was an even number they were assigned to the comparison group, otherwise they were assigned to the treatment group. While not perfectly random and not the ideal way of allocating participants to experimental groups, this pragmatic measure did have some strengths. It was quick to carry out during the registration process; it ensured that the probability of allocation to treatment was 0.5; the date of birth was verifiable so would remove any suspicion of researcher bias; and there are no a priori grounds for considering a six-figure date of birth could be linked to any factors that would

---

<sup>61</sup> An alternative approach would have been to exclude these participants entirely, because they were not on the original randomised lists. Given the challenge in recruiting sufficient numbers to meet the sample size target, this would likely have caused the trial to be insufficiently powered, as borne out by the fact that 31% of participants would have been turned away under this approach. Further, there was no reason to expect that recruiting those clients who were not on the original list would affect the representativeness or validity of the sample, as the issue largely lay with administrative processes. So, the best available alternative – on the day randomization – was adopted as a pragmatic solution.



systematically interfere with the treatment or with weight loss outcomes.<sup>62</sup>

Importantly, the randomisation checks reported below (Table 23) demonstrate that this practice did not skew the balance of baseline characteristics between treatment and comparison group. Tests also revealed that there was no association between treatment status and randomisation strategy ( $p=0.770$ ), and the arithmetic rule did indeed achieve a sensible one-in-two probability of treatment.

### **2.3. Randomisation balance check**

Initial characteristics between the two experimental groups are compared using t-tests for continuous variables and two-group proportion tests for dichotomous variables. When testing across multiple variables it is common to find some significant associations as a result of chance, particularly in smaller samples, even if randomization has been undertaken correctly. Glennerster and Takavarasha report on average one out of twenty variables will be unbalanced at the 95% confidence level and one out of ten will be unbalanced at the 90% confidence level (2013, p.151). Table 23 highlights that baseline variables are well balanced between treatment and comparison groups.<sup>63</sup> Other balance checks for categorical variables applying the Wilcoxon rank-sum test show no significant association of treatment assignment with group or tutor. A more detailed discussion of the baseline variables reported in Table 23 follows in section 3.

---

<sup>62</sup> A similar randomisation process is reported by Giné et al in their field experiment testing commitment devices on smoking cessation, where recruiters assigned participants on the spot using a simple arithmetic rule based on their date of birth (2010, p.210); and a further example is cited in Torgerson and Torgerson (2008, p.50).

<sup>63</sup> Based on hypothesis testing of covariate means across experimental groups. Two variables – exercise and initial weight – are mildly associated with treatment status ( $p<0.1$ ). As with outcome data, it is feasible to apply statistical techniques to correct for type I errors in multiple hypothesis testing during a balance check. Following a Benjamini-Hochberg correction to the significance thresholds, these variables are no longer significantly related to treatment status.

<b>Table 23: Randomisation balance check</b>		
Baseline Characteristic	Comparison N=100 (1)	Treatment N=97 (2)
Start weight (kg)	100 (15.0)	82.1 (12.4)
Body mass index	31.3 (3.8)	30.7 (3.64)
Age (years)	50.2 (15.4)	47.4 (14.3)
Female	0.83 (0.38)	0.84 (0.37)
Wellbeing (0-10)	6.3 (2.1)	6.4 (2.3)
Referred by: - self	0.36 (0.48)	0.39 (0.49)
- GP	0.3 (0.46)	0.34 (0.48)
- Other health professional	0.29 (0.46)	0.26 (0.44)
Attended introductory class (%)	0.65 (0.48)	0.68 (0.47)
Fruit and veg intake per day	3.8 (1.7)	3.7 (2.0)
Exercise sessions per week	1.63 (1.60)	1.3 (1.3)
Experienced major life changes	0.36 (0.48)	0.34 (0.48)
Sophisticated (%)	0.30 (0.46)	0.33 (0.47)
Other activities (%)	0.68 (0.47)	0.71 (0.46)
Myopic health attitudes (%)	0.49 (0.50)	0.59 (0.49)
Recruitment week (%) - <i>Week 1</i>	0.52 (0.50)	0.51 (0.50)
<i>Week 2</i>	0.34 (0.48)	0.34 (0.48)
- <i>Week 3</i>	0.14 (0.35)	0.15 (0.36)
Recruitment wave (%): - <i>Wave 1</i>	0.28 (0.45)	0.30 (0.46)
- <i>Wave 2</i>	0.32 (0.47)	0.30 (0.46)
- <i>Wave 3</i>	0.40 (0.49)	0.40 (0.49)

Notes: Standard errors in parentheses. BMI n=190, treated n=93 and comparison n=97; Fruit and veg intake n=138, treated n=71 and comparison n=67; Age n=193, treated n=96 and comparison n=97; Wellbeing n=190, treated n=92 and comparison n=98.

#### **2.4. Group dynamics and spillover effects**

While the syllabus is consistent across all groups, the fact remains that there will be differences across groups and group-based dynamics may play a key role in determining how effective the course is for people, how much they enjoy returning, and how well they are able to absorb the course content. The Research Design highlighted potential threats to internal validity from contamination and group spillover effects. Data on group level characteristics was therefore included in the dataset to serve as controls in the regression analysis assessing treatment effects, and also to allow for specific investigation into group spillovers.

First-hand observation during my participation in the Shape Up classes, discussions with group tutors, and follow-up interviews with participants all gave no evidence of cross talk between treated and untreated individuals. However, some interviewees raised the issue of group dynamics, in terms of friendships and a sense of shared purpose, using the network both to encourage and to seek encouragement. One participant remarked:

*“I was meeting other ladies there, and they were talking about their challenges and that makes your challenge feel it’s not so big. Also you had the opportunity to talk about different ideas of what they were doing in order to get fitter... The group was such a nice group because I felt like we all had something in common, and you got the support from the group as well. And I think that’s important because it motivates you to come to the group.” – female, age 68, ID 30068*

The excerpt highlights the potential for implicit group spillover effects, perhaps by creating virtuous circles of progress, motivation, encouragement, and further progress. While qualitative

data raises the possibility of their existing, there is no evidence from regression analysis (presented in section 6 below) that significant spillover effects were in play.

### **3. DESCRIPTIVE STATISTICS**

#### **3.1. Participant profile**

A typical participant was female, obese, and in their late forties. The remainder of this section summarises key baseline variables across the sample: weight and body mass index (BMI), gender, short-termism in health attitudes, and sophistication; the latter two are traits used for sub-group analysis later in the chapter. In addition, section 3.7 below considers baseline variables that might indicate forms of ‘commitment priming’, which form the basis for exploratory sub-group analysis in section 6. Other baseline variables reported in the balance check are summarised in the appendix.

#### **3.2. Weight and BMI**

Average starting weight is 84 kilograms, ranging from 54 to 139 kilograms. Average body mass index (BMI), calculated as weight/height<sup>2</sup>, is 31.0 ranging from 24.5 to 47.1.<sup>64</sup> The distribution across the standard four BMI categories is shown in Table 2 and Figure 3 below: a small minority have a healthy BMI (a score below 25), although the average amongst these 5 participants is at the top end of the cut-off point, suggesting these participants are at risk of becoming overweight (BMI between 25 and 30). In all, 98% of participants are carrying excess weight, meaning they have a BMI greater than 25. The modal category is obese (52% of participants) followed by overweight (39%). The histogram shows that the overweight are concentrated at the higher end of their BMI interval,

---

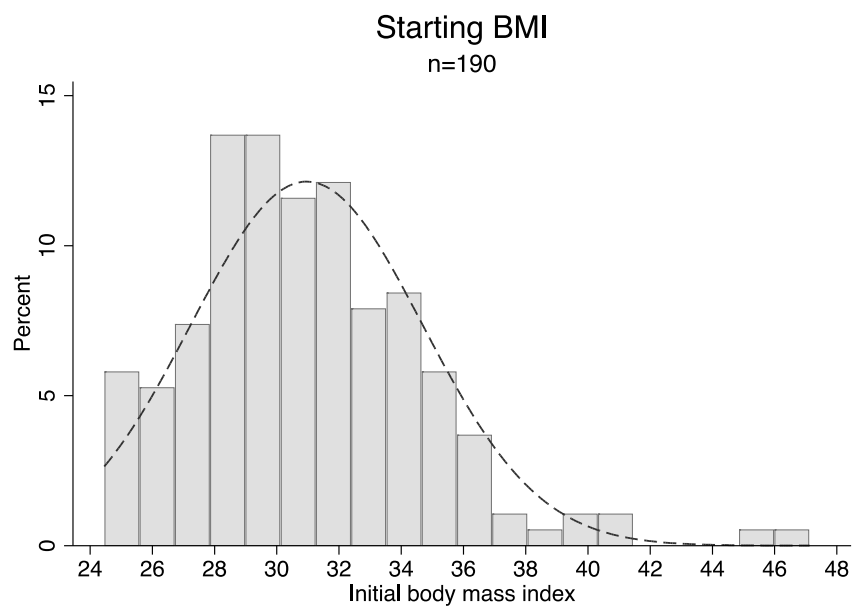
<sup>64</sup> Starting weight is available for all 197 participants. Due to administrative errors within Camden’s systems, height data was missing for 7 participants so BMI could not be calculated for these participants.

indicating these are individuals who are identified as being at risk of obesity (BMI over 30). A small proportion (6%) is severely obese (BMI of 40 or over).

	Mean	Range	SD	N
Initial weight (kgs)	83.7	53.7 – 139.4	13.8	197
Body Mass Index	31.0	24.5 – 47.1	6.5	190
	%	Mean	SD	N
Healthy, 18.5 < BMI < 24.99	2.5	24.7	0.2	5
Overweight, 25 < BMI < 29.99	39.1	28.0	1.3	77
Obese, 30 < BMI < 39.99	52.3	32.9	2.1	103
Severely obese, BMI > 40	6.1	42.8	3.3	12

*Notes: n = 197 for starting weight; n = 190 for BMI (see footnote 5)*

**Figure 29**



### **3.3. Comparisons with Food Monitor sample<sup>65</sup>**

These average characteristics are broadly similar to the Food Monitor sample, but a key difference is the much lower proportion of healthy BMI individuals in the Camden sample, as seen by the truncated distribution in Figure 29. The apparent cut-off at the upper threshold for normal BMI is explained by the eligibility criteria for the Shape Up classes, and demonstrates that the screening process was being administered effectively. Individuals with a lower BMI (well in to the normal range) would not have been referred to the programme by health professionals, and would not have been accepted on to the programme by Camden if they were self-referred. Another difference is that the Food Monitor sample had a closer balance between the obese and overweight groups, whereas Camden has a much higher proportion of obese participants. This, too, could be explained by the referral processes involved; and the fact that the Shape Up classes, being more intensive than the Food Monitor web tool, may appeal more strongly to those who feel in need of greater, personal support to address their weight management issues.

### **3.4. Gender**

The sample was made up of 164 women (83%) and 33 men (17%), in line with the Food Monitor experiment and many weight loss programmes who report a disproportionately female membership (Jolly et al. 2011, p.6). Further investigations into whether gender was correlated with any other characteristics revealed no significant associations; in particular, there is a good balance between treatment groups between men and women. The

---

<sup>65</sup> Baseline variables collected are broadly similar to those reported in the Food Monitor experiment (chapter 5), with the exception of some demographic variables: income, educational background, and employment. These socioeconomic variables were not included in the baseline survey for Camden participants in the interest of ensuring a streamlined registration process to fit with the recruitment processes in this experimental design. This experiment also excluded the discount rate question based on pre-testing with Camden staff and a class tutor, so comparable data on time inconsistency is not available in this chapter.

average age of participants was 49 years for both men and women (unlike the Food Monitor experiment, there is no differentiation of age between genders).

### **3.5. Short-termism in health attitudes**

Short-termism is expected to be a predictor of how well the commitment device works across people (chapter 3). The Food Monitor experiment measured short-termism through two proxy variables: time preference and myopia in health attitudes. The Camden experiment focuses solely on myopia in health attitudes. Just over half of participants reported myopic health attitudes (Table 25). In terms of the dual-self framework, these participants are most likely to have their doer sub-selves dominate their choices, exhibit time inconsistency, and fail to achieve their health goals. They are therefore hypothesised to benefit less from a commitment device, and this will be tested in the sub-group analysis in section 6.

**Table 25: Descriptive statistics: myopia**

Short-term health attitudes (n=197)	%	N
Myopic	54%	106
Far-sighted	46%	91

### **3.6. Sophistication**

The trial generates a novel proxy variable to capture self-awareness (chapter 3), by asking if participants had taken part in a weight management programme previously. The majority of participants had not done so (69%, n=135). Those who had done so (31%, n=62) cited Weight Watchers, Slimming World, and nutritional regimes such as Dukans Diet. The binary variable to indicate sophistication is used to test for heterogeneous treatment effects later in the chapter.

### **3.7. Commitment priming**

The Camden experiment differs from the Food Monitor trial because it allows for further testing of reputational commitment mechanisms that may be at work even before the clients enrol in the Shape Up programme (which may be present for Food Monitor clients, but not observable in that trial). These mechanisms are the referral route on to the programme, and the introductory session of the Shape Up classes, both of which may serve to prime participants on the importance of staying on track with their health goal, in other words priming them to stay committed. This priming may make the commitment contract more effective, and the Camden data allows this idea to be tested in two ways.

#### **3.7.1. Referral routes: GP, health practitioners, and self**

Participants came to the Shape Up programme in different ways and data on referral route was captured by Camden administrators. The most common routes were by self-referral (38%, n=74), followed by the GP (32%, n=63) or some other health practitioner (27%, n=54). Qualitative evidence fleshes out these findings further with interviewees asked what prompted them to join Shape Up when they did. Health fears were frequently cited, particularly in the context of discussions with GPs on issues such as diabetes or cholesterol. In these conversations, weight loss was often part of a wider strategy to improve their longer-term health, for example:

*“My main motivation is my health deterioration fears. I don’t want to end up with diabetes. I know that a year ago I was quite close. Now I’m quite a way off of it, which is good.”  
(male, age 50, id 30075)*



As well as GPs, many participants came to know of the Shape Up programme through other health professionals. For example, some participants with very high BMI scores had been part of other health initiatives by Camden to address severe weight management issues, and having progressed through those programmes were then referred to the Shape Up programme as a follow-on step. Other participants mentioned the NHS Health Check initiative, which is targeted at those who have recently reached the age of 40. More than a quarter of participants referred themselves to the programme. Sometimes this was prompted by family encouragement, and at other times by life events such as a wedding:

*“ [It was] just the feeling like I’m done being fat, and...I want nice clothes, I want... And part of it was we are planning our wedding in September, and I want to look, you know, I want to look nice. (female, age 35, id 11550)*

One severely obese interviewee explained her rationale as follows:

*“...it was a feeling that, if I didn’t try this, I was going to go down a worse road and I had to at least give it a chance, and give myself a chance. And it was almost a desperation.” (female, age 60, id 40028)*

The method of referral into the programme may be of greater significance if some routes involved commitment features that prime participants in advance of receiving the commitment contract itself. Allen et al (2015) highlighted the reputational commitment that some individuals felt towards their GP when they were given free access to a commercial weight loss programme. If similar feelings arose in the Camden trial participants when referred by health professionals, this could create a positive interaction with the commitment contract: signing the commitment contract reinforces and adds to an existing commitment (the doctor’s referral) and brings about stronger adherence to the programme. Interview feedback suggested this could indeed have taken place:

*“I went into my 40s, and the GP gave me this pass for free to enjoy the Shape Up programme, so I called. And the lady explained it to me, ‘are you ready to commit yourself to exercise and stuff like that?’, I said ‘yes!’, it’s what I’m looking for...” (male, age , id 30030)*

Arriving through a GP referral may also shed light on the individual’s (largely unobservable) innate motivation. A client referred by their GP rather than through a self-referral might have less motivation overall (they were ‘sent’ to the programme, rather than ‘signing up’ for themselves); or perhaps is more likely to be time-inconsistent, procrastinating about their health status until the doctor tells them what to do. For these individuals, the contract might plug a useful commitment gap, in line with planner-doer theory, leading to a larger, positive treatment effects for this subgroup. Having data on the variation in referral route meant it was possible to explore this potential heterogeneity pathway (reported in section 6). Binary variables controlling for referral route are also incorporated as control variables in the statistical model for average treatment effects.

### **3.7.2. Attending the introductory Shape Up class**

The qualitative data highlights that participants generally came on to the programme with a sense of purpose, and a clear notion that success would involve losing weight. Having decided to enrol on the programme, however, not everyone joined at the same stage. The 11-week programme is divided into an introductory session (week 0), nine substantive topics (weeks 1-9), and a final class on reflection and evaluation (week 10). Camden strived to ensure that all clients were booked in time for the introductory session, but in reality 66.5% (n=131) did so, with the rest joining from week 1 (26%, n=51), week 2 (7%, n=14) and in one case week 3 (0.5%). This variation across the sample may have substantive implications.

The introductory session does not begin nutrition and lifestyle coaching in earnest, but focuses on social introductions and team building; building rapport with peers and the tutor; understanding the format of the class (such as weekly weigh-in with the digital scales); explaining that Shape Up is about understanding themselves and their habits (“it’s a lifestyle, not a diet” is a common refrain from tutors); strongly encouraging clients to attend every week and complete the whole programme; and highlighting incentives to do so (such as free gym passes and membership).

All of this may introduce another source of ‘commitment priming’, with clients who attended the week 0 class experienced a stronger degree of reputational commitment to the group and the tutor, and in this sense ‘primed’ before receiving the commitment contract itself. Alternatively, it may also be the case that both comparison and treated participants felt a sense of commitment, which arguably makes it more difficult to detect treatment effects from the contract alone. Camden’s administrative records indicate whether participants joined attended in week 0, and this data is used to create a binary variable as a proxy for commitment priming.

Later analysis (see section 6.3.3 in this chapter) will explore whether commitment features like referral route and attending the Introductory session exert an additive effect on the commitment contract’s influence; or whether, as suggested by the Food Monitor results, individuals experience a form of commitment saturation and no further benefits from the treatment are found because other commitment features have already brought about behaviour change.

## **4. OUTCOME DATA**

This section provides an overview of distribution and average outcome data, discussing weight loss and participation in the Shape Up programme in turn; attrition; and strategies for mitigating threats to internal validity.

### **4.1. Weight loss data**

Weight loss outcomes were gathered from class registers at the end of each course, measured as weight change as a percentage of initial weight. Ideally, the final weight loss outcomes would be taken from week 10 class registers, but participants were not always able to attend the final session. Given the practice for Camden to treat week seven as a marker to identify completers, final weight loss outcomes were taken from the last reading available during weeks seven to ten. Follow-up with participants who failed to show up to the final classes aimed to secure a self-reported, recent weight reading. The exercise was successful for 11 participants and improved the completeness of outcome data, bringing down the number of missing observations to 36 and the attrition rate to 18%, in line with other health behaviour change studies (Elobeid et al. 2009, p.5; Chapman et al. 2015, p.731).

### **4.2. Attrition on end weight readings**

Attrition was anticipated in the research design based on information from Camden on drop-out rates for the Shape Up programme – indeed this motivated their interest in examining completion rates and attendance as outcomes affected by the commitment contract. Data from weeks 7-10 indicates fewer attritors in the treatment group, but this difference is not statistically significant ( $p=0.169$ ). On this basis, the main outcome variable that will be used in all statistical analysis for the remainder of this chapter is weight loss as a percentage of initial weight drawn from weeks 7-

10. Two alternative outcome measures are discussed in the appendix (section A22) and used for robustness checks (section A23).<sup>66</sup>

**Table 26: Attrition patterns across experimental groups at 7-10 weeks**

Missing outcomes	All sample (1)	Comparison (2)	Treatment (3)	p-value (2) = (3)
%	18.3	22.0	14.4	0.169
N	36	22	14	

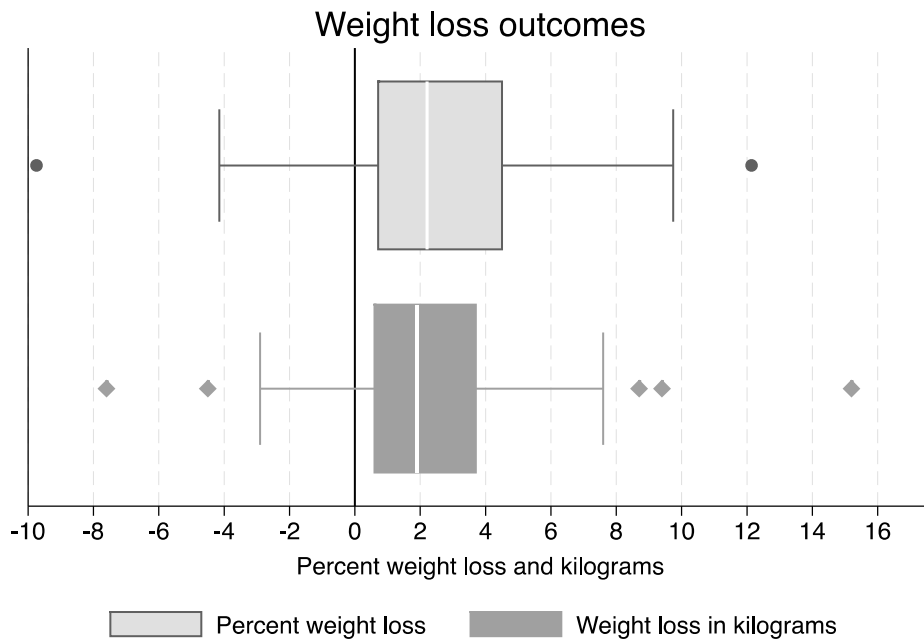
### 4.3. *Weight loss average outcomes and outliers*

The distribution of weight loss outcomes is summarised in Figure 30, measured as percentage weight loss and also in kilograms for illustration. While the median weight loss is only slightly above 2%, the boxplot has long whiskers reaching much higher values of 10% and beyond. The spread also demonstrates that success was not universal: 18% of participants gained weight during the programme (n=22).<sup>67</sup> The boxplot highlights a small number of outliers, identified as those who are more than 3 times the inter-quartile range above or below the edges of the box. These five outliers (two at the weight gain end of the spectrum, and three at the weight loss end) were investigated further to assess whether the weight change trajectories was plausible, and to rule out any data inputting error by myself (for full details see appendix). Verification of the outcome data by returning to the Camden sources, and triangulation of the specific cases with tutors provides sufficient assurance that these outliers are benign, and are therefore included in all analysis below.

<sup>66</sup> Alternative outcome measures include: the ‘last observation carried forward’ (LOCF) that is common in weight loss studies and effectively ensures no attrition because every participant has at least one observation to roll forward; and a narrower window of weeks nine and ten for final weight readings, with inverse proportionality weighting to account for differential attrition.

<sup>67</sup> Figures are similar to those reported in the Food Monitor experiment, where the refund and financial commitment groups lost 2.5% of initial body weight, and 14% of those aiming to lose or maintain their weight actually gained weight.

Figure 30



#### 4.4. Attendance and completion data

The secondary outcome variable relates to health behaviour in terms of sustained participation in the 11-week weight loss programme, measured using attendance rates and completion rates. Chapter 2 pointed out that health behaviours are often equated with exercise in the literature on commitment devices (Prestwich et al. 2012; Royer et al. 2015), and this dissertation aims to broaden the investigation of behavioural impacts to consider other, relevant health behaviours. The Food Monitor trial examined self-monitoring behaviour using the digital calorie-counter tool, and the Camden trial will focus on attendance at the Shape Up programme as a measure of sustained adherence to the weight loss principles and self-monitoring through weekly weigh-ins.

Data was again gathered from class registers, which were maintained on a weekly basis by tutors. Completion status (a binary variable coded 0 or 1) is based on whether the participant attended at least 7 classes. Attendance rates are calculated by the number of

sessions attended as a proportion of total sessions the individual could have attended.<sup>68</sup> Both attendance and completion variables had a complete dataset (no attrition, n=197). Table 27 presents summary statistics on participation rates.

**Table 27: Average participation rates**

	Mean	SD	Range
Attendance	0.69	0.25	0.1 - 1
Completion	0.73	0.44	0 - 1

By design, attrition was not a threat to the attendance and completion data, and neither were measurement errors. Table 27 reports the lowest attendance rate was 10%, which occurred in four cases where a participant attended a Shape Up class for the first time in week 1, coinciding with recruitment for the research project, but did not return to any further classes. The highest attendance rate was 100%, achieved by 22 participants (11%), with no difference across treatment groups (n=11 in each group).

The follow-up interviews offered an opportunity to probe reasons for dropping out or low attendance. Many participants cited illness such as a cold or flu, caring responsibilities for family, or work commitments crowding out their personal time. Some participants felt that by missing out on a few sessions they had fallen behind, and were planning to re-join a fresh class at the next opportunity. Some participants suggested the class was ‘not for them’, perhaps because of the tutor’s style, the challenge of getting to the venue at the regular time slot, or because the content itself was either not sufficiently challenging, or did not meet their expectations on physical activity (with some classes having too much and others too little).

---

<sup>68</sup> The denominator varied in some instances if, for example, a participant joined the course not in the introductory session (week 0) but in week one or two, and therefore had a maximum possible attendance of ten or nine sessions respectively. Some groups had condensed courses of ten weeks, for example group 24, which was due to conclude on Boxing Day and instead concluded the week earlier.

#### **4.5. Qualitative data**

Qualitative insights are incorporated throughout the results sections following, with the main objective (in this chapter) to contextualise the average and heterogeneous treatment effects uncovered through regression analysis. Data was gathered through 24 semi-structured interviews held soon after their Shape Up groups concluded and their final weigh-in readings were recorded. As discussed in chapter 4, interviewees were selected through convenience sampling, with an open recruitment policy of accepting willing trial participants without any selection rules, until a reasonable sample size was achieved. To provide additional opportunities for willing participants to engage, they were offered a choice of face-to-face meetings or over-the-phone. This flexible approach allowed for the target number of interviews (set at approximately 20 in the Research Design) to be exceeded.

The main disadvantage is that “there is no way to tell what wider population the sample group represents or how the sample might differ from other potential samples” (Tansey 2007, p.769). This trade-off was judged acceptable because the main aim of the interviews was to gather new, rich detail on the behaviour change endeavours and experiences of trial participants, those most closely involved with the process of interest, without seeking to generalise to wider populations. Ultimately, the interviews were usefully balanced between treatment and comparison groups: they spanned the full range of initial BMI categories, took place over all three waves of the trial, included people aged 29 to 74, with varied weight loss performance from 0.9 kg gain to 6.4 kg loss, and of both genders (see Appendix for topic list and summary of interviewee characteristics). However, selection issues are unlikely to have been fully overcome; for example, those responding were more likely to be fluent in spoken English, and had not dropped out early on and been lost to follow-up. While recognising these limitations, the exercise nevertheless



provides valuable and novel insights into the experiences of applying a commitment contract to place alongside the statistical results, and offers a test bed for the combination of follow-up interviews within the trial for future commitment device research.

Transcription and coding was conducted with NVivo for Mac (version 11) in two phases. In line with Hsieh and Shannon's definition of directed qualitative content analysis (2005, p.1286), codes were defined before and during data analysis, derived from theory and the data itself. The first phase applied a preliminary coding scheme (see chapter 4) to three interview transcripts to test the utility of the codes created in NVivo. These initial codes sought to contextualise treatment effects using three meta-themes: how did the interviewee explain their weight loss outcomes? How helpful did they find the Shape-Up course? And for treated individuals, how did they use the contract?

The initial coding exercise highlighted the need for a clearer unitisation policy, and an expansion of certain codes to better capture the nuance and complexity of participant reflections. For example, reflections on the contract referred both to its salience in their memory, and to its utility. Further, reflections on weight loss performance led to four distinct issues from not remembering the weight loss target, to changing behaviours, changing attitudes (which may not necessarily have been accompanied by a concrete change in behaviours), and wider discussion of circumstances and challenges that may have thwarted early good intentions to lose a significant amount of weight. A refined coding scheme of five meta-themes and 11 sub-themes was developed and applied to all 24 transcripts (see Table 28 below). The resulting analysis is woven through the following sections wherever it illuminates the research questions around average and heterogeneous treatment effects, either to corroborate or challenge the statistical analysis.

**Table 28: Coding scheme for interview data**

<b>Theme</b>	<b>Sub-themes</b>	<b>Definition</b>	<b>Examples</b>
<b>1. Initial motivation</b>	Health issues	Health issue or health status, doctors appointment, medical advice or referral	<i>“my main motivation is my health deterioration fears. I don’t want to end up with diabetes”</i>
	Broader changes in lifestyle	Wider changes to home or work environment or responsibilities	<i>“I went into my 40s, and the GP gave me this pass for free to enjoy the Shape Up programme”</i>
<b>2. Explaining weight loss outcomes</b>	Awareness of weight loss performance	Can they recall their last weigh-in and how it relates to their 5% target?	<i>“No, to be honest I wasn’t really keeping track. I think it was something like 93?”</i> <i>“I think it was about 5 kilos... I definitely met the 5% target”</i>
	Behaviours the changed over the programme	Any behaviours they initiated, substituted, reduced or increased	<i>“I started using a pedometer”</i> <i>“My portion sizes were too big...I’ve reduced the size of my portion...I weigh food”</i>
	Attitudes that changed over the programme	Any change in attitude or opinion about diet, exercise, or wider lifestyle choices	<i>“Regular eating has revolutionised my way of thinking... rather than feeling I had to sneak food, I could actually eat it”</i>
	Challenges in meeting weight loss goal	How easy or difficult was it to change habits, stay on track with the programme? How do they explain their weight loss result?	<i>“I have a problem establishing a regular pattern, both in the, you know, regular eating pattern that the course stressed, and in the regular exercise pattern”</i>

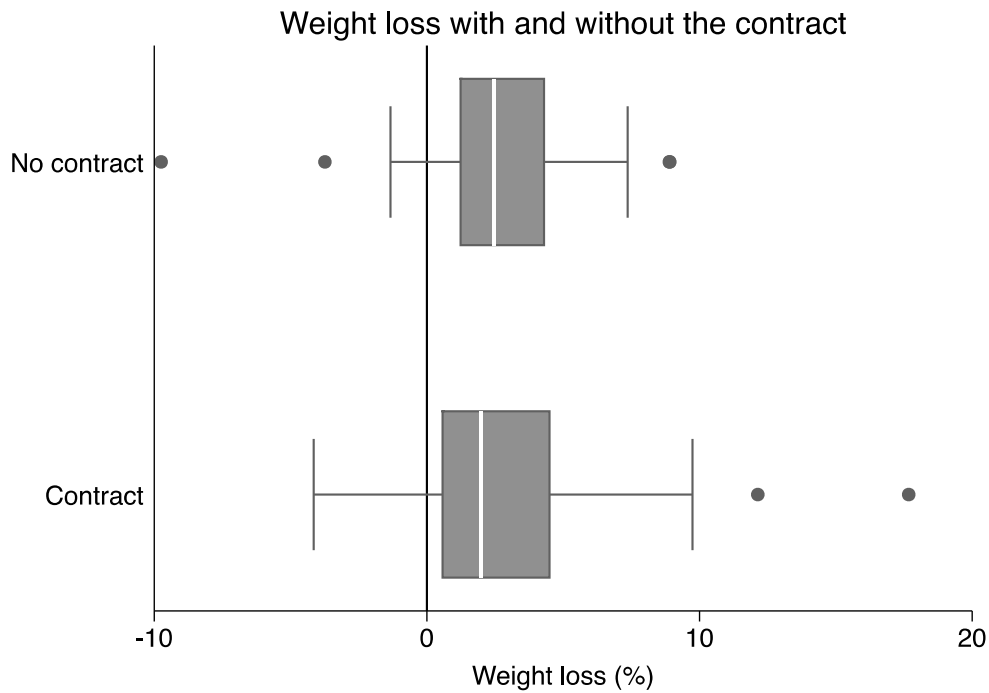
<b>Theme</b>	<b>Sub-themes</b>	<b>Definition</b>	<b>Examples</b>
<b>3. Commitment contract</b>	Saliency	Did they remember the contract? Where did they put it and who else may have seen it? Did they discuss it with anyone?	<i>“Yes I remember [the contract]”...it’s in my desk” “To be honest with you it wasn’t on the top of my mind, no”</i>
	Usefulness	Opinions on how useful or effective the contract was in encouraging behaviour change and sticking with the programme	<i>“[The contract] was an element of [my] self discipline”</i>
<b>4. Evaluating the Shape Up classes</b>	Negative feedback	Feedback on tutor, content of classes, or administration by central Camden team	<i>“I don’t think the class was really useful to be honest, but just having the regular meet-ups and weigh-ins was helpful, it helped keep you on track”</i>
	Positive feedback		
<b>5. Outlook</b>	Outlook	Any mention of future plans, goals, or intentions	<i>“The changes I’m making, I can continue with and I can still make some more changes as well” “I’m going to go round and book myself into the gym”</i>

## **5. AVERAGE TREATMENT EFFECTS**

### **5.1. Weight loss outcomes: comparison of means and graphical evidence**

With the control group losing 2.6% of their initial body weight against the treatment group's 2.8% ( $p=0.677$ ), the commitment contract appears to have made no impact on weight loss outcomes. Hypothesis tests confirm there is no statistical association between treatment status and weight loss outcomes (on any measure). Triangulating with a measure of how many participants met their 5% weight loss target: overall 17% met this target, with no significant differential across experimental groups; in the comparison group, 16% achieved the target, and in the treatment group 19% ( $p=0.637$ ). Figure 31 provides further visual evidence, with the median weight loss lines set very close together across the two boxes. The boxplots further highlight the wider range of weight loss outcomes amongst the treatment group, foreshadowing considerable heterogeneity of treatment effects.

**Figure 31**



**5.2. Weight loss regression model**

In a final test of the causal effect of commitment contracts, OLS regression is used to estimate the average treatment effect, or intent-to-treat estimate, applying the statistical model set out in chapter 4 (reproduced below).<sup>69</sup> The statistical model includes a set of explanatory variables with robust standard errors clustered at the individual level, as set out in the following equation:

$$[17] \quad Y_i = \alpha + \beta^c \cdot C + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

$Y$  is end weight (kgs). Treatment status is captured by dummy variable  $C$ , where  $C=1$  if the participant was offered a commitment contract. The OLS estimator for  $\beta^c$  provides the average treatment

<sup>69</sup> Chapter 4 also set out an equation without baseline covariates, however the results from this approach had such low explanatory value (very low R-squared) that their contribution to the analysis here is trivial. For transparency the results are reported in the appendix section A.23.

effects for the commitment contract.  $W$  is a series of baseline covariates  $J$ , with coefficients  $\gamma_j$ . These coefficients offer statistical association with outcome variable  $Y$ , and cannot be used to infer causality.

Baseline control variables ( $W$ ) include a set of individual characteristics to control for gender, age, and starting weight. Three further variables aim to control for individual motivation and effort: exercise sessions per week, and binary variables on whether they have experienced a major life change recently, and whether they take part in other activities to meet their weight loss goals. A binary variable on whether the individual has been on a previous weight loss programme is used to capture sophistication. Finally, the Healthy Foundations Segmentation categories are used to identify myopic health attitudes.

While many variables chosen for this analysis mirror those used to analyse the Food Monitor experimental results in chapter 5 (such as gender, age, life changes, and myopic health attitudes), this model has been tailored to the Camden experiment through the group and administrative variables, and the commitment priming variables.<sup>70</sup> Two variables are used to control for different degrees of potential ‘commitment priming’: the referral route on to the programme, using binary variables for GP and health practitioner referrals compared with self-referral; and whether the participant attended the introductory Shape Up session (in week zero).

A key feature of the Camden trial is the role of group dynamics and potential spillovers, as discussed in section 3 earlier. Four variables are included to capture group-level characteristics: the size of the group; amongst those the proportion who were offered a

---

<sup>70</sup> Wellbeing is not included in the reporting here due to a small number of missing observations, which reduce the size of the regression sample. Robustness checks in the annex show that including wellbeing makes no substantive difference to the results.

commitment contract; proportion of attritors in the group; and whether the group met during weekdays or outside of normal working hours. These variables aim to capture various group-level issues, for example the potential effect on participants in a group that may have had a higher than average (50%) treatment concentration. Finally, a series of administrative variables are used to control for differences in the recruitment week, tutors, and the wave of the trial they participated in.

### **5.3. *Weight loss regression results***

The treatment coefficient is negative and indicates the commitment contract caused end weight to be 0.47 kg lower in the treatment group on average, relative to the comparison group. Measured as a percentage of initial weight, the contract caused additional weight loss of 0.56%; and relative to the average sample mean of 2.72% weight loss, this implies a 21% increase. But, with Cohen's *d* calculations generating an effect size of 0.07, the magnitude of the commitment contract's impact on weight loss is small both in absolute terms and in relation to the effect sizes drawn from the literature in Chapter 2. Further, the treatment coefficient fails to meet conventional benchmarks for statistical significance in hypothesis testing ( $p=0.282$ ). This finding holds across a range of weight loss measures and model specifications (see appendix section A23). Overall, the regression results tell us that the commitment contract had no significant effect on weight loss, on average; refuting hypothesis 1.

**Table 29: Can commitment contracts boost weight loss?**

Commitment contract	-0.466 (0.282)
Starting weight	0.971*** (0.000)
Female	-0.211 (0.796)
Age	-0.039* (0.020)
Myopic health attitudes	0.224 (0.687)
Experienced major life changes recently	0.172 (0.782)
Other activities pursued to lose weight	0.792 (0.265)
Sophisticated	0.225 (0.645)
Exercise	-0.259 (0.091)
Daytime slot on weekdays	1.020 (0.168)
Proportion of group attritors	4.148 (0.215)
Proportion of group members in study	0.413 (0.855)
Proportion treated in group	-0.161 (0.118)
Attended week 0	-0.614 (0.166)
GP-referred	-0.563 (0.277)
Referred by other health practitioner	-0.290 (0.650)
Other referral route	-4.103 (0.128)
Tutor 2	1.491 (0.112)
Tutor 3	-0.301 (0.810)
Tutor 4	0.333 (0.806)
Tutor 5	0.378 (0.809)
Tutor 6	-1.796 (0.411)
Tutor 7	0.664 (0.599)
Tutor 8	1.273 (0.151)
Recruited in week 2	0.479 (0.519)
Recruited in week 3	-0.540 (0.597)
Recruited in wave 2	0.489 (0.507)
Recruited in wave 3	1.516 (0.175)
N	158
R <sup>2</sup>	0.974

*Notes: OLS regression on end weight outcomes. Base category 'self-referral' for referral routes, 'tutor 1' for tutors, and 'week 1' and 'wave 1' for recruitment variables. P-value in parentheses.*

A few baseline covariates are positively correlated with weight loss, although there can be no causal inference.<sup>71</sup> In terms of individual characteristics, starting weight was positively correlated with end weight, and older participants were more likely to lose weight; but there is surprisingly little significant association with other individual covariates or potential commitment priming variables.

<sup>71</sup> These variables are illuminated for discussion purposes only and not statistical inference, and as they do not form part of the pre-specified hypotheses they are not subjected to multiple comparison corrections. See appendix for robustness checks.



Group level variables are not statistically significant, but the coefficients are suggestive of some interesting group dynamics. Firstly, sessions that take place during the day mid-week are negatively associated with weight loss outcomes. Secondly, those with a higher group attrition rate report considerably lower weight loss. While this is not a statistically significant finding, the sign of the coefficient runs counter to the idea that attrition favours the more successful participants coming through to record their endline weight readings; rather, this appears to pick up on the idea that if group morale declines, participants are both less likely to attend and also less likely to succeed in their weight loss. On balance, results do not suggest any significant group spillovers, and provide reassurance against this possible threat to the validity of the treatment effects.

