Essays in Applied Microeconomics: Behaviour and Well–being

A thesis presented for the degree of

Doctor of Philosophy in Economics



Jessica Milligan

Lincoln College

University of Oxford

Trinity Term

Academic Year 2020–2021

Declaration

I certify that this thesis, submitted for the degree of Doctor of Philosophy in Economics at the University of Oxford, is my own work except where otherwise clearly stated.

I confirm that Chapter 1 is solely my own work.

Chapter 2 is co–authored with Séverine Toussaert (Department of Economics, the University of Oxford). Séverine Toussaert led the design of the experiment and the drafting of the Pre–analysis plans. I contributed to the study design and drafting of the Pre–analysis plans. I was wholly responsible for the building of the surveys in Qualtrics, the recruitment and communications with study participants, the data cleaning, preparation, analysis in Stata and for writing this thesis chapter.

I confirm that Chapter 3 is solely my own work. An earlier version was previously submitted and examined in partial fulfilment of the degree of MPhil in Economics at the University of Oxford.

This thesis is approximately 60,000 words.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Ian Crawford, for your willingness to take on a newbie Economist and unfailing patience and support as I have tackled the challenges of the DPhil. Thank you for being so cheerful, encouraging and generous with your time whilst sharing your knowledge and experience with me.

A special word of thanks goes to Dr Séverine Toussaert. What a privilege it was to work as your research assistant, and now to have you a mentor, co-author and marathon running coach. It has been an unparalleled opportunity to conduct the running experiment together: plenty of learning, hard work, highs, lows and kilometres under our trainers along the way.

My gratitude extends to the Department of Economics at the University of Oxford for having faith that a student with a degree in Modern Languages might complete a MPhil in Economics, and for the rigorous (and taxing) training provided. Thank you to the Economic Social and Research Council for the generous funding, especially the extension to support me in spending 3 months at the World Bank and then again when the nurseries closed during the pandemic. For not only funding my studies and research, but also for providing the community where I have met dear friends, I thank Lincoln College.

I am grateful to Dr Abi Adams–Prassl, Dr Jan–Emmanuel De Neve and my MPhil thesis examiners for insightful comments along the way. Your suggestions were of great help in developing the ideas presented in this work. I also thank the attendees of seminars and workshops in Oxford, and at the Centre for Economic Performance (London School of Economics), for your interest and questions.

Thank you to the inspiring and fierce women in my family: my mother, sister, cousin, aunts and grandma. You have all taught me through your examples to work hard and reach for the stars. To my father, for showing me determination and instilling in me your adventurous and competitive spirit. To my husband, for your belief in me and your partnership as we take on new challenges together. And to my son, Émile, for overflowing my world with love: what a joy to discover the world anew with you.

Thesis Abstract

This thesis explores challenges related to measuring well–being and behaviour, presents novel ways to handle such data and offers some evidence from analysing these types of data which has implications for welfare. Across the three chapters I have two main objectives. Firstly, I present and justify methods to handle data on unobservable feelings, such as satisfaction, motivation or pride. Secondly, the research questions are connected through their preoccupations with understanding human welfare.

In Chapter 1, I ask what importance (weightings) do individuals attach to different areas (domains) of life in their overall life satisfaction? In my answer I tackle a fundamental question for economists regarding reported satisfaction data: if we do not wish to assume such data are inter–personally comparable, how can we extract information from them (such as domain weightings)? Existing work uses regression: pooling data across individuals in order to identify average weightings, hence making the inter–personal comparability assumption. I instead use a Maximum Entropy criterion function, which can find a solution for each individual separately: I can identify J domain weightings from a single report of J domain satisfactions and overall satisfaction, and hence find individual–specific weightings.

In Chapter 2, jointly with Séverine Toussaert, we investigate the effectiveness of incentives that emerge from a perceived threat to one's image, triggered by another's assessment: an ego threat. We study this in the context of a road–running race. We ask runners to tell us about their time goal for a race and then tell them we doubt they can achieve their goal, or we offer them a reward (a voucher) if they achieve it, or both, or neither. We find that the ego threat message boosts goal achievement by 12 percentage points, but no effect of the voucher. We look for evidence for the mechanisms through which the message may work, and try to evaluate the welfare effect of the treatment.

In Chapter 3, I present a method to quantify the ordinal differences between two welfare measurement techniques: traditional utility–based Revealed Preference orderings versus Subjective Well–being orderings (from non–incentivized self–reports of well–being). As a proof of concept, I apply the method to an international dataset, and also outline ideal datasets to which this method could be applied for future research on quantifying differences between well–being measures.

Contents

Declaration ii

Acknowledgements iii

Thesis Abstract iv

Thesis Introduction viii

1 What drives well–being: a Maximum Entropy approach 1

1.1 Introduction . . . 4

1.2 Theory and Method . . . 12

1.2.1 Typical satisfaction data in household surveys . . . 12

1.2.2 Modelling overall satisfaction . . . 13

1.2.3 Solving the model . . . 17

1.2.4 Justification of Maximum Entropy for Subjective Well–being data . . . 21

1.2.5 Applying Maximum Entropy to the full model . . . 23

1.3 Data . . . 25

1.3.1 The British Household Panel Survey and sample selection . . . 25

1.3.2 Sample descriptives . . . 27

1.4 Results . . . 34

1.4.1 Weights and relative importance . . . 34

1.4.2 Weights and natural groupings of people . . . 39

1.5 Discussion . . . 45

2 On the motivational power of pride: a field study with participants in a road–running event* 50

2.1 Introduction . . . 52
2.2 Experimental Setup . . . 59
2.2.1 Study design . . . 59
2.2.2 Study context and set–up . . . 63
2.2.3 Sample size and balance tables . . . 66
2.3 Results . . . 68
2.3.1 Behavioural change: race performance . . . 69
2.3.2 Mechanisms . . . 73
2.3.3 Welfare . . . 78
2.4 Discussion . . . 82
2.4.1 Study findings . . . 82

3 If you're happy, can you show it? Non–parametric estimation of the consistency between revealed preference rankings and a well–being measure 90