Amongst the administrative variables, it appears to make no difference whether a participant registered for the trial earlier on (say, week 1) or a bit later (week 3); and there were no associations between weight outcomes and tutor.

#### **5.4. Discussion: why no effect on weight loss?**

In summary, the commitment contract does not improve weight loss outcomes on average. While the evidence is suggestive that participants offered the commitment contract lose slightly more weight on average than those in the comparison group, this difference is not statistically significant. In response to hypothesis 1, these results indicate that a commitment contract is unable to affect weight loss. So why was the effect of the commitment contract on weight loss so weak?

### ***5.4.1. Mild design***

One explanation is that for a commitment device to effectively overcome the challenges of inertia, status quo bias, entrenched lifestyles and habits, and the grinding task of maintaining willpower on a daily basis over several weeks, would need to exert a sizeable psychological tax. It is not clear that the commitment device design (*d*) in this experiment did, or even could, exert sufficient influence ( $\theta$ ). Health behaviour change is a complex and difficult process, and weight management has been a lifelong struggle for many of the participants in the study. The design used here for the contract was perhaps simply not intensive enough to accelerate weight significantly. This bears out the discussion in chapters 2 and 3 that the intensity of the commitment device will determine its effectiveness on changing behaviours, supporting hypothesis 2.

### ***5.4.2. Strong performance from the comparison group***

It may be the case that the comparison group simply performed better than expected and eroded the possibility of finding a treatment effect with a mild commitment device. The Camden trial is not alone in finding a zero average treatment effect of commitment device, with research by Chapman et al (2015) echoing the result in a test of a personal rule for exercise behaviours. The study compared participants who were asked to draw up an exercise plan and display it somewhere prominent, with a control group who were given information about the health benefits of exercise and asked to display that. While both groups experienced positive change in exercise frequency, the difference between groups was not significant. The Camden trial tests the effects of a commitment contract within a health programme that may (and is certainly aiming to) generate much larger behaviour change. Perhaps there is less scope for the contract to exert further benefits amongst those who take advantage of the Shape Up course; put differently, the comparison group in this

trial make so much progress that the bar is set higher than expected for the treated individuals to outperform them.

### ***5.4.3. Low adherence***

A third explanation relates to adherence – another key factor determining the overall intensity of the commitment device and an important predictor of its effectiveness. The contract was offered to participants with the advice of keeping it somewhere they would see it on most days, but the take-up of this advice was patchy. Qualitative data from follow up interviews confirmed that many treated individuals did not keep the contract somewhere visible, and in some cases it never emerged from their envelope after signing it (Table 30). Others remembered it and reviewed it on occasion, but only amongst a number of other handouts relating to the programme, and with no evidence that it held any special significance to them. These individuals were coded for low adherence.

In contrast, some interviewees embraced the contract, keeping it in visible and salient places, and talking about it with friends and family; they were coded for high adherence. The interview sample does not allow for generalisations or wider inference, but these insights corroborate the argument that adherence is key to commitment device effectiveness.

Although some participants spoke highly of the contract's value to them personally, it seems fair to conclude that the intensity of the treatment was relatively low, and would not therefore have generated a sizeable psychological tax on the doer sub-self. The very nature of the contract being administered early in the programme, by an unfamiliar facilitator instead of the tutor, and as a one-off intervention is also likely to have lowered the strength of the treatment. This feature is in contrast to many commitment device studies, where regular follow-up by the research team continually

maintains the salience of the commitment device amongst participants.<sup>72</sup>

The results reported here make a valid contribution, then, by isolating the effect of the commitment contract without the additional prompts and external influences that may generate a positive bias on the average treatment effect, leading to results that are an artefact of the trial rather than being generalizable to real world policy settings. A further contribution is hinted at by the variation in adherence shown in Table 30, which supports the idea put forward in chapter 3 that the degree of adherence to the commitment device is a potentially important heterogeneity pathway. This is explored further in chapter 7.

**Table 30: Use of commitment contracts by the treatment group**

<p>Low adherence</p> <p>N = 8 47% of 17 treated interviewees</p>	<p><i>“I didn’t have a chance to open it. I completely forgot about using it.” (id 30034, male, age 29)</i></p> <p><i>“Now that you’ve reminded me I can remember. I don’t necessarily remember signing it now.” (id 11407, female, age 45)</i></p>
<p>High adherence</p> <p>N = 9 53 % of 17 treated interviewees</p>	<p><i>“I put it on my fridge.” (id 11411, female, age 60)</i></p> <p><i>“When I look at that card, I automatically remember when I’m going to the kitchen, it’s just focused on eating smaller portions and different types of food, what I’ve learned on the course. So I thought that kind of helps, because when you’re having a full day, you never sort of remember these things but it sort of keeps you focused.” (id 40031, female, age 45)</i></p>

<sup>72</sup> Volpp et al. 2008

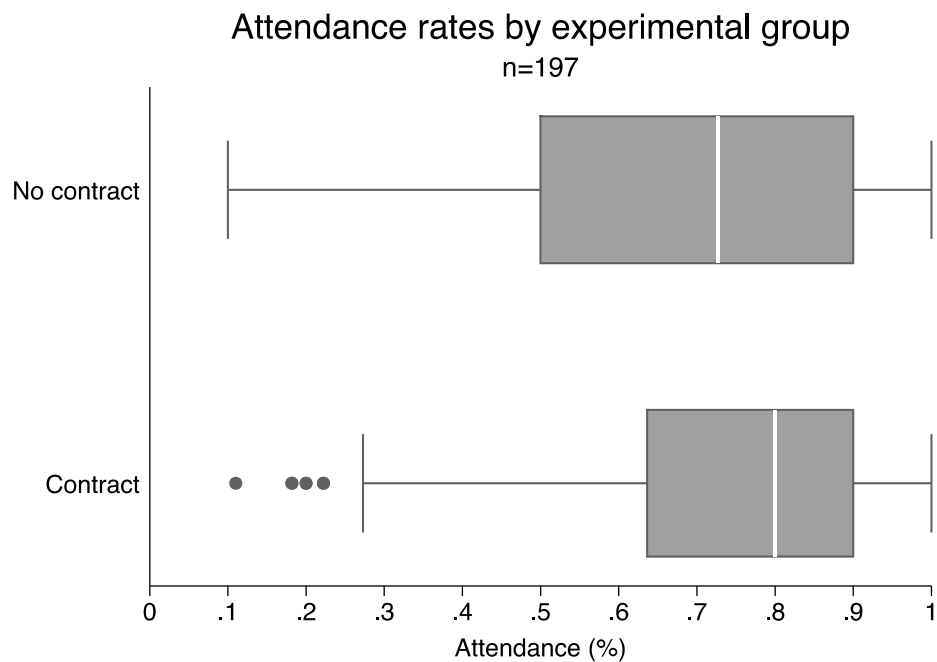
**5.5. Participation outcomes: comparison of means and graphical evidence**

Across the sample as a whole, participants attended 69% of classes, with 73% of participants qualified as completers. For both outcome variables, treated individuals had stronger outcomes: those offered the contract recorded higher attendance by 4 percentage points and higher completion rates by 9 percentage points. However hypothesis testing does not show this association as statistically significant at conventional levels (see Table 31).

<b>Table 31: Participation rates across experimental groups</b>			
	Comparison mean (n=100)	Treatment mean (n=97)	p-value H <sub>0</sub> : equality of means
Attendance	0.67	0.71	0.258
Completion	0.69	0.77	0.188

A boxplot further illuminates the distribution of attendance rates (Figure 32) with a notable contrast between median values and the inter-quartile range across experimental groups. Participants with very low attendance (10-20%) in the treatment group are classed as outliers, with median attendance at 80%.

Figure 32



### 5.6. Participation regression results

Regression analysis relies on the same statistical model set out earlier, with controls for individual and group-level covariates, commitment priming, and administrative variables, and average treatment effects uncovered through an OLS estimator on attendance rates and Probit estimator on completer status (see Table 32). The commitment contract improves attendance by 6% and completion rates by 14%, in line with hypothesis 1, but the effect is statistically significant only for completion outcomes ( $p=0.033$ ). In terms of comparable effect sizes, Cohen's  $d$  calculations suggest these are modest effects of 0.16 on attendance and 0.19 on completion rates.

**Table 32: Can commitment contracts boost attendance and completion rates?**

	Attendance (1)	Completion (2)
Commitment contract	0.062 (0.106)	0.481* (0.033)
Female	0.093 (0.080)	0.439 (0.142)
Age	0.004** (0.002)	0.017* (0.028)
Initial BMI overweight	-0.038 (0.746)	0.379 (0.569)
Initial BMI obese	-0.069 (0.559)	-0.015 (0.982)
Initial BMI severely obese	-0.028 (0.843)	-0.396 (0.625)
Exercise	0.022 (0.113)	0.167 (0.074)
Experienced life change	0.041 (0.264)	0.468 (0.069)
Other activities for weight loss	-0.068 (0.086)	-0.454 (0.072)
Sophisticated	-0.024 (0.560)	-0.196 (0.452)
Myopic health attitudes	0.035 (0.347)	0.302 (0.195)
Referral by GP	-0.041 (0.367)	-0.228 (0.408)
Referral by other health practitioner	0.000 (0.997)	-0.226 (0.457)
Attended Shape Up week 0	0.140** (0.001)	0.752** (0.002)
Daytime slot on weekdays	-0.022 (0.724)	0.311 (0.402)
Number of participants in study	-0.006 (0.407)	-0.033 (0.444)
Proportion of treated individuals	-0.201 (0.282)	-0.409 (0.716)
Recruited in Shape Up week 2	-0.015 (0.855)	-0.114 (0.794)
Recruited in Shape Up week 3	0.068 (0.409)	0.374 (0.490)
Participated in wave 2	0.024 (0.722)	0.017 (0.962)
Participated in wave 3	0.018 (0.798)	0.225 (0.607)
Tutor 2	0.035 (0.635)	0.611 (0.142)
Tutor 3	0.124 (0.060)	0.673 (0.071)
Tutor 4	0.012 (0.917)	0.668 (0.301)
Tutor 5	0.105 (0.130)	0.496 (0.298)
Tutor 6	-0.172 (0.290)	-1.558 (0.081)
Tutor 7	0.027 (0.773)	0.490 (0.480)
Tutor 8	0.031 (0.606)	0.398 (0.312)
Observations	192	192
R <sup>2</sup>	0.129	0.193

Notes: OLS regression on attendance rates, Probit regression for completion rates. Pseudo R-squared reported for Probit regression. Group attrition variable excluded due to risk of simultaneity bias with dependent variables. Probit coefficient on treatment variable implies marginal effect of 0.142. p-values in parentheses: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

Table 32 demonstrates that age and exercise are positively associated with both outcomes, and women are more likely to attend and complete than men. Groups that meet during the day mid-week have higher attendance and completion rates. No other group variables are significant, further supporting the argument above that group-level dynamics are not significant in the quantitative data.

One notable finding concerns the variable ‘other activities’, which would have been expected to exhibit a positive correlation with attendance and completion. Somewhat surprisingly, the opposite is true, with a negative impact on both outcomes ( $p=0.076$  and  $p=0.048$  respectively). This finding could be suggestive of time constraints due to a busy schedule of activities; perhaps the additional activity creates a moral license to skip the occasional class; or these individuals may have found an alternative means of meeting their goal (for example spending more time on physical activity) and no longer feel the need to take part as regularly in Shape Up.

### ***5.7. Discussion: why is the contract effective in raising attendance but not weight loss?***

In summary, the commitment contract causes a significant increase in attendance and completion rates (supporting hypothesis 1). However, this begs the question, when the commitment device is able to shift health behaviours, why have health outcomes not followed? The primary reason may be that the commitment device is strong enough to change simpler, more concretely defined behaviours (‘don’t skip today’s class’), but is too mild to effect more complex changes that require multiple, complementary and sustained actions (‘avoid high fat foods for three months’). The finding is intuitive, and confirms the Analytical Framework’s prediction (hypothesis 2) that design features determine how effective a commitment device might be; and the Literature Review’s argument that stronger commitment devices, particularly those



involving monetary stakes, exert the largest effects on complex change processes such as weight loss. Prestwich et al (2012) offer a rare example of a study that examines the effects of a commitment device on both behaviours and health outcomes; they report that a stronger (more public) version of a personal rule is able to improve both exercise frequency and weight management.

### **5.8. Summary: research question 1**

The Camden trial makes a contribution to the thesis by offering clear answers to research question 1. The commitment contract can change health behaviours, improving attendance (6%) and completion rates (14%) at a public weight management programme. However, the design of the commitment contract was too weak to exert a significant effect on weight loss itself. Qualitative evidence suggests the contract was popular amongst some participants, but easily forgotten by others, helping to explain the weak effects on weight loss. Together, this analysis provides evidence in favour of hypotheses 1 and 2, nuancing the expectations of commitment devices based on the complexity of the desired changes. The next section turns to sub-group analysis.

## **6. HETEROGENEOUS TREATMENT EFFECTS**

### **6.1. *Three heterogeneity pathways: sophistication, short-termism and commitment priming***

Research question 2 asks whether a commitment device may work more effectively for some individuals than others in promoting both. In line with pre-specified analysis plans in chapter 3, this section considers two possible sources of heterogeneous effects – sophistication and myopia – on weight loss and participation.

Sophistication was hypothesised to interact positively with commitment devices (hypothesis 3). Assuming that self-awareness increases from having taken part in a previous weight loss programme, this baseline covariate offers a useful proxy for sophistication. To capture myopia, health attitudes are used to identify a short-termist sub-group. Chapter 3 hypothesised that although they would have most to gain from a commitment device to correct this myopia, they were perhaps least likely to stay committed to a commitment device and reap these potential benefits in practice. On balance, myopic individuals were expected to benefit less from the commitment contract than a more far-sighted comparison group (hypothesis 4).

In addition, exploratory analysis is undertaken to better understand how two further heterogeneity pathways based on the concept of commitment priming (introduced earlier in this chapter, see section 3.7). The aim is to test whether the commitment contract works better or worse for those who experience some form of commitment priming, either by their referral route on to the programme, and if they attended the introductory session or not. In line with existing literature, GP-referral is singled out as a potential driver of commitment priming.

## 6.2. Regression model

Regression analysis incorporates the same baseline model as section 5, now incorporating interaction effects between the treatment offer and specific traits.<sup>73</sup> The statistical model is described by equation 18 below, and the regression results present the combined effect of the coefficients on the treatment and treatment-trait interaction variable for brevity (see appendix section A26 for full results). This combination of  $\beta^c + \beta^{tr}$  yields the average treatment effect within the subgroup, also known as the conditional average treatment effect (CATE) (Gerber & Green 2012, p.296). A positive and significant effect on the CATE indicates that the contract worked particularly well for this sub-group of participants.

$$[18] \quad Y_i = \alpha + \beta^c \cdot C + \beta^{tr} \cdot C \cdot Trait_i + \sum_{j=1}^J \gamma_j \cdot W_{ij} + \sigma \cdot S + \varepsilon_i$$

As discussed in chapter 4, to avoid the risk of type I errors inherent in multiple hypothesis testing, analysis reports both the original p-values and those corrected using the Benjamini-Hochberg technique (see appendix section A27). For brevity, Table 33 focuses on those variables that appear to have a statistical significance using uncorrected p-values, and reports whether the p-value withstands the more stringent significance threshold under the correction procedure (Fink et al. 2014; Coppock 2015).

---

<sup>73</sup> For the model testing weight loss, starting weight is included as a baseline covariate.

### **6.3. Regression analysis**

#### **6.3.1. Short-termism: myopic health attitudes**

The results suggest close association between commitment device effectiveness and myopic health attitudes, for behaviour change outcomes. In contrast to the predicted relationship, the positive sign on the myopic attitudes term confirms that those with a more short-term or negative health outlook benefit more from having a commitment device, bringing about greater attendance and raising the probability of completing the Shape Up course ( $p=0.07$  in both cases). No effects are found for weight loss outcomes (see Table 33 panel A).

#### **6.3.2. Sophistication**

There are significant sub-group effects for attendance and completion outcomes. Those who have undertaken previous weight management programmes respond particularly well to the contract: sophisticated participants have positive treatment effects for attendance and completing the Shape Up course (and withstand the Benjamini-Hochberg correction procedures). This confirms the expected relationship set out in hypothesis 3 between sophistication and commitment device effectiveness.

<b>Table 33: Do commitment contracts work differently across sub-groups?</b>		
	<b>CATE (1)</b>	<b>Where <math>p &lt; 0.05</math>, is <math>p &lt;</math> corrected threshold?</b>
<b>Panel A: Weight loss</b>		
Myopic health attitudes	-0.774 (0.464)	-
Sophistication	-0.755 (0.527)	-
Commitment priming: GP referral	-0.277 (0.845)	-
Commitment priming: week 0	0.147 (0.870)	-
<b>Panel B: Attendance</b>		
Myopic health attitudes	0.159 (0.066)	-
Sophistication	0.260** (0.006)	Yes
Commitment priming: GP referral	0.200 (0.063)	-
Commitment priming: week 0	-0.061 (0.344)	-
<b>Panel C: Completion</b>		
Myopic health attitudes	0.840 (0.073)	-
Sophistication	1.69** (0.004)	Yes
Commitment priming: GP referral	1.19* (0.045)	No
Commitment priming: week 0	-0.109 (0.813)	-

*Notes: OLS regressions recover CATE estimates in panels A and B, Probit regression in panel C. Full regression results and coefficient values in appendix. P-values in parentheses, checked against corrected Benjamini-Hochberg thresholds, see appendix for details.*

### **6.3.3. Commitment priming: GP referral**

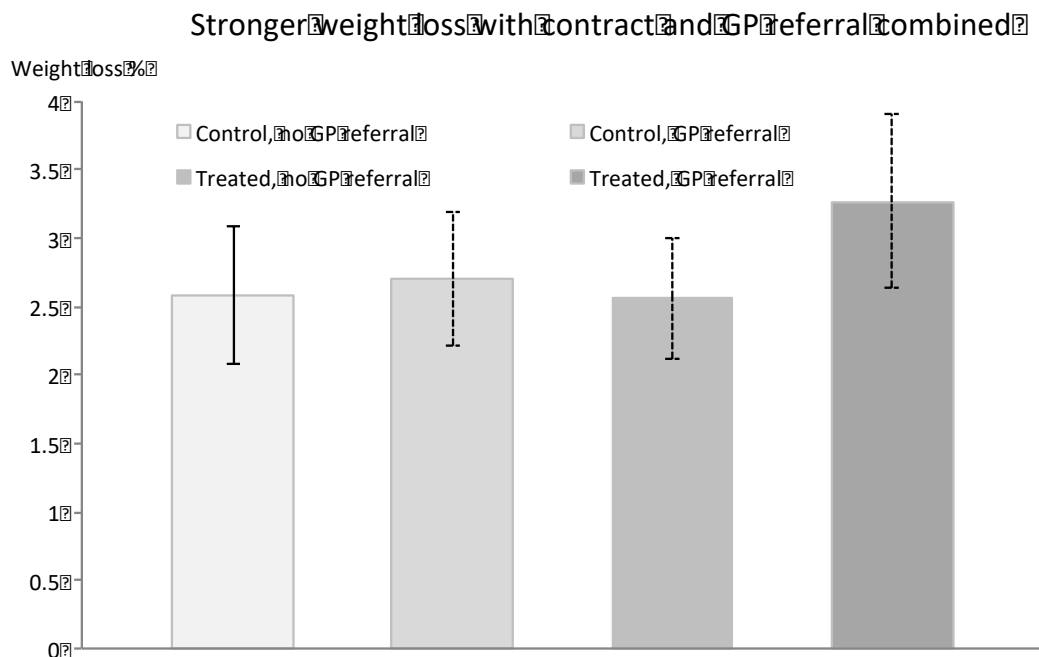
General Practitioner referral was argued to act as a double dose of commitment in earlier discussion, and Table 33 provides a test of this prediction. The regression analysis finds a positive treatment effect for the treated sub-group who are GP-referred, but

only for completion rates, and this effect does not withstand corrections for multiple hypothesis testing ( $p=0.045$ ).

Weight loss performance is expanded across referral subgroups in Figure 33 below, which shows no significant differences; however there is a suggestion from the graphical evidence that the effect of one commitment is almost conditional on the other. Perhaps because the two effects are relatively mild on their own, and only when combined is there a modest behavioural effect. Participants who were neither referred by a GP nor received the contract registered 2.6% weight loss. Those who were both GP-referred and treated, in contrast, registered 3.3%. The difference recorded here is not statistically significant, but future research isolating this particular issue could examine whether commitment contracts are an effective, additive policy instrument in this way.

What is apparent is the contrast with the Food Monitor experiment, which suggested commitment overload when additional layers of commitment were applied to an existing commitment strategy. In the Camden trial, despite there being a potential a double dose of reputational commitment (with the GP, and then with oneself), the contract remained mildly effective, and certainly not counterproductive in the way of the reputational coach treatment in the previous field experiment.

Figure 33



**6.3.4. Commitment priming: early motivational class**

Regression analysis of average treatment effects justified further investigation into the potential effects on the commitment device from attending the introductory Shape Up class. Table 33, however, finds no significant effects, refuting the idea that this early motivational class acts as a commitment primer, interacting with the commitment contract to deliver behaviour change.

However, a closer look at the attendance rate of the four sub-groups suggests there may yet be some noteworthy implications. In particular, Table 14 raises the prospects of the contract addressing commitment gaps that might arise early on and affect outcomes throughout the trial; and also provides evidence of commitment saturation, echoing findings from the Food Monitor trial.

Firstly, attending the introductory session appears to be sufficient for encouraging higher attendance amongst the participants. But, for those who did not attend the week 0 class, the contract appears to be a very useful substitute, boosting attendance from 53% to 71% ( $p=0.014$ ). This finding hinges on a comparison of small samples, and should be interpreted with caution; nevertheless, it appears that the contract had a strong and positive effect on attendance rates for the small sub-group of people who could not attend the introductory session, plugging the commitment gap that might have otherwise ensued.

**Table 34: Commitment contracts substitute for other commitment elements**

Mean attendance rates	Missed week 0 (1)	Attended week 0 (2)
No contract	53% (n=35)	75% (n=65)
Contract	71% (n=31)	72% (n=66)
Crude effect of contract	+ 18% ( $p=0.014$ )	- 3% ( $p=0.361$ )

In contrast, for those who did attend the early motivational session, going on to receive the commitment contract may have reduced attendance slightly (3%). This is signalled by the negative CATEs (panels B and C in Table 33), and the fall in crude attendance rates of 3% shown in Table 34. With the caveat that neither are statistically significant, these figures may be evidence of commitment saturation, as discussed in the Food Monitor trial: after a certain point, additional commitment strategies do not induce greater accountability and may even backfire. Too much reputational commitment, it appears, can be bad for behaviour change; and this echoes findings in related literature on personal rules (Verhoeven et al. 2013).



#### **6.4. Summary: research question 2**

The analysis above confirms the existence of heterogeneous treatment effects of commitment contracts. Robust evidence confirms hypothesis 3 that sophisticated individuals benefit more from the contract in terms of health behaviours. The data is strongly suggestive that myopia interacts with the commitment contract, but in the opposite direction to that expected by hypothesis 5. Further heterogeneity pathways are explored in relation to commitment priming, which raise implications for theory and future research. In line with results for average treatment effects, heterogeneous treatment effects are not found on weight loss outcomes, and this is likely to be a result of the treatment being too mild to exert a discernible influence (as argued above), compounded by smaller sub-group samples used for heterogeneity analysis that make it more difficult to uncover statistically significant treatment effects.

## **7. DISCUSSION: WIDER IMPLICATIONS**

This section briefly considers broader lessons arising from the experiment, beginning with how the empirical results speak to the existing literature on commitment devices, particularly in relation to effect sizes; then discussing ways in which the research design could be improved in future; and finally considering what the results mean for programme interventions and policy makers.

### **7.1. *Effect sizes in the literature***

With such marginal weight loss differences between experimental groups, a much larger sample size would be needed to detect a statistically significant effect. Nevertheless, the finding of no significant effect on weight loss impact is an important contribution to the commitment devices literature, which often cites more dramatic weight loss differentials between comparison and treatment groups. Cases such as Chapman et al (2015), the Camden trial and the Food Monitor trial should help future research designs produce more accurate ex ante sample size calculations for weight loss outcomes. The findings from the Camden trial speak to the criticism raised by Paloyo et al (2015) that existing studies often bundle the commitment device intervention amongst a number of other practices (such as intensive researcher contact) that make it difficult to isolate the effects of the commitment device. The advantage of this trial is that it does isolate the effects of the contract, and suggests that the contract alone is not powerful enough to leverage significant weight loss.

### **7.2. *Improving future research***

The research design has various avenues for improvement. The promise of future heterogeneity analysis – for example to delve more deeply into commitment priming – would call for a larger sample in future research. The combination of qualitative methods is

a strength of the study, and future research designs could incorporate more systematic follow up with treated individuals to better develop a measure of treatment intensity that could be quantitatively modelled; and probability sampling to allow for some extrapolation of findings. Within the constraints of partnership with a programme delivery agency, it would be important to ensure improved coverage of administrative data, and more tracking of attritors. Ideally the latter would take the form of intensive data follow-up with all clients who miss key weight readings; a more realistic alternative would be a second round random sampling strategy to provide unbiased estimates of the weight loss trajectory of participants who drop out.

### **7.3. *How to improve commitment contract design***

In a comprehensive behavioural programme such as Shape Up, the prospect of a mild commitment contract leveraging further progress on weight loss is limited. However, amongst some participants the contract was both popular and effective. Taking account of the qualitative insights implies that commitment contracts can work, but design matters. Three lessons can be learned from participant experiences analysed here. Firstly, future contracts could be administered by the tutor themselves to build on the reputational commitment that may already exist in that relationship, and to augment the importance of the contract (rather than it being seen as just another “piece of paper”).

Contracts could be reaffirmed during the 11-week programme to boost the strength of the treatment, and to test whether the contract is more effective as the goal’s deadline approaches. It could be the case that in the final weeks, as participants become more aware of the gap between their current weight and their 5% weight loss target, the commitment contract is something they reach for or rely on more to improve their focus and meet their goal.

Allowing participants to tailor their contracts somewhat, perhaps by inserting their own goals alongside the 5% weight loss target may also improve its relevance and salience. Even the simple act of choosing the format of the contract may help – many participants referred to the commitment contract in this trial as a “certificate”, and for some this was an attractive feature, but not for others.

## **8. CONCLUSIONS**

The chapter set out to present and discuss the findings of a second field experiment testing causal effects of commitment devices on health behaviours and weight loss. It has made four important contributions to the thesis: firstly, it has provided robust empirical evidence to answer the research questions; secondly, it generates useful lessons to inform the theoretical framework; thirdly it raises new questions for the scholarly debate on commitment devices; and finally it offers new insights for policy makers. The concluding remarks below summarise these contributions, and then highlight the remaining lines of enquiry that frame the next and final chapter on results and analysis.

### **8.1. Contribution to the research questions**

In response to research question 1, the trial demonstrates that commitment contracts can bring about desired improvements in health behaviours, in line with planner-doer theory. Although there was no significant impact on weight loss performance, the contract boosted participation in a public weight loss programme: the probability of completing the Shape Up course increased by 14% ( $p < 0.05$ ), and mean attendance was 6% higher ( $p < 0.10$ ). While these results on health behaviour imply modest treatment effects (Cohen's  $d$  of 0.19 and 0.16 respectively), the improvements are nonetheless valuable for both participants and service providers in behavioural weight management programmes where the syllabus continues to build on self-reflection and information over several weeks. The findings demonstrate that commitment contracts can help bring clients back week after week for regular and systematic self-monitoring and self-reflection, which have been shown to deliver health benefits and support sustained lifestyle changes.

Research question 2 asked whether commitment devices effect behaviour change in diverse ways across different people, and the Camden trial provides considerable evidence to indicate they do. In support of hypothesis 3, sophisticated participants benefitted more from the commitment contracts than others, with higher attendance and completion rates ( $p < 0.05$ ). Hypothesis 5 suggested a commitment device will work less effectively for those with myopic and fatalist attitudes towards their health. The evidence from the Camden trial opposes this view, with myopic participants benefitting more from receiving a contract ( $p < 0.10$ ). As with the average treatment effects on weight loss, no significant results are found from the sub-group analysis.

## **8.2. *Lessons for the theoretical framework***

On balance, the trial has borne out the predictions of the planner-doer model, both in terms of average and heterogeneous treatment effects. Significant effects are found on health behaviour. Testing of two pre-specified sub-group analyses has yielded fresh insights on heterogeneous treatment effects based on sophistication and myopic health attitudes.

The model predicted effects on weight loss as well as behaviour, but this has not been borne out in the weight loss data. However, the results remain consistent with the Analytical Framework, which can explain the weak ATE on weight loss as reflecting the commitment device exerting a very weak tax on the doer's over-consumption, as set out in proposition 4. The observed weakness of  $\theta$  is argued to be a consequence partly of low adherence ( $\lambda$ ), and partly of mild commitment device design ( $d$ ), as set out in Propositions 5 and 6. New questions were raised by the trial about the fit between the empirical data and the theoretical model, particularly on the model's implicit assumption of a linear increase in commitment generating a linear improvement in desired outcomes;

this and other issues for further analysis in the next chapter are discussed further in sections 8.3 and 8.5 below.

### **8.3. *New directions for future research***

Chapter 6 has reported a number of innovations from the Camden trial relative to the existing literature on commitment devices. The analysis has successfully woven in qualitative alongside quantitative evidence to provide a nuanced picture of treatment effects. New data has been generated by successfully operationalizing two theory-driven variables: sophistication and myopia. Investigation into health behaviour has broadened beyond the conventional focus on exercise to consider sustained participation in an 11-week weight loss programme. Exploratory analysis has introduced the concept of commitment priming; and found new evidence of commitment saturation, that corroborates chapter 5's findings on commitment overload. Both sets of results point to threshold and ceilings in how well commitment devices work.

The Camden trial created the opportunity to examine the potential effects of commitment priming through two mechanisms: arriving on the Shape Up programme via a GP referral, and taking part in an early motivational Shape Up class that emphasised the importance of regular attendance all the way through the programme. Both analyses, while exploratory, suggest the idea of threshold effects at work – as hurdles and as ceilings. The GP referral and commitment contract may be too mild on their own, but when combined they are able to exert a discernible effect on behaviour (the positive interaction clears a hurdle). In contrast, the contract or the introductory Shape Up class are better characterised as substitutes for one another. If someone attended the introductory class, the contract offered little more improvement in participation; but if someone missed that class, the contract plugged a commitment gap they would otherwise have faced over the course of the programme.

The analysis suggests new hypotheses around commitment thresholds and commitment saturation for future research.

#### **8.4. *Insights for public health programmes***

This experiment drew fresh conclusions on the effectiveness of a reputational commitment device in a public health setting. It is unlikely to offer significant traction as a stand-alone intervention to promote weight loss. While average treatment effects on weight loss are weak, there is scope for the design of the commitment contract to be improved and tailored to participants in programmes such as Shape Up. For example, it could be reaffirmed regularly with tutors, whom many participants described having a reputational commitment towards, in order to strengthen the intensity of the intervention. Those who arrive through the GP referral route may stand to benefit further from signing a contract, as might those who would be at risk of dropping out because they have not experienced some other commitment elements in the programme (such as the introductory Shape Up class). Easily observable characteristics around myopia and sophistication could be used in health initiatives to leverage the benefits of commitment contracts for greater participation and repeat attendance.

#### **8.5. *An agenda for chapter 7***

The results and analysis from chapters 5 and 6 so far leave three avenues of enquiry remaining. Firstly, qualitative analysis from both Food Monitor and Camden trials highlight a clear difference in approach amongst treated individuals, with some participants actively embracing the contract, while for others no specific actions are taken, and the commitment device may fade from memory. A deeper exploration of qualitative data in the next chapter will aim to make further contributions to the scholarly debate by shedding light on how adherence can be identified, why this diversity arises, and



whether this translates into stronger behaviour change and weight loss outcomes amongst those with high adherence.

Secondly, there is further scope for analysing qualitative data on sophistication, to supplement the proxy variable ‘experience of previous weight loss programmes’ applied thus far. The Camden trial found evidence to support the theorised heterogeneity of treatment effects amongst the sophisticated, and qualitative data promises to either corroborate or challenge this finding.

Thirdly, is there any evidence of the kind of internal planner-doer interactions that the analytical framework has assumed? Could it be the case that the planner-doer lens is not an apt characterisation of all Shape Up clients, and it is this mismatch that explains why the commitment device did not work as expected for weight loss outcomes? Chapter 7 continues to analyse qualitative evidence from both experiments, with the aim of drawing conclusions to these questions.

---

BLANK PAGE

---

## **Chapter 7**

### **RESULTS AND ANALYSIS (3):**

#### **Adherence, Planner-Doer**

#### **Interactions, and Sophistication**

---

## **1. INTRODUCTION**

The Research Design explained how the two field experiments sought to combine qualitative and quantitative methods, with three stated aims. The first was to contextualise and delve into the statistical results, gain a more nuanced understanding of the behaviour change process, and provide a fuller answer to the two research questions. This aim was prioritised in chapters 5 and 6 by weaving in qualitative analysis to discuss the treatment effects uncovered by statistical analysis. The remaining two aims are the focus of this third and final results chapter.

The second aim was to generate new data to further investigate heterogeneity pathways that are notoriously difficult to quantify: adherence to commitment devices and sophistication. A third aim was to search for tangible evidence of internal planner-doer interactions of the nature predicted by Thaler and Shefrin's dual-self model (chapters 2 and 3); which to date has been tested only in relation to its predictions and not its underlying assumptions. The endeavour takes on new importance in the context of the experimental results (chapters 5 and 6) that provided mixed support for the hypotheses derived from the planner-doer theory.

To begin, this chapter examines new evidence for heterogeneity based on how well the individual adheres to the commitment device beyond the moment of taking it up. The Analytical Framework highlighted both the importance of adherence (labelled  $\lambda$ ) for generating a psychological tax large enough to effect behaviour change; and the challenge of gauging  $\lambda$  with a single, quantitative measure. Section 2 introduces novel qualitative data from both field experiments that distinguishes between treated individuals who maintained strong adherence to the commitment contract (high- $\lambda$ ), and those who did not (low- $\lambda$ ), allowing for a comparison of baseline characteristics and health outcomes between

the two sub-groups. The exercise sheds new light on adherence as a heterogeneity pathway.

The chapter then critically reviews whether people behave in the way that the planner-doer model asserts they will, concentrating on two building blocks of the Analytical Framework: internal, strategic, interactions between sub-selves, and sophistication. The first, and arguably most novel, idea distinguishing Thaler and Shefrin's model within the dual-self tradition is that the individual can be understood as a combination of two sub-selves, a far-sighted 'planner' concerned with long run health, and a myopic 'doer' interested only in current gratification. Yet no studies have tried to unpack and test this assumption; indeed these assumptions have been critiqued as purely metaphorical (Frederick et al. 2002). Section 3 makes a contribution to the literature by analysing interview transcripts from the Camden trial for any such evidence as participants grapple with their personal weight loss challenges.

The Analytical Framework is also predicated on the idea that an individual experiences self-control problems leading to time inconsistency – recall Paul's weight management dilemma in chapter 3. If he is aware of this, and recognises the distance between his actual and preferred weight, he is likely to demand a commitment device. The Research Design discussed the difficulties in pinning down this concept of sophistication in quantitative field experiments, and noted that few studies have tried to do so. Section 4 presents a novel coding scheme to the Camden interview data to search for evidence of sophistication, the existence of other personal commitment strategies employed by the participants, and self-reported behaviour change that can be linked to such commitment strategies.

Section 5 concludes by drawing attention to three contributions to the thesis. Firstly, adherence to the commitment

device does appear to be a key factor determining its effectiveness. Qualitative data sheds new light on this heterogeneity pathway, and further explains why commitment devices yielded a large range of outcomes in the Food Monitor and Camden trials, spanning both positive and negative effects. This has practical implications for the design of commitment devices: if they can be easily put aside, they are unlikely to change behaviours.

Secondly, evidence of planner-doer strategic interactions is found in one-third of interviewees from the Camden trial, indicating that the planner and doer framework is more than just a metaphor for the inner jostling between immediate and delayed gratification. This raises questions about how widely commitment devices can be expected to play a role in health behaviour change, if they are only applicable to a rarefied sub-population that face the internal tussle between planner and doer when making health choices.

Thirdly, while the commitment devices offered as treatments in the Food Monitor and Camden trials did not have universal appeal, many Camden interviewees mentioned their own commitment strategies, finding ways to frustrate their future selves in order to achieve their health goals. In other words, there is a sizeable demand for commitment devices in the context of health behaviour change, as predicted by the Analytical Framework. This prediction arises from the planner-doer model's logic of needing to constrain future choices because without this constraint, the individual may stray from their (planner's) preferred actions. To meet this demand for commitment devices many people developed customised strategies to address their unique temptations and challenges, and these personal rules fit neatly in the taxonomy of commitment devices set out in chapter 2.<sup>74</sup>

---

<sup>74</sup> See Table 1, page 40.

## **2. INVESTIGATING ADHERENCE WITH QUALITATIVE DATA**

### **2.1. Adherence as a determinant of commitment device effectiveness**

An important feature of the analytical framework is the parameter  $\theta$ , which was explained in chapter 3 as a ‘psychological tax’ on the doer sub-self that brings about behaviour change. As such,  $\theta$  is a measure of how effective the commitment device is in changing behaviours. chapter 3 argued that  $\theta$  is partly determined by the design of the commitment device ( $d$ ), with financial commitment devices expected to exert a stronger influence than reputational commitment devices; by individual characteristics as reviewed in chapters 5 and 6 ( $\tau$ ); and by the degree to which the individual adheres to the commitment device ( $\lambda$ ) – do they embrace it in their behaviour change regime? Do they maintain its salience beyond the initial point of take-up? It is these questions that this section focuses on.

The experiments examine two different reputational commitment device designs: a reputational commitment in the form of a pledge to a family member or friend (the Food Monitor coach), and a reputational commitment in the form of a written pledge to oneself (the Camden contract).<sup>75</sup> The treatments were offered in a uniform way, however the manner in which they were taken up and applied varied considerably across participants. In the Food Monitor trial, data showed varying intensity of coach involvement for the reputational treatment group, and interviews from the Camden trial exposed different interpretations and visibility of the commitment contract. The analysis now examines these variations in greater

---

<sup>75</sup> The Food Monitor trial also tested a financial commitment in the form of a premium payment, but this is not comparable here as it involved a one-off decision to subscribe to the service. Intensity of the financial commitment device could vary through the size of the payment, as discussed in chapter 5, but that is defined as a design feature ( $d$ ) not adherence.

detail, using qualitative data to identify when the salience of the commitment device was sustained, and categorising treated individuals into low-adherence (low- $\lambda$ ) and high-adherence (high- $\lambda$ ) groups.

## **2.2. *Food Monitor: is degree of coach involvement associated with outcomes?***

### **2.2.1. *Dataset and coding scheme***

The Food Monitor trial offered the reputational commitment device to 118 participants and was accepted by 48 (40% compliance). As set out in the experimental design, some coaches were emailed at the end of the trial as a follow-up to investigate the coach experience and to triangulate the participant's weight loss efforts. Of the 30 coaches contacted, 16 were successfully reached, giving an overall response rate of 33% and a 53% response rate amongst those contacted.