3.1 Introduction . . . 93
3.2 Methodology . . . 100
3.2.1 Tastes, behaviours and revealed preferences . . . 100
3.2.2 Measurement of ordinal differences: UK example . . . 103
3.2.3 Measurement options . . . 104
3.2.4 Choosing a measure . . . 106
3.2.5 Benchmarking and inference . . . 107
3.3 Data . . . 108
3.3.1 Data requirements and sample selection . . . 108
3.3.2 International Comparison Program data . . . 111
3.3.3 World Database of Happiness data . . . 112
3.3.4 Data descriptives . . . 113
3.4 Results . . . 120
3.4.1 Within–country analysis . . . 120
3.4.2 Within–year analysis . . . 124
3.5 Discussion . . . 128

Thesis Conclusion 131

References 133

Appendices 142

Thesis Introduction

The three chapters in this thesis are connected by a common interest: interpreting information about human behaviour and well–being. I seek to explore the challenges related to measuring well–being and behaviour, present novel ways to handle such data and offer some evidence from analysing these types of data which has implications for welfare.

There are two main objectives across the chapters. Firstly, the presentation and justification of methods to handle data on unobservable human feelings, such as reported satisfaction levels, motivations to reach a goal or pride in an achievement. Working with such data involves consideration both of measurement issues and appropriate econometric methods to extract information from them. As detailed below, each chapter must confront these challenges in the answer to their separate research questions.

Secondly, the research questions themselves are connected by their fundamental preoccupations with understanding human welfare: in Chapter 1, I ask what matters most to individuals for their well–being? in Chapter 2, if we intervene to assist people to achieve their goals, and we are successful for some but not others, is this a good intervention? And in Chapter 3, do we get conflicting signals of when people are better or worse off, depending on our choice of measurement?

Working with subjective data: measurement and methods

In Chapter 1, I ask what importance do individuals attach to different areas of their life in their overall life satisfaction? I answer this question using a dataset containing individuals' reported satisfaction levels in both their life overall and in 8 different domains (areas) of their life, such as their house, income or social life. I model overall satisfaction as a weighted sum of these domain satisfactions, and hence the precise empirical question is: what are the weightings (importance) attached to each domain? In my answer, I directly tackle a fundamental question for economists regarding the use of such Subjective Well–being (SWB) data: if we do not wish to assume that such data are inter–personally comparable, because of heterogeneity in interpreting the response scales, how can we extract information (such as domain weightings) from such data? Existing work that estimates the weightings using regression of

overall satisfaction onto domain satisfaction, pools the data across individuals in order to identify the parameters; this relies on the assumption that satisfaction reports are indeed inter–personally comparable. In the method I present, using a Maximum Entropy (ME) criterion function, I do not need this assumption because the ME procedure finds a solution for each individual separately: I can identify 8 domain weightings from a single report of 8 domain satisfactions and overall life satisfaction. ME is a technique from Information Theory. This chapter is primarily concerned with the presentation and justification of using the ME approach when working to identify well–being related parameters using data like SWB data, when we do not wish to pool reports across participants because of inter–personal comparability concerns.

In Chapter 2, which is a joint research project with Séverine Toussaert, we investigate the effectiveness of incentives that emerge from a perceived threat to one's image, triggered by another's assessment: an ego threat. We survey participants before and after a running race. We ask runners to tell us about their time goal for the race and then tell them that we doubt they can achieve their goal, or we offer them a reward (a voucher) if they achieve it, or both, or neither. We find evidence that a message expressing doubt that a runner will meet their goal on average boosts goal achievement by 12 percentage points, but we find no evidence of an effect of the voucher. A measurement challenge is to identify the psychological channels through which the message worked and to understand why they are stronger for some groups than others. We use a set of different survey design tools to answer these questions: measures of pre–race beliefs, self reported confidence and commitment, together with post–race choices, recall questions and self–evaluations related to expressions and rejections of feelings of pride. In doing so, we look for empirical evidence to help us understand how our message provoking people to "prove us wrong" works to helps some to achieve their goals, and how they (and those who did not succeed) feel and act afterwards. In this chapter, I report the study design and results for the first data collection, but I also acknowledge the fragility of our present understanding of the psychological treatment's channels and effects. This has motivated a replication of the study, with more participants to increase statistical power and also some study design updates. In doing so, this chapter speaks directly to one of the key preoccupations of this thesis: the challenges of measuring and interpreting data on unobservable human feelings, such as belief or pride in oneself.

In Chapter 3, I present a way to quantify the difference in our understanding of welfare, if we use two different measurement techniques. It is often argued that traditional welfare economics would be greatly enriched if we had access to more comprehensive and cardinal measures of well–being. Happiness, or specifically SWB data has been proposed as an alternative. So the question arises: how different is our understanding of well–being if SWB data rather than traditional utility–theory based measures are used? To answer this question, I use consumption data across years and countries

to rank years (within–country) and countries (within–year) by their Revealed Preference (RP) relation. I use SWB data (from non–incentivized self–reports of well–being) to build equivalent rankings for the same years and countries. I then conduct tests for the similarity of the RP and SWB orderings. As with Chapter 1, the presentation and justification of a novel approach to working with such data is the key contribution of this chapter. The application of the method to an international dataset, with associated assumptions regarding the aggregation of both consumption and SWB data, is used as a proof of concept, but I also outline ideal datasets to which this ordinal comparison method could be applied for further research on how to quantify the differences between well–being measures.

Across the three chapters, I therefore engage with the scrutiny of measures and analysis of intangible feelings: well–being weights and levels in Chapter 1 and optimisim / pessism and pride / disappointment in Chapter 2. In Chapter 3, I suggest a way to quantify differences in the ordinal information carried within different measures, specifically ones that aim to measure overall well–being.

Welfare analysis with subjective data

The second common thread across the chapters is the focus on the welfare implications of each research question. In Chapter 1 the main question is what weights should be attached to different domains of life to summarize overall well–being. But the rich ME output — which is a set of weighting parameters recovered separately for each individual in my dataset — enables a wide range of secondary analyses. I take the weightings and interrogate their patterns, finding (1) gender and age heterogeneity in ordinal and cardinal domain importance and (2) evidence for natural groupings of people (according to their weightings) which are predictable by observable characteristics. These findings could enable the tailored construction of overall well–being indices for different societal groups and opportunities to target welfare policy in different domains of life for groups whose priorities differ.

In Chapter 2, we present evidence that the message treatment (where we express doubt that a runner will reach their time goal for the race) may have induced more runners to meet their goal successfully. A welfare question hence arises: is this a positive intervention? For those who otherwise might not have met their goal, do the psychological mechanisms inducing the behavioural change detract from the good feelings of success? For those that did not meet their goal, do they feel worse for failing given they were unable to prove us wrong? We consider two measures related to welfare — decision to donate vouchers to charity, and reported pride — to evaluate the welfare costs or gains to runners who participated in the study. Overall, we find evidence that those who met their goal have slightly lower levels of self-reported pride if they were exposed to the challenging message. Those who had this message were also more likely to donate the value of the vouchers they won to charity than

those who had not seen the message doubting their achievement: evidence of possible negative image concerns associated with having been challenged. The results are suggestive that the self–pride gains of achieving a goal may be cancelled out if you had to prove someone wrong. Note that the sample sizes in these comparisons are small, and so effect sizes may not be robust, which is in part the reason for the planned replication of the study. In particular, the replication will facilitate robust investigation of heterogeneity as well as inclusion of additional measures of welfare, so that we can interrogate this aspect of the research in greater depth.

In Chapter 3, when I apply the proposed method to country–level data, I find that within–country RP and SWB orderings are consistent over time, whilst within–year (cross country) orderings are not. This divergence in ordinal rankings when cross–country welfare is examined points to a potential hurdle in using SWB measures to compare international welfare: the results in this chapter speak to a possible cultural difference in interpreting SWB scales and hence evaluating differences between countries.

Outline

In Chapter 1, I begin by highlighting the importance of investigating the weightings of different domains of life in overall well–being. I then suggest why reported satisfaction data may not be comparable across people, which motivates my alternative method to answer the question. In Section 1.2 I lay out the proposed model for reporting SWB: an individual's overall satisfaction in each period is the sum of their domain satisfactions, each weighted by its relative importance. I lay out the theory of ME, provide a simple example of how ME can be applied to SWB data, suggest its advantages in the SWB context, and finally specify the precise method used to implement the ME procedure in the panel dataset used. Section 1.3 describes the data used for the analysis, and discusses the choice of procedure for dealing with missing domain satisfactions, which are fairly common in the dataset. Section 1.4 first describes the overall results and then highlights the gender and age heterogeneity in ordinal and cardinal domain importance. I then present the results of further analysis: evidence for natural groupings of people (according to their weightings), which are predictable by observable characteristics. In Section 1.5, I discuss these results and suggest next steps for this research.

Chapter 2 begins with an outline of our general research question: what is the motivation that emerges from perceived threats to one's (self or social) image triggered by others' assessments. We discuss the application to our particular context of road running races. In Section 2.2 we present the experimental set–up: we outline the design of the study, discuss the treatment messages and associated ethical considerations, give details on the context, recruitment and timeline, and explain the choice of sample size. In Section 2.3, we first present the key analysis of the effect of the message treatment on

goal attainment for our study participants. We then explore possible mechanisms through which the message may have had an effect, or not had an effect, and finally try to quantify the welfare impact of participation in the study. In Section 2.4 we discuss the study findings and in doing so motivate the replication of the study.

In Chapter 3, I outline the differences in style and conceptual understanding of the well–being that utility–based measures and SWB based measures have, which motivates the research question: how different is our understanding of well–being when using one approach versus the other. In Section 3.2 I present a test of cyclical consistency in choice data — the Generalized Axiom of Revealed Preference (GARP); this test is used to establish whether the consumption dataset provides evidence for countries having stable preferences over time, and whether preferences may be homogeneous across countries. I then set out the method used to generate the ordinal comparison results: I discuss potential measures of difference between the two preference orderings, and choose the Kendall τ statistic. Section 3.3 describes the dataset. Section 3.4 reports the results. The GARP tests find evidence that each country has stable preferences over time, and likewise evidence that countries have homogenous preferences in each time period. The comparison of the RP orderings with SWB orderings per country find that these are similar; in the within–year analysis I find slightly weaker evidence of similarity. Section 3.5 discusses potential interpretations of these results and suggests that one explanation could be that there are different interpretations of the SWB scale by different nations. I outline potential extensions for this research.

In the conclusion to the thesis, I connect the findings across the chapters, and state the key next steps for this work.

Chapter 1 | What drives well–being: a Maximum Entropy approach

Chapter Abstract

This chapter applies a new method to reported satisfaction data to answer the question: what importance (weighting) does an individual attach to different life domains when reporting their satisfaction overall? I model overall life satisfaction as a weighted sum of domain satisfactions; using a Maximum Entropy (ME) criterion function I can identify 8 domain weightings from a single report of 8 domain satisfactions and overall life satisfaction. Other work, in estimating the weightings using regression of overall satisfaction onto domain satisfaction, pools the data across individuals in order to identify the parameters; this relies on satisfaction reports being inter–personally comparable which is a contentious assumption. I do not need this assumption, as the ME procedure finds a solution for each individual separately. This rich output enables a wide range of secondary analyses. I take the weightings and interrogate their patterns, finding (1) gender and age heterogeneity in ordinal and cardinal domain importance and (2) evidence for natural groupings of people (according to their weightings) which are predictable by observable characteristics. These findings could enable the tailored construction of overall well–being indices for different societal groups and opportunities to target welfare policy in different domains of life for groups whose priorities differ.