Coaches were asked if they were aware that they had been named as a 'coach', indicating their awareness of the trial more generally. Thirteen were aware they were coaches (81% of those responding).<sup>76</sup> Coaches were also asked to describe their level of involvement on a five-point scale: none, minimal, moderate, active and very active. These responses allow for a proxy measures of the intensity of the reputational treatment as applied by the participant: coaches with active or very active involvement indicate this was a highly salient treatment (high- $\lambda$ ) for those participants; conversely, those coaches who reported zero, minimal or moderate involvement indicate the treatment will have had lower salience amongst those participants. With this proxy measure in place, the coach follow-up

---

<sup>76</sup> Suggests that 19% did not reach out to their coach at all, which may indicate naiveté in taking up this commitment device as discussed in chapter 3.



survey identifies a high- $\lambda$  group of nine participants and a low- $\lambda$  group of seven participants.

The model asserts that the impact on the health behaviour depends on the strength of  $\theta$ , which is dependent on the size of  $\lambda$ .<sup>77</sup> The test of this assumption is the difference in self-monitoring and weight loss performance between the two groups. This small-n dataset cannot be used to undertake regression analysis or hypothesis testing, but may offer some useful directions for future research through simple mean comparisons of baseline characteristics and outcomes.

### ***2.2.2. Adherence and baseline characteristics***

The contrast between low and high adherence groups (Table 35) suggests that those who embrace their reputational commitment device are also somewhat more likely to be older, report slightly less healthy initial behaviours around diet and exercise, and have more short-termist and negative health attitudes. However there is no difference in starting BMI.

### ***2.2.3. Adherence and outcomes***

In terms of outcomes, the high adherence group appear to lose more weight in the short term (2%, compared to the low adherence group's 0.4%) and the medium term (3%, compared to 0.5%). Due to attrition there are three or fewer observations for the low- $\sigma$  group so this data must be read with caution; further the sample is non-randomly selected and is used for descriptive not causal inference, recognising that the data may not be representative of the full sample. However, it does provide tentative evidence that those who embraced the reputational commitment device were more successful in their weight management. The same is not true of self-monitoring

---

<sup>77</sup> See equation 7, chapter 3.

behaviours, which are identical at four weeks. Adherence does not affect the probability of having a missed weight observation at 4 weeks, although at 12 weeks the high-adherence group appear to be slightly more likely to still be using the Food Monitor tool.

**Table 35: Participant characteristics by adherence groups (Food Monitor)**

	Low- $\lambda$ (n=7)	High- $\lambda$ (n=9)
Female	100%	100%
Age	30s	30s and 50s
Initial BMI	29	29
Fruit and vegetable intake	4.3	2.9
Exercise sessions	3	1.2
HFS: doer	71%	78%
Weight loss (%) 4 weeks	0.4%	1.9%
Weight loss (%) 12 weeks	0.5%	3.0%
Self-monitoring over 4 weeks	23	23
Drop out at 4 weeks	57%	56%
Drop out at 12 weeks	71%	67%

*Notes: Modal group shown for age. Weight loss outcomes at 4 weeks n=3 for low adherence and n=4 for high adherence; at 12 weeks n=2 for low adherence and n=3 for high adherence.*

### **2.3. Camden: is adherence associated with outcomes?**

#### **2.3.1. Dataset and coding scheme**

Qualitative information from the Camden interview transcripts is used to identify the different ways in which participants applied their commitment contracts (as discussed briefly in chapter 6, where Table 30 presented selective excerpts arising from the coding exercise detailed here). Among the 24 interviews (12% of full sample), seven participants did not receive the contract and 17 did. A simple coding system was used to classify these 17 treated individuals as either high- $\lambda$  or low- $\lambda$ . The high adherence individuals are those who engaged more closely with the commitment contract, perhaps reporting that they looked at it every week or even every day if it was placed in their kitchen. They may have discussed it with others and

reflected on it even without a visual prompt.<sup>78</sup> Applying the coding criteria set out in Table 36, nine of the treated interviewees were recorded as high- $\lambda$ . In contrast, the other eight treated interviewees reported they had not taken the contract out of the envelope, or that it had been lost among other Shape Up papers and handouts; overall that it had failed to make a strong impression. For these participants, the commitment contract had low salience, and they exhibited low adherence to it.

Limitations to statistical analysis using this dataset are its small size (n=24), likely selection bias and challenges to extrapolating to the wider sample, due to interviewees being selected not at random but based on convenience sampling. For example, it is possible that those who were happier with their experiences on the Shape Up programme and/or their final weight loss outcomes were more willing to share their experiences, and are therefore over-represented in my sample.<sup>79</sup> Conversely, some of those who dropped out of the programme cited dissatisfaction with the course content or tutor (Chapter 6) in brief follow-ups to gather missing outcome data; none of the attritors were involved in the more in-depth interviews. It is plausible, then, that while there appears to be a 50/50 balance between low and high adherence participants, in the sample as a whole the high adherence participants could have been a minority. High-adherence participants, then, could be a smaller sub-group

---

<sup>78</sup> One participant referred to the contract as a “certificate”, and interpreted it as a reward she would give herself only at the end of the Shape Up programme. Although the contract was not on show during the experiment, in the interview she indicated she was mindful of the contract and intended to put it on display that very day, as the interview took place immediately after her final weigh-in and she was pleased with the results: *“I promised myself I would wait, I wouldn’t miss any lectures, which I didn’t do. I’d work hard at it, and when I have actually achieved what the certificate says, then I’m going to stick it on my wall”* – ID 30102, female, age 61.

<sup>79</sup> Participants who responded positively to requests for follow-up interviews were also more likely to be comfortable communicating in English; and more willing and/or able to spare the time.

than implied by the interview dataset, but it is not possible to gauge this conclusively.<sup>80</sup>

<b>Table 36: Coding high- and low-adherence in the Camden trial</b>		
Intensity	Coding criteria	Excerpts from transcripts
High- $\lambda$	<ul style="list-style-type: none"> <li>▪ Looked at contract most weeks</li> </ul>	<i>“Whenever I opened the fridge it’s just right in front of me.” – ID 11411, female, age 60</i>
	<ul style="list-style-type: none"> <li>▪ Placed in a visible spot at home</li> </ul>	<i>“You’ve got something in writing that’s just there ... keeps reminding you if you do sort of slip up.” – ID 30047, female, age 74</i>
	<ul style="list-style-type: none"> <li>▪ Discussed with family or friends</li> </ul>	
	<ul style="list-style-type: none"> <li>▪ Remembered and reflected on it without visual prompt</li> </ul>	<i>“I talked to some of my friends about this, yeah. And family.” – ID 30102, female, age 61</i>
Low- $\lambda$		<i>“I left it with the notes I got from the course.” – ID 30011, female, age 55</i>
	<ul style="list-style-type: none"> <li>▪ Did not remember contract</li> </ul>	
	<ul style="list-style-type: none"> <li>▪ Did not take it out of envelope</li> </ul>	<i>“I kept it in the envelope, and it’s downstairs in my kitchen somewhere.” – ID 30068, female, age 68</i>
	<ul style="list-style-type: none"> <li>▪ Left it with Shape Up papers</li> </ul>	
	<ul style="list-style-type: none"> <li>▪ Saw it once during programme</li> </ul>	<i>“I looked at it halfway through the course, and now it’s kind of buried under a bunch of papers.” – ID 40028, female, age 60</i>

<sup>80</sup> This issue is relevant largely because it tells us something about how likely high adherence is, and so how useful this heterogeneity pathway might be for practical targeting purposes. From a theoretical standpoint, it being a small or larger subgroup is less relevant, while from a methodological perspective it indicates the dataset may be less representative of the wider participant pool, and due caution should then be applied to descriptive inference.

### **2.3.2. Adherence and baseline characteristics**

While the size and selection of the sample precludes sophisticated quantitative analysis, a simple comparison of key baseline characteristics is suggestive of the two groups having some distinct features (Table 37). There are no men in the high adherence group. Participants in the high- $\lambda$  group tend to be older, and are more likely to be self-referred. They report doing less exercise at the baseline and a slightly lower consumption of fruit and vegetables than the low- $\lambda$  group, raising the prospect that those who rely less on the contract have a somewhat healthier set of habits at the baseline. However high- $\lambda$  participants report more control over eating habits.

In other respects, the groups are broadly comparable. Starting body mass index is 31.4 in both cases, at the low end of the obese range; and the profile of BMI categories is broadly similar. Finally, those in the low- $\lambda$  group are more likely to report a long-term and positive outlook on their health, while the high- $\lambda$  group report short-term and negative health attitudes. This last finding is of some surprise, but indicates that a commitment contract may be seen as unnecessary or redundant by those who had a strong inner drive to address their weight concerns; whereas for those who perhaps were more likely to feel negative or short-termist, the commitment contract was perhaps viewed as a useful way to externalise their motivation and stay focused on the goal.

### **2.3.3. Adherence and outcomes**

Table 37 also reports outcomes. The small sample size precludes hypothesis testing. However, the results are suggestive that interviewees who reported stronger engagement with the commitment contract did better in terms of attendance (9% higher) and completion (14% higher). They also reported stronger weight loss (18% higher) using the later outcome measurements (from Shape Up

weeks 9-10) that has two fewer observations from the high adherence group.

**Table 37: Participant characteristics by adherence groups (Camden)**

	Low- $\lambda$ (n=8)	High- $\lambda$ (n=9)
Female	75%	100%
Age	47	53
Initial BMI	31	31
Self-referred	25%	44%
GP-referred	38%	33%
Fruit and vegetable intake	4.3	3.7
Exercise sessions	1.5	1.2
Control over eating habits	2.4	3.1
Attended Introductory session	75%	89%
HFS: doer	38%	67%
Weight loss (%) weeks 7-10	2.9%	3.0%
Weight loss (%) weeks 9-10	3.4%	4.0%
Attendance	79%	86%
Completion	88%	100%

*Notes: Self-reported responses on 'control over eating habits' on a 5-point Likert scale, where 1 = disagree strongly and 5 = agree strongly. HFS doer includes Unconfident Fatalists and Live for Today's who tend to be more short-termist and have more negative health attitudes, and excludes Health Conscious Realists and Balanced Compensators. Diet variable means drawn from 8 available responses from each group; all other variables have 8 responses from low- $\lambda$  and 9 from high- $\lambda$ . Weight loss at weeks 9-10 had 7 responses from the high- $\lambda$  group and 8 from the low- $\lambda$  group.*

Using an alternative weight loss variable that takes outcomes from week 7 (used in Chapter 6) ensures all participants have an end weight reading, and the means are no different across the two adherence groups. This is in line with the discussion in Chapter 6 that those who persevere with the course until the end are more likely to report higher weight loss, and perhaps having the contract and ensuring its salience played a role in this. But because of the nature of the attrition patterns in the final fortnight of the Shape Up programme, potential bias precludes statistical inference. For the purposes of this section, the evidence can be read as suggestive that high adherence encourages higher attendance and completion; and appears to be positively associated with higher weight loss also.

## 2.4. Summary

The evidence is indicative that adherence to a commitment device,  $\lambda$ , is positively associated with health behaviour change and outcomes. In the Food Monitor trial those reporting high adherence also reported higher weight loss. In the Camden trial, those reporting high adherence also reported higher attendance and completion of the weight loss course. Small samples and some missing endline data entail cautious interpretation of the point estimates, but the overall trend appears clear.

A further insight is that traits and adherence interact in ways the Analytical Framework did not fully predict. In both datasets, low and high adherence groups have some distinctive baseline characteristics. These findings offer consonance with chapter 3's modelling of individual traits as potential predictors of adherence (see Figure 8). However, the high adherence interviewees were more likely to report negative and short-termist health attitudes. While this is consistent with statistical analysis of the Camden experiment – which reported that myopic participants benefitted from the commitment contract – the findings run counter to the predictions made in chapter 3.

The qualitative findings from the Camden experiment also offer a contrast with statistical analysis reported in the Food Monitor experiment (see chapter 5 Table 21) that suggested those with myopic health attitudes benefitted more when neither financial nor reputational commitment device was applied (the refund group). The triangulation exercise thus supports the existence of a complex causal nexus between traits ( $\tau$ ), design features ( $d$ ) and adherence ( $\lambda$ ) to determine the effectiveness of the commitment device ( $\theta$ ); unwinding the idea of simple, linear, heterogeneous pathways.

### ***3. THE PLANNER-DOER TUSSLE: STRATEGIC INTERNAL INTERACTIONS***

Under-investment in good health is, according to the planner-doer model, a result of the planner losing the battle to prioritise long-term health over more immediate gains from procrastination and self-indulgence. Critique of dual-self theories suggests this idea of two sub-selves is merely a metaphor for how individuals behave. As argued in chapter 2, the neuroeconomics literature (Mcclure et al. 2004; Camerer et al. 2005) provides grounds for expecting that competing sub-selves are more than metaphor, but no previous work has investigated whether such internal bargaining takes place in the way the planner-doer model characterises. Can qualitative evidence from research participants corroborate this notion of an internal dialogue between far-sighted planner and myopic doer?

#### ***3.1. Developing a coding scheme to uncover planner-doer interactions***

This section analyses Camden interview transcripts for any evidence of such planner-doer interactions. Participants were not asked directly if they viewed their personal weight loss challenges as an internal tussle between their doer and planner sub-selves, so qualitative content analysis was applied to interpret the transcripts and investigate whether evidence of such thought processes exists. A preliminary coding scheme (chapter 4) was distilled through a process of iteration, going backwards and forwards between the original planner-doer model (Thaler & Shefrin 1981) and a small selection of interview transcripts. The resulting two coding criteria (see Table 38) were then applied to all interview records.

The first criterion examined whether the individual reported competing objectives around their health and lifestyle. For example, cravings for certain foods which they found tasty and satisfying, but



which they knew were unhealthy and likely to impede their weight loss progress. This disjoint between short run gains (gastronomic satisfaction) and long run gains (healthier body) is precisely the tussle put forward in the planner-doer model. The second coding criteria searched for references to some internal rewards or penalties, some reminder to oneself of the costs or benefits of a certain course of action, designed to influence the choices made. The underlying logic here is that the planner has to resort to some form of strategic bargaining in order to curb the doer's natural tendencies; and conversely the doer may need to justify flouting the long-term plan.

**Table 38: Evidence of planner-doer sub-selves**

Theme	Coding criteria	Example
Internal tussle between planner and doer sub-selves within the individual	<ul style="list-style-type: none"> <li>• Reference to competing objectives or desires on health and lifestyle</li> </ul>	<p><i>I used to buy things like sugar ... when I knew I shouldn't take sugar, because sugar is not helpful. So I promised myself 'well that's just for visitors'. Or biscuits: 'that's just for visitors'. After a while I saw that I didn't have no visitors but they're gone (laughs)!" – ID 30002, female, age 61)</i></p>
Internal strategic interaction with oneself	<ul style="list-style-type: none"> <li>• Anticipated reward or penalty from pursuing a certain behaviour</li> </ul>	<p><i>"I do still have times when [I would] really like a bar of chocolate or a bag of crisps (laughs). But I'm more conscious now, my mindset is 'no, don't do it [interviewee name], because you know you'd just end up putting on more weight' " – ID 30035, female, age 45</i></p> <p><i>"I would have been totally satisfied to have reached 9 stone on my own, but now I'm under, I can see that I can have a bit of Christmas food, and 9 stone will still be there." – ID 30027, female, age</i></p>

### **3.2. *Discovery of planner-doer individuals***

This exercise discovered some evidence of participants struggling to stay on course with the weight loss plan due to salient temptations in everyday life. In total, 12 references were coded as showing evidence of planner-doer sub-selves, from eight interviewees (33% of all interviewees). Evidence of planner-doer interaction is observed, then, in a minority of participants.

The planner-doer individuals span the full range of BMI categories from normal (24.5) to severely obese (40), with a majority of individuals having received the commitment contract (seven treated and one untreated). The treated individuals were more likely to have embraced the commitment contract (five were coded high- $\lambda$ ). They are all female, likely to have myopic health attitudes (6 of the 8), and are slightly older than the average Camden participant at 53 years.

They all succeeded in losing some weight, and with a mean weight loss of 4.0 kg they are considerably more successful than both the sample as a whole (mean weight loss 2.3 kg) and the others in the interview pool (mean weight loss 2.1 kg). While caution is required with statistical testing in such a small sample, it appears that those coded positively for planner-doer themes performed better than the other interviewees. They were all completers, but this is not surprising as only one of the 24 interviewees did not attend enough sessions to be called a completer. More tellingly, those who were coded positively for planner-doer themes had 90% attendance, compared to 80% for the others ( $p=0.072$ ); and 5.4% weight loss compared to 2.4% amongst the rest of the interview pool ( $p=0.059$ ).<sup>81</sup>

---

<sup>81</sup> Hypothesis tests were not planned for this qualitative content analysis, which aims to uncover new descriptive insights. However, where hypothesis tests are done, the associations are very close to the 5% statistically significant level.

### **3.3. *Planner-doer tussles laid bare***

Despite this apparent weight loss success, these interviewees reported a number of self-control challenges that fit with the planner-doer predictions. There are various examples of participants 'giving in' to some impulse, which implicitly points to the idea that part of them does not want to do so (planner), but is overpowered by other internal desires (doer). For example, one participant made reference to a failed attempt at maintaining a food diary while on holiday. She explained that while she made the effort of taking it with her, and even starting to complete it in the morning, she gave up for the rest of the day as a sort of gift to herself. This was a way of giving herself a break from self-monitoring because she was on holiday, almost as if the planner was (perhaps, begrudgingly) releasing the doer from the usual obligations; allowing short run pleasure to override health concerns (the participant was severely obese with an initial BMI of 40).

Some participants had developed certain strategies to manage their competing desires. One participant pointed out the importance of deciding her meal order before arriving at the restaurant, hungry and distracted, so as to ensure she stayed on track with the diet she was following. Such examples were sometimes described in a form that makes the internal dialogue explicit. For example, one participant knew that snacks shared at the workplace could be hard to turn down. She was already trying to use the stairs at work once a day, and wrapped this strategy in to a negotiation with herself:

*“[At work] I am making myself walk up the stairs. And especially if somebody’s brought in treats (laughs), you know you do at work. You think ‘right, if you have the cake, you have to walk up the stairs at least twice more’.” – ID 30064, female, age 51*

In two instances participants suggested a form of self-manipulation to ensure they continued with a plan of action they would otherwise shy away from. The same participant quoted above recalled there were 2 days between her seeing the Shape Up programme advertised and joining in the first session:

*“I think that suited me better, whereas if I had to dither about and it wasn’t starting for a couple of weeks, I would have just talked myself out of it, saying ‘oh yeah, you can do it on your own, you’d be fine’, you know.” – ID 30064, female, age 51*

This excerpt suggests that her natural tendency would be to procrastinate and avoid signing up to a structured health programme, but acting quickly before her doer sub-self could take control meant that she did join Shape Up, and indeed completed the programme. Another participant explained that she actively sought a form of mild self-deception in order to undertake more exercise:

*“For me, to run around a room is really boring. But if I can take an hour-long class of, I don’t know, badminton or belly dancing or something like that, it feels like fun and not exercise. But I think the result is equally good. So that’s kind of, in my mind, how I plan to trick myself into more physical activity, by making it enjoyable.” – ID 40028, female, age 60*

The exercise uncovers plausible evidence of strategic internal reasoning, and supports the notion of a planner and doer sub-self competing for control of an individual’s health behaviour. This evidence is found in one-third of cases, eight of the 24 interviewees, and demonstrate the diverse ways in which participants employed threats (“*you have to walk up the stairs at least twice more*”); bargains (“*when I have actually achieved what the certificate says about it, then I’m going to stick it on my wall*”); chicanery (“*I plan to trick myself into more physical activity*”); and “strategic self-frustration” (Schelling 1984, p.4) (“*when you go in the shops, don’t buy the things that you know are no good for you*”).

### **3.4. *Why are planner-doer individuals a minority?***

The fact that the planner-doer coded references do not emerge more frequently across the interview pool could be explained in two ways. It may be the case that the coding scheme is sensitive to an individual's natural way of speaking and reflecting on their weight loss journey, and although many others may experience the planner-doer tussle, they do not vocalise it in a way that would be captured by the coding scheme. Men, for example, may have a different experience of addressing impulses around food and exercise, and therefore describe it in a way that was not picked up by the two-pronged coding scheme applied here.

Alternatively, it may be that the planner-doer model applies only to a minority of individuals, and perhaps it is these individuals who are more familiar with reflecting on their health choices that tend to use a dual-self type of narrative to describe their experiences. This would imply that the planner-doer model could be applied only selectively across people to describe a health behaviour problem; which, in turn, indicates that the prescribed solution of pre-commitment may only apply to a sub-population who are indeed facing a planner-doer tussle that creates the self-control problem.

This explanation is backed up by a small number of cases where participants simply did not see the problem as one of (the doer's) inaction on 'good' health behaviours, but of (the planner's) information on what the good behaviours were. One participant joined the Shape Up programme not to lose weight but to learn new exercises that she could do at home. For her it was not a question of willpower but simply learning something new. Another participant (ID 20063) readily admitted: "I didn't really have a proper [weight loss] target. It was more to go and learn how things work", referring to nutrition and healthy lifestyle tips in the Shape Up programme. In these cases, clearly the planner-doer framework would not be the

appropriate lens to view behaviour change challenges, and no demand for commitment devices would be expected.

### **3.5. Summary**

The evidence is the first of its kind to make explicit the internal planner-doer tussles put forward by Thaler and Shefrin (1981) and used in the Analytical Framework for this thesis. The findings argue against the critique that the dual-self model is mere metaphor. According to the lived experiences of the Camden trial participants, the concept of competing sub-selves jostling to have the final say on health and lifestyle choices is a meaningful phenomenon, and can have tangible effects on behaviour and health outcomes. Two credible reasons why planner-doer interactions were not noted across a larger proportion of respondents are if time-inconsistency was not a factor for their weight management, and if participants were not sophisticated about their time-inconsistency problem. This question of demand of external commitment aids, and the implied sophistication underpinning it, is the subject of the following section.

#### **4. SOPHISTICATION AND DEMAND FOR COMMITMENT DEVICES**

In their original paper, Thaler and Shefrin highlight that “pre-commitments will occur primarily for those goods whose benefits and costs occur at different dates” (1981, p.398), much like health investment. This implicitly requires that individuals are aware of their tendency to under-invest in their health, and so take up a commitment device in anticipation of their self-control problem. Indeed, the definition of commitment devices as a voluntary and strategic tool to bring about behaviour change is predicated on self-awareness of self-control problems (Bryan et al. 2010). This trait is termed sophistication (O’ Donoghue & Rabin 1999, p.104), but the literature does not offer clear criteria for identifying degrees of sophistication as opposed to naiveté (see chapter 2). Sophistication is often assumed rather than investigated in the literature on commitment devices, because of this methodological challenge of pinning down the concept in quantitative terms. Statistical analysis in chapter 6 tested a new proxy variable for sophistication, and this chapter makes a further contribution to the literature by testing for qualitative evidence of sophistication.

##### **4.1. *Developing a coding scheme to uncover overt and implicit sophistication***

Two content analysis exercises were undertaken: searching for overt, and for implicit, evidence of sophistication. The coding scheme is based on two assumptions. Firstly, a sophisticated person will demonstrate some awareness that their future preferences will not be the same as their present preferences. Secondly, a sophisticated person will identify the need for some commitment strategy to address the chance they will go off track with their health plan in the

future.<sup>82</sup> The coding scheme applied to Camden interview transcripts (see Table 39) allows for exploration of whether and how participants apply their own commitment devices in their everyday lives, to uncover new qualitative evidence of sophistication.

#### **4.2. *Discovery of overtly sophisticated individuals***

Considering overt evidence of sophistication, 16 references were uncovered in the Camden interview data, drawn from 10 different interviewees. These participants knew that, despite their best intentions, they were likely to need some additional aid to stay on track. One participant had reflected extensively on how her workplace demands made regular eating difficult, and so used a food diary to plan her meals and snacks, and prompt herself to take fruit and healthy snacks to her office. This was a core principle of the Shape Up programme, and during those weeks she described taking in a bag of fruit to ensure she was eating throughout the day. After the programme, this habit lapsed, and she found that she would then be so caught up in her work that she would not eat until late afternoon, which raised the risk of blood sugar fluctuations and overeating to compensate. To put herself back on track with regular eating, she decided:

*“I’ve started keeping my diary again, which I’d stopped. I thought I was keeping it, then I thought ‘I don’t need to keep it because I know what I’m doing’ and as soon as I stopped keeping it I realised I didn’t know what I was doing... It’s not really enough to rely on yourself, you kind of have to have a plan I think. Because your mind is too unreliable, or at least my mind is too unreliable.” – ID 11407, female, age 46*

---

<sup>82</sup> As discussed in chapter 2’s taxonomy of commitment devices, where personal rules can include commitments and resolutions to oneself, driven by the desire to maintain “self-reputation” (Benabou & Tirole 2004, p.849).



**Table 39: Coding sophistication in the Camden trial**

Theme	Coding criteria	Excerpts from transcripts
Overt evidence	<ul style="list-style-type: none"><li>Awareness that their future preferences will not be the same as their present preferences, and that some impulse will need to be constrained in order to stay on track with their weight loss goal</li></ul>	<p>“ Well, I knew, that when I went to functions where there were refreshments available, I know that’s my weakness because it’s free and it’s nice.” – <i>ID 30027, female, age</i></p> <p>“I really feel like... moving more, getting out and doing more physical activity, is key to my weight loss success. And yet I needed some kind of structure so I would start doing that. And having a place to go and a certain time every week to go is what helped me get off the couch and go do it.” – <i>ID 40028, female, age 60</i></p>
	<ul style="list-style-type: none"><li>No particular interest or demand in locking oneself into a course of action</li></ul>	<p>“It’s not really so much a question of willpower. Sticking with the exercise is never a problem, and I eat that way.” – <i>ID 30043, female, age 57</i></p>
Implicit evidence: Demand for commitment strategies	<ul style="list-style-type: none"><li>Mention of ‘personal rules’ as commitment strategies</li><li>References to Shape Up group and/or tutor as the source of commitment</li></ul>	<p>“I sit down once a week on the weekends and [think] ‘ok, we’re going to have this, this, this for dinner; I’m gonna do this for breakfast; I’m gonna do this for lunch’. And then taking that and making a grocery list from that, and you know when I go to the grocery store, if I can go without my husband (laughs), I only buy what’s on the list.” – <i>ID 11550, female, age 35</i></p> <p>“I think just having the regular meet-ups and weigh-ins was really helpful. Because that kind of kept you on track.” - <i>ID 30008, female</i></p>

Regular exercise was raised as another challenge, as was snacking. One interviewee showed a considerable degree of self-reflection on the environmental and emotional prompts that often led to unhealthy behaviour:

*“I think my biggest problem and still it’s something that I’m still working on is the, you know, the internal triggers... if I’m upset, it’s like the first thing that pops into my head, is, you know, to get something to eat. It’s the comfort thing for me. And if I’m bored, you know if I’m home watching TV or something, suddenly I get, you know, munchie cravings and stuff like that. And... (long pause) that’s been my hardest part to work towards.” – ID 11550, female, age 35*

Taking a closer look at the ten participants coded as sophisticated reveals overlap between those tagged as sophisticates and those who reported planner-doer interactions (five of the ten planner-doer individuals are sophisticated), which is intuitive. The sophisticates are entirely female, 90% of them received a contract, and mostly they ensured the contract remained salient throughout the trial.

Turning to a comparison of outcomes, average weight loss amongst the sub-group of (qualitative) sophisticates is 3.7%, which is higher than the rest of the interview pool (3.1%), and the full sample (2.7%). However, weight loss across the sophisticates varied widely: two individuals gained a minor amount of weight, while four individuals each lost more than four kilograms. Self-awareness, then, does not guarantee weight loss success, although those who were more self-aware were more likely to register higher weight loss outcomes.

### ***4.3. Comparing qualitative and quantitative measures of sophistication***

The analysis also offers some comparison with the quantitative measure of sophistication identified in Chapter 4: whether the individual had previous experience of a structured weight loss programme. The argument made there was that someone who had such experience could be expected to understand themselves and their approach to health behaviours reasonably well, including their weaknesses, and so could be called a sophisticated person. The alternative was to be a naïve person, unable to predict they were at risk of failure because of the doer sub-self dominating their health choices. The baseline variable ‘previous programme’ was used to capture the individual’s status as sophisticated or not, and used in chapters 5 and 6 to explore whether commitment devices had a differential impact for those who were sophisticated. Chapter 4 hypothesised that a sophisticated person was more likely to benefit from a commitment contract, because they were more likely to anticipate the value of curbing future impulses in order to stay on track with the weight loss target, and these results were borne out by the sub-group analysis with the Camden dataset (chapter 6, table 33 pg 267).

However, the qualitative measure cannot be directly compared to the quantitative because it is being elicited after the trial. The Shape Up programme specifically tries to encourage greater self-knowledge, with the hope that if difficult situations could be anticipated, then they could be avoided or mitigated. The interviews capture how sophisticated participants after the programme, and it is plausible that this would be different to how they were before the programme, when they may have been offered the contract. Of the ten sophisticates identified in the interviewees, four had not taken part in a previous weight loss programme, and were identified as non-sophisticated in the statistical analysis. It is plausible that these

participants grew more self-aware because of the Shape Up course, as they took on board new information about internal and external triggers, and were encouraged to reflect on the good and bad weeks they recorded.

#### **4.4. *A shallow evidence base for overt sophistication***

Given the behavioural components of the Shape Up programme, it is a puzzle that more participants were not recorded as sophisticated at the end of the trial – 14 of 24 interviewees showed no evidence of sophistication (58%), and yet they too would have been counselled to improve their self-knowledge. Of the 10 participants tagged as sophisticates (42%), for seven there was only a single reference that fit within the coding scheme; for the remaining three participants, there were two or more references coded. All in all, this suggests a rather weak evidence base for sophistication amongst the sample of interviewees.

What might explain this? As suggested earlier, the coding scheme relies on qualitative interpretation of the participants' reflections, and the interview schedule did not directly ask about their self-awareness of health behaviours; indeed, could not have done without risk of leading questions generating biased answers (particularly as it may be seen as a virtue to have a good understanding of oneself). It is plausible that many others are more self-aware than the transcripts provide evidence of, but chose to answer questions more directly and with fewer examples that shed light on their sophistication. Nonetheless, the exercise was useful in highlighting that participants sometimes cite awareness of their self-control problems, providing overt evidence of sophistication. The discussion now examines whether people imply evidence of sophistication, by setting up their own commitment strategies.

#### **4.5. *Implicit evidence of sophistication: demand for commitment strategies***

The Literature Review presented a typology of commitment devices, ranging from the informal to the formal. At the informal end were personal rules to follow a certain routine or action in order to meet the health goal. These are not as formalised as the commitment treatments provided in the Food Monitor and Camden trials. During the course of the interviews, participants were invited to share examples of any personal strategies they used to help stay on track with good behaviours they had identified they wanted to switch to, and the transcripts were coded using the three criteria set out in Table 39. This exercise found 47 references to personal strategies from 22 of the 24 interviewees, and the remainder of this section discusses these personal strategies organised in four categories: rules relating to grocery shopping, time management, digital tools, and the Shape Up programme itself.

##### **4.5.1. *Personal rules for buying groceries***

The first cluster of these rules relate to the way grocery shopping is done. Table 39 provides a quotation from one participant who used a pre-written grocery list to avoid buying snacks and unhealthy food in the supermarket, unknowingly in line with research that advocates shopping lists as commitment devices (Au et al. 2013). Another participant highlighted that she would avoid the aisle with certain foods (“Kettle crisps”) and two-for-one offers to pre-empt the urge to buy unhealthy food. In a similar vein, another participant talked about avoiding a certain fast food chain she was fond of:

*“One of the changes I’ve learned is ‘avoid processed foods’. And I used to like KFC a lot (laughs). I used to like KFC a lot. But what I’ve learned here is... it’s actually common sense. Fill your tummy at home, so when you go past the KFC, that*

*temptation isn't too strong (laughs)! So I learned that here and I'm practising it." – ID 30002, female, age 61*

#### **4.5.2. Personal rules for time management**

A second cluster of responses highlighted that a busy life prevented them from making time for healthy meals and exercise. To address this, participants mentioned various simple rules, such as getting off the bus one stop early to walk more; or booking classes or personal trainer sessions at the leisure centre to create structure and pre-commitment. One interviewee described her strategy to make sure she was eating well despite a busy schedule:

*"The main thing for me is to plan what I'm going to eat next day. Because what was happening was I was dashing around, very busy person, and coming in and thinking 'right, I want to eat something, what's the quickest thing I can eat?'. But now I plan something so it's defrosted, it's got to be eaten today, and I will make a proper meal of it. So my planning has improved." – ID 30027, female, age 67*

#### **4.5.3. Using digital tools to support commitment**

A third set of responses highlighted the value of pedometers and smartphone apps to track diet and exercise and reinforce the personal strategies they put in place. One participant explained:

*"I bought new electronic scales to be more precise, and I'm using apps on my phone to measure my walking, you know to get to the 10,000 steps a day, and to measure what I eat, to keep an eye on my nutritional intake. So yeah, the apps have been fundamental. Beyond the behavioural changes, the nutritional changes, the apps have been very useful in getting to control myself a lot." – ID 30034, male, age 29*

An interesting feature of this interview was that the participant found this, and other personal strategies, very useful as a means of self-discipline and bolstering habits-in-the-making; and

was intent on keeping up these rules well after the Shape Up classes had ended. Yet he was not at all keen on the commitment contract and gave it no thought after taking it home. The stark contrast in appreciation for some commitment strategies and not others is discussed further below.

#### ***4.5.4. The Shape Up programme as a commitment strategy***

The research design for the Camden trial (Chapter 4) highlighted that the Shape Up programme itself could be perceived as a form of commitment device, with reputational commitment formed to the group and tutor early on. This idea is borne out in the qualitative data, with 12 participants citing the Shape Up group (either tutor or peers) as having a positive effect on their coming back week after week, and trying harder to achieve their weight loss goals:

*“I think the commitment, to me, was getting there. I was probably more committed to Mike actually, the group leader ... I think he was very positive as a role model... I’ve booked myself on the [follow up] now, so there’s another weigh-in point. And I think that I was a bit embarrassed thinking ‘oh if I get to the weigh-in and I’ll be the same weight as I was when I left 6 weeks ago. That’s not very good is it?’... So I’d better make some more effort to get to the next point without turning up saying ‘oh it’s the same’, or even, perhaps more awfully, ‘I weigh even more than I did before I started!’ (laughs).” – ID 11407, female, age 45*

#### **4.6. Commitment saturation**

Where such strong reputational commitment had already been forged, transcripts provide evidence of commitment saturation: if the group offered a keen sense of commitment and external accountability, the contract sometimes had no further value to participants. The participant quoted directly above only vaguely

recalled the commitment contract when prompted. One lady who stated the contract did not influence her at all (“I didn’t give it a thought to tell you the truth”) later mentioned her feelings of commitment to the class and the tutor. Similarly, a male participant who forgot about the contract entirely referred to the class as a useful “external check” on his efforts:

*“[It was] not just information. It was also useful to have this kind of regular weigh-in, discussion about your weight, you know the effect of having to report to someone else as well as to yourself.” – ID 30034, male, age 29*

Taken together, this suggests that for some participants, the commitment device needs to be externalised, and the contract did so only partially. While it was a visual reminder, a kind of persistent speech bubble made physical, it was no substitute for having another person checking on your weight loss efforts. This corroborates the Analytical Framework’s prediction that design matters (*d*), that the stronger designs (with public commitment rather than private) may have more leverage to influence behaviour change; and underscores the importance of ensuring that the intensity of the psychological tax ( $\theta$ ) is appropriately pitched.

#### **4.7. Where commitment is not sought**

This narrative of the Shape Up classes as commitment strategies themselves bolsters the idea that participants are in demand of additional, external support for the planner sub-self’s health goals. But two respondents did not indicate such demand (8% of interviewees), and it is perhaps not a coincidence that neither were driven by their own motivation to lose weight. One lady appeared to be simply following doctor’s orders to attend. The other was upfront that weight loss was not her concern, rather she was looking for information about exercises she could do at home. Unlike the



discussion above, she was clear that it was not self-discipline she was seeking but a very specific knowledge gap:

*“Just that I wanted to learn some more exercises that I could do by myself every day without going to [an exercise] class... It’s not really so much a question of willpower. Sticking with the exercise is never a problem, and I eat that way... I have strategies that I use and I’ve used always because that’s how I grew up.” – female, ID 30043, age 57*

These counter examples further contextualise the low average treatment effects on weight loss reported in chapter 6, and supports the Analytical Framework’s expectation that commitment devices appeal to a selective sub-population (proposition 3). In situations where health behaviour change is not an explicit goal, where information rather than time inconsistency is the primary barrier to change, or where a person is sufficiently naïve to have not identified a need for a commitment device, then a commitment contract would not be a useful or necessary aid.

#### **4.8. Summary**

Qualitative data gathered alongside the trial provide novel support for the concept of sophistication and demand for commitment devices, which are so often assumed in the literature but rarely uncovered using the lived experiences of trial participants. This is a contribution to the literature, which has so far struggled to pin down the role of sophistication (Royer et al, 2015) in commitment devices. While there are fewer examples of overt sophistication, there is robust evidence that people are implicitly aware of their self-control problems and want commitment devices to keep themselves on track, which they often develop in the form of personal rules for diet and exercise that fit around their unique home and work lives.

Further, the data triangulates with the concept of commitment saturation, by showing that where a strong reputational commitment has already arisen within the Shape Up course, a further self-reputational commitment from the contract was not always needed or wanted. This entrenches the idea of diverse individual preferences, and the need for precise tailoring of commitment contract designs to fit these preferences well.

## **5. CONCLUSIONS**

In a culmination of results and evidence from two field experiments, this chapter set out to provide deeper descriptive and analytical richness to the research questions: providing new data for triangulation against the theoretical framework and contextualising the statistical findings. The analysis has rested on small datasets ( $n < 30$ ) drawn from non-probability samples, with likely self-selection from respondents, and results have therefore been interpreted with due care. Future research would benefit from larger samples, further application of the proxy variables that have been tested here, and replication of the coding schemes to test their validity in other health behaviour and commitment device settings. Recognising these limitations, it remains the case that this chapter provides credible and innovative qualitative contributions for the research questions, and more widely for planner-doer theory.

### **5.1. Adherence**

The Analytical Framework argued that a theory of commitment devices must take into account the individual's adherence to the strategy as a key determinant of how intensely the commitment strategy is experienced, in order to understand and predict how effective it will be. This chapter demonstrated that it is possible to develop proxy measures of adherence,  $\lambda$ , with qualitative data. Drawn from both field experiments, this data is the first of its kind in the commitment device literature, and the first such combination of qualitative analysis within a quantitative field experiment. The analysis makes a contribution to the scholarly debate both in terms of methods and findings.

The data shows clear variation in adherence to the treatments. Taking up a commitment device makes for a highly personalised experience, with a series of voluntary decisions over time to maintain

its salience (or not), to publicise it to others (or not), and to be guided by it (or not). In answer to research question 2, adherence is a promising heterogeneity pathway that can explain why commitment devices influence some people more positively than others. Food Monitor clients who maintain active rapport with their coaches reported losing more weight. Camden participants who embraced the contract also tended to lose more weight by the end of the Shape Up programme. High adherence individuals are somewhat distinct at the baseline, and this opens new avenues for research: if it is possible to identify in advance the characteristics that predict adherence, commitment devices can be targeted at those who are more likely to embrace them, and more likely to benefit from them.