Chapter Notation

- $i \in \{1, ..., n\}$ = individual
- $j \in \{1, ..., J\}$ = life domain *("he" health; "in" income; "ho" house; "pa" partner; "jo" job; "so" social life; "la" amount of leisure time; "ls" how leisure time is spent)*
- $t \in \{1, ..., T\}$ = time period
- $U_{i,t} \in \{1, ..., 7\}$ = *reported* overall life satisfaction, subject to a quantisation (rounding) error
- $u_{i,t,j} \in \{1, ..., 7\}$ = *reported* satisfaction in domain j, subject to a quantisation (rounding) error
- $\omega_{i,t,j} \in [0, 1]$ = quantisation (rounding) parameter
- $\phi_{i,t,j} \in \{-\frac{1}{2}, \frac{1}{2}, 1\}$ = re–scaler for quantisation (rounding) parameter $\omega_{i,t,j}$
- $\epsilon_{i,t,j}$ = quantisation (rounding) error, rescaled from parameter $\omega_{i,t,j}$
- $U^*_{i,t} = U_{i,t} + \epsilon_{i,t} \in [1, 7]$ = *true* overall life satisfaction
- $u^*_{i,t,j} = u_{i,t,j} + \epsilon_{i,t,j} \in [1, 7]$ = *true* satisfaction in domain j
- $p_{i,j} \in [0, 1]$ = weighting (importance) on satisfaction in domain j in overall satisfaction
- $h(\cdot)$ = entropy, a measure of information or uncertainty

1.1 | Introduction

Motivation: A new method for satisfaction data

Well–being is multi–dimensional. An important research avenue is how to aggregate its relevant dimensions into a summary well–being indicator.[1] Devising and collecting indicators of subjective well–being (SWB) across a range of domains of life is a first step. The question then arises: how important are these domains relative to each other? That is, what weights should be attached to different domains of life to summarize overall well–being? Understanding these weightings — and particularly any differences across people — has potential for policy: (1) informing the construction of well–being indices whose weightings for domains are sensitive to sub–populations' characteristics and (2) proposing well–being policies for different sub–populations, to obtain maximum return on investment by targetting them to the areas of life that matter most for each group.

Micklewright comments that for summary indicators the default of equal weightings is hard to resist, but agreement on this is far from unanimous (2001).[2] Researchers look to empirical evidence to justify alternative choices of weightings. Benjamin et al. design and implement a stated preference survey and pool respondents' answers to choice–scenario questions in order to estimate relative marginal utilities of various SWB components (2014). Van Praag, Frijters and Ferrer–i–Carbonell regress overall satisfaction onto domain satisfactions, and interpret the size of the regression coefficient on each domain as its relative importance for predicting overall satisfaction (2003).

Empirical evidence for estimations of domain weightings is often collected by asking a question such as "How satisfied are you with [X]?" and providing a 1 to 7 scale for the respondent to circle the number reflecting their satisfaction level with the well–being dimension X.[3] The key consideration when using such data is: are responses inter–personally comparable, given that respondents are each left to make their own interpretation of the scale? Regression of overall satisfaction onto domain satisfactions to estimate weightings (the regression coefficients) implictly assumes that the data are inter–personally comparable because it involves pooling responses across respondents to be able to determine the parameters.

[1] This is highlighted as an important research avenue by well–being economists, for instance see Diener et al. (2017) or Stiglitz, Sen and Fitoussi (2009). Dolan and White (2007) and Clark et al. (2019) call for policy makers to make better use of multi–dimensional indicators in public policy, rather than indicators designed to measure overall well–being directly.

[2] For example, the Human Development Index (HDI) assigns equal weights to the income, literacy and life expectancy components to obtain the summary HDI indicator. The Micklewright comments are within the recorded discussion at a CASE LSE Workshop: *A discussion of approaches to monitor the Government's strategy to tackle poverty and social exclusion*, recorded in Hills (2001). The manuscript of the paper is in Florence (not directly consulted for this Chapter).

[3] For the precise wording of such questions, see the text in Section 3 of this chapter.

This chapter presents a method that recovers appropriate weights for a set of domains of life within total SWB *without assuming that reported satisfactions are inter–personally comparable.* As a simple analogy to demonstrate why the inter–personal comparability assumption may not be valid: if a survey respondent were asked to report the temperature of their environment, and were given a unitless scale of -500 to +500, they might interpret the units as Kelvin (K), Celsius (^oC), Fahrenheit (^oF), Rankine (^oRa) or Réaumur (^oRe), and report their measurement accordingly. Room temperature is typically 21^oC which is 69.8^oF, $294.15K$, 529.47^oRa and 16.8^oRe: if the group of respondents used a variety of units, but all responses were assumed to be in the same unit, the responses at typical room temperature would appear to be extremely different actual temperatures. If the researcher wanted to compare differences across respondents, a difference of 1 unit could be 1 degree change in any of the 5 units, which does not measure the same jump in temperature: a change of 1^oC is a change of $1K$, $\frac{9}{5}^oF$, $\frac{9}{5}^oRa$ or $\frac{4}{5}^oRe$.

Returning to the satisfaction survey context, when there are no common units at all of "degrees of satisfaction" and indeed where the best (scale item 7) and worst (scale item 1) satisfaction levels are imagined differently across individuals, the situation is even more complicated. It is as if a respondent were asked to write down in any temperature unit the coldest and hottest rooms they imagine could exist, and asked to assess their own room in relation to these. Given this, it is hard to justify the assumption of inter–personal comparability of satisfaction reports.[4]

This chapter therefore proposes a new method to use reported satisfaction data without requiring the assumption of inter–personal comparability: I ask what weightings does *each individual* attach to their satisfactions in different life domains when they report their satisfaction with life overall? I recover the weightings per–person, and therefore do not pool responses across people, avoiding the need for comparability.[5] This method uses the Maximum Entropy (ME) criterion function to identify the weights, together with the simple modelling assumption that overall satisfaction is the weighted arithmetic mean of domain satisfactions.

Once I have the per-person weightings, I use patterns across the weightings to identify different groups of individuals. Individuals within each group share some observable characteristics. This information could help targeting welfare policies to particular groups within societies. For example, I find a group whose common observable traits include that they are not likely to be working, they do not have children nor a co–habiting partner: this group put over half the available weight on Social life.

[4] See the more detailed discussion below on the inter–personal comparability or not of satisfaction reports, in this introduction's subsection on literature and contribution.

[5] Note that the method does pool responses across domains, and hence I am assuming cross–domain comparability of satisfaction per individual. Also, the version of the method used in this chapter pools responses for an individual across time, and hence does assume inter–personal comparability for an individual. There is more on these assumptions in Section 2.