## **5.2. *Planner-doer tussles and sophistication***

Sections 3 and 4 presented novel qualitative evidence on the underlying assumptions of the theoretical framework, triangulating positively with neuroeconomics evidence in support of dual-self modelling. An original coding scheme was used to identify planner-doer internal interactions from interviews. This, too, is the first data of its kind in the published literature, and offers new support for the characterisation of inter-temporal health choices as a contest between the short-sighted doer sub-self and the far-sighted planner. Evidence of this dichotomy driving health behaviour and decisions is found among one-third of interviewees in the Camden trial. Those who referred to this kind of mental landscape were all women, who tended to have maintained the salience of the contract if offered one, and reported higher weight loss outcomes.

Notably, the other two-third of interviewees gave no indication that planner-doer tussles were at the heart of their health decision-making. In a few of these cases, the motivation to join the weight loss programme was a lack of information rather than a lack of self-control to make the behaviour changes. The planner-doer framework

is not applicable for modelling behaviour change amongst these individuals, and the commitment device could not be expected to effect change – this helps explain the low effect size reported in Chapter 6. The methodology used to elicit information about planner-doer issues could be used in public health programmes to identify whether the constraint is self-control (as opposed to lack of information); and on this basis target patients and clients for whom commitment devices offer appropriate help and promises greatest impact.

Another central concept to the planner-doer framework is that of sophistication, and this too has long eluded attempts to be measured and investigated as a source of heterogeneous effects. Qualitative analysis of the interviews allowed for evidence of sophistication to be probed more delicately, and uncovers several examples of participants being highly self-aware of their self-control problems and why weight loss is so challenging for them. It also found evidence of strong demand for personalised commitment strategies to help make healthier choices, often tailored to the micro-circumstances of the individual's home and work lives. This draws to a close the results and analyses for the dissertation. Chapter 8 offers Conclusions on the research questions and thesis as a whole.

---

BLANK PAGE

---

# **Chapter 8**

## **CONCLUSIONS**

---

## **1. WHAT DID THE DISSERTATION SET OUT TO ACHIEVE, WHY, AND HOW?**

### **1.1. Two research questions and their motivation**

Chapter 1 framed two research questions: can commitment devices change health behaviours and promote weight loss, and do commitment devices work differently across different people? These questions were motivated by three stylised facts: firstly, investments in health can easily fall into the behavioural trap of time inconsistency; secondly, the proportion of overweight and obese people in England has remained stubbornly high at two-thirds of the population; and thirdly, there is growing demand for behavioural nudges to aid weight management but a relatively small evidence base on how best to apply them.

Preventative health measures are essentially a series of intertemporal choices between future and present utility: effort is required now to reap long run benefits and avoid ill health, poor wellbeing, and in the most extreme cases, reduced life expectancy. In the context of weight loss, the preventative measures might involve dietary restraint or increased physical activity in order to manage weight over the long run and avoid future health problems such as cancer, diabetes and heart disease.

Theory has long recognized time inconsistency as a puzzling phenomenon (Strotz 1955). The dominant explanation for why people do not follow through on their decisions is the idea that time preference is overly skewed towards the present time at the expense of the future (O' Donoghue & Rabin 1999). As Dupas notes, there are many reasons why people lack sufficient health investment, including financial constraints and poor access to information; but there remains a scenario where "even though people would like to... adopt healthy behaviors in the long run, they might not be willing to



sacrifice consumption or pleasure today” (2011, p.22). Present bias, a root cause of time inconsistency, affects health behaviour (Fan & Jin 2013).

Dual-self theories, which characterize the individual as being made up of two sub-selves, highlight that making the ‘right’ inter-temporal choice can be extremely challenging due to “conflicting internal preferences” (Benabou & Tirole 2004, p.894). For inter-temporal health choices, in the context of such self-control problems, it is all too easy to prioritise the immediate rewards over the distant: the result is an under-investment in health and poor health outcomes.

Alongside the scholarly backdrop described, the policy context for this thesis is dominated by the fact that conventional policies to address overweight and obesity (for example through education and information) are not having a sufficient impact on behaviour, and the political appetite for intervention through taxes and regulations appears to have reached its peak. Policy makers, service providers and individuals are increasingly turning to the power of small, everyday changes to deliver large improvements in health: behavioural public policy to complement the conventional health education and regulation policy toolkit (Loewenstein et al. 2012).

Researchers have examined the impact of a multitude of nudges designed to counter an obesogenic environment, such as reducing plate size, curtailing the visibility of snacks, and labeling food with calorie information (Wansink 2013; Wansink et al. 2016; Liu et al. 2014). Amongst this array of behavioural policies, commitment devices have also received some attention, and are increasingly employed in the weight management sector in the form of public pledges, deposit contracts, and personal rules and plans (Volpp et al. 2008; Nyer & Dellande 2010; Relton et al. 2011; Prestwich et al. 2012).

## ***1.2. Gaps identified in the literature***

This broad summary of the theoretical and policy context explains how the research questions arose, and in particular how the focus came to rest on health behaviours for weight loss. In reviewing the literature on commitment devices for health behaviours, Chapter 2 highlighted a number of under-researched issues on which this dissertation aimed to generate new evidence. Much of the scholarly debate focused on a type of financial commitment device known as a deposit contract, and while there was recognition of reputational commitment devices whose cost was primarily psychological (Bryan et al. 2010), there was relatively little empirical evidence on personal rules within this category of commitment devices. The focus of the dissertation therefore was placed on these under-attended interventions: a financial commitment device that involves a premium payment to lock in a service or arrangement, and two reputational commitment devices that rely on the idea that a mild public pledge or a promise to oneself generates the psychological tax required to effect behaviour change.

Although dual-self theories have been extended and developed since Thaler and Shefrin's planner-doer model was first introduced, there remained some fundamental gaps. Theory and empirical literature clearly points to the likelihood of heterogeneous effects, for example take-up of commitment devices in studies was often low indicating selective appeal to the target population (Giné et al. 2010); and results within treatment groups showed highly diverse effects of the commitment device (John et al. 2011). Yet the literature has offered few answers on what these heterogeneity pathways are or how they could be used to improve targeting of public programmes. Sub-group analysis is too often confined to demographic characteristics rather than investigating the behavioural underpinnings for why people respond differently to commitment devices. Indeed, no studies of commitment devices have explicitly

applied the planner-doer model to add the much-needed “psychological texture” to confidently understand the observed behaviour change. The thesis set out to address all of these gaps, while recognizing that other questions of interest – around the welfare impacts of commitment devices, for example – were beyond the bounds of feasibility.

### **1.3. *A new Analytical Framework and innovative Research Design***

A first step in answering the research questions was to establish an Analytical Framework (chapter 3) built on Thaler and Shefrin’s planner-doer theory. The dual-self characterisation of a myopic doer and far-sighted planner were applied specifically to the health behaviours related to weight loss outcomes. Planner-doer theory has not previously been set out to this level of clarity in a health behavior context. A particular contribution to the literature was to unpack and formalise heterogeneity of commitment device effects arising from three factors: different designs of commitment devices, individual traits, and individual adherence to the commitment device, as discussed in further detail below. Chapter 3 set out six propositions implied by the model, with six associated hypotheses (see Table 40 below), to frame the research design.

Randomised controlled trials had been used to test the effect of commitment devices on health behaviour change, as discussed in Chapter 2, and their advantages in uncovering unbiased estimates were well known. While observational studies on commitment devices have added new insights – for example on malaria prevention and weight loss (Tarozzi et al. 2009; Relton et al. 2011) – they were unable to definitively answer the question of whether the commitment device caused the improved health behaviour and outcomes. The desire to uncover causal effects of commitment

devices in this thesis motivated the research design choice of field experiments.

The innovation, relative to the literature on commitment devices, was to place the randomised controlled trial at the heart of a mixed methods approach. The aim was to complement the advantages of robust causal inference with a superior understanding of contextual factors and the lived experience of using a commitment device to change behaviours. Chapter 4 articulated this research design for two field experiments, both undertaken with existing weight management service providers in a real-world context; and both aiming to draw on novel, qualitative data alongside the quantitative.

There were three specific objectives to this mixed methods approach. Firstly, qualitative data would offer a fresh approach to gathering data on heterogeneity based on individual adherence, directly supporting the answers to research question 2. Secondly, qualitative evidence would be used to contextualize and triangulate with the statistical results on average treatment effects and heterogeneous treatment effects. Thirdly, qualitative methods were uniquely placed to shed light on whether the planner-doer framework is an apt theory for health behaviour change, by searching for any evidence of the theory's underlying assumptions on internal tussles, sophistication, and the ensuing demand for commitment devices.

Chapters 5 and 6 described the implementation and results of the two trials, with further qualitative results and analysis presented in Chapter 7. The research findings from these three chapters are summarized in the next section, organized along the lines of the six hypotheses and their overall implications for the two research questions. Table 5 is reproduced (from chapter 3) to guide this discussion (see next page). The remainder of this chapter considers the evidence for the planner-doer theory, generalizability and

limitations of the study (section 3), contributions made to the literature (section 4), and implications for future research (section 5). The chapter closes with final remarks and concludes the thesis has delivered on its aims and provided interesting and robust answers to the questions framed.

**Table 40: Research Questions, Selected Propositions from the Model and Hypotheses <sup>83</sup>**

<i>Research Question</i>	<i>Model's Prediction</i>	<i>Hypothesis</i>
RQ1: Can commitment devices change behaviour to promote desired health policy outcomes?	A commitment device can change health behaviours and deliver desired weight loss. <i>(Proposition 2)</i>	1. A commitment device will generate positive average treatment effects on weight loss and health behaviours.
	A commitment device that generates more costs acts as a more severe tax on the doer's consumption and brings about greater effects. <i>(Proposition 4)</i>	2. A more intense commitment device design will generate larger average treatment effects on weight loss and health behaviours.
RQ2: How does the effect of a commitment vary across people?	Effectiveness will depend on the individual's traits and adherence <i>(Propositions 5 and 6)</i>	3. A commitment device will work more effectively for more self-aware individuals.
		4. A commitment device will work less effectively for individuals with short-termist and myopic attitudes.
		5. A commitment device will work more effectively for individuals who embrace the commitment device more fully.

<sup>83</sup> Reproduced from chapter 3, page 103

## **2. RESEARCH FINDINGS**

### **2.1. Hypothesis 1: A commitment device will generate positive average treatment effects on weight loss and health behaviours**

Neither trial provides evidence of significant, positive average treatment effects from the commitment devices on weight loss. The reputational commitment devices – the commitment contract in the Camden trial and the coach treatment in the Food Monitor trial – had no effect on weight loss in the short run, and actually reduced weight loss at 12 weeks. The limited commitment condition in the Food Monitor trial did not show any negative impact on weight loss, which would have been expected if the financial commitment device were exerting a positive average effect on weight loss. In sum, these findings run counter to much of the published research on alternative commitment device designs such as deposit contracts and public pledges. However, the findings are consistent with the limited available evidence on milder commitment strategies (Chapman et al, 2015); with the exception of the negative treatment effect from the reputational commitment device, which is the first such finding to the best of my knowledge.

Going in to the findings in more detail, the commitment contract treatment may simply have been too mild an intervention to generate additional weight loss effects, for two reasons. Firstly, because there were already commitment elements built in to the Shape Up programme (for example accountability towards group members and the tutor) and there was less scope or need for the commitment contract to address time inconsistency problems on average; and secondly, because of the complex nature of weight management problems, as highlighted in the baseline survey and interviews. Qualitative follow up in the Camden trial highlighted a number of wider life issues that arose during the trial that may have

swamped any positive effect from the commitment devices, including ill health, caring responsibilities, and work pressures. For many participants across both trials, the contract may have been too mild a design ( $d$ ) to generate a large enough psychological tax on the doer ( $\theta$ ) to bring about weight loss in the face of countervailing forces such as challenging and changeable life and health circumstances.

In terms of the financial commitment device, as discussed in chapter 5, there was little difference in weight loss performance between those continuing to pay the monthly subscription fee (experiencing a financial commitment device) and those who received a refund (experiencing limited commitment in the short term). In crude terms this points to zero effect from the financial commitment, but an alternative interpretation suggests there is a positive effect from the commitment that is not undermined by refunding the money. Willingness to pay the financial premium was an upfront payment for staying committed to weight loss goals, and this was not easily dismantled. Taking away the financial obligation for a short time period, as the limited commitment treatment did, had no significant effect on weight loss performance because the psychological tax had already been established, and the commitment already formed; the influence of this commitment was sustained even through the offer (and acceptance) of a refund. Further testing of the effects of financial commitment devices from paying a premium could create a comparison group in some alternative way that did not entail a short-term refund to further unpack this mechanism.

The addition of the coach to the existing financial commitment condition did not generate the expected positive treatment effects. On the contrary, the data suggests that the coach treatment may have had a negative effect over 12 weeks. As discussed in chapter 6, this result may be linked to low compliance and low implementation fidelity amongst the treatment group, as well as pointing to more fundamental issues of commitment overload, and the mismatch of



treatment design features with the digital preferences of Food Monitor clients.

While the commitment devices showed no effect on weight loss in general, they did have significant effects on health behaviours. The commitment contract improved participation in the weight loss programme by raising attendance by 6% and completion rates by 14%. These results imply a Cohen's *d* effect size of 0.16 and 0.19 respectively. The implied benefits in the form of improved self-monitoring, practical knowledge, self-awareness and group support mean that increased participation is an important outcome for the service providers. Commitment contracts are, then, an effective way of increasing regular attendance at public health programmes that rely on repeat visits from participants to have maximum impact on health outcomes. A separate question of why the increased participation did not automatically lead to significantly increased weight loss was not directly addressed by this project, but is arguably an important avenue for future research and evaluation of Camden's weight loss programmes.

The Food Monitor trial generated more complex results on average health behaviour change, measured in terms of online self-monitoring. The reputational commitment device has zero effect on self-monitoring, but the limited commitment condition had a significant and positive effect. The refund offer appears to have spurred on health self-monitoring, contrary to what was hypothesized, and this finding remains a puzzle in two ways. Firstly, why did the increased self-monitoring not lead to a related increase in weight loss amongst the refund group? Secondly, why would the refund have encouraged greater use of the Food Monitor tools? These questions remain unresolved, and point to interesting future avenues for research to understand how to encourage greater self-monitoring through digital health tools, and also how to convert self-monitoring into desired health outcomes.

**2.2. Hypothesis 2: A stronger commitment device will generate larger average treatment effects on weight loss and health behaviours**

The dissertation aimed to test this hypothesis in two ways. Firstly, the Food Monitor trial allowed for a comparison of three commitment conditions in increasing size of implied psychological tax. Secondly, the effect sizes implied from both trials were to be compared to the prior literature on weight loss, as summarized in the Literature Review.

Applying the first test, the Food Monitor trial results refute the hypothesis. The increase in psychological tax implied by the overlaying of reputational commitment to an existing financial commitment would have been expected to leverage a larger, positive effect on weight loss, but the opposite is true, with the coach group performing worse in the medium run and no better in the short run than those in the limited and financial commitment groups. The earlier discussions highlighted the role of commitment overload, and there remains the possibility that artefacts of the trial in the form of low compliance and low adherence may mask the true impact of the coach treatment. The fact remains, however, that the adding up of commitment elements did not lead to an unambiguously superior outcome for weight loss and behaviour change.

Further, the statistically insignificant difference between the limited and financial commitment conditions underscores the idea that a simplistic interpretation of the psychological tax based on the stakes is unwarranted. The size of the refund was not associated with weight loss performance or self-monitoring behaviour. Making the premium payment was the key step towards cementing commitment; being offered a short term refund did not dismantle that sense of commitment, implying that once the financial commitment has

already been voluntarily made, adding or taking away small financial stakes is irrelevant to the psychological tax an individual experiences.

These findings highlight the need for a nuanced understanding of ‘stronger’ and ‘weaker’ commitment devices. The intensity of the psychological tax exerted by a commitment device cannot be accurately understood on a linear scale depending on the nature of the stakes (reputational or monetary), as implied by Bryan et al who distinguish between soft and hard commitments based on whether psychological or monetary costs are at risk (2010, p.672). This was the basis of the commitment device intensity spectrum set out in Chapter 2, and effect sizes drawn from published literature broadly fit this rule of thumb. For example, Chapter 2 highlights larger effect sizes arising from deposit contract interventions (Volpp et al, 2008, John et al), while more modest effects are reported from public pledges and personal rules made with a partner (Nyer & Dellande 2010; Prestwich et al. 2012).

The results of the Camden and Food Monitor trials showed that other design features come to the fore, including the appropriateness of the stakes for the population, and the possibility of too much commitment causing unintended negative effects (discussed further below). These factors suggest the stylized spectrum of commitment devices set out in Figure 1 of the Literature Review is overly simplistic.

### **2.3. Summary: research question 1**

In answer to research question 1, the trials show that the relatively mild forms of commitment device tested here can change narrowly-specified behaviours (attending a class, using a calorie counter), but are ineffective as a solo intervention to bring about weight change. This finding plays a useful role in grounding the expectations of what commitment devices as a behavioural tool can achieve in the weight loss sector: arguably they are best seen as an aid to support time-bound goals that rely on specific actions, rather than a complex behaviour change process or an ambitious goal such as 5% body weight loss.

How do these findings speak to the literature? The average treatment effects reported in chapters 5 and 6 challenge much of the published literature (summarised in Figure 34), which has hitherto reported much larger and positive effects (notably Volpp et al, 2008, and John et al, 2011).<sup>84</sup> The disparity in the graph below is partly explained by the fact that the Camden and Food Monitor trials deliberately focused on under-researched commitment devices – a premium payment and two reputational commitments – that are milder in form than those that have been widely discussed in the scholarly debate. Following this argument, the findings bear out the idea that different commitment device types (as set out in chapter 2) give rise to varying degrees of the psychological tax ( $\theta$ ) that is critical for delivering health behaviour change. The results are consistent with the idea established in the literature that the stronger the commitment device, the more likely it is to generate sizeable effects on behaviour and outcomes. However, the Camden and Food

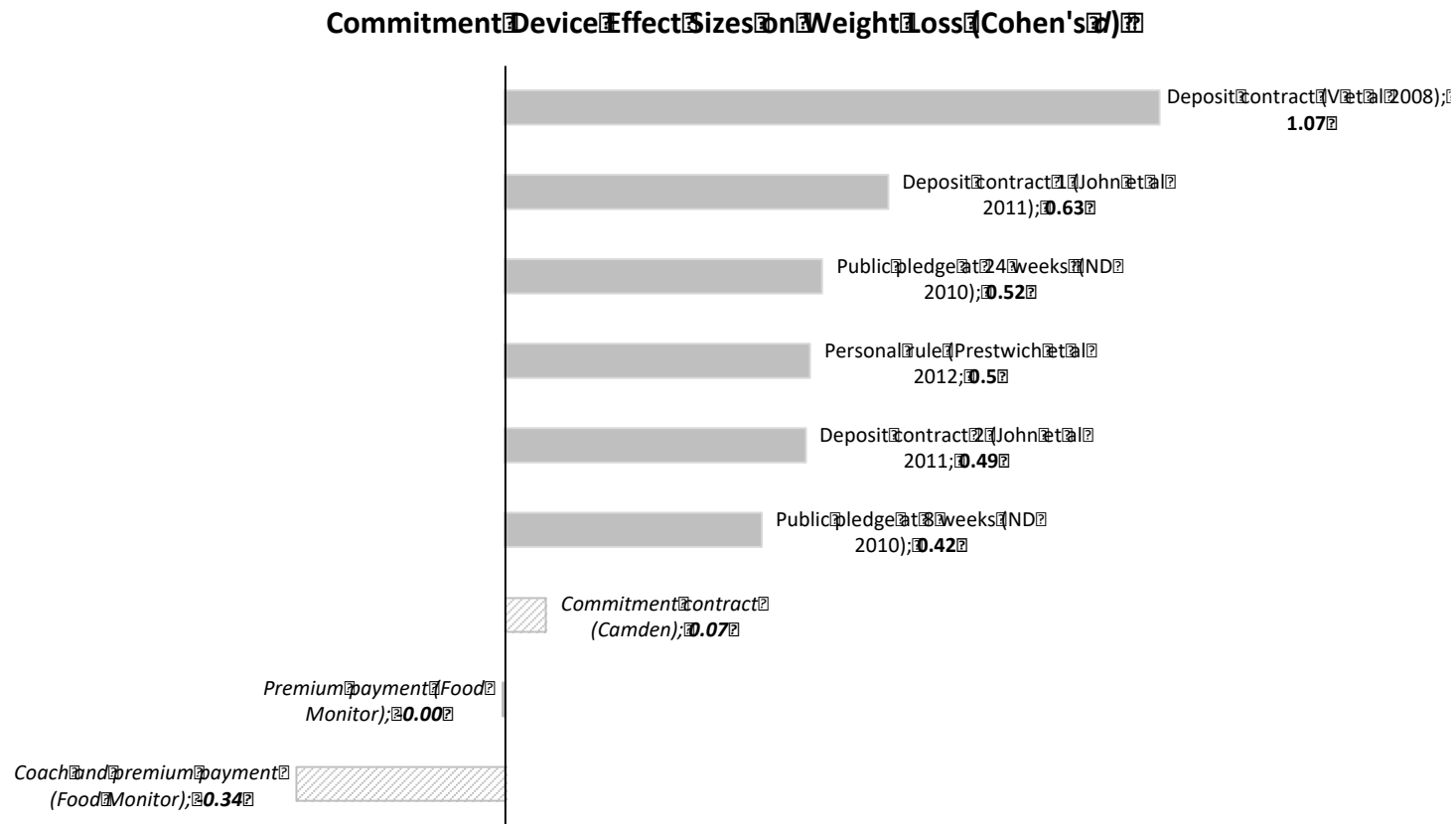
---

<sup>84</sup> Using Cohen's *d* effect sizes for published literature as reported in chapter 2, with effect sizes for Camden and Food Monitor trials calculated using the same method. Where more than one set of results was available, studies are cited twice (for example Nyer and Dellande 2010 report results at 8 and 24 weeks and have two separate bars on the graph). Results are not standardized across timeframes (for example John et al consider outcomes at 32 weeks, while the Food Monitor trial is at 4 weeks) so data should not be treated as meta-analytic. Figure 34 is purely illustrative of wide-ranging reported effect sizes across commitment device types.

Monitor trials add nuance to this finding, pointing to the need for more granularity in examining what makes a commitment device stronger or milder: design features around its form (digital or physical), its medium (online or personal), and its salience (how easily forgotten is it) are equally if not more important than just the stakes involved (money, or reputation).

Figure 34 also corroborates the criticism leveled by Paloyo et al (2014) that financial deposit contract studies have tended to bundle the intervention other features. Figure 34 may be pointing to the much milder treatment effects arising when commitment devices are isolated, and left to operate without frequent researcher involvement and external prompts that maintain the salience of the weight loss goal and costs of renegeing on it.

Figure 34



#### **2.4. Hypothesis 3: A commitment device will work more effectively for more sophisticated individuals**

Sophistication was tested both quantitatively and qualitatively in the Camden trial. Heterogeneity regression analysis relied on a proxy for self-awareness based on whether the participant had taken part in a weight loss programme previously. The hypothesis was predicated on sophisticated individuals being more aware of their self-control challenges, and would therefore seek to apply the commitment device more carefully to bring about behaviour change and weight loss. More naïve participants, in contrast, might have under-estimated their own need for the external commitment aid, and subsequently benefitted less from it.

Robust results emerged on behaviour change: the sub-group of sophisticated individuals registered higher attendance ( $p=0.006$ ) and completion ( $p=0.004$ ). No association was found between sophistication and weight loss using quantitative data, but qualitative analysis identified that the sub-group of sophisticates reported average weight loss of 3.7%, higher than the 3.1% among other interviewees and the 2.7% of the wider sample, suggesting some association between sophistication and weight loss.

The analysis underscored the difficulty of developing an operational and unbiased measure of sophistication. While there were few overt examples of sophistication in the qualitative content analysis, convincing qualitative evidence – from 47 references to personal commitment strategies covering 92% of interviewees – highlighted that people are aware of their self-control problems and want commitment devices to keep themselves on track, which they often develop for themselves in the form of tailored personal rules for diet and exercise that fit around their home and work lives. All in all this evidence suggests that sophistication is an important trait for future research into behavioural aspects of weight management.

**2.5. Hypothesis 4: A commitment device will work less effectively for those with short-termist attitudes towards their health**

The second individual trait expected to generate heterogeneous treatment effects was the degree of short-termism in the individual's outlook, based on the assumption that impatience runs on a spectrum, and present bias is not binary. The greater the preference for current gains over future gains, the less effective the commitment device was expected to be in changing health behaviours.

Short-termism was operationalized through two contrasting measures. The first was a customized measurement to proxy the cost of waiting for a delayed payoff. The higher the cost of waiting, the greater the implied degree of present bias. A negative interaction between the commitment device and degree of present-bias was confirmed in the Food Monitor weight loss results, with the reputational plus financial commitment device causing lower weight loss amongst those with a higher degree of present bias. No further associations were found for self-monitoring behaviour or short run weight loss.

The second measure used a health attitudes survey instrument to identify short-termism in relation to health behaviours in particular. This sub-group were expected to benefit less from the commitment devices, because it was plausible that their doer sub-self was more influential in decision-making, and so could be more resistant to commitment devices that aimed to rein in current gains in favour of the long term.

Findings from heterogeneity analysis across both trials showed no significant association between commitment devices, health attitudes, and weight loss outcomes. Myopic health attitudes



could not explain the large variation in weight loss readings within experimental groups, but did explain some part of the observed health behaviour change, and this finding is new to the literature.

The expected relationship is borne out in the Food Monitor trial where the refund, which represented the relaxing of the commitment, was especially effective among those with negative and short-termist health attitudes ( $p=0.079$ ). This insight is challenged by the results from the Camden trial, where participants with myopic health attitudes were more likely to benefit from the contract and participate more actively in the Shape Up classes ( $p=0.07$  in both cases). This finding refutes the hypothesis but offers new information to the scholarly debate; and raises the possibility that those with a more negative and short-termist outlook may stand to gain more from a commitment device in some settings.

Taken together, the contrasting findings for short-termism across different commitment device interventions suggests an interaction between the design of the commitment device and health attitudes. Not only do design features and individual traits matter on their own in determining heterogeneity of treatment effects, as set out in chapter 3 (equation 7), they also interact amongst themselves, generating further complexity to the causal mechanisms underpinning commitment devices. This emerging evidence on the interaction between health attitudes and commitment device design features is new to the literature on commitment devices, and prescribes a refinement to the Analytical Framework presented in chapter 3. Future research designs will need to embrace qualitative methods and allow for precise sub-group comparisons of appropriate scale to develop a deeper understanding of how commitment devices can be best designed and targeted for optimal effect.

**2.6. Hypothesis 5: A commitment device will work more effectively for individuals who embrace the commitment device more fully**

The third factor proposed to affect  $\theta$ , the magnitude of the commitment device effect, is adherence to the commitment device,  $\lambda$ . In the case of the commitment contract, high adherence would mean the individual recalled the contract and used it as a visual reminder during the Shape Up programme; while in the Food Monitor trial, adherence to the coach treatment would imply that the named coach would be aware of the trial and kept informed of weight loss progress. Qualitative data enabled exploratory analysis in both trials, starting with a coding scheme to identify low and high adherence participants (see chapter 7), to evaluate whether differences in adherence explained differences in commitment device treatment effects.

In both cases, the hypothesis appears to be borne out: those who maintain the salience of their commitment device tend to report strong outcomes. High adherence participants in the Food Monitor trial appear to be more successful with weight loss. In the short term the high adherence group loses 2% of initial weight compared to the low adherence group's 0.4%; and at 12 weeks registers weight loss of 3%, compared to 0.5% amongst the low adherence group. No difference is noted in self-monitoring behaviour, suggesting that some other factors underpin the causal mechanism.

Camden participants are also easily coded into low and high adherence groups, with some distinctive features. The high adherence group is entirely female, older and self-referred, and far more likely to report negative or short-termist health attitudes. High adherence participants registered stronger participation in the Shape Up programme with 9% higher attendance and 14% higher completion rates, but no significant difference in weight loss outcomes.

What do these findings imply for the application of commitment devices? The Analytical Framework is correct to specify adherence as a key factor determining the effectiveness of the commitment device on behaviour change. Commitment devices will be effective only when they are embraced and their salience is sustained. If it is the case that those who face more challenging personal circumstances for their weight loss journeys (such as ill health, work or family pressures) are also less likely to adhere, this offers insights into the wider context of weight loss challenges, and the potential role for commitment devices. The implication is that commitment devices may be more effective when they are taken up in a conducive context, when it is relatively easy to stick to the good behaviours without wider, complex life circumstances that overpower the best of intentions; if the latter circumstances prevail, weight loss efforts will be markedly more difficult and less likely to succeed, even with a commitment device.

## **2.7. Summary: research question 2**

In answer to research question 2, the dissertation delivers a robust body of evidence that commitment devices work differently across individuals. In support of the propositions arising from the Analytical Framework, the findings show that individual traits and adherence are important. Not only do commitment devices have selective appeal as demonstrated in the literature to date, they also work best for particular sub-populations. There are many ways to dissect a target population, and the dissertation has shed light on two key traits – sophistication and short-termism – and shown they merit further research as informative heterogeneity pathways. Importantly, the design of the commitment device matters greatly, as do the interactions between design and adherence (a more appealing commitment device will remain salient, while an unappealing one is more likely to be forgotten) and the interactions between design and individual traits (those with myopic health attitudes performed better with a visually salient contract, but also responded positively to relief from the premium payment). These findings represent new contributions to the literature in terms of methods, theory and evidence, which are elaborated in later sections.

### **3. GENERALISABILITY AND LIMITATIONS OF THE STUDY**

The dissertation has arrived at sound conclusions on a number of issues, providing convincing answers to the two research questions. There remains, of course, ways in which the research design could have been improved. The following section reviews four key areas: sample size and power, attrition, qualitative analysis, and the external validity of findings.

#### **3.1. Sample size for weight loss outcomes and heterogeneity analysis**

Firstly, the trials would ideally have had larger sample sizes, built on calculations with more realistic assumptions: on attrition rates; on weight loss impacts of reputational commitment devices; and on sub-group analysis. On the first issue, the sample size calculations did not account for potential attrition (flagged in chapter 5). While there may not have been an easy solution to the problem of attrition even if it had been identified in advance – it may not have been feasible to target a 50% larger sample due to constraints with the partner firm – future research should pre-empt the erosive effects of attrition on effective sample size by building in a relatively high degree of attrition, particularly as the time span for outcome data grows (attrition at 12 weeks was higher than at 4 weeks).

In a similar vein, the weight loss differentials based on commitment device interventions can also be made more realistic by assuming lower effect sizes than those most commonly cited in the literature. The trial presented in this thesis arguably serves as such a pilot, and can support decision making for future research efforts. This would ensure that experiments are sufficiently powered to detect modest impacts of commitment devices on complex processes

such as weight loss outcomes as well as more tightly defined behavioural outcomes.

### **3.2. *Attrition strategies***

Secondly, while attrition was anticipated in both trials, the extent was larger than expected particularly with Food Monitor, where many participants used the weigh-in tool infrequently. As an example of a trial design feature that might encourage improved outcome data reporting, weight loss self-reports could be prompted through personal email and text reminders for users of digital weight management services like Food Monitor, as was done in the Volpp et al (2008) study that incorporated daily contact with the researchers. However, this would be feasible only in an alternative study where self-monitoring behaviour was not measured as an outcome, and where the external prompts are seen as part of the treatment itself. Neither of these features were attractive for the research design presented in this thesis, which aimed to isolate the effects of the commitment device, and to study those effects as they related to self-monitoring.

Instead, this dissertation has made robust use of ex post statistical techniques to work around the attrition problem. The experience has underscored the value of ex ante techniques to minimize missing outcome data, particularly where the nature of missingness remains unobservable.<sup>85</sup> Where there is no scope for triangulation with administrative data, nor for incentivizes to generate more regular participant self-reports, and constraints to large-scale follow up participants directly, a second-round sampling strategy that makes maximum use of a small number of randomly assigned follow-ups would be a great advantage. This would have involved randomly selecting attritors for follow up efforts, and using data recovered from these efforts as a basis for modeling the missing

---

<sup>85</sup> The pre-analysis plan did not expand on techniques to address attrition, and this is a lesson learned for future analysis plans.

data from the full set of attritors (Gerber & Green 2012). In the Food Monitor case, such follow up was not feasible due to the preferences of the partner organization. One lesson that emerges is the importance of forging agreement with all stakeholders in advance of the trial to allow for a follow up stage, although this may well be more feasible with public agencies (such as Camden) than private sector firms (such as Food Monitor).

### **3.3. *Qualitative analysis for behaviour change***

The contrast between Camden and Food Monitor highlights the invaluable depth and nuance that qualitative follow up can offer, particularly when the quantitative results report contrary findings that do not fit with theory. The negative effect of the reputational commitment device at 12 weeks is a puzzle, particularly overlaid with a zero effect at 4 weeks. There are unresolved questions here about how the treatment was interpreted, and why weight loss performance seemed to go off track in the two months after the treatment period. The larger qualitative component in the Camden trial came about partly as a lesson learned from Food Monitor, and partly due to greater flexibility from the research partner. The Camden partners were content for me to interact directly with clients, and this allowed for much more in-depth probing of how the contract was received and applied. The interviews gave rise to unique insights on adherence and the concept of commitment saturation.

While the Camden trial made stronger and more timely use of the qualitative data, including pre-testing the contract design with experienced Shape Up tutors, both trials could have investigated how treatments were interpreted by clients in a pre-testing phase. This would have allowed, for example to identify whether the refund was viewed as a relaxing of the premium payment, leading to clients feeling less pressure to make maximum use of the Food Monitor service; or, alternatively, if the refund was seen as a gift or incentive

to use the service further. Qualitative pre-testing with clients themselves would have helped to explain whether they were interpreting the online treatment messages as intended. A key constraint that prevented this exercise was the reticence of Food Monitor to allow direct interaction with their fee-paying clients, and the lack of resources in terms of time and money for a stand-alone pilot phase. The qualitative work done in advance of the trials therefore focused on experts rather than the target population, but ideally future research would involve both sets of people.

### **3.4. *External validity***

A common criticism of randomized controlled trials is their inability to travel to alternative contexts, with the applicability of findings confined to the policy and programme circumstances that the trial took place in (Deaton & Cartwright 2016). To some extent this study is answerable to this critique, which can be decomposed into three specific aspects: the programme context, the design of the intervention, and the population of interest.

The type of weight management programmes that have hosted the trials are fairly widespread across the UK. Many public health authorities offer similar group programmes as the Shape Up course in Camden; and the Food Monitor website is available across the UK and beyond, with similar digital health tools offered by competing firms reaching millions more people worldwide hoping to better manage their weight. In other words, the programme context is far from niche, and even if the findings only apply in such weight management contexts, there remains a large pool of opportunities for commitment devices to be further tested and applied.

The criticism against generaliseability strikes most closely at the second aspect, with findings arguably extending only to very similarly designed commitment device interventions in other health programmes. This is borne out by the discussion in section 2, which



highlights the complexity of the causal mechanisms, and the fact that commitment devices cannot be treated as a single entity. For example, this dissertation has highlighted an important distinction between premium payments and deposit contracts, both of which can be understood as financial commitment devices but which operate in different ways and with different effect sizes. It is right, therefore, to be cautious about how well the findings on the commitment devices tested here – the commitment contract, coach, and premium payment – can be used to predict the performance of different commitment devices such as public pledges and deposit contracts, even in a similar setting. Further, the studies here have offered new insights on heterogeneity of treatment effects in particular, but good science demands that these novel findings be subject to replication and verification before external validity can be claimed.

It would be claiming too much to say that the findings could immediately be applied to other health behaviours, such as smoking cessation. The trials conducted here were concerned with weight management, and the population of interest was obese and overweight people accordingly. Commitment devices have been shown to be acceptable and effective for significant sub-populations of overweight and obese people. Even if the results can only be claimed to apply to these sub-populations, this remains a sizeable and important target group in the context of rising obesity in the UK.

What of the qualitative results? These were not designed to offer generaliseability in the same way as the field experiments; rather, “generalization in qualitative research usually takes place through the development of a theory that not only makes sense of the particular persons or populations studied, but also shows how the same process, in different situations can lead to different results” (Maxwell 1992, p.293). In this sense, with the qualitative research offering robust evidence to support the theoretical framework – confirming the selective appeal of commitment devices, shedding

light on the use of personal rules, new insights on the validity of assuming planner-doer internal interactions for at least a minority of participants, and affirming the varying degrees of sophistication in understanding one's health behaviours and preferences – the thesis as a whole is better equipped to claim that the findings can travel beyond the confines of the trials presented here.

Overall, no single field experiment can or should be used to extrapolate statistical findings, and the trials presented in this dissertation are no different. Replication is important to either confirm or refute the exploratory analysis, and the novel and contrary findings, reported in chapters 5, 6 and 7 (McDermott 2011, p.34). Within the caveats discussed above, the research presented here is, however, arguably of wider relevance, and can be used to inform and test interventions for other health behaviours and other contexts. For example, the theoretical foundations for the heterogeneity analysis provides a solid basis for expecting commitment devices to interact with personal traits in other weight management contexts, or if applied to other behaviours that rely fundamentally on intertemporal choice, where planner-doer tussles are most likely to arise.

#### **4. CONTRIBUTIONS TO THE SCHOLARLY DEBATE**

Despite the challenges in implementing and analyzing the field experiments, this dissertation makes contributions both for the scholarly debate and for policy. It presents a new analytical framework derived from planner-doer theory to explain health behaviour change, and offers fresh evidence that the planner-doer theory is more than merely a “metaphor” for behaviour. The field experiments add new evidence on how commitment devices work in practice for weight management, contributing both to the scholarly debate on these interventions, and offering fresh insights for public health programme design. Further, the combined quantitative-qualitative design represents an innovation in health behaviour change field experiments. These contributions are expanded below.

##### **4.1. Contributions to theory**

###### **4.1.1. A new Analytical Framework for health behaviour change**

In the 36 years since the publication of Thaler and Shefrin’s planner-doer theory, scholarly developments have tended to focus on empirical applications for savings and health behaviours (Ashraf et al. 2006; Giné et al. 2010; Dupas & Robinson 2013), and theoretical extensions through game theory (Bénabou & Pycia 2002; Fudenberg & Levine 2006) or the development of specific concepts such as sophistication (O’ Donoghue & Rabin 1999). Relatively recently, the dual-self framework was used to model the problem of food over-consumption (Ruhm 2012). Commitment devices have been understood as a natural solution arising from the planner-doer tussle (Bryan et al. 2010); but no work, to the best of my knowledge, had been undertaken to apply and extend the planner-doer framework to commitment devices for health behaviour.

This dissertation fills this gap (chapter 3) by presenting an original formal model of the planner-doer framework. Key predictions on both the demand for and effect of commitment devices on health behaviour and weight management were made explicit, and these propositions tested formally through the research design of two field experiments. The framework also provides the most detailed analysis to date of how heterogeneity of commitment device effects can be expected to arise.

How useful has the analytical framework proved? It was especially valuable in creating a research design brief, by highlighting the need for a mixed quantitative-qualitative approach, and clarifying three categories of heterogeneity pathways to be tested: individual traits, adherence, and design features of the commitment strategy. Arguably, many features and predictions of the framework were confirmed by the trials, including the fundamental notion of a dual-self individual, with competing desires for the short and long term (see next section). The first proposition of the framework states that an individual will identify their need for a commitment device, and the findings suggested that this is often true: not only did a sizeable portion of participants take up the commitment devices offered (albeit less for the coach treatment than the contract), the interviews highlighted that many people applied their own personal rules as commitment strategies to aid their weight management.