For these individuals, interventions to enhance the opportunities for social gatherings and meet–ups might provide the best return on overall well–being. Individuals in the group which are Leisure–activities focused, who are also likely to be in single–person household and be older, would benefit from investment into community groups offering leisure–activities suitable for the elderly.

Overview of the method

I use an unbalanced panel dataset extracted from the British Household Panel Survey (BHPS) which includes reported satisfaction in 8 domains of life and in life overall.[6] Each individual reports satisfactions in up to 8 periods in the panel. I assume that each individual has a set of 8 domain weightings that are stable over time. That is, they attach the same importance to each domain of life each time they evaluate their overall life satisfaction. Then in each period, an individual's overall satisfaction is the weighted arithmetic mean of their domain satisfactions.[7] This model is very similar in approach to the conceptual framework proposed by Benjamin et al., who construct an index of Personal Well–Being (PWB) made up of "fundamental aspects of well–being" with each aspect being weighted by the individual's estimated marginal rate of substitution for that aspect (2014).[8]

Specifically, for each individual ($i \in \{1, \ldots, n\}$), the goal is to identify their set of stable weighting parameters $p_{i,j}$ (one for each domain, $j \in \{1, \ldots, J\}$), such that when these are multiplied by their true satisfaction in period t, $u^*_{i,j,t}$, the sum equals their true overall satisfaction in period t, $U^*_{i,t}$:

$$U^*_{i,t} = \sum_j p_{i,j} u^*_{i,j,t}$$

The sum of the weights is one, $\sum_j p_{i,j} = 1$, and the weights are all non–negative, $p_{i,j} \geq 0 \; \forall j$. Hence, there are $T+1$ equations in this system, plus J inequality constraints on the weights. When $T+1 < J$, recovering the importance attached to each domain of life when reporting overall SWB *per individual* is an ill–posed inverse problem: there are an infinite number of solutions.[9]

The weights are non–zero numbers that sum to one, and in all respects they are like probabilities.[10] It

[6] See Section 3 for the full description of the BHPS subsample used in this chapter. Figure 1.3 gives an overview of the selection criteria for inclusion in the subsample.

[7] In the full model, I also assume that when an individual reports their satisfaction, this report is subject to a quantisation error due to the fact that the true satisfaction level (any real number) has to be reported as a whole number on the 1 to 7 scale. The full details of the model and discussion of the assumptions are in Section 2.

[8] Note that in this weighted mean all domains count in the aggregation. There are of course plausible alternatives, for instance a *min max* procedure, if an individual's overall well–being were driven by their feelings in the domain in which they are most unhappy.

[9] T is the number of periods that the individual makes satisfaction reports. In the BHPS, an individual reports satisfactions in up to 8 periods, hence $1 \geq T \geq 8$.

[10] They satisfy the three Kolmogorov axioms: non–negativity (each domain weight is non–negative), the probability over the event space is one (summing over all the domain weights gives us a total of one), and the probability of two independent

is possible therefore to re–think this problem of choosing the weights as isomorphic to the problem of choosing a probability distribution. To choose one set of weightings (i.e. one probability distribution) out of the infinite set that could solve the problem requires a principled way of choosing the "best" one. I use the principle of Maximum Entropy (ME).[11] I briefly outline the ME principle below, and then give detailed information on the principle, the justification of its application to this context and the precise method in Section 2.

The ME principle looks for a solution which conserves the information that is available and does not add any additional information: choose weightings that both solve the weighted sum and also best reflect the level of uncertainty remaining after the individual's reports. The uncertainty in a system is measured by *entropy*. In terms of a probability distribution over potential outcomes, the greatest uncertainty occurs when there are equal probabilities on every outcome: *maximum entropy* is achieved when there is no information about which outcome is more likely than another. If all events are equally likely, all component probabilities are equal (uniform distribution): ME is the "least informative'" probability distribution, because we could not be any more uncertain about an outcome. The ME criterion can be interpreted as an extension of LaPlace's Principle of Insufficient Reason: that in the absence of any relevant evidence, agents should distribute their credence (or 'degrees of belief') equally among all the possible outcomes under consideration. Shannon's measure of entropy, used in this chapter, is indeed maximised when all outcomes are equally likely and minimised when the probability distribution is "most informative": i.e. it is known that one particular outcome will occur with probability 1 (the distribution is degenerate on a particular outcome). In Bayesian terms the uniform (ME) distribution is the simplest non-informative prior. In the SWB problem, the most uncertain distribution of weights across the domains is the one that attaches equal weight to each domain. When there are 8 domains, this would therefore be $\frac{1}{8}$ on each domain: the criterion function selects the set of weights, $p_{i,j}$, that keeps each weight as close to $\frac{1}{8}$ as possible, whilst also adhering to the constraint that overall satisfaction is a weighted sum of the domain satisfactions in every period.

Once I have the ME weightings for each individual, I combine these with information about the individual and their household such as age, household size or income. Whilst the satisfaction reports are not assumed to be inter–personally comparable across people, once I have the weightings, these are inter–personally comparable and hence differences in weighting patterns for people with different characteristics can be investigated using regression and other techniques. In the second part of Section 4 I use k–means clustering analysis to investigate whether there are groups of individuals who attach

events is the sum of the probabilities (the combined weighting on domains j and $j+1$ is simply $p_j + p_{j+1}$.) Intuitively, we can think of the set up as like a binomial distribution, where each domain's satisfaction level (the event) has an associated weighting (probability).

[11] Note that if $T + 1 \geq J$, there is either no solution, or there is a unique solution, and the ME principle will find that solution.

similar patterns of weightings to each other, and then describe the characteristics of these groups.

Literature and Contribution

The application of the ME principle to SWB data is, as far I know, novel. The ME approach has previously been used in empirical economics research, applied to problems where partial or incomplete data result in trying to recover too many parameters relative to the data available: these are known as ill–posed problems.[12] Golan, Judge and Karp show how to apply ME to estimate two types of dynamic problems, and illustrate the method with Monte Carlo experiments (1996). Golan outlines the use of information theory methods in econometrics (2006).