The analytical framework also unpacked the concept of  $\theta$ , the preference modification parameter in Thaler and Shefrin's original model. In doing so, the framework highlighted that the psychological tax that could be weaker or stronger, and effect behaviour change accordingly. The qualitative analysis of interviews corroborated this assumption, with people often describing feelings and experiences akin to a psychological tax on their consumption behaviour once they had taken up a commitment strategy (chapter 7).

The framework identified three broad heterogeneity pathways – design of the commitment device, individual traits, adherence and put forward a simplified functional form to describe their relationship (Equation 7 from chapter 3 is repeated here for ease of reference):

$$[7] \quad \theta = f(d, \tau) \cdot \lambda$$

The results discussed in section 2 above highlight the interaction effects between design features, personal traits and adherence to the commitment device once it is taken up, offering clear support for this functional form.

The empirical findings also call for some refinements to the framework, such as the assumption that the psychological tax increases monotonically with commitment, and behaviour change is effected to a greater degree in the same (positive) direction. Rather, the results discussed in chapters 5 and 6 suggest the possibility of commitment saturation, commitment overload, and thresholds for these phenomena varying across overlaid layers of commitment. For example, the addition of the coach to those already paying a premium payment had a negative effect on weight loss in the Food Monitor trial, with one possible explanation being a commitment overload and participants being ‘switched off’ from their health goals. In contrast, exploratory evidence from the Camden trial suggested that receiving a GP referral and then a contract delivered a larger behavioural effect than either commitment element on its own. The simplified linear relationship between commitment,  $\theta$ , and outcomes may not always hold, and this is an important area for future research (see section 5 below). On balance, however, the Analytical Framework has provided clarity, transparency, and a robust basis for theorizing health behaviour change using the planner-doer framework, and has served the thesis well.

#### ***4.1.2. Evidence for planner-doer theoretical assumptions***

A final set of findings provides credible and novel support for some of the key assumptions underlying the planner-doer framework relating to the characterisation of dual sub-selves, the concept of sophistication, and the predicted demand for commitment strategies.

An innovative coding scheme was developed to identify the strategic internal grappling between planner and doer sub-selves in chapter 7. Qualitative analysis suggests a minority of participants do exhibit such internal tussles, offering the first such evidence that the ‘metaphor’ of dual-self theory can be operationalised as a practical reality. It also emerged that the framework, and its foundations in the intertemporal self-control problem, may not be applicable to all participants seeking weight management help. Where the key barrier was information or wider circumstances preventing them from taking the right actions, the planner-doer lens was clearly not the appropriate framework, and commitment devices not the right solution. These findings help explain the low average treatment effects, and underscore the need for precise targeting of commitment devices towards those aiming to overcome self-control problems to achieve health behaviour change (see figure below).

The Analytical Framework implicitly incorporated this in chapter 3 by expressing Paul’s time inconsistency problem as one rooted in willpower and inaction, not a lack of health information, nor some physical inability to act on that information. Insights from trial participants entrench the idea that the planner-doer theory is not universally applicable to all behaviour change contexts; rather it offers a sound explanation and solution for a subset of cases where healthy behaviour is constrained by willpower, and time inconsistency is a result of self-control problems. This clarification, that a mainstream public health programme may have only a

minority of clients who require a commitment strategy, helps explain the low average treatment effects on weight loss, and calls for more precise targeting of commitment devices.

Evidence of sophistication in weight management behaviours also emerged from the Camden interviews, with participants highlighting a high degree of self-awareness of when, where, and how they felt tempted to trade off their future health goals for some momentary gain. Chapter 7 provided many examples of how they would then cajole, threaten, and trick themselves into staying on track, providing further support for the idea of a planner sub-self trying to instill discipline on a doer sub-self's wayward actions. Finally, these conversations often led to examples of personal commitment strategies that made this self-discipline more likely to withstand daily temptations: changing work and home routines to fit in exercise and good diet more easily, avoiding shops and aisles in the supermarket to remove the source of the temptation altogether, and using new digital tools to improve self-monitoring and provide regular feedback on health outcomes and behaviour.

Given the low and weak average treatment effects, it is appropriate to question whether the theory is fit for purpose. The evidence discussed in chapter 7 provides a robust body of evidence that the fundamental assumptions of the planner-doer theory, and their predictions on demand for commitment devices as set out in the Analytical Framework's propositions 1, 2 and 3, are valid. Explanations for the statistical results are found, instead, in the issues discussed above, including the importance of design features, targeting of the appropriate population in need of commitment devices, and the interaction with individual traits.

## **4.2. *New evidence on how commitment devices work***

### **4.2.1. *Financial commitment devices***

To the literature, this research adds new evidence that financial commitment devices are not all the same, with premium payments exerting different effects to deposit contracts. The former appear to rely on strong innate motivation, and the commitment element is ‘sticky’, so a temporary removal does not erode the underlying will to achieve behaviour change. Where there is a premium payment in place to serve as a financial commitment device, it is not parting with the money that is important but the willingness to do so. Once this psychological commitment has been made, the return of that money has little effect, and this may explain why prior literature has found that people pay ‘not to go to the gym’ (DellaVigna and Malmendier, 2006). This is a critical distinction between premium payments and deposit contracts, with the latter very much operating on the principle that the individual does not want to lose the money, with loss aversion driving the behaviour change process. The finding also has potential implications for a related literature on financial incentives for health, suggesting that such incentives, particularly if they are in the form of refunds, may be less effective where an individual is already accustomed to the idea of paying for something.

### **4.2.2. *Commitment overload and commitment saturation***

The research highlights for the first time the prospect of commitment overload. Rather than there being a monotonic and linear relationship between commitment and behaviour change, too much commitment can have unintended negative consequences. The existence of a commitment threshold is indicated by the negative treatment effect from the coach treatment overlaid on the premium



payment commitment device (chapter 5). At some previously unidentified threshold, additional commitment appears to have a counterproductive effect. This finding speaks to the literature on implementation intentions, and the potential negative effects from an overload of planning towards a goal (Verhoeven et al. 2013), and provides new evidence that a similar principle applies to commitment devices.

The related idea of commitment saturation is further supported by the Camden trial, where the scope for commitment aids only work up to a certain point and beyond this point had no further effect (unlike the Food Monitor trial, there was no evidence of negative effect beyond the threshold). Specifically, participants who attended the introductory session of the Shape Up programme benefitted less from subsequently receiving a commitment contract. Having already received a form of commitment priming from group tutors, there was little impact left for the commitment contract to make and the treatment effect was weak. In contrast, those who missed the introductory session faced a commitment gap, tended to participate less and experienced lower weight loss. If they received the contract they fared much better, indicating that the contract plugged the commitment gap, and allowed catch up with those who had attended the introductory session.

### **4.3. *Innovations in methodology***

The Literature Review highlighted the lack of qualitative analysis in field experiments as a key drawback to understanding the psychological processes underpinning health behaviour change. Various studies that had found positive and significant average treatment effects did not undertake qualitative analysis to understand the causal effects reported by statistical models, to understand how the commitment device was experienced, to what extent it was salient in the minds of participants, and whether it could be overtly linked to successful (or unsuccessful) behaviour change. Despite positive treatment effects in many of the studies reported in the Literature Review, they often registered wide variation in weight loss experiences, and qualitative follow up might have shed light on this heterogeneity.

To fill the gap and provide more granular understanding of the statistical findings, this thesis set out to combine qualitative analysis with the core field experiment methodology to investigate causal effects of commitment devices. Chapters 5, 6 and 7 have shown that quantitative and qualitative methods can be successfully combined within a field experiment, with significant value added from generating new data, corroborating and challenging statistical findings, and framing new research hypotheses. In particular, qualitative analysis has uniquely allowed for generating a new proxy variable for sophistication and adherence to commitment devices, and for investigating the veracity of the planner-doer characterisation for intertemporal choice. In terms of process, qualitative data has helped to assess the fidelity of the field experiment design (White 2013). For example, it corroborated the expectation of no contamination across treatment and control groups in the Camden trial, with no evidence from the interviews suggesting that participants in the control group were aware of the contract. This

supports the conclusion in chapter 6 that this potential threat to validity was not, ultimately, of concern to the analysis.

While rare, it is not unique to incorporate qualitative analysis and field experiments within a mixed methods design; but this thesis goes beyond much of the literature in ensuring that findings from both methods and datasets are closely integrated for interpretation (Lewin et al. 2009, p.5). In line with recommendations for integrating mixed methods data, it is possible to triangulate the findings by identifying areas of consonance, dissonance and corroboration between the qualitative and quantitative analysis (O’Cathain et al. 2010), and this has been woven organically throughout chapters 5 to 7 and summarized in chapter 7. A good example of corroboration in the Camden trial is the qualitative evidence that suggests where a strong reputational commitment was forged with tutors or group members, a further self-reputational commitment from the contract was not appealing. This provides further evidence of commitment saturation, which helps to contextualise the low average treatment effect discovered in the regression analysis and highlights new heterogeneity pathways for future research.

## **5. IMPLICATIONS FOR POLICY AND FUTURE RESEARCH**

### **5.1. Fresh insights for policy makers**

For policy makers, the dissertation ultimately finds that commitment devices can work, but in specific circumstances. They are not a universally applicable solution, but work best for a selective sub-population who face a time inconsistency problem, are sophisticated enough to recognize their self-control problems, and are not already overloaded with external commitments. So how can commitment devices best be harnessed for public health programmes? Figure 35 maps a range of scenarios for policy makers who are addressing weight management issues.

#### **5.1.1. Know your target audience**

The first fact to establish is what is the key constraint to addressing excess weight, which requires in-depth understanding of the individual and their personal barriers to behaviour change. Is it a lack of information, or some life circumstances that prevent action from being taken? Or is it an issue of willpower and sustaining motivation in order to follow through on a course of action? If the former, a commitment device is likely to be ineffective: it is the wrong solution for the problem. If the latter, there are a menu of potential options, including commitment devices.

As set out in the Analytical Framework there are understandable reasons why commitment devices would not be taken up, despite their being an appropriate solution to bind future choices. A commitment strategy is inherently costly (to one's freedom, for example), and it would require certain motivational thresholds to encourage an individual to opt in. The Camden trial highlighted that despite the wide availability of information on healthy living, information gaps remain a barrier to making healthy choices (chapter

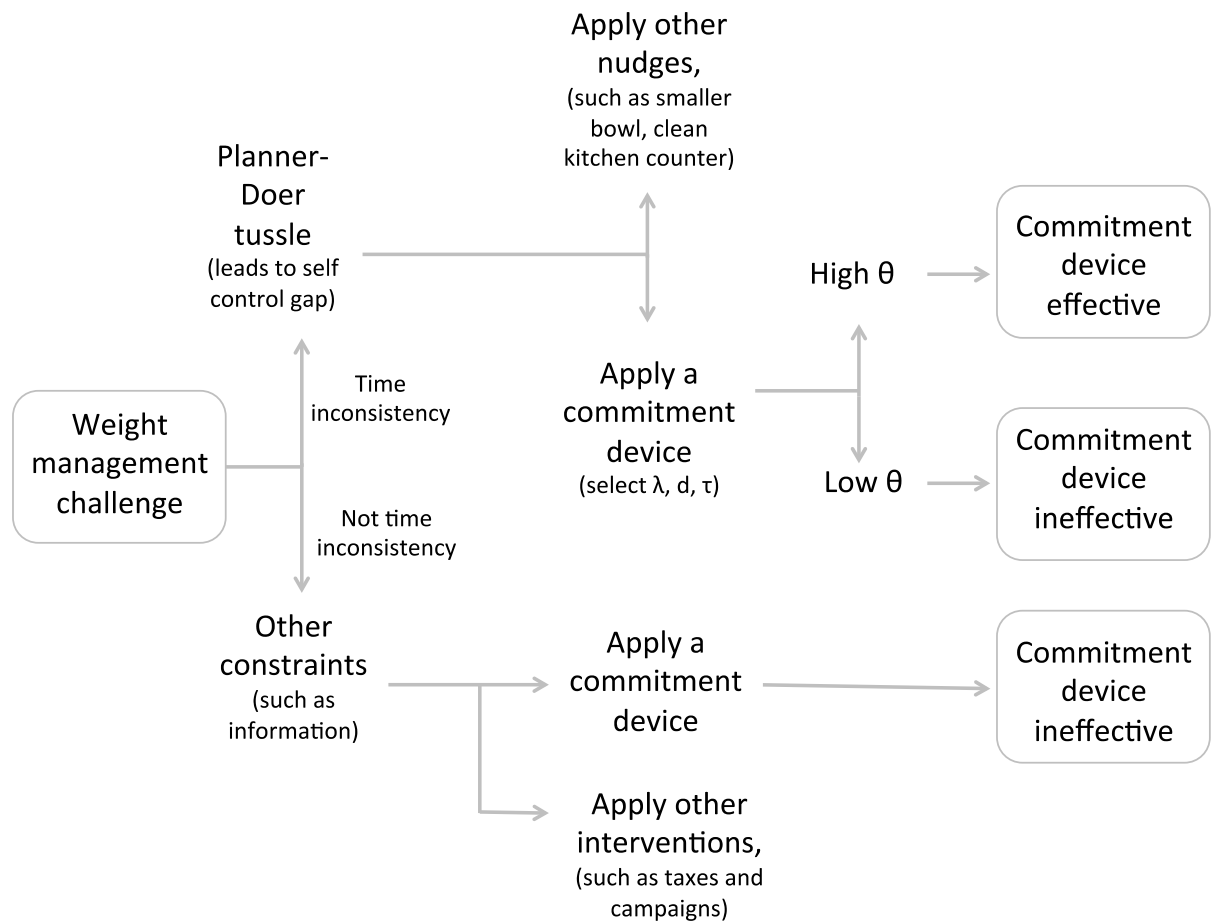
7), and not everyone who is overweight or obese is time-inconsistent due to self-control issues. The advice to policy makers is clear: commitment devices are applicable to a sub-set of individuals who are not following through on their choices due to internal conflict over short run versus long run benefits.

### ***5.1.2. Tailor the commitment device design to the individual***

Figure 35 highlights that once a commitment device is identified as the appropriate and preferred intervention, in-depth understanding is once again required in order to design a commitment strategy that is tailored to the individual's unique needs and preferences. Individual traits and health motivations should be considered, whether it is money or reputation at stake, how public to make the commitment, and how frequently the strategy will be monitored and reaffirmed to ensure its salience. All these features emerged as key issues from the Camden and Food Monitor trials.

While cheap, they do require careful consideration of targeting and design, as sub-optimal features can have adverse consequences for adherence and the goals themselves. The suggestion of personal coach for Food Monitor clients may have jarred with those who specifically chose an online financial commitment device to reflect their preferences. The best way to ensure the commitment device is designed appropriately is perhaps to co-create it, finding ways to tailor it to the individual's preferences: are the stakes reputational, financial, or both? Is it a digital commitment device or one made in person? How public should it be, shared across social media or with just one trusted health professional?

Figure 35: A guide to applying commitment devices for health behaviour change



Adherence, too, is critical. The Camden trial highlighted the importance also of returning to the commitment device regularly, perhaps re-affirming it or revising it, to ensure it is fit for purpose and not a drag on the behaviour change process. Learning the lessons of the Shape Up trial, partners at Camden have expressed a desire to consider motivational interviewing involving the commitment contracts in their new phase of weight management programmes.

A one-size-fits-all approach will not work with commitment devices, but simple steps to apply commitment devices in the context of individual self-reflections or counseling could pay dividends. Figure 2 reiterates that these features – individual traits ( $\tau$ ), design features of the commitment device ( $d$ ), and adherence to the commitment device ( $\lambda$ ) – all play a role in determining the effectiveness of the commitment device ( $\theta$ ).

***5.1.3. Commitment devices are better able to change simple, discrete behaviours rather than deliver complex outcomes***

With sound targeting, positive interaction between individual traits and design, and sensitivity to salience and adherence, the commitment device could leverage effective behaviour change. But there is also the chance that it will deliver weak behaviour change, and hence it cannot be relied upon in isolation to tackle complex weight management issues. Like many other behavioural interventions, commitment devices are best seen as part of a range of actions to support healthy living. Where they work well, individuals cite increased motivation, self-discipline, and enthusiasm for their health goals; and this explains why many individuals set up personal commitment devices in various small ways that collectively can add up to behaviour change and better health. While reputational commitment devices have smaller effects than deposit contracts, they

are cheaper and easier to implement, and extremely adaptable to any situation or health behaviour.

## **5.2. *Future research directions***

Heterogeneity of commitment device effects does indeed exist, and many new avenues for enquiry appear promising. This dissertation made progress in operationalizing sophistication, and more work remains to be done. This research is one of few attempts to apply the Healthy Foundations Segmentation model to health behaviour change research, and it warrants further application with a larger sample size to further understand the differences across the five motivation groups. This may prove to be particularly useful in targeting the individuals that would benefit most from commitment devices.

The findings on the importance of design and tailoring commitment devices to the individual suggest value in an experiment to test the benefit of co-creating a commitment device in a programme setting, to identify how it affects both adherence, and also outcomes. Other topics relating to commitment devices that were beyond the scope of this dissertation – for example, on the impact of commitment on wellbeing – remain open to future research also.

Exploratory heterogeneity analysis highlights several findings that merit further research. The Camden trial offered a rare test of the role of GP referral as a commitment element in weight loss programmes. Following the work of Allen et al (2015), who argued that GP referrals to a free NHS weight loss programme created a sense of financial and moral obligation to attend, the Camden study highlighted the potential importance of GPs as a source of external accountability – in the language of this thesis, a source of reputational commitment – for health behaviour change. In contrast



to the above discussion on commitment saturation and overload, it appears that some prior reputational commitment experienced by those referred to the Shape Up programme by a GP interacts positively with the commitment contract to oneself. Receiving either a GP referral or a contract does not lead to significant weight loss difference relative to the sub-group who received neither; but those who received *both* outperformed the rest of the sample. The finding rests on weak statistical association, but is a promising avenue for further research given the findings of a nascent literature on primary care referrals. In the context of potential commitment overload and saturation, the question is: when are additional commitments most useful, when might they be ineffective, and when might they have adverse consequences?

## **6. CLOSING REMARKS**

Advances in the scholarly debate on behavioural biases are increasingly filtering through to policy design, with phenomena such as present bias and the ensuing time inconsistency now well established in the literature both theoretically and empirically. Far from being perceived as abstract peculiarities, it is now taken for granted that these issues cannot be ignored in the design of health interventions.

Behavioural insights and nudge theory has been popularised in recent years, offering a number of ways in which these biases can be taken account of in policy design, and promising new solutions to protracted policy challenges in a diverse range of fields from pensions savings to smoking cessation. Against a menu of interventions including defaults, messaging and priming, commitment devices have re-emerged as practical measures that can play a role in supporting behaviour change (Dolan et al. 2012; Oliver & Ubel 2014).

The empirical findings from this thesis support the more cautious approach advocated by some scholars, that commitment devices offer welcome improvements in health at the margin, but by no means offer a silver bullet for the obesity crisis (Loewenstein et al. 2012; Liu et al. 2014; Oliver & Ubel 2014). Rather, commitment devices can be seen as a complement to ongoing public health interventions, and may hold particular value as a refinement to programmes requiring sustained participation over many weeks in order to deliver the intended health benefits. Precise targeting of individuals facing the internal tussle between short term and longer term payoffs, those who cannot easily overcome the “tyranny of the moment” (Wansink 2013) to stick with a plan for longer term health and wellbeing, will generate the greatest impact of commitment devices; as will careful design of commitment devices to the unique

preferences of the individual, thereby avoiding the negative effects associated with commitment overload.

Commitment devices are not a panacea, but should remain of interest to policymakers and service providers searching for relatively cheap and easily administered improvements to conventional weight loss efforts. Together, the statistical findings and qualitative evidence offer robust support for the planner-doer framework as a lens to view and explain the time inconsistency problem, and as a basis for supporting those who identify the right course of action but have difficulty completing it. The dissertation has shown that reputational commitment devices in particular may be insufficient to generate sizeable weight loss outcomes on average, but certain sub-groups will benefit from improved self-monitoring behaviours and participation in public weight loss programmes. These findings represent a significant and original contribution to the scholarly debate on commitment devices and behavioural public policy.

---

## **APPENDIX**

## **A1: COHEN'S *d* CALCULATIONS USED IN TABLE 2**

Cohen's *d* formula where average outcomes and standard deviations are known for treatment groups:

$$d = \frac{x_t - x_c}{s_{pooled}}$$

Where:

$$s_{pooled} = \sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c}}$$

And:

*x* refers to average outcomes

*n* refers to number of participants

*t* and *c* subscripts refer to treatment and control groups

Cohen's *d* formula where standard deviations are not known but F statistics are known:

$$d = \sqrt{F \left( \frac{n_t + n_c}{n_t n_c} \right) \left( \frac{n_t + n_c}{n_t + n_c - 2} \right)}$$

Where:

*F* refers to the F test statistic

*n* refers to number of participants

*t* and *c* subscripts refer to treatment and control groups

Online effect size calculators have been used to verify author's calculations of effect sizes: <http://www.uccs.edu/~lbecker/>

## **A2. HEALTHY FOUNDATIONS SEGMENTATION MODEL**

In order to operationalize a wide range of motivational constructs applied in health psychology, I employ the Healthy Foundations Segmentation (HFS) model of health motivation (Williams et al. 2011). This segmentation approach relates to the suggestion in the literature on commitment devices that different “consumer types” respond differently (Giné et al. 2010, p.229). The HFS tool was commissioned by the UK Department of Health to create a better understanding of how different people, organised into five motivation groups or segments, respond to health campaigns and services; with the broader aim of designing appropriate and appealing health interventions for a target sub-population. An appealing feature of the HFS tool is that it extends traditional segmentation based solely on demographic characteristics by incorporating attitudinal and psychological factors.

The model was developed through a rigorous process underpinned by theory. Beginning with a general review of the literature, the authors identified 17 different constructs that had been shown to affect health behaviour, and for which reliable measurement indicators were available (such as ‘anticipated regret’, ‘fatalism’, ‘behavioural intentions’, and ‘self-regulation’). Collectively their measurement scales added up to 98 question items, which were condensed to 19 questions in the final survey after various stages of field-testing. The questionnaire and segmentation allocation algorithm categorises respondents to one of five different segments that sit on a spectrum of low to high motivation. These segments are called: Unconfident Fatalists, Hedonistic Immortals, Live for Todays, Health Conscious Realists and Balanced Compensators (see Table 1).

Extensive qualitative testing of the segments was also undertaken through 52 focus groups and 45 immersive interviews. Qualitative analysis highlighted that belonging to one of these segments is not fixed over the life course, with some participants identifying that while they may have been Live for Todays in the past, events in their life changed their attitudes and beliefs and they grew to identify with Balance Compensators (Smith et al. 2011, p.32).

Drilling down into the HFS questionnaire further, eight constructs show a statistically significant association with being overweight or obese (Smith et al. 2011). Some of these appear to offer a sound fit with the planner-doer framework, and may help operationalize the theory. For example, some constructs indicate the individual has an active planner sub-self ('self-efficacy', 'health locus of control', 'intention to lead healthy lifestyle', 'goal-setting'); and others might indicate a dominant doer sub-self ('short-termism', 'health locus of control', and 'risk-taking'). A sense of having control over their own health is high amongst Balanced Compensators and Health Conscious Realists. Short-termist attitudes are particularly strong amongst Live for Today's and Unconfident Fatalists relative to the other segments. Live for Today's are "most likely to be resistant to change and don't acknowledge that their behaviour needs to change" while Unconfident Fatalists "know that their health is bad and they should do something about it, but feel too demotivated to act" (Smith et al. 2011, p.22).

Application of the HFS as a module in the Health Survey for England 2012 demonstrates a clear association between the motivational group and the kinds of health behaviours and outcomes recorded, with the more motivated groups reporting more preventative health behaviours (good diet and exercise) and better health outcomes (lower obesity rates). The Unconfident Fatalists most likely to be obese, have the lowest subjective wellbeing score, were most likely to have a poor diet, and have the lowest levels of exercise. The report concluded they were the "least healthy group and an important focus for behaviour change interventions" (Robinson 2012, p.18).

The HFS makes it feasible to develop a stronger understanding of heterogeneous treatment effects by operationalizing health motivations, which to the best of my knowledge has not yet been done in field experiments on commitment devices. The 19-item survey and allocation model identify the HFS category an individual belongs to, and it is then relatively straightforward to test for heterogeneous effects with commitment devices in a field experiment setting. Although an individual may evolve from one segment to another over time, the HFS questionnaire provides a valid snapshot of a person's health motivations and attitudes at a

given time, and can be applied as a baseline variable to understand heterogeneity of commitment device effects.

Placed on a spectrum of low to high motivation, the Live for Today's and the Unconfident Fatalists are the more negative groups with lower motivation to change their behaviours, and a sense that following a healthy lifestyle will not be easy. They are more fatalistic, believing that their actions are unlikely to have an impact on their own health, which could plausibly be linked to a lower appetite for the self-denial required to rein in consumption (as set out in proposition 2 above). They are also most likely to have short-termist views on their health, which could signal a dominant doer sub-self. These factors reduce the value of  $\tau$  and so individuals in these motivation segments are less likely to benefit from a commitment device: low values of  $\tau$  constrains the size of  $\theta$ , implying lower effectiveness on behaviour change and health outcomes. In contrast, Balanced Compensators and Health Conscious Realists from the higher end of the motivation spectrum are more likely to benefit from commitment devices.





**A3. CONSORT 2010 CHECKLIST  
OF INFORMATION TO INCLUDE WHEN REPORTING A RANDOMISED TRIAL**

<b>Table A.1: CONSORT CHECKLIST: FOOD MONITOR EXPERIMENT</b>			
<i>Section/Topic</i>	<i>Item No</i>	<i>Checklist item</i>	<i>Reported on page No</i>
<b>Title and abstract</b>			
	1a	Identification as a randomised trial in the title	<b>(i), 105, 183</b>
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	<b>183-237 (Ch 5)</b>
<b>Introduction</b>			
Background and objectives	2a	Scientific background and explanation of rationale	<b>115-116</b>
	2b	Specific objectives or hypotheses	<b>103, 184</b>
<b>Methods</b>			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	<b>121-131</b>
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	<b>n/a</b>
Participants	4a	Eligibility criteria for participants	<b>126</b>
	4b	Settings and locations where the data were collected	<b>121-22, 126-27</b>
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and	<b>122-23</b>

		when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	<b>131-34</b>
	6b	Any changes to trial outcomes after the trial commenced, with reasons	<i>n/a</i>
Sample size	7a	How sample size was determined	<b>124-25, 392-94</b>
	7b	When applicable, explanation of any interim analyses and stopping guidelines	<b>125, 188</b>
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	<b>126</b>
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	<b>129</b>
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	<b>129</b>
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	<b>126, 129</b>
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	<b>129</b>
	11b	If relevant, description of the similarity of interventions	<b>123</b>
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	<b>131-34</b>
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	<b>133-34</b>
<b>Results</b>			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	<b>189</b>
	13b	For each group, losses and exclusions after randomisation, together with reasons	<b>207-8</b>
Recruitment	14a	Dates defining the periods of recruitment and follow-up	<b>188-89</b>

	14b	Why the trial ended or was stopped	<b>188</b>
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	<b>194</b>
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	<b>194</b>
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	<b>212, 215-16, 221-22</b>
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	<b>n/a</b>
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	<b>225-28</b>
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	<b>n/a</b>
<b>Discussion</b>			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	<b>154-96, 233</b>
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	<b>174-79, 234</b>
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	<b>217-30</b>
<b>Other information</b>			
Registration	23	Registration number and name of trial registry	<b>119</b>
Protocol	24	Where the full trial protocol can be accessed, if available	<b>119</b>
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	<b>119</b>

**Table A.2: CONSORT CHECKLIST: CAMDEN EXPERIMENT**

<i>Section/Topic</i>	<i>Item No</i>	<i>Checklist item</i>	<i>Reported on page No</i>
<b>Title and abstract</b>			
	1a	Identification as a randomised trial in the title	<b>(i), 105, 239</b>
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	<b>239-95 (Ch 6)</b>
<b>Introduction</b>			
Background and objectives	2a	Scientific background and explanation of rationale	<b>115-16</b>
	2b	Specific objectives or hypotheses	<b>103, 240</b>
<b>Methods</b>			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	<b>137</b>
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	<b>n/a</b>
Participants	4a	Eligibility criteria for participants	<b>135</b>
	4b	Settings and locations where the data were collected	<b>433</b>
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	<b>136-37</b>
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	<b>142-46</b>
	6b	Any changes to trial outcomes after the trial commenced, with reasons	<b>n/a</b>

Sample size	7a	How sample size was determined	<b>137-38, 395-96</b>
	7b	When applicable, explanation of any interim analyses and stopping guidelines	<b>243</b>
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	<b>141-42, 246, 397</b>
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	<b>141, 246</b>
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	<b>141-42, 246</b>
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	<b>141-42, 246</b>
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	<b>141-42, 246</b>
	11b	If relevant, description of the similarity of interventions	<b>n/a</b>
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	<b>143-45, 267</b>
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	<b>144-46, 281</b>
<b>Results</b>			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	<b>243, 245</b>
	13b	For each group, losses and exclusions after randomisation, together with reasons	<b>243-44</b>
Recruitment	14a	Dates defining the periods of recruitment and follow-up	<b>244</b>

	14b	Why the trial ended or was stopped	<b>244</b>
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	<b>248</b>
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	<b>271, 277</b>
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	<b>267, 271, 276-77</b>
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	<b>277</b>
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	<b>280-87</b>
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	<b>n/a</b>
<b>Discussion</b>			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	<b>154-69, 288</b>
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	<b>174-79, 359-61</b>
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	<b>272-90</b>
<b>Other information</b>			
Registration	23	Registration number and name of trial registry	<b>119</b>
Protocol	24	Where the full trial protocol can be accessed, if available	<b>119</b>
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	<b>119</b>

#### **A4. INTERVIEW TOPIC LIST FOR CAMDEN PARTICIPANTS**

Purpose: to explore weight loss experiences and reflections on the programme; identify specific behaviours that changed, if any; examine how the commitment contract was applied and how useful the individual found it; explore the use of personal commitment strategies; and generate insights into whether the behaviour of participants fits the planner-doer theoretical framework.

##### *The weight loss programme*

Can you recall your final weigh-in record?

Prompt if not able to. Did you meet the 5% target weight loss?

How did you feel about the result?

Were there any high or low points over the Shape Up programme that stand out in particular?

##### *Behaviour change*

What prompted you to join the group at that time?

Did any specific behaviours change around food or exercise?

##### *Wellbeing*

How would you describe your overall life satisfaction on a scale of 0 to 10?

Do you recall your initial score?

Why do you think there has been a change/no change?

##### *Commitment Contract*

(If treated) Can you recall the commitment contract you were offered? What did you do with it when you took it away?

Was there anything you did to try and lock yourself in to staying on track with your goals?

## **A5. SAMPLE SIZE CALCULATIONS: FOOD MONITOR**

### **Food Monitor trial baseline estimate**

The outcome variable of interest is self-reported weight loss in lbs over a period of 4 weeks, which is the period that the treatment corresponds to. For now I am assuming that participants aim to lose weight. Medical advice states individuals can safely lose 1-2lb per week. The removal of the financial commitment (refund group) is expected to reduce weight loss, and the reputational commitment (coach group) is expected to raise weight loss.

The sample size calculation requires assumptions for the mean outcome for the control group and for the treatment group, and standard deviations for both groups. I assume the control group will lose 2lb over 4 weeks with a standard deviation of 5. I assume for simplicity the treatment groups will experience equivalent magnitudes of change but in different directions. The refund treatment group will lose 0lb weight and the coach treatment group will lose 4lb. Both treatment groups are assumed to have a standard deviation of 3. For a woman weighing the UK mean of 170lb, under these assumptions the coach treatment is expected to generate 1.1% weight loss, which is arguably a conservative assumption.

I am assuming that both treatments are being compared to the financial commitment group and not to each other. This is a conservative assumption, as the differences in mean weights are likely to be smaller in relation to the control group (as hypothesised), and therefore the required sample size is expected to be higher.

#### Parameters

$Mean_0 = 2$  (i.e. the comparison group sheds 0.5 lb per week over 4 weeks)

$Mean_1 = 0$  (i.e. the refund group loses zero weight over 4 weeks)

$Mean_2 = 4$  (i.e. the coach group shed 1lb per week over 4 weeks)

$SD_0 = 5$  (i.e. the standard deviation for the control group is 5 lb, with 2/3 of all control group members being somewhere in the range -0.5 to 4.5 lb, allowing for some to gain weight or stay where they were)

$SD_1 = 3$  (i.e. 2/3 of those in the refund group lie in the range -1.5 to +1.5 lb weight loss)



$SD_2 = 3$  (i.e. 2/3 of those in the coach group lie in the range 2.5 to 5.5 lb weight loss)

#### Stata output

For treatment group 1:

```
. sampsi 2 0, sd1(5) sd2(5) power(0.9)
```

Estimated sample size for two-sample comparison of means

Test Ho:  $m_1 = m_2$ , where  $m_1$  is the mean in population 1 and  $m_2$  is the mean in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.9000
m1 = 2
m2 = 0
sd1 = 5
sd2 = 5
n2/n1 = 1.00
```

Estimated required sample sizes:

```
n1 = 132
n2 = 132
```

For treatment group 2:

```
. sampsi 2 4, sd1(5) sd2(5) power(0.9)
```

Estimated sample size for two-sample comparison of means

Test Ho:  $m_1 = m_2$ , where  $m_1$  is the mean in population 1  
and  $m_2$  is the mean in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.9000
m1 = 2
m2 = 4
sd1 = 5
sd2 = 5
n2/n1 = 1.00
```

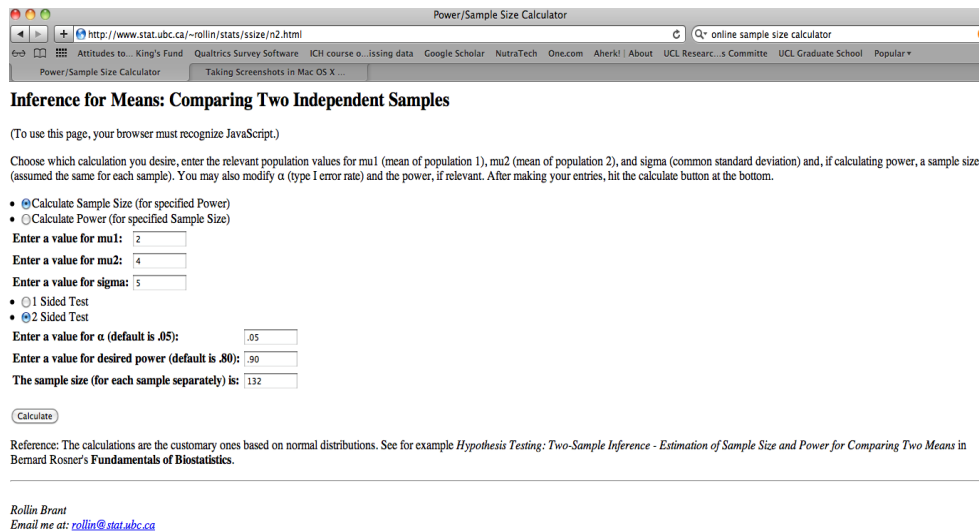
Estimated required sample sizes:

```
n1 = 132
n2 = 132
```

These results demonstrate the equivalence of the sample size estimations for both treatments, and sensitivity analysis therefore refers to the treatments together.

## Sensitivity tests

Further calculations varied the influence of treatment (lower and higher), and the variability in weight loss outcomes amongst the treatment groups. The table in chapter 5 provides a prescriptive sample size per group, and with three experimental groups yields a total target sample of 364.



Power/Sample Size Calculator

http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

online sample size calculator

### Inference for Means: Comparing Two Independent Samples

(To use this page, your browser must recognize JavaScript.)

Choose which calculation you desire, enter the relevant population values for  $\mu_1$  (mean of population 1),  $\mu_2$  (mean of population 2), and  $\sigma$  (common standard deviation) and, if calculating power, a sample size (assumed the same for each sample). You may also modify  $\alpha$  (type I error rate) and the power, if relevant. After making your entries, hit the calculate button at the bottom.

- Calculate Sample Size (for specified Power)
- Calculate Power (for specified Sample Size)

Enter a value for  $\mu_1$ :

Enter a value for  $\mu_2$ :

Enter a value for  $\sigma$ :

- 1 Sided Test
- 2 Sided Test

Enter a value for  $\alpha$  (default is .05):

Enter a value for desired power (default is .80):

The sample size (for each sample separately) is:

Reference: The calculations are the customary ones based on normal distributions. See for example *Hypothesis Testing: Two-Sample Inference - Estimation of Sample Size and Power for Comparing Two Means* in Bernard Rosner's *Fundamentals of Biostatistics*.

Rollin Brant  
Email me at: [rollin@stat.ubc.ca](mailto:rollin@stat.ubc.ca)

The analysis demonstrates the very high sensitivity to small changes in the underlying assumptions. This, coupled with a lack of data on what actual weight loss is with the Nutracek product, suggests that the sample size calculations should be interpreted with care. It is difficult to know in advance whether the participants will be mainly of a healthy BMI looking to maintain their weight, or high BMI looking to lose weight intensively over the summer. These estimates, while necessary and important, should therefore be used as a guide. In the range of estimates generated, it is posited that the baseline estimate is the most reasonable, requiring 132 participants per group and 396 in total. However, with the externally imposed cap on financial treatment assignment of  $n=100$ , this will mean recruitment will continue until  $n=364$ . Even with  $n=100$ , treatment group 1 could generate statistically significant results under scenario 3, where standard deviation around weight loss is lower. It therefore remains a viable treatment in the experiment.

## A6. SAMPLE SIZE CALCULATIONS: CAMDEN

### Camden trial baseline estimate

As before, the outcome variable of interest is weight loss measured in lbs. In this trial, outcomes are measured at the end of the 11-week Shape Up programme. I assume a relatively conservative baseline weight loss for the control group of 0.5 lbs per week, leading to 5.5 lbs net weight loss over the 11-week Shape Up programme. I further assume the treatment group outperforms the control group by a conservative 1lb at the final weigh-in. With these mean outcomes and assumed standard deviation of 1lb in both groups, calculations imply that each group should contain 85 participants, yielding a total sample size of 170.

#### Parameters

Mean<sub>0</sub> = 5.5 (i.e. the comparison group sheds 0.5 lb per week over 11 weeks)

Mean<sub>1</sub> = 6.5 (i.e. the contract group loses an additional 1lb over 11 weeks)

SD<sub>1</sub> = SD<sub>2</sub> = 2 (i.e. the standard deviation for both groups is 2 lb, with 2/3 of all control group members being somewhere in the range 1 to 3 lb)

#### Stata output

```
. sampsi 5.5 6.5, sd1(2) sd2(2)
```

Estimated sample size for two-sample comparison of means

Test Ho:        m1 = m2, where m1 is the mean in population 1  
                  m2 is the mean in population 2

Assumptions:

```
alpha = 0.0500 (two-sided)
power = 0.9000
  m1 = 5.5
  m2 = 6.5
  sd1 = 2
  sd2 = 2
n2/n1 = 1.00
```

Estimated required sample sizes:

```
n1 = 85
n2 = 85
```

### **Sensitivity tests**

A range of alternative sample size estimates were created by varying the parameters in two ways: firstly, the treatment group were allowed to register different weight loss differentials relative to the control group, with moderate and high values of 1.5 lbs and 2 lbs greater than the control group's 5.5 lbs net weight loss. Secondly, higher variability in weight loss outcomes was built in to the calculations by setting standard deviation in both treatment and control groups at 3lbs around their respective mean values. For contrast, a low standard deviation of 1lb was also tested. The full spectrum of results is set out in chapter 6. As with the earlier trial, the sensitivity analysis shows a wide range of potential sample sizes, and with uncertainty around the precise values of weight loss and standard deviation, the exercise has to be treated with a degree of caution. Nevertheless, these estimates provide a basis for planning the experiment around a target sample of 170, based on two equally sized experimental groups of 85 participants.

## **A7. SAMPLE DO-FILE OF ADVANCE RANDOMISATION EXERCISE**

Stata do-file dated January 2016

\*\*\* Randomisation code Jan 19 2016 \*\*\*

\* Dataset with n=43 ID codes beginning 40001 to mark ids imported from Excel \*

\* Provides random assignment for 43 clients on register for groups 25-27, all beginning in January \*

\* Exercise logged \*

sort id

set seed 19012016

gen random = uniform()

sort random, stable

gen treat = 0

replace treat = 1 if \_n <= \_N/2

summarize treat

sort treat

by treat: list id

## **A8. INFORMED CONSENT FIELDWORK DOCUMENTS**

The Information Sheet was made available in hard copy in the Camden experiment. The form below was used in wave 3 of the fieldwork hence refers to the time period Sept 2015 – March 2016. Previous waves were identical except for the time period. Similar text was used at the outset of the online survey in the Food Monitor experiment with appropriate references to the Food Monitor service.

### **Information Sheet for Participation in Research Studies**

**We would like to invite you to participate in this research project.**

Title of Project: **Commitment to weight loss**

This study has been approved by the UCL Research Ethics Committee (Project ID Number): **4518/003**

It is part of a PhD research project based at UCL's School of Public Policy.

Name                      Manu Savani and Professor Peter John

Work Address        29/31 Tavistock Square, London, WC1H 9QU

Contact Details    [m.savani.12@ucl.ac.uk](mailto:m.savani.12@ucl.ac.uk) (PhD Candidate)

+44 7775 835 448

[peter.john@ucl.ac.uk](mailto:peter.john@ucl.ac.uk) (Principal Researcher)

**You will be given a copy of this information sheet.**

#### **Details of Study:**

This study looks at what strategies help people achieve their personal goals. In collaboration with Camden Active Health, we hope to understand which strategies might be particularly effective in supporting people to reach and maintain their preferred weight. The potential benefits of this project include helping people to successfully change their behaviour, improve their health, and boost overall wellbeing.

We are recruiting participants who are fully registered on the Shape Up programmes from September 2015 to March 2016.

Taking part in this project will involve filling in a short survey. You may then be offered a new strategy to support your weight loss journey. Beyond that, please just use the service as you normally would. At the end of the programme you might be asked for your views and experiences of the programme, and you can choose at the time if you would like to share them.

Your weekly weigh-ins will be used to inform the research project. It would also be helpful to understand how the weight loss journals are used. Any information you share will not be identifiable back to you.

To ensure we have a full picture of your progress, we may contact you if you miss two consecutive group meetings. We ask that you provide contact details for your preferred mode of contact for this reason. These will not be passed on to any third parties, and will be destroyed at the end of the research project. We will contact you up to two times with no response, and we will then assume you have left the programme.

All data will be handled with the strictest confidentiality, and for research purposes only. Your data will not be passed on to any third parties. At the end of the research project, we will disseminate results with Camden Active Health. Any results that are published will maintain fully the anonymity of study participants.

It would be very helpful for the smooth running of the project if your participation in this research project is not discussed in the group setting, as it might interfere with data quality and confidentiality.

There are no major risks identified in this project, but please do not hesitate to contact us if you want to discuss any issues. You should also feel free to discuss with family, friends or your doctor at any time.

If you decide to take part you will be given this information sheet to keep, and will be asked to sign a consent form.

It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. If you do decide to take part you are still free to withdraw at any time and without giving a reason.

Feel free to contact us if you have any specific questions.

**All data will be collected and stored in accordance with the Data Protection Act 1998.**

Thank you for your time,

Manu Savani

## **Informed Consent Form for Participating in Research Studies**

**Please complete this form after you have read the Information Sheet.**

Title of Project: **Commitment to weight loss**

This study has been approved by the UCL Research Ethics Committee (Project ID Number): **4518/003**

Thank you for your interest in taking part in this research. Before you agree to take part, the person organising the research must explain the project to you. The information sheet aimed to do this.

If you have any questions arising from the Information Sheet, please contact Manu Savani: [m.savani.12@ucl.ac.uk](mailto:m.savani.12@ucl.ac.uk). If at any time you would like to contact a second researcher, please get in touch with the project's Principal Researcher Professor Peter John at UCL's Department of Political Science: [peter.john@ucl.ac.uk](mailto:peter.john@ucl.ac.uk).

You will be given a copy of this Consent Form to keep and refer to at any time.

### **Participant's Statement**

I

- have read the notes written above and the Information Sheet, and understand what the study involves.
- understand that if I decide at any time that I no longer wish to take part in this project, I can notify the researchers involved and withdraw immediately.
- consent to the processing of my personal information for the purposes of this research study.
- understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.
- agree that the research project named above has been explained to me to my satisfaction and I agree to take part in this study.
- I understand that the information I have submitted will be published as a report. Confidentiality and anonymity will be maintained and it will not be possible to identify me from any publications.
- I agree that my non-personal research data may be used by others for future research. I am assured that the confidentiality of my personal data will be upheld through the removal of identifiers.
- I understand that the researchers may contact me if I miss two consecutive group meetings, and am providing details for my preferred mode of contact here:  
\_\_\_\_\_.

Signed:

Date:



## **A9. BASELINE SURVEY<sup>86</sup>**

**UCL Research Project 4518/003  
Survey ID:**

**Thanks for agreeing to take part!**

**Kindly fill in this survey as part of the registration process.**

*It should take about 4-5 minutes to complete.*

*The first section is about your weight loss plans, and the second is about your attitudes.*

*We would like to get your fullest response, but you can skip questions if you would rather not answer.*

*Your data will be stored securely and anonymously.*

*Feel free to ask if anything is unclear.*

---

<sup>86</sup> This survey was provided in hard copy for Camden participants. Very similar content was used in the online Food Monitor baseline survey, but with 2 key differences: the online survey asked additional demographic questions (income, educational background, number of children at home and job status), and included a behavioural economics question to elicit time preference (cost of waiting).

## Section 1: Your weight loss regime

1. Have you taken part in a weight loss programme before?

- Yes
- No

If 'yes', could you give examples:

2. Would you say you have experienced any big changes at home or work recently?

- Yes
- No

If you want to say a bit more about this, please use the space below to describe it in your own words:

3. Overall, how satisfied are you with your life nowadays, on a scale of 0 to 10?

Here, 0 means very dissatisfied, and 10 means very satisfied.

4. Other than your participation in Shape Up, is there anything else you do to work towards your target weight?

- Yes
- No

If 'yes', could you give examples:

Section 2: Your attitudes and perceptions

This section asks about your attitudes and perceptions about your health. These questions are taken from a Department of Health survey.

5. Here are some statements that other people have made. Please tick one circle to show how much you agree or disagree with each of them:

	Disagree strongly	Disagree	Disagree slightly	Neither agree nor disagree	Agree slightly	Agree	Agree strongly
<i>I feel good about myself</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I get a lot of pleasure from taking risks</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I generally focus on the here and now rather than worry about the future</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I learn from my mistakes</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

6. Here are some things that other people have said they would like to have over the course of their lives. Could you tell me how important each of them is to you personally:

	Not at all important						Very important
<i>To have money, wealth and possessions</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>To have an image that others find appealing</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

7. Here are some more statements that other people have made. Please choose one to show how much you agree or disagree with each of them:

	Disagree strongly	Disagree	Disagree slightly	Neither agree nor disagree	Agree slightly	Agree	Agree strongly
<i>Following a healthy lifestyle is an effective way to reduce my chances of becoming ill</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>If you don't have your health you don't have anything</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>There is nothing more important than good health</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I'm very involved in my health</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I am in control of my own health</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>The main thing which affects my health is what I personally do</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>If a person is meant to get ill, it doesn't matter what a doctor tells them to do, they will get ill anyway</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<i>I intend to lead a healthy lifestyle over the next 12 months</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8. For you, would leading a healthy lifestyle be...

	Very difficult						Very easy
<i>Please tick one circle</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

9. How much control do you believe you have over whether or not you lead a healthy lifestyle over the following year?						
	No control					Complete control
<i>Please tick one circle</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. For you, would leading a healthy lifestyle be...						
	Not enjoyable					Enjoyable
<i>Please tick one circle</i>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11. Which one of these best describes your view:						
If I don't lead a healthy lifestyle, my health could be at risk:						
<input type="radio"/> In the next 12 months <input type="radio"/> In the next few years <input type="radio"/> In the next 10-20 years <input type="radio"/> Much later in my life <input type="radio"/> Not at all						
12. Compared with other people of your age, how likely do you think it is that you will get seriously ill at some point over the next few years?						
<input type="radio"/> I am much MORE likely to get seriously ill than other people of my age <input type="radio"/> I am a little more likely <input type="radio"/> No more or less likely <input type="radio"/> I am a little less likely <input type="radio"/> I am much LESS likely to get seriously ill than other people of my age <input type="radio"/> I already have a serious illness						

**Thank you, that's the end of the survey!**  
**Please give your form back to the researcher to complete your registration.**

## A10. DATA GATHERED THROUGH FOOD MONITOR SYSTEMS

<b>Table A.3: Data reports from Food Monitor</b>		
<i>Category</i>	<i>Variable</i>	<i>Frequency</i>
Weight	Self-reported weight entry	As new entry is input, ordered by date
	Initial target weight	Closest available prior to registration date
	Target weight during project	As new entry is input, ordered by date
	Calorie benchmark linked to initial target weight	Closest available prior to registration date
	Food and exercise diary input	Summarised by day
Account Usage	Number of log-ins per day	Summarised by day
Account Type	Cost of monthly subscription	Once, at time of registration
	Start date of subscription	Once, at time of registration
Verifications	Height in client profile	Once, at time of registration
	Gender in profile	Once, at time of registration

**A11. REPUTATIONAL TREATMENT TAKE-UP DECISION  
(ONE-SIDED NON-COMPLIANCE)**

**Table A.4**

<b>Table .7 What determines whether a coach is nominated?</b>	
Starting weight (kg)	-0.004 (0.812)
Overweight	0.395 (0.374)
Obese	-0.031 (0.957)
Severely obese	-0.347 (0.737)
Weight loss target at 4 weeks (% of initial weight)	-0.031 (0.731)
Exercise sessions per week	-0.047 (0.273)
Fruit and vegetable intake	-0.035 (0.589)
Experienced major life changes recently	0.476 (0.120)
Other activities pursued to lose weight	0.757 (0.090)
Female participant	0.564 (0.161)
Short term health attitudes	0.656* (0.032)
Cost of waiting 1 month for £10 payoff	0.006 (0.601)
Number of children at home	-0.200 (0.175)
Phase 1 randomisation	1.636* (0.029)
August	0.299 (0.463)
September	0.568 (0.300)
October	1.565 (0.079)
November	0.810 (0.358)
Aged over 40	-0.241 (0.425)
Low income	-0.306 (0.516)
Observations	109
PseudoR <sup>2</sup>	0.185

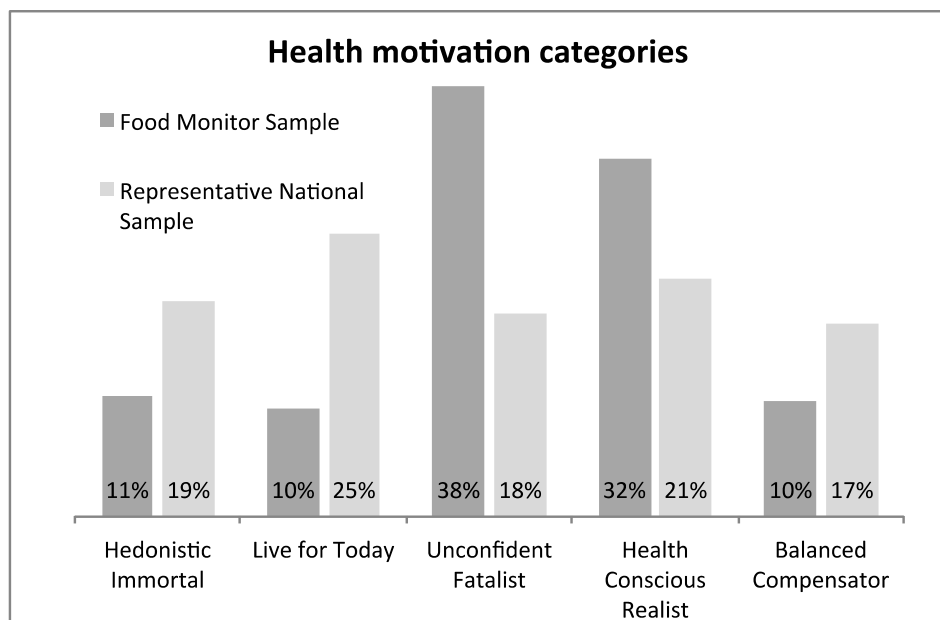
*Notes: Probit estimates, p-values in parentheses\* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001*

## A12. FOOD MONITOR BASELINE VARIABLES

### Health motivation

Figure A.1 below highlights two differences in the distribution of motivation segments between the Food Monitor sample and a nationally representative sample (Williams, 2008). Trial participants are somewhat polarised between a low-motivation group called the Unconfident Fatalists, and a high-motivation group called the Health Conscious Realists. Unconfident Fatalists are most likely to be obese (Robinson 2012), and this is borne out in the Food Monitor baseline data, where the mean BMI is highest for this group at 33, compared to 26 for Balanced Compensators. The concentration of participants in the Unconfident Fatalist group indicates that the Food Monitor service is particularly appealing to those who are obese and overweight, and who are seeking external support. The over-representation of Health Conscious Realists might be explained by the fact that subscription to the Food Monitor service represents a financial commitment. Individuals in this segment are likely to be more conscious of the need to pursue good health behaviours, and the act of signing up to Food Monitor is about making a positive investment for their health.

Figure A.1





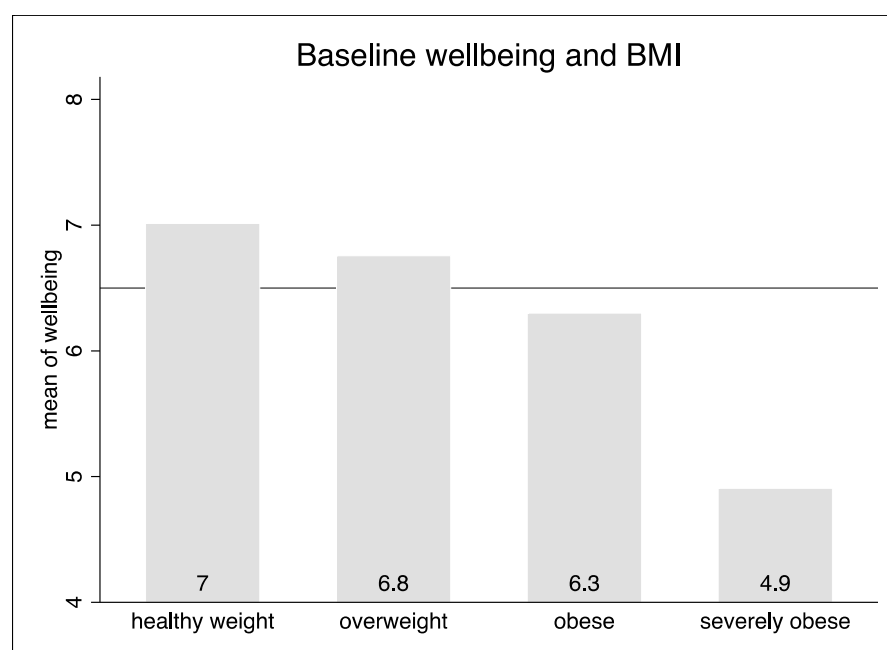
### Weight loss goals

The size of weight loss targets increased with BMI as would be expected, but as a percentage of initial body weight there is little difference across BMI groups with the exception of the severely obese whose weight loss targets are a little lower. This may reflect prior experience and realism about the difficulty in losing weight, particularly for those who are severely obese and may experience greater challenges to making lifestyle changes.

### Subjective wellbeing

Subjective wellbeing was reported on a scale of 0 to 10 with a mean value of 6.5. There were 13 missing observations (3.5%) to this question. Perhaps unsurprisingly, there is a negative association between baseline wellbeing and BMI, implying that those who were closer to their healthy weight at the start of the trial were more satisfied with life as a whole.

Figure A.2



### Diet and exercise

Dietary quality is measured by the number of fruit and vegetable portions consumed the previous day, which is a popular and recognisable rule and offers a useful comparison against wider samples such as those in the Health Survey for England. While most people are aware that the recommended daily intake of fruit and vegetables is five per day, the

majority of the sample (63%) currently consume less than this benchmark. The average intake is four portions a day, with a small minority of individuals reporting more than 10 per day. This is broadly comparable with nationwide data that reports adults consumed on average 3.6 portions of fruit and vegetables in 2013 (HSE 2013).

In order to find out how active respondents were, the baseline survey asked how many exercise sessions they had undertaken during the previous week. The question was deliberately worded openly, with text boxes to provide more detail on the kind of exercise and activity that individuals counted towards their weight loss goals. The Food Monitor tool allows for a large range of activities – from walking the dog and cleaning the house to mountain biking and swimming – to be counted towards the net daily calorie goal, as part of self-monitoring daily consumption. Where no specific number of exercise sessions was given, the text was read and coded in numerical format. For example, one respondent stated they “went swimming five times last week and went to the gym once”, and this was coded as six exercise sessions.

The majority of respondents (75%) reported undertaking some exercise in the previous week. Exercise varied from low-intensity activities such as walking to more high-intensity activities such as boxing and indoor group cycling. Among the 25% who reported zero exercise sessions, some ill health as preventing exercise the previous week: “none because I have a had a nasty cough and cold”; while others pointed to more serious and longer-term health issues: “currently suffering from prolapsed disc and associated nerve pain making exercise impossible”. Another respondent reported, “I suffer with social anxiety and agoraphobia”. The amount of exercise might be linked to motivation, but the qualitative responses here warn that zero exercise may not necessarily indicate poor motivation, rather they may point to the existence of more challenging physical issues that prevent rapid or possibly any weight loss in the short run. The data serves as a useful grounding in the realities of weight loss efforts, which might face the added challenge of difficult personal circumstances.

### Other activities to lose weight

The baseline survey asked whether they took part in any ‘other activities’ in order to pursue their weight loss goals. The majority of participants said yes (87%), citing exercise regimens such as walking and gym classes, other self-monitoring tools such as pedometers and running apps, and dietary regimens such as Slimming World and the 5:2 diet. For these individuals, the Food Monitor app was an additional tool employed alongside a number of other lifestyle changes in order to achieve their target weight. Some individuals mentioned more involved activities such as hypnotherapy and personal training, indicating they are willing and able to pay for additional services as an investment in their weight loss goals. These people are relatively well off and able to afford weight loss aids alongside their Food Monitor subscription (see income statistics below). The smaller group of people who said they did nothing else might simply be more honest, or they may be less motivated on their weight loss goal.

The overweight and obese are most likely to be pursuing other activities (90% and 87% respectively), while those with a normal weight are less likely (85%) and those who are severely obese are least likely (76%). It is plausible that for those in this latter category the number of options available for further activities may be more limited, and it may also be that motivation to undertake these activities is considerably lower because of perceived and actual physical and mental challenges involved. Examined through the lens of health motivation, it is not surprising to note that Live for Today's and Unconfident Fatalists – the two groups that are most short-termist in their outlook – were less likely to report ‘other activities’ (77% and 80% respectively).

### Life changes

One-third of participants reported experiencing a major life change recently, such as a change in their own health status (“I have been told I am ‘pre-diabetic’ and MUST make lifestyle changes”); a change in the health status of a family member (“My husband had a heart attack. He had surgery and had a stent fitted. Although this happened to him and not me, it has made me think about our lifestyles, the way we eat and exercise.”); or a change in job (“I have moved from a site-based role to an office based role. Less walking around is now involved”). It is notable that only 19% of

Balanced Compensators cited such change, suggesting that these individuals were more likely to take up commitment devices such as Food Monitor without any external push to do so. In contrast, the HFS categories who are likely to have a short-termist health outlook are much more likely to report a major life change (41% of Hedonistic Immortals, 40% of Live for Today's), perhaps indicating that only in response to large external impetus would they make a commitment to investing in their health.

#### Time preference: construction of variable to operationalise short-termism

Time preference, or the description of how individuals favour current payoffs versus future payoffs, is a central concept in the dual-self planner-doer model: it is the key characteristic that differentiates the planner from the doer, giving rise to the internal, intertemporal, tussles that explain why the best intentions to change health behaviours often fail. The literature review discussed empirical work that tried to operationalise time preferences through proxy measures such as a discount rate, to establish the link between high discount rates (a strong preference for current payoffs) and specific health behaviours.

Earlier chapters considered the role of time preference as a trait that may explain the extent to which commitment devices are effective across individuals, arguing that where the preference for a current payoff against a future payoff is greater, this indicates a bias in favour of present over future gains. This trait is associated with the doer sub-self, as set out in Chapter 3. Conversely, a milder preference for current payoffs at the expense of future payoffs is more consistent with the planner's outlook and utility function. Being able to identify the degree of patience around delayed payoffs would allow for testing the hypothesis that those with a stronger present bias are less likely to benefit from a commitment device.

The theoretical underpinnings for this thesis, unlike those of hyperbolic discounting models, do not require the discount rate parameter itself be calculated, and this means that the research design was free to find other workable measures for time preference. Instead of the time-consuming and laborious process of uncovering precise discount rates using a series of choices, the baseline survey instead employs a single

question designed to measure the cost of waiting for a modest cash sum (£10) for an additional 1 month and an additional 6 months relative to receiving that cash today. The additional amount required to delay receiving the cash sum is interpreted as the individual's cost of waiting. The spectrum of values generated is a proxy for patience: the higher the amount entered implying a higher degree of impatience.

The formulation of the question in this way aimed to balance the need for identifying an operational measure of time preference, with the need to prevent respondent fatigue and dropout, and ensure as high a number of completed surveys as possible to meet the sample size target and provide a robust basis for analysis. With only 13 missing values (3.5%) for the time preference question, this aim is judged to have been successfully met.

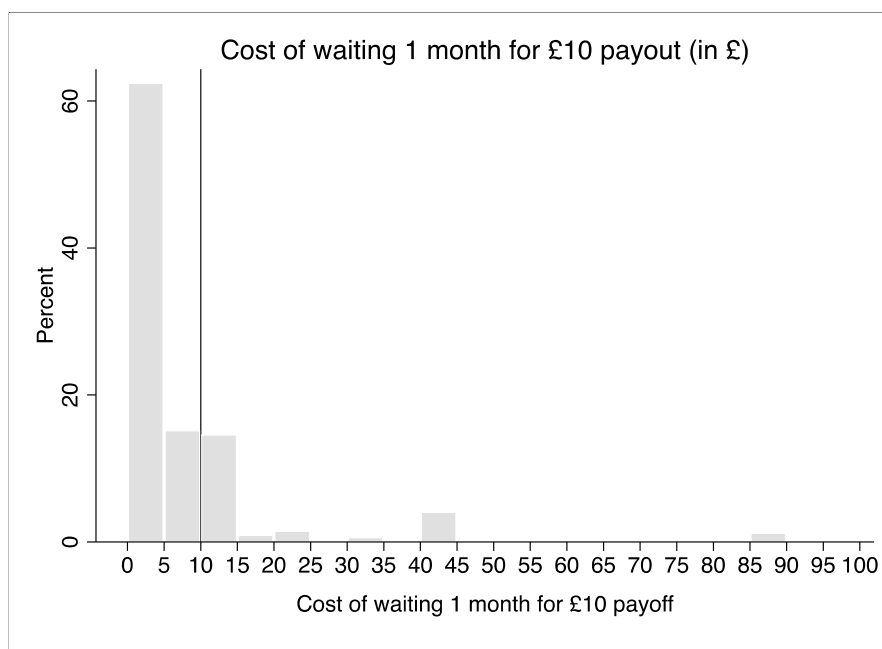
The question deliberately asks 'how much would you need to delay receiving the £10 today?', rather than 'how much *extra* would you need...', in order to avoid encouraging participants to ask for compensation when they may not otherwise demand it. In aiming for both brevity and to avoid leading the respondent, however, the survey question may have given rise to two different interpretations. The respondent was asked to enter the amount they would have to receive in order to accept the delay. This was asking for the total amount, not the additional amount to be added to the £10 original offer. Those respondents who entered 'zero' may imply one of two things. First, they may have been referring to the additional compensation for waiting an extra month; implying that their total payoff remains £10. Or, the more literal interpretation, is that they now are indifferent to the total payoff falling to £0. The former response indicates a high degree of patience and zero cost to waiting for the original £10 offered. The latter response indicates that the individual is giving the money back, rejecting the offer, and has a very highly negative discount rate. The latter seems implausible, and so in order to ensure that the 'zero' respondents are not dropped from the analysis, any sum entered that is below 10 has a further 10 added. This is another way of stating that the participant's original response referred to the additional sum required to delay, with the transformed variable now presenting the total payoff.

The transformed variable, 'compensation at 1 month' refers to the total payoff that participants would accept. To better understand the cost of waiting, which is the more relevant construct, another transformation to the variable extracts the original £10 offer that all participants hypothetically receive. Removing this original sum produces a variable, 'cost of waiting', that measures the excess cash required to make the delay acceptable, in other words the cost of waiting for the payout. The cost of waiting ranges from £0 to £90, with an average payoff value of £5.60 (sd 12.5). At 6 months, the required total payoff increases to an average £37.56 (sd 120), with values ranging from £0 to £990. There are two notable features of this data. Firstly, we would expect that the cost of waiting increases with the duration of the delay. In simple terms, if the average cost of waiting 1 month is £5.60, then scaled up to 6 months in a linear fashion would imply that the cost of waiting 6 months is £33.60, which is very close to the reported average over the longer run horizon. This implies the degree of patience, on average, remains roughly the same over both time horizons.

Secondly, the spectrum of costs highlights a wide range of time preferences across the sample, reflected in the large standard deviation around the mean values. Many individuals are at the patient end of the spectrum not requiring very much, if any, additional money to compensate them for the delay; while a small number of individuals at the impatient end demand large sums (such as £990 for a 6 month delay on £10).

These are useful insights, but there remains ambiguity around this variable, as those participants who responded '10' may have referred either to the total payoff (i.e. cost of waiting is 0) or to the excess required (i.e. cost of waiting is 10). As a robustness check, it is possible to derive a second variable to measure impatience. The continuous variable 'cost of waiting' is used to derive a categorical variable that identifies the subset of most impatient participants. The distribution of costs of waiting for an additional month are graphed in figure 6. A large majority (92%) of participants are patient, with a cost of waiting at or less than £10. The remaining 8% (n=28) reported a cost of waiting higher than £10, and this group is identified as 'impatient' (falling to the right of the reference line in figure 6). Among the impatient, the cost of waiting is on average £40, while among the patient the average is £2.60.

**Figure A.3**



This binary ‘impatient’ variable has operational clarity: regardless of how exactly the question was interpreted, these respondents remain at the high end of the spectrum of costs of waiting; and being a binary variable the precise value of the survey response is no longer needed. This removes the threat of downward bias from measurement error on the cost of waiting variable, and allows the model to analyse whether impatience as a trait is associated with the effectiveness of the commitment device on weight loss. However, the variable captures only a small set of people at the very top end of the spectrum, so is better characterised as identifying the highly impatient. Analysis in Chapter 5 therefore relies on both variables, with the continuous variable used in the results presented on average and heterogeneous treatment effects. Robustness checks later in the Appendix show that using the binary variable makes no difference to the emerging story on present bias.

<b>Table A.5: Time preference</b>			
<i>n=351, 97% of sample</i>	<i>Mean</i>	<i>Range</i>	<i>SD</i>
Cost of waiting 1 month (£)	5.6	0 – 90	12.5
Cost of waiting 6 months (£)	37.6	0 – 990	120.3
Impatient	0.08		
Cost of waiting 1 month among ‘impatient’	40.2	15 – 90	22.8
Cost of waiting 1 month among ‘patient’	2.6	0 – 10	3.7

## Age, Income, Education and Employment

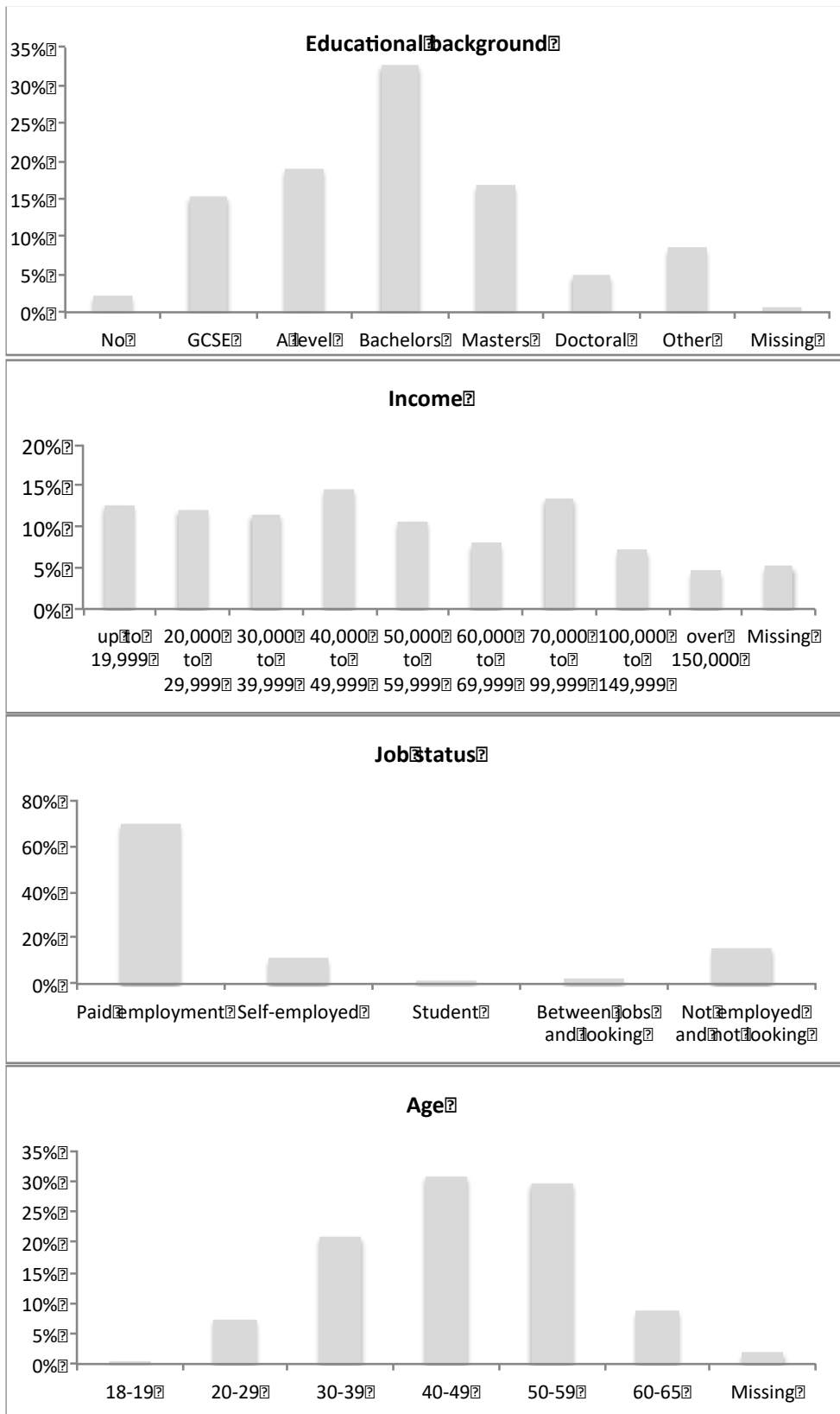
Age is captured in the online survey using categories from 18-19 years to 65. For the purposes of regression analysis, a binary variable is used to capture whether the individual is below 40 years, or equal to or greater than 40 years old.

<b>Table A.6: Age</b>		
<i>Age (n=357, 98% of sample)</i>	%	<i>N</i>
Under 30 (%)	8	29
30 – 39 (%)	21	76
40 – 49 (%)	31	112
50 – 59 (%)	30	108
60 – 65 (%)	9	32

<b>Table A.7: Income, Education and Employment</b>		
<i>Income (n=345, 95% of sample)</i>	%	<i>N</i>
<i>Of which:</i> up to £19,999	13	46
£20,000 - £29,999	13	44
£30,000 - £39,999	12	42
£40,000 - £49,999	15	53
£50,000 - £59,999	11	39
£60,000 - £69,999	8	29
£70,000 - £99,999	14	49
£100,000+	13	43
<i>Education (n=362, 99% of sample)</i>	%	<i>N</i>
<i>Of which:</i> No formal qualifications	2	8
Secondary education	35	125
Tertiary education	55	198
Other	8	31
<i>Employment (n=363, 99% of sample)</i>	%	<i>N</i>
<i>Of which:</i> In paid employment	69	252
Self-employed	11	40
Not employed and not looking	16	57
Looking for work	2	8
Student	2	6



**Figure A.4**



### **A13. INTERPOLATED OUTCOME DATA**

In a number of cases, participants who were missing at 4 weeks returned later in the trial and recorded a weight reading at 12 weeks (n=40). A simple interpolation of their weight loss trajectory based on the two points in time (baseline and 12 weeks) sheds light on what their likely 4-week weight reading would have been, had it been observed.

Amongst these 40 participants, 15 recorded weight loss and 25 recorded weight gain, with 0.2% average weight loss implied at the four-week stage. Counting the interpolated data points improves the dataset from 187 to 227 outcome observations. The assumption is that participants experience a smooth, indeed linear, weight loss trajectory over 12 weeks. This is unlikely to be true in all cases, with many participants recording an initial acceleration in weight loss that then reaches a plateau or decelerates. However, the assumption of linear progress is a conservative one – it is more likely to under- than over-estimate actual progress over the early weeks – and on this basis is judged to be a sensible way of getting the most from the available data.

The data argues against there being a one-to-one association between attriting and failing to lose weight, with a range of experiences from gaining over 2% in weight to losing almost 3%. The data also shows no relationship between those who attrited at 4 weeks and returned at the 12 week point based on treatment group ( $p > 0.1$ ). Those with limited commitment were under-represented in the sample of returning participants (20% of the 40 returners) but this was not statistically significant ( $p = 0.2613$  comparing financial and limited commitment groups). These findings are significant because they further support the idea that being missing (or observed) is irrespective of treatment status – a key assumption required for the Lee Bounds estimates reported in chapter 5.

**Table A.8: Outcome data for participants  
who were not observed at 4 weeks and observed at 12 weeks**

<i>ID</i>	<i>Reported weight loss at 12 weeks</i>	<i>Implied weight loss at 4 weeks (kg)</i>	<i>Implied weight loss at 4 weeks (%)</i>
186832	0.45	0.15	0.23
453023	-2.72	-0.91	-1.25
51214	0.00	0.00	0.00
369615	4.99	1.66	1.62
134665	2.27	0.76	0.88
316741	-3.17	-1.06	-1.61
384343	-0.91	-0.30	-0.40
401272	-5.90	-1.97	-1.26
447785	-2.72	-0.91	-0.84
188320	-0.91	-0.30	-0.37
20181	-1.36	-0.45	-0.59
373767	-1.13	-0.38	-0.54
44237	0.91	0.30	0.39
125504	0.23	0.08	0.10
55350	3.57	1.19	1.25
168721	0.00	0.00	0.00
111119	0.45	0.15	0.22
23948	1.82	0.61	0.85
42354	5.44	1.81	2.82
308611	-1.81	-0.60	-0.82
257709	0.45	0.15	0.23
141108	-3.35	-1.12	-1.54
195400	-1.36	-0.45	-0.48
62720	0.91	0.30	0.40
125798	4.28	1.43	2.02
311741	-0.45	-0.15	-0.24
162668	2.95	0.98	1.34
142376	-2.54	-0.85	-1.01
216995	-2.83	-0.94	-1.20
313425	-1.36	-0.45	-0.46
62506	-0.94	-0.31	-0.28
185529	-3.18	-1.06	-1.28
378093	1.36	0.45	0.56
52364	-4.54	-1.51	-2.31
356930	-0.91	-0.30	-0.32
19275	0.27	0.09	0.15
149200	-4.54	-1.51	-1.89
412389	-0.45	-0.15	-0.17
91292	-2.72	-0.91	-1.06
419936	-0.91	-0.30	-0.27

Linear weight change assumed, and weight change at 4 weeks is calculated as 1/3<sup>rd</sup> of weight change at 12 weeks. Negative number and percentage for weight loss implies weight gain, positive number denotes weight loss.

## A14. WEIGHT LOSS PERFORMANCE OUTLIERS (CH 5)

**Table A.9: Investigating outliers**

Panel A: Top 5% weight loss records at 4 weeks (kgs)				
<i>ID</i>	<i>Start (kg)</i>	<i>End (kg)</i>	<i>Weigh t loss</i>	<i>Comments</i>
465615	102.06	93.44	8.62	Computational error converting start weight of 15 stones, which equates to 95.25 kgs not 102.06 kgs. Correction made in baseline data setting start weight at 95.25kg and outcome data amended. Revised weight loss 1.81 kgs at 4 weeks, and 2.71 kgs at 12 weeks. Revised outcome used in analysis.
309549	113.85	107.04	6.80	Data verified as credible, no changes. Regular reporting of modest weight loss suggests reliable end weight reading.
456059	95.26	88.90	6.35	Data verified, no changes.
459882	134	127.91	6.09	Data verified, no changes.
406440	97.07	91.17	5.90	Data verified, no changes.
152037	105	99.34	5.66	Data verified, no changes.
416048	93.44	88.0	5.44	Data verified, no changes.
153785	80.74	75.3	5.44	Data verified, no changes.
409816	109	103.87	5.13	Data verified, no changes.

Panel B: Top 5% weight gain records at 4 weeks (kgs)				
<i>ID</i>	<i>Start (kg)</i>	<i>End (kg)</i>	<i>Weight gain</i>	<i>Comments</i>
424034	138.34	147.87	9.52	Suspected participant error in baseline reading of 305lbs, which is considerably lower than the 330lbs reading three days later. Closest available reading after baseline survey and registration date is taken as the new start weight. With start weight changed to 149.69kg, weight outcome revised to weight loss of 1.82kg before being used in analysis.
381343	82.55	85.28	2.72	Data verified as credible, no changes.
140195	70.31	72.57	2.27	Data verified as credible, no changes.
335603	70.31	72.57	2.27	Data verified as credible, no changes.
161621	63.96	65.77	-1.81	Data verified as credible, no changes.
415378	67.13	68.04	-1.81	Data verified as credible, no changes.
342954	82.10	83.91	-1.81	Data verified as credible, no changes.
34221	78.79	80.29	-1.59	Data verified as credible, no changes.
156792	68.0	69.40	1.40	Data verified as credible, no changes.