While the application of the ME principle to SWB data is novel, investigating the determinants of overall SWB is an extremely well–established and growing vein of literature. Dolan, Peasgood and White provide a comprehensive review of the findings of the literature on the determinants of SWB from 1990 to 2006 (2008): these papers use regression — with varying functional forms, regression techniques, and controls — to identify the coefficients on the drivers of well–being. The overall message of the review is that *on average* poor health, separation, unemployment and lack of social contact are associated with low SWB.[13] This is aligned with the average results in this paper: I find that on average the domains of health and social life are most highly weighted. The comparison of the ME findings in this chapter to the body of research on the SWB determinants is discussed in Section 5.

A key characteristic of the ME approach in this chapter is that I obtain a personalised set of weightings for every individual: the method allows for unrestricted individual–level heterogeneity. Senik highlights the fact that the drawback of existing work is the lack of controls for unobserved heterogeneity (2005). But, there is a strand of research that looks specifically for heterogeneity in the drivers of SWB and there is a range of methods employed to do so. For example, Winkelmann and Winkelmann allow for intercept heterogeneity, motivated by the "anchoring" effect on SWB reports that stems from individuals' optimism / pessimism in assessing their life satisfaction (1998). In general, this research

[12] Solving ill–posed inverse problems can also be approached using regularisation techniques: over–fitting – recovering too many parameters — is penalised using a loss–function. In machine learning, the aim is to avoid estimating a model in the training data that fits extremely well (by means of many parameters) but that then performs poorly in the test data. McNeish describes the use of a frequently used technique, Lasso, in behavioural science data (2015).

[13] The paper reviews only those papers that used a large dataset (including the BHPS) and who had overall SWB as outcome of interest. The total number of papers reviewed is 153. Since this review, more recent work echoes these findings. For instance, Boarini et al. use Gallup World Poll data to show that health, personal security and freedom of choice are important drivers of well–being, although these have a larger impact relative to income when regressed on affect measures of SWB rather than evaluative measures such as the satisfaction questions used in this chapter (2012). Table 1 in the paper compiles effect sizes across 10 domains of life as estimated in 17 recent papers on the determinants of SWB. See Boarini et al. for details (2012, 13). Clark, Layard and Senik list the key external determinants of happiness as income, work, community and governance, and value and religion whilst key personal features are mental health, physical health, family experience, education, and gender and age (2012).

controls for intercept heterogeneity using fixed or random effects models.[14] Closer to the approach in this chapter, Clark et al. allow for *slope* heterogeneity (2005): this paper estimates for whom does money buy the most happiness (i.e. which groups have higher / lower marginal utility of income). The researchers identify four classes of individuals by time invariant characteristics; these groups have a different intercept and slope to estimate the income / satisfaction relationship. In a similar vein, Binder and Coad use quantile regressions and show decreasing importance of income, health and social factors on SWB overall as happiness quantiles increase (2011). Looking at subsamples, Boarini et al. find relatively little difference in drivers of SWB between men and women, but do find substantial evidence for changing priorities over the life course (2012). Again, the results of this research strand compared to the ME results in this chapter are discussed in Section 5.

In comparison to the sets of parameter estimates obtained through the use of different regression techniques, when applied to the subsample of BHPS data the ME principle provides a rich 22,211 x 8 matrix of weightings. This provides the opportunity to look in great detail at what matters most for whom.[15] As well as describing the overall results and highlighting heterogeneity in both ordinal and cardinal information in the distributions of weights, in Section 4 I look for patterns in the weights across the population: I use k–means clustering to identify 6 types of people, grouped by similarities in their weightings, and then present descriptive results of which group an individual may belong to using observable characteristics. As the richness of the ME output is unique to the ME approach, the use of unsupervised learning techniques such as k–means clustering on domain importance data is novel, as far as I am aware. However, the matching of the groups to characteristics of the individuals within them is compared to some of the findings from regression specifications that allow for slope heterogeneity in the determinants of SWB in Section 5.

The key motivation for using the principle of ME is that by running the procedure person–by–person, the method can avoid the assumption of inter–personal comparability of SWB data. It is therefore well–suited to such data, because whether or not SWB reports are comparable is a contentious issue. Diener, Inglehart & Tay validate the SWB scales (2013). Diener notes that such scales (1) correlate with other measures of the same concept, (2) can predict future behaviour and (3) do correlate with non–self–reported measures (2013); this suggests that the assumption of inter–personal comparability is not unreasonable. There is a recent critique in Bond and Lang who test whether happiness ordered interval data can identify two groups' relative average happiness and suggest that standard

[14] Dolan, Peasgood and White highlight the difficulty of using Fixed Effects models when some SWB determinants have little or no variability within people (2008); but Ferrer–i–Carbonell and Frijters show that controlling for individual unobserved effects can impact the findings (2004). Therefore this vein of research is challenging when using regression techniques.

[15] This is connected to regression–based research on satisfaction in life domains and connection to other characteristics. See for example Easterlin and Sawangfa, who look at patterns within domain satisfaction data and overall happiness, in relation to socio–economic status, time, age and cohort (2007).

parametric results are reversed under transformations of the ordered interval happiness reports (2019). However, Kaiser and Vendrik show this reversability does not occur in any plausible circumstances (2019). Fabian presents a way to measure scale–norming (the different interpretations of the satisfaction reporting scale in surveys) and finds that in small sample of 147 students, 15% are scale–norming (2018). This small fraction may mean that the average results from regression analysis do not suffer greatly from the assumption that 100% of respondents have comparable reports. Indeed, in Section 5, I discuss how the similarity of the averaged results in this chapter to the findings in the existing SWB literature (that does make the inter–personal comparability assumption) suggests the assumption is perhaps non–damaging to regression estimation procedures.[16]

There are drawbacks to using SWB data that are not addressed through the use of ME. Firstly, there are concerns about accuracy of measurement of well–being via self–reports. Evidence for mis–measurement is mixed. For example, Dolan and Metcalfe find evidence of focusing effects on SWB reports using an experiment before and after a soccer match (2010). Schwarz and Clore show that mood has a large influence on survey responses (1983). Likewise, external factors such as seasons may affect the SWB reports. However, Diener reports evidence for the long–term reliability of SWB reports (1984). This chapter acknowledges the concerns with SWB mis–measurement, but only seeks to address one of them: through the inclusion in the model of a quantisation error, arising from the requirement that individuals report their satisfaction on a fixed scale of integers.