Panel C: Top 5% weight gain records at 12 weeks (kgs)				
<i>ID</i>	<i>Start (kg)</i>	<i>End (kg)</i>	<i>Gain (kg)</i>	<i>Comments</i>
243512	62.6	78.02	15.42	Highly implausible weight gain of 25% of initial body weight over three months. Reading at 12 weeks suggests weight of 172lbs following baseline of 144lbs, and every other reading in the range of 141 – 149lbs. This suggests an inputting error (perhaps on a mobile device) with the number '7' pressed in place of '4'. The nearest available reading to the original 12 week weigh-in was substituted, giving a revised end weight of 67.6kg. Revised weight gain at 12 weeks is 5kg, which remains large but plausible, and is used in analysis.
424034	138.3	147.0	8.62	Baseline error identified and corrected as per earlier table. Revised start weight of 149.69 kg implies weight loss of 2.73kgs at 12 weeks. Revised outcome used in analysis.
170226	67.1	75.3	8.16	Monthly weight readings showed large shifts, with 5% increase in first month, followed by a 5% decrease in second month, and 10% increase in following month. Volatile pattern and infrequent weigh-ins suggest the data cannot be trusted and could be subject to quite large measurement errors. This outlier is dropped from 12-week analysis, with robustness checks on results when it is included (see below).
401272	156.5	162.4	5.90	Readings show gap in self-monitoring from 30 August to 1 November, during which time weight readings jumped from 340lbs to 360lbs. This is large, but plausible if no active weight management was taking place. Later, as self-monitoring improves weight begin to fall slowly. If the individual's weight trajectory is closely tied to monitoring, the lack of weigh-ins over the two-month period is credibly associated with large weight gain. No revisions made.
316741	63.5	69.0	5.45	Discrepancy discovered between baseline survey reading of 63.5kg and same-day weigh-in of 65.8kg. Reading inputted on the website is plausibly more likely to be correct if it follows immediate weigh-in on scales. Start weight revised to 65.77kg, with weight gain at 12 weeks now 3.18kg.
398556	93.4	98.4	4.99	Data verified as credible, no changes.
52364	65.3	69.9	4.53	Data verified as credible, no changes.
149200	79.8	84.4	4.53	Data verified as credible, no changes.
353948	79.7	84.0	4.31	Data verified as credible, no changes.

No changes were made to any of the records in the top 5% of weight loss at 12 weeks beyond that already for ID 465516.

In this outlier investigation, five participants or 1% of the sample were identified as erroneous outliers, with possible inputting or measurement errors undermining the veracity of their outcome data. One of the inputting errors was mine (ID 465516). Three other observations were revised either at the baseline or endline using the closest available data points to eliminate the likely measurement error (ID 424034, ID 243512, ID 316741). One observation was deemed implausible and dropped from the main analysis (ID 170226).

Robustness checks reported in section A17 assess the impact of these decisions, by (a) leaving the original data as it was without any efforts to eliminate measurement error for ID 424034, ID 243512, ID 316741, and ID 170226 (see Table A18 below); and (b) using the cleaned data for these three participants and including ID 170226 (see Table A19 below).

## A15. WHAT DRIVES ATTRITION IN THE FOOD MONITOR EXPERIMENT?

**Table A.10**

<b>Table A.13: Probit regression on missing weight loss data</b>		
	At 4 weeks	At 12 weeks
Refund	-0.234 (0.232)	-0.184 (0.338)
Coach	0.117 (0.508)	0.037 (0.835)
Starting weight in kg	0.015 (0.067)	0.013 (0.098)
Overweight	0.221 (0.340)	-0.100 (0.662)
Obese	-0.139 (0.642)	-0.462 (0.127)
Severely obese	-0.642 (0.240)	-1.096* (0.045)
4-week weight loss target	0.020 (0.605)	-0.011 (0.782)
Fruit and vegetable intake	0.065 (0.062)	0.062 (0.096)
Exercise sessions per week	-0.080* (0.013)	-0.093** (0.002)
Experienced life changes	0.126 (0.431)	0.151 (0.351)
Other activities	0.429 (0.103)	0.473 (0.053)
Wellbeing	-0.071 (0.065)	-0.040 (0.301)
Female	0.566* (0.034)	0.661* (0.010)
Live for Today	-0.178 (0.605)	0.356 (0.289)
Unconfident Fatalist	0.054 (0.842)	0.377 (0.152)
Health Conscious Realist	0.182 (0.507)	0.442 (0.101)
Balanced Compensator	0.076 (0.819)	-0.044 (0.892)
Impatient	-0.005 (0.392)	0.001 (0.849)
Age any category significant?	No	No
Children at home	0.172 (0.296)	0.240 (0.136)
Phase 1	0.136 (0.763)	-0.065 (0.878)
Start date: - August	-0.026 (0.897)	-0.178 (0.382)
- September	-0.329 (0.248)	-0.346 (0.221)
- October	0.093 (0.863)	-0.254 (0.626)
- November	0.358 (0.560)	-0.125 (0.832)
Observations	324	324
Pseudo R <sup>2</sup>	0.082	0.085

A small number of baseline characteristics are associated with attrition (note: no adjustments made for multiple hypothesis testing). Female participants are more likely to drop out, and overweight participants are more likely to attrite early on. Those who report a higher wellbeing score are less likely to attrite, as are those who report taking more exercise. It might be expected that attrition varies with seasonal factors but there is no significant link with starting month. Demographic variables are separately tested using the rank sum test for categorical variables, and only one significant association is found with income: those at the top end of the income spectrum are much less likely to attrite.

Differential attrition along other lines that are not observed in the dataset could lead to biased estimates of the treatment effects. For example, it is plausible that those who dropped out may have gone off track with their weight loss efforts, and did not want to report weight gain. This scenario is arguably more likely than the alternative explanations that they are missing at random, or that they are so successful they no longer need the Food Monitor tool. If many dropouts gained weight in reality, their absence in the dataset could lead to upwardly biased treatment effect estimates, but this depends on there being a systematic difference in performance amongst the attritors by treatment group. The earlier discussion of participants who provided readings at 12 weeks but not 4 weeks offers useful insight here: that data suggested that on average those who dropped out at 4 weeks had close to zero weight loss, but with more people losing than gaining. The finding rules out a one-to-one association between weight loss outcomes and attrition.

At the 12-week stage, it is possible to analyse more explicitly the link between attrition and previous outcomes. The association between outcome data missing at 12 weeks and weight loss progress up to 4 weeks suggests that this relationship is not statistically significant ( $p=0.934$ ), and so attrition on the 12-week data can be treated as missing independent of potential outcomes, and conditional on covariates (as set out in Table 8). This characterisation is important as it warrants inverse probability weighting to address attrition.



## **A16. INVERSE PROBABILITY WEIGHTING (IPW) FOR 12-WEEK ANALYSIS**

### Step 1

What variables are associated with missing outcome data at 12 weeks?  
Finding: Significant variables at 10% level include exercise, otheractiv, female

```
xi: probit wt12miss limcom repcom startwt i.bmicat wltargpc fruitveg exercise  
lifechange otheractiv wellbeing female age30 i.hfscat impatient phase1 children,  
vce(cluster id)
```

### Step 2

Create variable to measure probability of being observed and test against the predictors of missingness.

```
gen ob12 = 0  
replace ob12 = 1 if wtl12pc != .
```

```
logit ob12 exercise female otheractiv, vce(cluster id)
```

```
predict probob12, p  
sum probob12
```

```
sort expgrp2  
by expgrp2: sum probob12
```

Finding: equation has Wald chi squared test statistic 11.11 and  $p < 0.011$  so jointly significant variables for attrition at 12 weeks

Finding: probob12 mean = 0.446, half of 12 week data missing, with probability of being observed 45% in limcom, 44% fincom 44% repcom \*

### Step 3

Use different probabilities to create inverse probability weights. Applying weights should yield different regression results - check by regressing wtl12 with treatment dummies for sample of available data

```
gen w12_ipw = 1/probob12
```

```
xi: regress wtl12pc limcom repcom [pweight = w12_ipw] if ob12 == 1, vce(cluster  
id)
```

```
xi: regress wtl12pc limcom repcom if ob12 == 1, vce(cluster id)
```

Finding: without weights limcom  $p = 0.824$ , repcom  $p = 0.187$ ; with weights fincom  $p$  value = 0.899 repcom  $p = 0.115$ , sig improves on repcom and R-squared improves marginally but still low, N=162 no change

### Step 4

Apply weights to full sample with covariates – used to generate results Chapter 5

```
xi: quietly regress wtl12pc limcom repcom female age40 i.bmicat i.hfscat  
impatient fruitveg exercise lifechange otheractiv i.startmonth phase1 if wlgol ==  
1 & id!=170226 [pweight = w12_ipw], vce(cluster id)
```

**A17. ROBUSTNESS CHECKS ON AVERAGE TREATMENT EFFECT REGRESSIONS IN CHAPTER 5**

**Table A.11: Average treatment effects on refund treatment**

	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
<b>Refund</b>	<b>-0.303</b> <b>(0.330)</b>	<b>-0.302</b> <b>(0.151)</b>	<b>-0.145</b> <b>(0.865)</b>	<b>-0.183</b> <b>(0.641)</b>	<b>-0.060</b> <b>(0.942)</b>
Starting weight (kg)	0.980*** (0.000)	0.993*** (0.000)	1.004*** (0.000)	1.000*** (0.000)	0.999*** (0.000)
Coach group	0.032 (0.924)	0.013 (0.948)	1.378 (0.143)	0.442 (0.280)	1.290 (0.148)
Female	1.011 (0.102)	1.111* (0.016)	3.231 (0.053)	2.504* (0.014)	3.558* (0.037)
Aged over 40	-0.359 (0.172)	-0.230 (0.185)	-1.352 (0.066)	-0.492 (0.170)	-1.357 (0.058)
Overweight	0.401 (0.289)	0.364 (0.169)	0.462 (0.636)	0.225 (0.662)	0.531 (0.578)
Obese	0.422 (0.488)	0.209 (0.590)	-0.959 (0.544)	-0.396 (0.590)	-0.728 (0.622)
Severely obese	0.550 (0.593)	0.282 (0.689)	0.431 (0.880)	-0.031 (0.983)	0.295 (0.918)
Myopic health attitudes	-0.302 (0.290)	-0.231 (0.213)	-1.172 (0.130)	-0.367 (0.307)	-0.918 (0.228)
Present biased	0.027** (0.010)	0.019** (0.002)	0.079* (0.025)	0.024* (0.038)	0.074* (0.025)
Exercise	-0.053 (0.314)	-0.047 (0.169)	-0.049 (0.711)	-0.064 (0.313)	-0.068 (0.618)
Experienced life changes	0.182 (0.491)	0.082 (0.671)	0.330 (0.650)	0.053 (0.884)	0.287 (0.686)
Other activities to lose weight	0.208 (0.550)	0.370 (0.143)	2.151* (0.030)	1.659** (0.005)	2.076* (0.049)
Low income	0.275 (0.545)	0.160 (0.558)	0.817 (0.399)	0.085 (0.866)	0.575 (0.536)
Recruited in August	0.544 (0.127)	0.329 (0.158)	-0.725 (0.478)	-0.081 (0.856)	-0.630 (0.535)
Recruited in September	0.527 (0.236)	0.306 (0.306)	-1.756 (0.130)	-0.484 (0.339)	-1.442 (0.223)
N	171	271	121	270	121
R <sup>2</sup>	0.992	0.995	0.967	0.981	0.967

Notes: Full OLS regression results on weight outcomes as set out in Chapter 5 equation 15a. Columns 1 and 3 present complete case analysis, columns 2 and 4 use the baseline observation carried forward where end weight is missing. Sample drawn from phase 1 only, where all three treatments were available. ATE for refund treatment group recovered through comparison with the monthly fee-paying group. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

<b>Table A.12: Average treatment effects on refund treatment</b>					
	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
<b>Coach</b>	<b>0.189</b> <b>(0.553)</b>	<b>0.205</b> <b>(0.296)</b>	<b>1.606*</b> <b>(0.040)</b>	<b>0.664</b> <b>(0.053)</b>	<b>1.632*</b> <b>(0.034)</b>
Starting weight (kg)	0.973*** (0.000)	0.985*** (0.000)	0.983*** (0.000)	0.987*** (0.000)	0.980*** (0.000)
Female	0.322 (0.613)	0.326 (0.477)	1.486 (0.380)	1.009 (0.233)	1.897 (0.248)
Aged over 40	-0.632* (0.037)	-0.316 (0.093)	-0.388 (0.580)	-0.148 (0.674)	-0.541 (0.426)
Overweight	0.181 (0.672)	0.192 (0.481)	0.661 (0.563)	0.269 (0.578)	0.843 (0.448)
Obese	0.894 (0.169)	0.412 (0.336)	-0.748 (0.639)	-0.378 (0.593)	-0.777 (0.603)
Severely obese	1.093 (0.322)	0.667 (0.399)	0.873 (0.729)	0.333 (0.797)	0.686 (0.780)
Myopic health attitudes	-0.036 (0.920)	-0.022 (0.922)	0.607 (0.504)	0.507 (0.207)	0.730 (0.397)
Present biased	0.017 (0.059)	0.010 (0.061)	0.050** (0.008)	0.023** (0.003)	0.045* (0.028)
Exercise	0.033 (0.624)	0.005 (0.897)	0.057 (0.691)	0.007 (0.909)	0.045 (0.764)
Experienced life changes	0.352 (0.311)	0.229 (0.298)	0.355 (0.691)	0.145 (0.701)	0.259 (0.765)
Other activities to lose weight	-0.488 (0.244)	-0.013 (0.963)	1.018 (0.317)	0.894 (0.084)	1.087 (0.320)
Low income	0.369 (0.451)	0.200 (0.490)	0.588 (0.556)	0.076 (0.868)	0.317 (0.747)
Recruited in August	0.164 (0.731)	-0.107 (0.659)	-2.142 (0.092)	-0.770 (0.107)	-2.002 (0.085)
Recruited in September	-0.366 (0.511)	-0.415 (0.187)	-3.450* (0.013)	-1.498** (0.003)	-3.402** (0.009)
Recruited in October	-0.008 (0.990)	-0.214 (0.573)	-1.358 (0.385)	-0.444 (0.481)	-1.371 (0.385)
Recruited in November	-0.916 (0.184)	-0.638 (0.117)	-2.436 (0.069)	-1.142 (0.052)	-2.133 (0.106)
N	145	245	106	244	106
R <sup>2</sup>	0.992	0.995	0.969	0.984	0.970

Notes: Full OLS regression results on weight outcomes as set out in Chapter 5 equation 15a. Columns 1 and 3 present complete case analysis, columns 2 and 4 use the baseline observation carried forward where end weight is missing. Sample drawn from phase 1 and 2 excluding the treatment with quota applied. ATE for coach treatment group recovered through comparison with the monthly fee-paying group. Robust standard errors clustered at individual level.

P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

**Table A.13: Average treatment effects on self-monitoring outcomes**

		OLS estimates
<b>Panel A:</b>	<b>Refund</b>	<b>4.861** (0.009)</b>
	Coach	-0.106 (0.955)
	Female	-7.772** (0.002)
	Aged over 40	0.976 (0.533)
	Overweight	-5.923** (0.005)
	Obese	-3.582 (0.077)
	Severely obese	-5.488* (0.040)
	Myopic health attitudes	0.954 (0.559)
	Present biased	-0.125* (0.042)
	Exercise	0.197 (0.495)
	Experienced life changes	1.491 (0.369)
	Other activities to lose weight	-2.151 (0.359)
	Low income	0.229 (0.923)
	Recruited in August	-3.247 (0.071)
	Recruited in September	-1.068 (0.643)
	N	278
	R <sup>2</sup>	0.126
<b>Panel B:</b>	<b>Coach</b>	<b>-0.304 (0.858)</b>
	Female	-2.243 (0.373)
	Aged over 40	-0.709 (0.676)
	Overweight	-4.595 (0.066)
	Obese	-1.327 (0.581)
	Severely obese	-6.308* (0.035)
	Myopic health attitudes	-1.902 (0.299)
	Present biased	-0.095 (0.135)
	Exercise	-0.036 (0.898)
	Experienced life changes	-0.744 (0.675)
	Other activities to lose weight	-2.080 (0.463)
	Low income	2.339 (0.362)
	Recruited in August	-2.500 (0.271)
	Recruited in September	0.883 (0.747)
	Recruited in October	-4.667 (0.128)
	Recruited in November	-2.732 (0.462)
	N	250
	R <sup>2</sup>	0.069

Notes: Full OLS regression results on weight outcomes as set out in chapter 5 self-monitoring results table column 2. Panel A recovers ATE for refund treatment using equation 15a, and panel B recovers ATE for coach treatment using equation 15b, both through comparison with monthly fee-paying group. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

**Table A.14: Regression analysis without covariates on weight outcomes**

	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
Panel A:					
Refund	-5.94 (0.069)	-6.19* (0.018)	-5.30 (0.216)	-6.06* (0.022)	-5.61 (0.174)
N	128	200	88	198	88
R <sup>2</sup>	0.0627	0.0275	0.0177	0.0263	0.0208
Panel B:					
Coach	-4.41 (0.128)	-1.77 (0.466)	-2.87 (0.415)	-1.40 (0.566)	-2.07 (0.556)
N	152	254	112	254	112
R <sup>2</sup>	0.014	0.002	0.006	0.001	0.003

Notes: Panel A sets out ATEs on limited commitment treatment, and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. ATEs based on end weight comparison with monthly fee-paying group. Sample of participants aiming to lose weight. 12-week data excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

**Table A.15: Regression analysis on participants aiming to maintain and lose weight**

	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
Panel A:					
Refund	-0.228 (0.430)	-0.258 (0.207)	0.060 (0.941)	-0.176 (0.650)	0.113 (0.887)
N	178	278	125	277	125
R <sup>2</sup>	0.992	0.995	0.967	0.982	0.968
Panel B:					
Coach	0.190 (0.538)	0.204 (0.291)	1.53* (0.049)	0.64 (0.062)	1.56* (0.042)
N	150	250	108	249	108
R <sup>2</sup>	0.992	0.995	0.969	0.9842	0.970

Notes: Panel A sets out ATEs on limited commitment treatment and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. 12-week data excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**Table A.16: Regression analysis using alternative time preference variable**

	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
Panel A:					
Refund	-0.199 (0.519)	-0.224 (0.287)	0.022 (0.980)	-0.061 (0.877)	0.121 (0.885)
Impatient	0.792* (0.045)	0.547* (0.039)	1.30 (0.231)	0.545 (0.314)	1.30 (0.234)
N	177	281	126	280	126
R <sup>2</sup>	0.992	0.995	0.966	0.981	0.967
Panel B:					
Coach	0.242 (0.437)	0.245 (0.206)	1.63* (0.032)	0.695* (0.040)	1.64* (0.028)
Impatient	0.628 (0.063)	0.296 (0.156)	1.42 (0.081)	0.535 (0.179)	1.243 (0.153)
N	152	254	112	253	112
R <sup>2</sup>	0.992	0.995	0.969	0.984	0.970

Notes: Panel A sets out ATEs on limited commitment treatment and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. Alternative time preference measure is a binary variable to capture most 'impatient' participants. 12-week data excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**Table A.17: Regression analysis with alternative outlier strategy (ID 170226 included)**

	12 weeks CC (1)	12 weeks BOCF (2)	12 weeks IPW (3)
Panel A:			
Refund	-0.453 (0.612)	-0.292 (0.474)	-0.264 (0.756)
N	122	271	122
R <sup>2</sup>	0.965	0.981	0.966
Panel B:			
Coach	1.50 (0.055)	0.608 (0.081)	1.58* (0.039)
N	107	245	107
R <sup>2</sup>	0.968	0.983	0.969

Notes: Panel A sets out ATEs on limited commitment treatment and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. Alternative outlier strategy includes excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**Table A.18: Regression analysis with alternative outlier strategy (raw data used)**

	4 weeks CC (1)	4 weeks BOCF (2)	12 weeks CC (3)	12 weeks BOCF (4)	12 weeks IPW (5)
Panel A:					
Refund	-0.377 (0.235)	-0.365 (0.090)	-0.097 (0.914)	-0.244 (0.536)	-0.051 (0.954)
N	171	271	121	270	121
R <sup>2</sup>	0.990	0.994	0.960	0.980	0.960
Panel B:					
Coach	0.213 (0.532)	0.162 (0.417)	1.47 (0.063)	0.566 (0.107)	1.51 (0.053)
N	14	245	107	245	107
R <sup>2</sup>	0.989	0.994	0.965	0.982	0.965

Notes: Panel A sets out ATEs on limited commitment treatment and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. 12-week data excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**Table A.19: Regression analysis using alternative estimator on self-monitoring outcomes**

		Poisson estimate
Panel A:	Refund	0.218* (0.010)
	N	278
	R <sup>2</sup>	0.069
Panel B:	Coach	-0.015 (0.861)
	N	250
	R <sup>2</sup>	0.042

Notes: Panel A sets out ATEs on limited commitment treatment and panel B on reputational commitment treatment, following equations 15a and 15b in chapter 5. Poisson coefficient represents increase in log unit under treatment condition. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**Table A.20: Complier average causal effects (CACE) on reputational commitment treatment**

	4 weeks CC (1)	12 weeks CC (2)	12 weeks IPW (3)	Self- monitoring
Refund	0.486 (0.526)	5.261* (0.038)	5.244* (0.032)	0.456 (0.907)
N	145	106	106	245
R <sup>2</sup>	0.992	0.961	0.963	0.080

Notes: Instrumental variables regression (2SLS) using standard covariates as instruments, with coach group assignment as an instrument for coach treatment. All other baseline variables included in line with earlier analysis, solely CACEs reported here for brevity. 12-week data excludes ID 170226. Robust standard errors clustered at individual level. P-values in parentheses\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ .

**A18. HETEROGENEOUS TREATMENT EFFECTS IN  
CHAPTER 5**

**Table A.21: Sub-group analysis on refund and coach treatments**

	Self-monitoring (1)	Weight 4 weeks CC (2)	Weight 12 weeks IPW (3)
<b>Panel A</b>			
Refund	3.939 (0.167)	-0.457 (0.343)	0.733 (0.568)
Refund x present bias	0.173 (0.153)	-0.016 (0.444)	0.028 (0.676)
Refund x myopia	0.087 (0.979)	0.394 (0.469)	-1.617 (0.263)
Starting weight	-	0.980*** (0.000)	0.998*** (0.000)
Coach	-0.067 (0.972)	0.048 (0.888)	1.251 (0.168)
Female	-8.179** (0.001)	1.094 (0.078)	3.307 (0.062)
Aged over 40	0.757 (0.633)	-0.338 (0.197)	-1.444* (0.043)
Overweight	-6.228** (0.004)	0.436 (0.263)	0.484 (0.618)
Obese	-3.702 (0.070)	0.449 (0.472)	-0.763 (0.609)
Severely obese	-5.537* (0.040)	0.637 (0.541)	0.066 (0.981)
Myopic attitudes	1.028 (0.620)	-0.487 (0.172)	-0.260 (0.783)
Present biased	-0.198* (0.018)	0.041* (0.011)	0.052 (0.182)
Exercise	0.178 (0.539)	-0.053 (0.317)	-0.067 (0.624)
Life changes	1.601 (0.337)	0.158 (0.562)	0.352 (0.626)
Other activities	-2.136 (0.357)	0.230 (0.521)	1.955 (0.069)
Low income	0.471 (0.841)	0.275 (0.550)	0.522 (0.569)
Recruited in August	-3.156 (0.078)	0.552 (0.120)	-0.650 (0.513)
Recruited in Sept	-0.944 (0.682)	0.507 (0.262)	-1.435 (0.238)
N	278	171	121
R <sup>2</sup>	0.131	0.992	0.968
<b>Panel B</b>			
Coach	0.131 (0.961)	0.626 (0.195)	2.819* (0.016)
Coach x present bias	-0.158 (0.179)	0.019 (0.429)	-0.002 (0.967)
Coach x myopia	0.776 (0.824)	-0.915 (0.129)	-2.164 (0.144)
Starting weight	-	0.972*** (0.000)	0.979*** (0.000)
Female	-2.482 (0.329)	0.284 (0.643)	2.028 (0.208)
Aged over 40	-0.830 (0.622)	-0.645* (0.033)	-0.543 (0.423)
Overweight	-4.299 (0.082)	0.178 (0.682)	1.015 (0.365)
Obese	-0.855 (0.720)	0.882 (0.181)	-0.408 (0.794)
Severely obese	-6.079* (0.045)	1.060 (0.335)	0.810 (0.743)
Myopic attitudes	-2.282 (0.342)	0.353 (0.447)	1.633 (0.142)
Present biased	-0.031 (0.647)	0.014 (0.196)	0.047 (0.055)
Exercise	-0.071 (0.800)	0.037 (0.570)	0.051 (0.729)
Life changes	-0.876 (0.624)	0.372 (0.285)	0.386 (0.647)
Other activities	-1.817 (0.524)	-0.590 (0.150)	1.016 (0.347)
Low income	2.482 (0.330)	0.361 (0.451)	0.207 (0.836)
Recruited in August	-2.373 (0.305)	0.224 (0.621)	-1.626 (0.149)
Recruited in Sept	0.697 (0.799)	-0.330 (0.539)	-3.081* (0.018)
Recruited in Oct	-4.513 (0.147)	0.035 (0.959)	-1.040 (0.531)
Recruited in Nov	-2.430 (0.524)	-0.877 (0.216)	-1.843 (0.179)
N	250	145	106
R <sup>2</sup>	0.075	0.992	0.970

Notes: OLS regression underpinning CATEs presented in chapter 5.



**A19. BENJAMINI-HOCHBERG CORRECTIONS FOR  
HETEROGENEOUS TREATMENT EFFECTS IN CHAPTER 5**

**Table A.22: Benjamini-Hochberg significance thresholds for selected sub-group findings**

Outcome	Trait	P-value	Revised threshold	P < revised threshold?
Weight at 12 weeks	Present bias	0.015	0.025	Yes

Notes: Benjamini-Hochberg significance threshold applied for those findings in chapter 5 that emerge as statistically significant. Corrected threshold assumes two hypotheses are being tested per model.

## A20. TUTORS AND RECRUITMENT DETAILS

**Table A.23: Shape Up groups**

Wave	Group	Recruitment	Tutor no.	Tutor name	Venue	Day	Time	Participants
1 (2014)	1	20 January	5	Mike	Armoury Gym	Monday	16:30	5
	2	20 January	3	Ian	Crowndale Centre	Monday	19:00	7
	3	25 January	2	Bianca	Camden Town Hall	Saturday	11:00	11
	4 <sup>87</sup>	4 February	6	Robbie	Kentish Town Library	Tuesday	10:30	0
	5	29 January	3	Ian	Swiss Cottage Library	Wednesday	10:30	8
	6	29 January	3	Ian	Swiss Cottage Library	Wednesday	12:30	6
2 (2014)	7	20 February	1	Augusto	Armoury Gym	Thursday	10:30	12
	8	11 March	4	Maria	Kentish Town Health Centre	Wednesday	18:00	4
	9	15 March	1	Augusto	Swiss Cottage Library	Saturday	11:00	4
	10	7 May	3	Ian	Swiss Cottage Library	Wednesday	10:30	8
	11	12 May	5	Mike	Armoury Gym	Monday	16:30	4
	12	17 May	2	Bianca	Camden Town Hall	Saturday	11:00	7
	13	3 June	4	Maria	Kentish Town Health Centre	Tuesday	18:15	5
	14	5 June	1	Augusto	Armoury Gym	Thursday	10:30	3
	15	7 June	1	Augusto	Swiss Cottage Library	Saturday	11:00	12
	16	2 July	8	Tim	Camden Town Hall	Wednesday	17:00	5
	17	16 July	3	Ian	Swiss Cottage Library	Wednesday	12:30	8
18	7 August	6	Robbie	Kentish Town Community Centre	Thursday	10:00	4	
19	7 August	2	Bianca	Crowndale Centre	Thursday	18:00	5	
3 (2015 / 2016)	20	23 September	7	Sharon	Swiss Cottage Library	Wednesday	12:30	9
	21	26 September	5	Mike	Swiss Cottage Library	Saturday	11:00	9
	22	30 September	2	Bianca	Kentish Town Health C	Wednesday	18:00	16
	23	15 October	8	Tim	Talacre Sports Centre	Thursday	19:00	14
	24	24 October	1	Augusto	Camden Town Hall	Saturday	11:00	6
	25	20 January	8	Tim	Swiss Cottage Library	Wednesday	10:30	6
	26	21 January	8	Tim	Talacre Sports Centre	Thursday	19:00	9
	27	23 January	1	Augusto	Camden Town Hall	Saturday	11:00	10

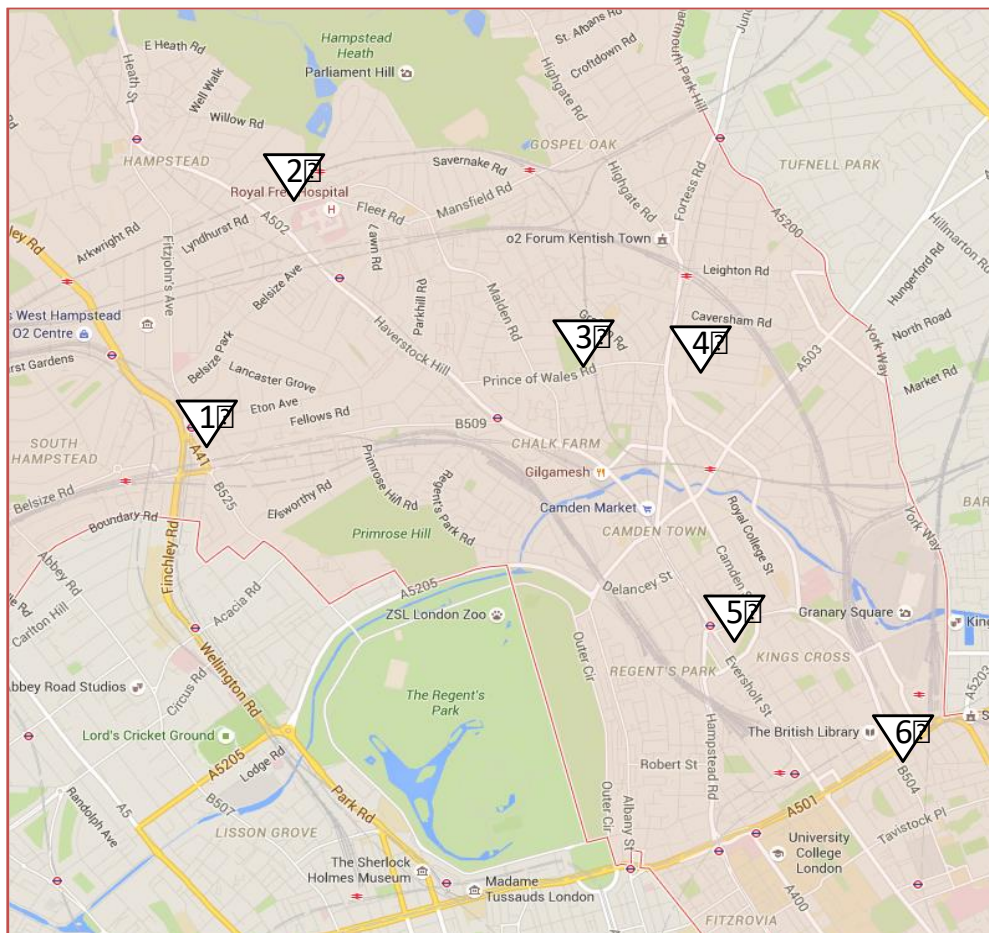
<sup>87</sup> Group discontinued for administrative reasons, 2 participants excluded.

## **Tutor introduction script**

- Let me introduce Manu Savani, a PhD student at UCL, who is working with us on a research project about commitment strategies and weight loss.
- Manu will be doing the weigh-ins today.
- She is recruiting participants for the research project.
- When you complete your weigh-in, she will give you an information sheet, and you can decide if you want to take part in the research project.
- It's entirely up to you. It won't affect your participation in the Shape Up programme.
- If you do want to take part, you'll need to spend a few minutes to register, and then just continue with the programme as usual.

## Fieldwork sites (Shape Up venues)

Figure A.5



**Table A.24: Fieldwork venue details**

Site number	Venue
1	Swiss Cottage Library, 88 Avenue Road, London NW3 3HA
2	The Armoury Gym, 25 Pond St, Belsize Park, London NW3 2PN
3	Talacre Sports Centre, Dalby Street, London NW5 3AF
4	Kentish Town Health Centre, 2 Bartholomew Rd, London NW5 2BX
5	Crowndale Centre, 218 Eversholt St, Kings Cross, London NW1 1BD
6	Camden Town Hall, Judd Street, London WC1H 9JE

## **A21. CAMDEN BASELINE VARIABLES**

### Health motivation

The Camden sample covered the full range of health motivation profiles, as did the Food Monitor sample. Participants were most likely to be Unconfident Fatalists or Health Conscious Realists – an interesting juxtaposition of the two ends of the motivation spectrum, low and high respectively. Balanced Compensators and Health Conscious Realists had lower average BMI scores (29.5 and 30.6), than the other groups who averaged over 31, in line with the predictions of the Healthy Foundations Segmentation model. Hedonistic Immortals, Live for Today's and Unconfident Fatalists are arguably those with the most short-term outlook on their health decisions, and they account for 54% of the sample. In terms of the dual-self framework, these participants are most likely to have their doer sub-selves dominate their choices, exhibit time inconsistency, and fail to achieve their health goals. This proposition will be tested in the heterogeneity analysis in Section x.

### Subjective wellbeing

190 participants reported wellbeing scores. Subjective wellbeing averaged 6.3 and ranged from 0 to 10, varying with BMI status. Those in the healthy BMI category reported a wellbeing score over 8, which fell to 6.4 for the overweight and 6.2 for the obese. Somewhat surprisingly, the severely obese had a slightly higher wellbeing score of 6.5. Wellbeing scores also varied with age, with older participants more likely to report a higher sense of satisfaction with life as a whole: the 48 participants aged 60+ reported wellbeing scores of 6.9, compared to 6.0 for the 55 individuals under 40.

### Life changes

A third of participants had experienced a major change in their life recently, including changes to work such as becoming unemployed, retiring, or starting a new job; or changes to family life, such as becoming a parent, experiencing a bereavement, or increasing caring responsibilities for family members. The data suggests a mild link between health motivation profile ( $p=0.092$  using a Mann-Whitney test) and whether they had experienced a recent life change: Balanced Compensators and Health Conscious Realists were less likely to cite a change (30% said yes) relative to Hedonistic Immortals (44%) and Live for Todays (53%). As in the Food Monitor experiment, this could be evidence that participants with more positive health motivations tend to sign up to the Shape Up programme in the absence of major life changes; in contrast, participants with a more short-term view are more likely to have experienced a major life change that may have then prompted them to take action on their weight management.

### Diet and exercise

Baseline information on diet and exercise was collected by Camden as part of their own 'starter survey', usually completed in the first week of the course. Dietary quality can be measured by the number of fruit and vegetable portions consumed the previous day. The average for England is 3.6 portions per day (HSCIC 2014), and the Food Monitor sample reported an average of 4 per day (Chapter 5). The Camden participants who completed this survey ( $n=138$ ) report an average of 3.7, ranging from 0.5 to 9 portions. Looking more closely at the data, 79% report consuming 5 or more pieces of fruit and vegetables, which is notably higher than the 37% who report the same in the Food Monitor sample. The starter survey also asked participants how many exercise sessions per week they usually undertook. Due to a number of missing responses to Camden's

survey, the trial's baseline survey responses on 'other activities to lose weight' were used to supplement the data. An augmented exercise variable reports that participants took part in 1-2 exercise sessions per week, which is lower than the Food Monitor participants who reported undertaking 3 sessions a week, and 31% took part in no exercise at all. These statistics paint a picture of the Camden sample being more conscious of dietary rules of thumb, such as the 5-a-day fruit and vegetable recommendation, but being far less physically active than the Food Monitor sample.

#### Other activities to lose weight

Most participants stated they were pursuing other activities alongside the Shape Up programme (69%). These included exercise at the gym, swimming, and more walking, and also new dietary habits such as managing portions and being more mindful of what foods they were eating. This leaves a sizeable proportion (31%) who did not report any complementary efforts to meet their weight loss targets. This could reflect a lack of information and ideas about how to do so, at that early stage in the Shape Up course; but it might also indicate a lack of motivation or belief that not much needed to be done.

## **A22. WEIGHT LOSS OUTCOME VARIABLES AND ATTRITION PATTERNS IN CHAPTER 6**

While Shape Up tutors encouraged all group members to attend each session to week 10, in a number of cases there was no weigh-in data for the final weeks: 81 participants missed the final 2 sessions of the programme. The analysis in chapter 6 takes a window of weeks seven to ten as a reasonable end point to derive weight loss data from class registers. There are two other measures that were considered, but judged sub-optimal relative to the weeks seven to ten outcome measure.

The first alternative is to apply the ‘last observation carried forward’ (LOCF) approach that is common in weight loss studies. It effectively ensures no attrition because every participant has at least one observation to roll forward; and a narrower window of weeks nine and ten for final weight readings. A second alternative is to accept a higher rate of attrition, and apply a narrower window for outcome data in weeks nine and ten of the Shape Up programme.

The LOCF outcome variable implicitly assumes that whatever the last reading, weight is maintained at that level through to the end of the 11 week programme. This assumption could be criticised on two grounds: some participants might continue to lose weight after they stop attending the programme (in part because of the contract), meaning the weight loss estimate is lower than the true weight loss achieved. Alternatively, some participants may gain weight if they stop attending the programme, which is not uncommon amongst those who have long struggled with weight management issues. The LOCF method in some cases rolls forward an outcome measure based on the very early weeks of the programme, and here in particular the assumption of zero weight change may be implausible, given the



history of participants in making steady and often unhealthy weight gain when left to their own efforts.

A comparison of implied mean weight loss across the three outcome variables (the baseline measure used in chapter 6 and the two alternatives discussed here), demonstrates the different pictures painted by the different measurement techniques. Data from weeks 7-10 suggest average weight loss of 2.7%, while LOCF data suggests average weight loss of 2.4%. In contrast, data from weeks 9-10 suggest somewhat greater weight loss of 3%. The assumption of no further weight change after the last available reading has the likely effect of a downward bias on average weight loss, generating conservative estimates of treatment effects. The downward bias would also affect with the variable measuring outcomes during weeks seven to ten but would reasonably be expected to be less pronounced; this expectation is borne out in Table A.25 below.

**Table A.25**

<b>Table A.27: Summary of Weight Loss Outcome Variables</b>					
% Weight Loss Variable	N	Missing	Mean	SD	Range
Weeks 9-10	127	70	3.00	3.4	-9.7 to 17.7
Weeks 7-10	161	36	2.72	3.1	-9.7 to 17.7
LOCF	197	0	2.39	3.0	-9.7 to 17.7

The longer the participants attended, the more weight was lost, and this might indicate that the Shape Up programme is effective. But a form of self-selection is likely at work here, with those staying on the course to the very end more likely to be those who could boast higher weight loss; and conversely those who are disappointed with slow or no progress more likely to drop out and not be counted at the end of the programme. Following through on this logic, data from weeks 9-10 is more likely to suffer from (upward) attrition bias: 70 missing outcomes indicates an attrition rate of 36% that could be largely driven by self-selection and other non-random factors.