With SWB domain research in general, there is discussion as to the comprehensiveness / exhaustiveness of the items considered as relevant domains. Easterlin and Sawangfa state that there is as yet no complete agreement on which domains of life are conceptually preferable, and the classification of life domains is a subject of continuing research (2007).[17] The main contribution of this chapter is the

[16] Two further comments:
(1) A work in progress by Benjamin, Debnam, Fleurbaey, Heffetz and Kimball "Correcting Subjective Well–Being Measures for Cross–Sectional Difference in Scale Use" considers corrections for SWB scale–use differences, and then presents regression results for corrected and non–corrected SWB measures on regressors for comparison. The abstract for this paper is on Jakina Debnam's website, last accessed in March 2021; I have enquired as to whether I could read the draft of this paper.
(2) There is also debate on intra–personal comparability, and inter–country comparability (Chapter 3's analysis is related to this). Considering inter–country comparability, Diener et al. review evidence for differences across cultural groups, U.S. states, and societies (2017): there are cross–cultural mean differences. In terms of intra–personal comparability, rather than adapting to shocks, it has been suggested that individuals *scale–norm*: as circumstances change over time, the meaning of the points on a person's scale change (Frederick and Loewenstein 1999). Note that in this chapter, I assume that SWB reports are intra–personally comparable as I do pool individuals' reports over time. In future work, I could also avoid the intra–personal comparison, as discussed in Section 5.

[17] Cummins provides a review on domains and suggests that SWB can be efficiently and comprehensively measured through the 7 "ComQol" (Comprehensive Quality of Life Scale) domains (Material, Health, Productivity, Intimacy, Safety, Community, Emotional) (1996). In Dolan, Peasgood and White, the authors review of the literature on the influencers of well–being and determine seven broad headings (2008): income, personal characteristics, social developed characteristics, how we spend our time, attitudes and beliefs towards self / others / life, relationships, the wider economic, social and political environment. The OECD *How's Life 2020: Measuring Well–being* report list the following ten domains: income and wealth; jobs and earnings; housing; health status; work/life balance; education and skills; social connections; civic engagement and governance; environmental quality; personal security (2020). The domains in this chapter are: health, income, job, partner,

application of the ME principle as a method to work with domain SWB reports; the addition or subtraction of domains is unlikely to effect the ordinal results of domain importance in the BHPS sample. Also, the method can be easily applied to any dataset with domain SWB reports, or indeed domain well–being indicators rather than self–reported satisfactions (these would need scaling to enable their use in the weighted sum model). Note that in this chapter, the weighted means model assumes separability of domain satisfactions. However, this is an assumption that could be relaxed and associtated weightings could still be recovered for interaction terms using the ME method outlined in this chapter.

Chapter organisation

In Section 2 I lay out the proposed model for reporting SWB: an individual's overall satisfaction in each period is the sum of their domain satisfactions, each weighted by its relative importance. I lay out the theory of ME, provide a simple example of how ME can be applied to SWB data, suggest its advantages in the SWB context, and finally specify the precise method used to implement the ME procedure in the BHPS panel. Section 3 describes the data used for the analysis, and discusses the choice of procedure for dealing with missing domain satisfactions, which is fairly common in the dataset. Section 4 first describes the overall results and then highlights the gender and age heterogeneity in ordinal and cardinal domain importance. I then present the results of further analysis: evidence for natural groupings according to weightings of types of people, predictable by observable characteristics. In Section 5, I discuss these results and suggest next steps for this research.

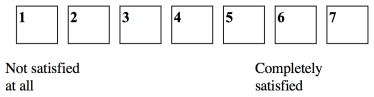
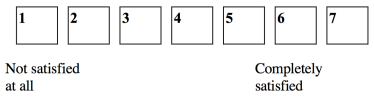
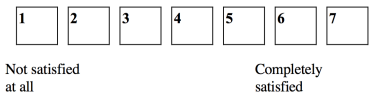
house, social life, how leisure time is spent and how much leisure time is available. These are missing the external–focused domains such as governance / environment, and self–introspective domains such as attitudes / beliefs.

1.2 | Theory and Method

This section presents the method used to recover the weights attached to different areas of life when combined into overall well–being. I first describe the typical structure of satisfaction data as collected in surveys. I then outline a simple model of how satisfaction in domains of life may be combined into an overall life–satisfaction: a weighted arithmetic mean. I assume that (1) an individual's weightings are static overtime, and (2) satisfaction reports are subject to a quantisation (rounding) error. In this model, when there are more domains than time periods, there are an infinite number of parameters that could solve the system of equations. I suggest maximising the entropy of the parameters to identify them. I lay out the theory of Maximum Entropy, demonstrate how it solves a simplified example, justify its use in the SWB context, and then specify the steps I use to apply it to the full SWB model.

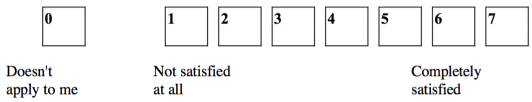
1.2.1 Typical satisfaction data in household surveys

In order to better understand well–being, household surveys collect reported satisfaction data for individuals: for example, the British Household Panel Survey (BHPS) collects data from individuals on both domain–specific and overall life satisfaction.[18] In the BHPS, the precise wording of the questions is as follows:

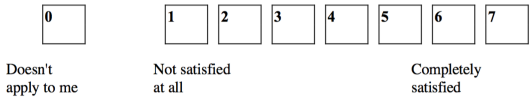
> (1) Here are some questions about how you feel about your life. Please tick the number which you feel best describes how dissatisfied or satisfied you are with the following aspects of your current situation.
>
> 1 = NOT SATISFIED AT ALL 7 = COMPLETELY SATISTFIED
>
> (a) Your health
>
> 
>
>
> (b) The income of your household
>
> 
>
>
> (c) Your house / flat
>
> 
>

[18] The BHPS was carried out by the Economic and Social Research Council (ESRC) UK Longitudinal Studies Centre and the University of Essex. This project collected survey responses on a yearly basis from 1991 to 2009 to investigate social and economic changes in Britain. There are more details on the dataset in Section 3.