Missing outcome data in weeks 9-10 is significantly associated with treatment status ( $p=0.012$ ). Data from weeks 7-10 demonstrates a similar pattern, with fewer attritors in the treatment group, but this difference is not statistically significant ( $p=0.169$ ). The implication is that overall attrition increased substantially in the final fortnight of the programme, also sharpening the contrast between experimental groups. Such differential attrition can be problematic, generating biased treatment effects by introducing a selection problem and unwinding the effects of random assignment to treatment.

**Table A.26**

<b>Table A.28: Attrition patterns across treatment groups</b>				
Missing weight loss outcomes	All sample (1)	Comparison (2)	Treatment (3)	p-value (2) = (3)
At 9-10 wks	35.5%	44.0%	26.8%	0.012
At 7-10 wks	18.3%	22.0%	14.4%	0.169

Attrition on weight loss data was anticipated in the Research Design (chapter 4), and even with follow-up efforts to mitigate missing outcomes it remains the case that attrition comes to 36% on data from weeks 9-10 and 18% on data from weeks 7-10. The latter, however, does not appear to suffer from problematic correlations with treatment status, with the negative association between treatment and attrition notwithstanding conventional hypothesis testing – see table A.27 below, which delves more deeply into the factors associated with attrition.

Probit regression highlights that missingness of outcomes in weeks 9-10 is significantly correlated with treatment status: those offered a commitment contract are less likely to attrite ( $p=0.002$ ). Older participants are less likely to attrite, as are those who attend the introductory Shape Up session. Attrition rates vary by tutor, with tutors 2, 3, 5 and 8 likely to discourage attrition more than tutor 1. This could reflect a number of circumstantial issues around venue and class timings, or may reflect level of effort expended by tutors in following up with class members who skip a week. Unspecified referral routes are most likely to predict attrition ( $p=0.003$ ), however this applies to a small minority of participants ( $n=6$ ).

In summary, complete case analysis using data from weeks 9-10 is likely to generate biased estimates of causal effects. Data from weeks 9-10 may only be justified for statistical analysis with corrective measures such as inverse proportionality weighting (IPW). Results with LOCF outcomes and appropriately weighted data from weeks 9-10 are reported in later appendices as part of robustness checks.

**Table A.27**

<b>Table A.29: What drives attrition in the Camden trial?</b>		
	At 9-10 weeks	At 7-10 weeks
Received contract	-0.808** (0.002)	-0.417 (0.089)
Female	0.012 (0.969)	-0.147 (0.665)
Age	-0.016* (0.038)	-0.009 (0.310)
Initial BMI overweight	-0.142 (0.820)	-0.613 (0.388)
Initial BMI obese	0.233 (0.714)	-0.436 (0.546)
Initial BMI severely obese	-1.988 (0.071)	-1.038 (0.262)
Exercise	-0.148 (0.074)	-0.199* (0.042)
Life change	-0.343 (0.173)	-0.327 (0.201)
Other activities for weight loss	0.168 (0.503)	0.210 (0.428)
Sophisticated	0.472 (0.054)	-0.067 (0.803)
Live for Today	-0.468 (0.449)	-0.744 (0.286)
Unconfident Fatalist	-0.745 (0.180)	-0.362 (0.538)
Health Conscious Realist	-0.926 (0.116)	-0.279 (0.648)
Balanced Compensator	-0.190 (0.765)	-0.298 (0.654)
Referral by - <i>GP</i>	-0.055 (0.836)	-0.106 (0.704)
- <i>other health practitioner</i>	-0.567 (0.052)	-0.209 (0.504)
- <i>unspecified</i>	2.415 (0.010)	1.435 (0.079)
Attended Shape Up week 0	-0.646* (0.011)	-0.797*** (0.001)
Daytime slot on weekdays	0.059 (0.875)	0.056 (0.886)
Number of participants in study	-0.031 (0.480)	0.048 (0.327)
Proportion of treated individuals	0.168 (0.875)	0.314 (0.777)
Recruited in week 2	0.344 (0.511)	0.057 (0.908)
Recruited in week 3	0.261 (0.635)	0.316 (0.571)
Participated in wave 2	-0.019 (0.961)	-0.136 (0.714)
Participated in wave 3	0.881 (0.078)	-0.131 (0.803)
Tutor 2	-0.793 (0.066)	-0.398 (0.390)
Tutor 3	-1.446*** (0.001)	-0.632 (0.140)
Tutor 4	-1.894** (0.006)	-0.782 (0.238)
Tutor 5	-1.585*** (0.001)	-0.768 (0.141)
Tutor 6	0.781 (0.417)	0.628 (0.437)
Tutor 7	-1.138 (0.079)	-0.208 (0.782)
Tutor 8	-0.952* (0.018)	-0.321 (0.494)
N	192	192
Pseudo $R^2$	0.292	0.208

**A23. ROBUSTNESS CHECKS ON AVERAGE TREATMENT EFFECT REGRESSIONS IN CHAPTER 6**

**Table A.28: Alternative model specification for ATE estimates**

Panel A: Using alternative outcome measures			
	Weight loss % at weeks 7-10	Weight loss % using LOCF	Met the 5% weight loss target
Contract	0.566 (0.274)	0.499 (0.231)	0.187 (0.393)
N	158	192	192
R <sup>2</sup>	0.198	0.167	0.094
Panel B: Without covariates			
	Weight (end weight in kg)	Attendance	Completion
Contract	-3.80 (0.086)	0.040 (0.257)	0.254 (0.189)
N	161	197	197
R <sup>2</sup> /Pseudo-R <sup>2</sup>	0.019	0.007	0.008
Panel C: Including wellbeing covariate			
	Weight (end weight in kg)	Attendance	Completion
Contract	-0.459 (0.339)	0.065 (0.104)	0.515 (0.028)
N	153	186	186
R <sup>2</sup> /Pseudo-R <sup>2</sup>	0.975	0.273	0.016

Notes: Across all panels, columns 1 and 2 use OLS regression and column 3 probit regression. In Panel A, column 1 measures weight loss as a % of starting weight, using data from weeks 7-10; while column 2 applies the same measure and carries forward the last available observation. Panels B and C column 1 use end weight in kg as the outcome variable. Panel B includes only the treatment variable and no other covariates, and panel C applies the full model presented in chapter 6 along with a wellbeing variable.

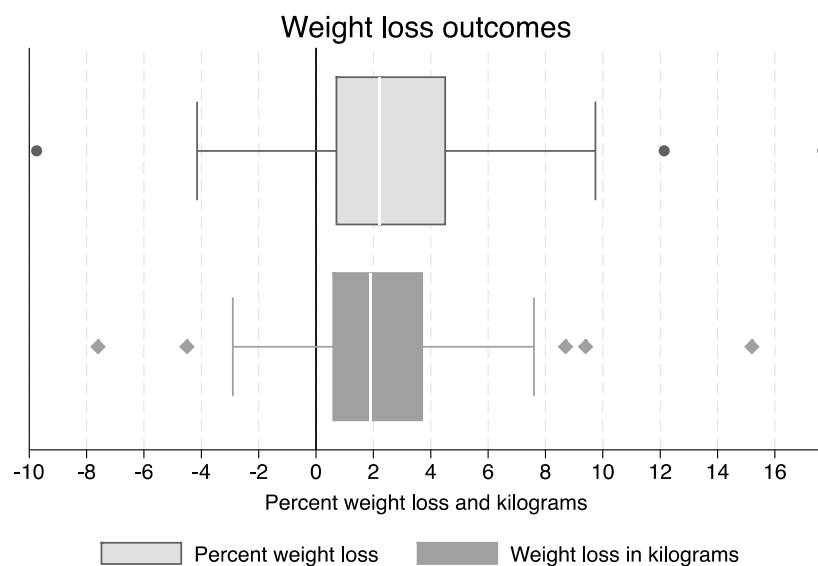
P-values in parentheses \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

## A24. WEIGHT LOSS PERFORMANCE OUTLIERS

The boxplot of weight loss outcomes highlights five outliers: two at the weight gain end of the spectrum, and three at the weight loss end. Table A.29 below provides a careful checking of each of the five cases to search for measurement or inputting errors, and assess plausibility of the outcomes registered. Verification of the outcome data by returning to the Camden sources, and triangulation of the specific cases with a Camden tutor provides sufficient assurance that all five outliers are benign, and are therefore included in all analysis below.

It is notable that the two participants who gained more than 4 kilograms were at the lower end of the health motivation spectrum. The participant with the highest weight gain of 7.6 kg is a wheelchair user who would have found it much more challenging than others to meet the physical activity recommendations of the Shape Up programme. All of the three participants who lost more than 8 kg were Health Conscious Realists, and two were male. There are no apparent patterns relating to tutor or treatment status for any outliers.

Figure A.6



**Table A.29: Analysis of outliers in weight loss performance**

---

1.	ID	Weight	BMI	HFS	Attendance	Gender	Age	Group	Wave
	11483	Gained 4.5 kg	32.4	HI	36%	Male	38	3	1

---

*Notes:* Individual was gaining weight on the programme, and left in week 5. Subsequent follow up provided a self-reported measurement that implied a considerable but plausible weight gain, given the low attendance rate, lack of progress during the programme, and drop out.

---

2.	ID	Weight	BMI	HFS	Attendance	Gender	Age	Group	Wave
	40004	Gained 7.6 kg	28.7	UF	82%	Female	32	25	4

---

*Notes:* Individual was recently confined to a wheelchair and would have found it very challenging to meet the physical activity requirements of the programme. Despite good attendance at the classes, this degree of weight gain is plausible considering the individual was still adjusting to her new life and disability, during registration mentioned the difficulty in finding public facilities for activity and socialising, and the negative health attitudes reported at baseline.

---

3.	ID	Weight	BMI	HFS	Attendance	Gender	Age	Group	Wave
	11538	Lost 15.2 kg	35.8	HCR	82%	Female	43	9	1

---

*Notes:* Individual reported weight change from 86kg to 70.8, moving from obese to the cusp of a healthy BMI, and implying 1.4kg weight loss per week. This is plausible given the positive health motivation and regular attendance.

---

4.	ID	Weight	BMI	HFS	Attendance	Gender	Age	Group	Wave
	30034	Lost 8.7 kg	28.2	HCR	73%	Male	29	21	3

---

*Notes:* Plausible outcome, as males tend to lose weight more rapidly, and interview follow up indicated this individual made several lifestyle changes and was deeply inspired by his tutor.

---

5.	ID	Weight	BMI	HFS	Attendance	Gender	Age	Group	Wave
	40045	Lost 9.4 kg	40.4	HCR	80%	Male	40	26	4

---

*Notes:* Individual appeared highly motivated at registration, reflected in HFS profile. Plausible weight loss given the individual was severely obese at baseline, and belonged to a group with physical activity incorporated into classes.

---

## A25. SUMMARY OF INTERVIEWEES IN CAMDEN FIELD EXPERIMENT

**Table A.30: Summary of Interviews and Basic Characteristics**

Interviewee	Participant ID	Interview date	Gender	Age	Initial BMI	Weight loss (kg)	Contract	Group	Medium
1	11411	11 April 2014	F	60	31.8	-0.4	1	1	In person
2	11407	14 May 2014	F	45	30.3	4.5	1	1	In person
3	11550	17 May 2014	F	35	45.5	1.1	1	8	In person
4	11549	20 May 2014	F	33	25.5	-0.4	1	9	Phone
5	20031	21 July 2014	F	34	34.0	1.9	0	12	Phone
6	20063	24 Oct 2014	F	50	25.4	-0.1	1	15	Phone
7	30002	25 Nov 2015	F	61	29.3	1.6	1	20	In person
8	30008	1 Dec 2015	F	40	28.2	3.8	0	20	Phone
9	30009	1 Dec 2015	F	68	29.7	-0.9	0	20	Phone
10	30011	25 Nov 2015	F	55	35.0	1	1	20	In person
11	30026	1 Dec 2015	F	61	27.3	4.9	0	21	Phone
12	30027	7 Dec 2015	F	67	24.5	7.6	1	21	Phone
13	30029	1 Dec 2015	F	48	29.1	0	1	21	Phone
14	30030	2 Dec 2015	M	40	28.9	4.2	1	21	Phone
15	30034	17 Dec 2015	M	29	28.2	8.7	1	21	Phone
16	30035	3 Dec 2015	F	45	26.2	5.9	0	21	Phone
17	30043	9 Dec 2015	F	57	27.1	3.3	0	22	Phone
18	30047	7 Dec 2015	F	74	34.9	6.4	1	22	Phone
19	30064	9 Dec 2015	F	51	31.1	3.7	1	22	Phone
20	30075	18 Dec 2015	M	50	35.9	6.0	0	23	Phone
21	30102	21 Dec 2015	F	30	31.3	0.2	1	24	Phone
22	30068	21 Dec 2015	F	68	34.0	5.5	1	23	Phone
23	40031	5 April 2016	F	45	29.2	-0.6	1	27	Phone
24	40028	12 April 2016	F	60	40.0	2.4	1	26	Phone

**A26. HETEROGENEOUS TREATMENT EFFECTS IN  
CHAPTER 6**

**Table A.31: Full regression results for sub-group analysis**

	Weight	Attendance	Completion
Contract	-0.171 (0.857)	0.160 (0.082)	1.052* (0.032)
Treat x myopia	-0.603 (0.567)	-0.001 (0.988)	-0.213 (0.634)
Treat x sophistication	-0.585 (0.542)	0.099 (0.184)	0.633 (0.201)
Treat x GP referral	-0.106 (0.913)	0.040 (0.588)	0.137 (0.781)
Treat x attended week 0	0.317 (0.744)	-0.221** (0.009)	-1.161* (0.014)
Starting weight	0.961*** (0.000)	-	-
Female	-0.340 (0.712)	0.089 (0.073)	0.412 (0.173)
Age	-0.038* (0.021)	0.004* (0.010)	0.015 (0.065)
Overweight	1.495 (0.427)	-0.039 (0.677)	0.262 (0.684)
Obese	1.634 (0.396)	-0.080 (0.407)	-0.157 (0.811)
Severely obese	2.170 (0.299)	-0.045 (0.702)	-0.584 (0.456)
Exercise	-0.286 (0.071)	0.019 (0.176)	0.154 (0.105)
Experienced major life changes recently	0.200 (0.763)	0.044 (0.232)	0.470 (0.064)
Other activities pursued to lose weight	0.808 (0.292)	-0.061 (0.122)	-0.434 (0.091)
Sophisticated	0.505 (0.401)	-0.054 (0.311)	-0.401 (0.261)
Myopic health attitudes	0.544 (0.402)	0.038 (0.441)	0.459 (0.159)
GP referred	-0.510 (0.443)	-0.053 (0.380)	-0.296 (0.463)
Referred by other health practitioner	-0.172 (0.789)	-0.007 (0.879)	-0.292 (0.339)
Other referral route	-3.978 (0.179)	-0.112 (0.292)	0.022 (0.974)
Attended week 0	-0.768 (0.277)	0.242*** (0.000)	1.288*** (0.000)
Daytime slot on weekdays	1.009 (0.189)	-0.011 (0.860)	0.398 (0.305)
% of group attritors	4.125 (0.259)	-	-
Group members in study	-0.154 (0.160)	-0.007 (0.328)	-0.040 (0.382)
% treated in group	0.280 (0.910)	-0.185 (0.301)	-0.421 (0.700)
Recruited in week 2	0.438 (0.587)	-0.017 (0.839)	-0.126 (0.781)
Recruited in week 3	-0.663 (0.537)	0.106 (0.197)	0.667 (0.234)
Recruited in wave 2	0.481 (0.546)	0.029 (0.661)	0.066 (0.854)
Recruited in wave 3	1.515 (0.206)	0.047 (0.507)	0.470 (0.295)
Tutor 2	1.457 (0.128)	0.051 (0.496)	0.687 (0.135)
Tutor 3	-0.419 (0.758)	0.152* (0.029)	0.837* (0.032)
Tutor 4	0.295 (0.831)	0.017 (0.890)	0.691 (0.297)
Tutor 5	0.446 (0.792)	0.084 (0.192)	0.385 (0.406)
Tutor 6	-1.689 (0.408)	-0.170 (0.353)	-1.681 (0.088)
Tutor 7	0.769 (0.550)	-0.021 (0.816)	0.209 (0.762)
Tutor 8	1.400 (0.118)	0.033 (0.581)	0.321 (0.420)
Observations	158	192	192
R <sup>2</sup> / Pseudo R <sup>2</sup>	0.975	0.299	0.220

*Notes: OLS regressions in columns 1 and 2, probit regression for completion outcomes in column 3. P-values in parentheses.*



**A27. BENJAMINI-HOCHBERG CORRECTIONS FOR  
HETEROGENEOUS TREATMENT EFFECTS IN CHAPTER  
6**

**Table A.32**

**Table A.36: Benjamini-Hochberg corrected p-value thresholds**

Outcome	Trait x contract combined coefficient in rank order for testing	P-value (* if <0.05)	Revised threshold	P < revised threshold?
Attendance	1 Sophistication	0.006**	0.013	Yes
	2 Myopic health attitudes	0.066	0.025	No
	3 GP referral	0.063	0.038	No
	4 Commitment priming	0.344	0.050	No
Completion	1 Sophistication	0.004**	0.013	Yes
	2 GP referral	0.045*	0.025	No
	3 Myopic health attitudes	0.073	0.038	No
	4 Commitment priming	0.813	0.050	No

No corrections reported for weight loss heterogeneity analysis as original results all have p-values > 0.05.

## REFERENCES

- Alan, S. & Ertac, S., 2015. Patience, self-control and the demand for commitment: Evidence from a large-scale field experiment. *Journal of Economic Behavior & Organization*, 115, pp.111–122.
- Allen, J.T., Cohn, S.R. & Ahern, A.L., 2015. Experiences of a commercial weight-loss programme after primary care referral : *British Journal of General Practice*, (April), pp.248–255.
- Alós-ferrer, C. & Strack, F., 2014. From dual processes to multiple selves: Implications for economic behavior. *Journal of Economic Psychology*, 41, pp.1–11.
- Angelucci, M. & Di Maro, V., 2015. *Program Evaluation and Spillover Effects*,
- Angrist, J.D. & Pischke, J.-S., 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*, Oxford: Princeton University Press.
- Ashraf, N., Karlan, D. & Yin, W., 2006. Tying Odysseus to the Mast : Evidence from a Commitment Savings Product in the Philippines. *The Quarterly Journal of Economics*, 121(2), pp.635–672.
- Au, N. et al., 2013. The cost-effectiveness of shopping to a predetermined grocery list to reduce overweight and obesity. *Nutrition & Diabetes*, 3(6), pp.1–5.
- Augurzky, B. et al., 2012. Does Money Burn Fat? *Ruhr Economic Papers*, pp.1–44.
- Aveyard, P. et al., 2016. Screening and brief intervention for obesity in primary care: a parallel, two-arm, randomised trial. *The Lancet*, 388(10059), pp.2492–2500.
- Baird, S. et al., 2014. Designing Experiments to Measure Spillover Effects. *World Bank Policy Research Working Paper*, 6824(March).
- Bénabou, R. & Pycia, M., 2002. Dynamic inconsistency and self-control : a planner – doer interpretation. *Economics Letters*, 77, pp.419–424.
- Benabou, R. & Tirole, J., 2004. Willpower and Personal Rules. *Journal of Political Economy*, 112(4), pp.848–886.
- Bohm-Bawerk, E. V., 1890. *Capital and Interest: A Critical History of Economical Theory*, London: Macmillan and Co.
- Boutelle, K.N. et al., 1999. How Can Obese Weight Controllers Minimize Weight Gain during the High Risk Holiday Season? By Self-Monitoring Very Consistently. *Health Psychology*, 18(4), pp.364–368.
- Boutron, I., John, P. & Torgerson, D.J., 2010. Reporting Methodological Items in Randomized Experiments in Political Science. *The ANNALS of the American Academy of Political and Social Science*, 628(1), pp.112–131.
- Brady, H.E., Collier, D. & Seawright, J., 2010. Refocusing the Discussion on Methodology. In H. E. Brady & D. Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield, pp. 15–31.
- Brune, L. et al., 2015. *Facilitating Savings for Agriculture: Field Experimental Evidence from Malawi*,
- Bryan, G., Karlan, D. & Nelson, S., 2010. Commitment devices. *Annual Review of Economics*, 2, pp.671–698.
- Burger, N. & Lynham, J., 2010. Betting on weight loss ... and losing: personal gambles as commitment mechanisms. *Applied Economics Letters*, 17(12), pp.1161–1166.
- Butland, B. et al., 2007. Tackling Obesities : Future Choices – Project

- report. *Government Office for Science*, pp.1–161.
- Butryn, M.L. et al., 2007. Consistent Self-monitoring of Weight: A Key Component of Successful Weight Loss Maintenance. *Obesity*, 15(12), pp.3091–3096.
- Camerer, C., Loewenstein, G. & Prelec, D., 2005. Neuroeconomics: How Neuroscience Can Inform Economics. *Journal of Economic Literature*, 43(1), pp.9–64.
- Cavaliere, A., Marchi, E. De & Banterle, A., 2014. Healthy – unhealthy weight and time preference . Is there an association ? An analysis through a consumer survey ☆. *Appetite*, 83, pp.135–143.
- Chapman, J., Campbell, M. & Wilson, C., 2015. Insights for Exercise Adherence From a Minimal Planning Intervention to Increase Physical Activity. *Health Education & Behavior*, 42(6), pp.730–735.
- Charmaz, K., 2011. Qualitative Interviewing and Grounded Theory Analysis. In J. F. Gubrium & J. A. Holstein, eds. *Handbook of Interview Research*. Sage Publications, Inc., pp. 675–694.
- Christakis, N.A. & Fowler, J.H., 2007. The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, 357(4), pp.370–79.
- Coppock, A., 2015. 10 Things You Need to Know about Multiple Comparisons. *EGAP Methods Guides*. Available at: <http://egap.org/methods-guides/10-things-you-need-know-about-multiple-comparisons> [Accessed September 28, 2016].
- Costa-Font, J. et al., 2013. Understanding healthy lifestyles: The role of choice and the environment. *Applied Economic Perspectives and Policy*, 35(1), pp.1–6.
- Crutzen, R., 2010. Adding effect sizes to a systematic review on interventions for promoting physical activity among European teenagers. , pp.6–10.
- Deaton, A. & Cartwright, N., 2016. *Understanding and Misunderstanding Randomised Controlled Trials*,
- DeCuir-Gunby, J.T., Marshall, P.L. & McCulloch, A.W., 2011. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. *Field Methods*, 23(2), pp.136–155.
- Department of Health, 2011. *Healthy Lives , Healthy People : A call to action on obesity in England*,
- Dolan, P. et al., 2012. Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, 33(1), pp.264–277.
- Downs, J.S., Loewenstein, G. & Wisdom, J., 2009. Strategies for promoting healthier food choices. *American Economic Review*, 99(2), pp.159–164.
- Duflo, E., Glennerster, R. & Kremer, M., 2007. *Using randomization in development economics research: a toolkit*,
- Duflo, E. & Kremer, M., 2008. Using Randomization in Development Economics: A Toolkit. In *Handbook of Development Economics*. p. 68.
- Dunning, T., 2008. Strengths and Limitations of Natural Experiments .? ? r hofdat. *Political Research Quarterly*, 61(2), pp.282–293.
- Dupas, P., 2011. Health Behavior in Developing Countries. *Annual Review of Economics*, 3(1), pp.425–449.
- Dupas, P. & Robinson, J., 2013. Why don't the poor save more? Evidence from health savings experiments. *American Economic Review*, 103(4), pp.1138–1171.
- Dziura, J.D. et al., 2013. Strategies for dealing with Missing data in clinical

- trials : From design to Analysis. *Yale Journal of Biology and Medicine*, 86, pp.343–358.
- Elobeid, M.A. et al., 2009. Missing Data in Randomized Clinical Trials for Weight Loss : Scope of the Problem , State of the Field , and Performance of Statistical Methods. *Plos One*, 4(8), pp.1–11.
- Fan, M. & Jin, Y., 2013. Obesity and Self-control: Food Consumption, Physical Activity, and Weight-loss Intention. *Applied Economic Perspectives and Policy*, 36(1), pp.125–145.
- Fink, G., McConnell, M. & Vollmer, S., 2014. Testing for heterogeneous treatment effects in experimental data : false discovery risks and correction procedures. *Journal of Development Effectiveness*, 6(1), pp.44–57.
- 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.
- Freedman, D.A., 2010. On Types of Scientific Inquiry: The Role of Qualitative Reasoning. In H. E. Brady & D. Collier, eds. *Rethinking Social Inquiry*. Lanham: Rowman and Littlefield, pp. 221–236.
- Fudenberg, D. & Levine, D.K., 2006. A Dual-Self Model of Impulse Control. *American Economic Review*, 96(5), pp.1449–1476.
- Gaines, B.J. & Kuklinski, J.H., 2011. Treatment Effects. In J. N. Druckman et al., eds. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, pp. 445–458.
- Gerber, A.S. & Green, D.P., 2012. *Field Experiments: Design, Analysis and Interpretation*, Norton.
- Gine, X., Karlan, D. & Zinman, J., 2010. Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation. *American Economic Journal: Applied Economics*, 2(October), pp.213–235.
- Glanz, K. et al., 1998. Why Americans eat what they do: Taste, nutrition, cost, convenience, and weight control concerns as influences on food consumption. *Journal of the American Dietetic Association*, 98(10), pp.1118–1126.
- Glennerster, R. & Takavarasha, K., 2013. *Running Randomized Evaluations: A Practical Guide*, Princeton University Press.
- Gollwitzer, P.M., 1999. Implementation Intentions: Strong Effects of Simple Plans. *American Psychologist*, 54(7), pp.493–503.
- Graham, J.W., 2009. Missing data analysis: making it work in the real world. *Annual review of psychology*, 60, pp.549–576.
- Guest, G. & Johnson, L., 2006. How Many Interviews Are Enough ? An Experiment with Data Saturation and Variability. , 18(1), pp.59–82.
- Gul, F. & Pesendorfer, W., 2004. Self Control, Revealed Preference and Consumption Choice. *Review of Economic Dynamics*, 7(2), pp.243–264.
- Halperin, S. & Heath, O., 2012. *Political Research: Methods and Practical Skills*, Oxford.
- Halpern, S.D. et al., 2015. Randomized Trial of Four Financial-Incentive Programs for Smoking Cessation. *New England Journal of Medicine*, 372(22), pp.2108–2117.
- Halpern, S.D., Asch, D. a. & Volpp, K.G., 2012. Commitment contracts as a way to health. *British Medical Journal*, 344, pp.e522–e522.
- Health Social Care Information Centre, 2015. HSE 2014 obesity and overweight adult trends.
- Heatherton, T.F. & Baumeister, R.F., 2013. AUTHORS ' RESPONSE and Future Failure : Past , Present , , 7(1), pp.90–98.
- Heckman, J. et al., 2000. Substitution and dropout bias in social experiments: a study of an influential social experiment\*. *Quarterly*

- Journal of Economics*, 115(2), pp.651–694.
- Hesse-Biber, S., 2013. Thinking Outside the Randomized Controlled Trials Experimental Box: Strategies for Enhancing Credibility and Social Justice. *New Directions for Evaluation*, 138(Summer), pp.49–60.
- Hill, J.O., 2006. Understanding and addressing the epidemic of obesity: An energy balance perspective. *Endocrine Reviews*, 27(7), pp.750–761.
- Hill, J.O., Wyatt, H.R. & Peters, J.C., 2012. Energy Balance and Obesity. *Circulation*, 126(1), pp.126–132.
- HSCIC, 2014. *Health Survey for England - 2013: Trend tables*, HSCIC.
- Hsieh, H.-F. & Shannon, S.E., 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, 15(9), pp.1277–1288.
- Hutchesson, M.J. et al., 2016. Enhancement of Self-Monitoring in a Web-Based Weight Loss Program by Extra Individualized Feedback and Reminders : Randomized Trial Corresponding Author : *Journal of Medical Internet Research*, 18(4), pp.1–11.
- Imbens, G.W. & Wooldridge, J.M., 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), pp.5–86.
- Imison, C. et al., 2016. *Delivering the benefits of digital health*, London.
- John, L.K. et al., 2011. Financial incentives for extended weight loss: A randomized, controlled trial. *Journal of General Internal Medicine*, 26(6), pp.621–626.
- Johnson, F. & Wardle, J., 2011. The association between weight loss and engagement with a web-based food and exercise diary in a commercial weight loss programme: a retrospective analysis. *The international journal of behavioral nutrition and physical activity*, 8(1), p.83.
- Jolly, K., Lewis, A. & Nsper, N., 2011. Comparison of range of commercial or primary care led weight reduction programmes with minimal intervention control for weight loss in obesity : Lighten Up randomised controlled trial. *British Medical Journal*, 343, pp.1–16.
- Kahneman, D., 2003. Maps of Bounded Rationality : Psychology for Behavioral Economics †. *American Economic Review*, 93(5), pp.1449–1475.
- Kiesler, C.A. & Sakumura, J., 1966. A Test of a Model for Commitment. *Journal of Personality and Social Psychology*, 3(3), pp.349–353.
- Kinder, D.R., 2011. Campbell ’ s Ghost. In J. N. Druckman et al., eds. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, pp. 525–530.
- Kunreuther, H. & Weber, E.U., 2014. Aiding Decision Making to Reduce the Impacts of Climate Change. *Journal of Consumer Policy*, 37(3), pp.397–411.
- Laibson, D., 1997. Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, 112(2), pp.443–477.
- Laibson, D., 1994. Mental Accounts, Self-Control and Intrapersonal Principal-Agent Problem. *Mimeo*.
- Lee, D.S., 2002. Trimming for Bounds on Treatment Effects with Missing Outcomes. *NBER Technical Working Paper Series*, (June), pp.1–18.
- Levy Paluck, E., 2010. The Promising Integration of Qualitative Methods and Field Experiments. *The ANNALS of the American Academy of Political and Social Science*, 628, pp.59–71.
- Lewin, S., Glenton, C. & Oxman, A.D., 2009. Use of qualitative methods alongside randomised controlled trials of complex healthcare interventions: methodological study. *BMJ (Clinical research ed.)*, 339, p.b3496.

- Little, R.J. & Rubin, D.B., 2000. Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes : Concepts and Analytical Approaches. *Annual Review of Public Health*, 21, pp.121–145.
- Liu, P.J. et al., 2014. Using behavioral economics to design more effective food policies to address obesity. *Applied Economic Perspectives and Policy*, 36(1), pp.6–24.
- Loewenstein, G. et al., 2012. Can behavioural economics make us healthier? *Bmj*, 344(May), p.e3482.
- Loewenstein, G., 1996. Out of Control : Visceral Influences on Behavior. *Organizational Behavior and Human Decision Processes*, 65(3), pp.272–292.
- Manski, C.F., 1989. Anatomy of the Selection Problem. *The Journal of Human Resources*, 24(3), pp.343–360.
- Maxwell, J.A., 1992. Understanding and Validity in Qualitative Research. *Harvard Educational Review*, 62(3), pp.279–300.
- Mcclure, S.M. et al., 2004. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science*, 306(October), pp.503–508.
- McDermott, R., 2011. Internal and external validity. In J. N. Druckman et al., eds. *Cambridge Handbook of Experimental Political Science*. Cambridge Books Online, pp. 27–40.
- Miller, D.T. & Prentice, D.A., 2013. Psychological Levers of Behaviour Change. In E. Shafir, ed. *The Behavioral Foundations of Public Policy*. Princeton University Press, pp. 301–309.
- Moody, A., 2013. Adult anthropometric measures, overweight and obesity. *Health Survey for England 2012*, Chapter 10, pp.1–39.
- Ntuk, U. et al., 2014. Ethnic specific obesity cut-offs for diabetes risk: cross-sectional study of 490,288 UK Biobank participants. *Diabetes Care*, 37(9), pp.2500–2507.
- Nyer, P.U. & Dellande, S., 2010. Public Commitment as a Motivator for Weight Loss. *Psychology & Marketing*, 27(1), pp.1–12.
- O’ Donoghue, T. & Rabin, M., 1999. Doing it now or later. *American Economic Review*, 89(1), pp.103–124.
- O’Cathain, A. et al., 2010. Three techniques for integrating data in mixed methods studies. *Bmj*, 341(sep17 1), pp.c4587–c4587.
- O’Cathain, a et al., 2013. What can qualitative research do for randomised controlled trials? A systematic mapping review. *BMJ open*, 3(6).
- Oliver, A. & Ubel, P., 2014. Nudging the obese: a UK-US consideration. *Health economics, policy, and law*, (April), pp.1–14.
- Paloyo, A.R. et al., 2014. the Causal Link Between Financial Incentives and Weight Loss: an Evidence-Based Survey of the Literature. *Journal of Economic Surveys*, 28(3), pp.401–420.
- Parrott, R.L. et al., 1998. Communicating to Farmers About Skin Cancer. The Behavior Adaptation Model. *Human Communication Research*, 24(3), pp.386–409.
- Peterson, N.D. et al., 2014. Dietary self-monitoring and long-term success with weight management. *Obesity (Silver Spring, Md.)*, 22(9), pp.1962–7.
- Pigou, A.C., 1932. Desires and Satisfactions. In *The Economics of Welfare*. London: Macmillan and Co.
- Prestwich, A. et al., 2012. Randomized controlled trial of collaborative implementation intentions targeting working adults’ physical activity. *Health Psychology*, 31(4), pp.486–495.
- Relton, C., Strong, M. & Li, J., 2011. The ‘ Pounds for Pounds ’ weight loss financial incentive scheme : an evaluation of a pilot in NHS Eastern and Coastal Kent. *Journal of Public Health*, 33(4), pp.536–542.

- Robinson, C., 2012. Healthy Foundations Segmentation. In *Health Survey for England 2011*. HSCIC, pp. 1–38.
- Rogers, T. et al., 2015. Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science and Policy*, 1(2), pp.33–41.
- Rogers, T., Milkman, K.L. & Volpp, K.G., 2014. Commitment devices: using initiatives to change behavior. *JAMA : the journal of the American Medical Association*, 311(20), pp.2065–6.
- Royer, H., Stehr, M. & Sydnor, J., 2015. Incentives, Commitments and Habit Formation in Exercise: Evidence from a Field Experiment with Workers at a Fortune-500 Company. *American Economic Journal: Applied Economics*, 7(3), pp.51–84.
- Ruhm, C.J., 2012. Understanding overeating and obesity. *Journal of Health Economics*, 31(6), pp.781–796.
- Schelling, T.C., 1984. Self-Command in Practice, in Policy, and in a Theory of Rational Choice. *The American Economic Review*, 74(2), pp.1–11.
- Schulz, K.F., Altman, D.G. & Moher, D., 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. *Open Medicine*, 4(1), pp.E60–E68.
- Smith, A. et al., 2011. *The Healthy Foundations Lifestage Segmentation Model: Research Report No 2: The qualitative analysis of the motivation segments*,
- Smith, A., 1790. The Theory of Moral Sentiments. In *The Theory of Moral Sentiments*. London: A. Millar.
- Smith, R.M., 2002. Should We Make Political Science More of a Science or More about Politics? *PS: Political Science and Politics*, 35(2), pp.199–201.
- Spiegelman, B.M. et al., 2001. Obesity and the Regulation of Energy Balance. *Cell*, 104(4), pp.531–543.
- Starr, M.A., 2014. Qualitative and Mixed-Methods Research in Economics: Surprising Growth, Promising Future. *Journal of Economic Surveys*, 28(2), pp.238–264.
- Strotz, R.H., 1955. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3), pp.165–180.
- Stubbs, R.J. et al., 2015. Weight outcomes audit in 1.3 million adults during their first 3 months' attendance in a commercial weight management programme. *BMC Public Health*, 15:882, pp.1–13.
- Sutton, R., 2012. Adult anthropometric measures, overweight and obesity. *Health Survey for England 2011*, Chapter 10, pp.1–39.
- Tangney, J.P., Baumeister, R.F. & Boone, A.L., 2004. High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of personality*, 72(2), pp.271–324.
- Tansey, O., 2007. Process Tracing and Elite Interviewing: A Case for Non-Probability Sampling. *PS: Political Science and Politics*, 40(4), pp.765–772.
- Tarozzi, A. et al., 2009. Commitment Mechanisms and Compliance with Health-Protecting Behavior: Preliminary Evidence from Orissa, India. *American Economic Review*, 99(2), pp.231–235.
- Tarrow, S.A., 2010. Bridging the Quantitative-Qualitative Divide. In H. E. Brady & D. Collier, eds. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield, pp. 101–110.
- Tauchmann, H., 2014. The Stata Journal. *The Stata Journal*, 14(4), pp.884–894.
- Thaler, R.H., 2016. Behavioral Economics : Past , Present , and Future †. *American Economic Review*, 106(7), pp.1577–1600.
- Thaler, R.H. & Benartzi, S., 2004. Save More Tomorrow™: Using

- Behavioral Economics to Increase Employee Saving. *Journal of Political Economy*, 112(S1), pp.S164–S187.
- Thaler, R.H. & Shefrin, H.M., 1981. An Economic Theory of Self-Control. *Journal of Political Economy*, 89(2), pp.392–406.
- Thaler, R.H. & Sunstein, C.R., 2003. Libertarian Paternalism. *American Economic Review*, 93(2), pp.175–79.
- Thaler, R.H. & Sunstein, C.R., 2008. *Nudge: Improving decision about health, wealth and happiness*, London: Penguin Books.
- Thalheimer, W. & Cook, S., 2002. *How to calculate effect sizes from published research: A simplified methodology*,
- Torgerson, D.J. & Torgerson, C.J., 2008. *Designing Randomised Trials*, Basingstoke: Palgrave Macmillan.
- Valente, C., 2011. Household Returns to Land Transfers in South Africa : A Q-squared Analysis. *Journal of Development Studies*, 47(2), pp.354–376.
- Verhoeven, A.A.C. et al., 2013. Less is more : The effect of multiple implementation intentions targeting unhealthy snacking habits. *European Journal of Social Psychology*, 43(August), pp.344–354.
- Della Vigna, S. & Malmendier, U., 2006. Paying Not to Go to the Gym. *The American Economic Review*, 96(3), pp.694–719.
- Volpp, K.G. et al., 2008. Financial Incentive–Based Approaches for Weight Loss. *JAMA: The Journal of the American Medical Association*, 300(22), pp.2631–2637.
- Wansink, B., 2013. Turning Mindless Eating into Healthy Eating. In E. Shafir, ed. *The Behavioral Foundations of Public Policy*. Princeton University Press, pp. 310–328.
- Wansink, B., Hanks, A.S. & Kaipainen, K., 2016. Slim by Design : Kitchen Counter Correlates of Obesity. *Health Education & Behavior*, 43(5), pp.552–558.
- White, H., 2013. The Use of Mixed Methods in Randomized Control Trials. *New Directions for Evaluation*, 138(Summer), pp.61–73.
- Wilkinson, N. & Klaes, M., 2012. *An Introduction to Behavioral Economics* 2nd ed., Palgrave Macmillan.
- Williams, B. et al., 2011. The Healthy Foundations Lifestage Segmentation: Research Report No.1: Creating the Segmentation using a quantitative survey of the general population of England. , (1), pp.1–312.
- Yu, Z., Sealey-Potts, C. & Rodriguez, J., 2015. Dietary Self-Monitoring in Weight Management: Current Evidence on Efficacy and Adherence. *Journal of the Academy of Nutrition and Dietetics*, 115(12), pp.1931-1933-1938.