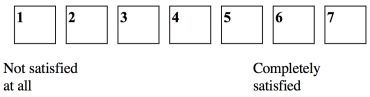
(d) Your husband / wife / partner



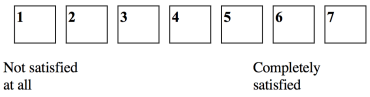
(e) Your job (if in employment)



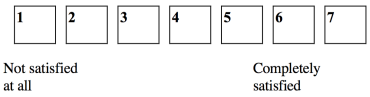
(f) Your social life



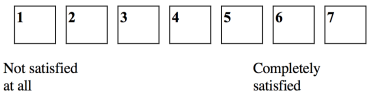
(g) The amount of leisure time you have



(h) The way you spend your leisure time



(2) Using the same scale, how dissatisfied or satisfied are you with your life overall?



Data of this sort therefore have the following structure: each person, $i \in \{1, ..., n\}$, in a set of time periods, $t \in \{1, ..., T\}$, reports their overall life satisfaction, $U_{i,t}$.[19] They also report their satisfaction in J different domains of their life, $u_{i,t,j}$. In the BHPS, individuals are asked about 8 life domains ($J = 8$): health, income, house, partner, job, social life, how much leisure time they have and how their leisure time is spent. The BHPS is an unbalanced panel and across the 18 waves, only 8 waves included questions about satisfaction in different domains of life; individuals therefore report domain satisfactions up to 8 times ($1 \leq T \leq 8$).[20]

1.2.2 Modelling overall satisfaction

I seek to identify the importance that different individuals place on the satisfaction in different areas of life when they report their overall satisfaction. With a data structure such as the BHPS in mind,

[19] I use balanced panel notation in the exposition to reduce the notational load of using T_i, which would be the notation to indicate this is an unbalanced panel, with the total time periods varying from person to person. Writing T is without loss of generality: the ME application allows for an unbalanced or balanced panel.

[20] See Section 3 for the full description of the BHPS subsample used in this chapter. Figure 1.3 gives an overview of the selection criteria for inclusion in the subsample.

I model an individual's overall satisfaction as the sum of their domain satisfactions, each weighted by its relative importance. Overall satisfaction is therefore a weighted arithmetic mean of domain satisfactions:[21]

$$U^*_{i,t} = \sum_j p_{i,t,j} u^*_{i,t,j} \tag{1.1}$$

Where: each person, i, in each time period, t, has an overall satisfaction $U^*_{i,t}$; the importance they place on domain j is $p_{i,t,j}$; $u^*_{i,t,j}$ is their satisfaction with domain j; $*$ denotes *true* satisfaction level, rather than *reported* satisfaction, which I argue below is subject to a quantisation (rounding) error due to survey response constraints. Note that this model assumes that satisfactions across each domain for an individual are comparable. This seems reasonable, as the way in which a person intuitively interprets the scale for the first question of this sort is likely to be consistent across subsequent questions.

In practice, one can think of this model as representing the process that an individual might follow when thinking about how to answer a question on their overall life–satisfaction level: (1) Think about which aspects of life matter to them (the set of domains); (2) Consider how satisfied they are with each of those aspects ($u^*_{i,t,j}$); (3) Decide how these satisfaction levels are combined into their overall life satisfaction ($p_{i,t,j}$, leading to $U^*_{i,t}$). The model is consistent with the idea that reported satisfaction is a measure of *experienced*, rather than *decision* utility. A weighted mean of satisfaction across life domains (some of which may have more immediate salience than others) makes up overall satisfaction: this combines both notions of *total–utility* (moment based) and *remembered–utility* (retrospective assessments).[22] When described via the process above, the model is a bottom–up model: life satisfaction arising as a result of satisfaction in domains of life, and this follows the flow of the survey questions (domain questions first, then overall).[23]

[21] There are three broad types of means that could have been chosen: (1) **Arithmetic**: Sensitive to extreme values / outliers. Given the responses are bounded by the satisfaction scale, extreme values / outliers do not pose a problem in this sort of data. Note however that we cannot compute this mean for those individuals for whom the $U_{i,t}$ value is outside of the set of domain–specific values $u_{i,t,j}$ (these individuals are removed from the sample: see Section 1.3 for details). If a missing domain satisfaction is assigned a satisfaction of 0, this domain simply drops out of the weighted mean, and hence $U_{i,t}$ is made up only of the non–missing domain satisfactions. Indeed, as is detailed in the following section, I decide to treat missing domain satisfactions as 0 (i.e. simply allow that domain to drop out of the equation). (2) **Geometric**: Useful if values vary exponentially or in the case of a skewed distribution. There could be an argument for using such a mean in this context: there is often a positive skew in reported satisfaction data. However, the geometric mean cannot be computed if any values are 0 or negative, and so in cases where a respondent has put "not applicable" for a particular domain, I would be forced to exclude the individual's report that year. (3) **Harmonic**: Often used when the values are ratios or rates. Useful to mitigate the impact of large values and increase the impact of small values. Here, we do not have any reason to weight domains with less satisfaction more highly than those with greater satisfaction. The same restriction as per the geometric mean applies here: inability to compute the harmonic mean should any domain be assigned a value of 0.

[22] See the Office of National Statistics (ONS) publication on methodology to measure subjective well–being, in particular the section on salience (2011). See Chapter 37 of Kahneman and Tversky for an overview of the components of experienced utility and comparison with decision utility (2000).

[23] See Heady, Veenhoven and Wearing on top–down versus bottom–up theories of SWB, where top–down takes life satisfaction overall as the influencer for the domain satisfactions (1991) .

Applying the reported satisfaction data from the BHPS panel to the model in equation (1.1), per individual there is one such life satisfaction equation in each time period, hence 8 unknown weights per time period.

Assumption 1: domain weightings are static over time

I make the simplifying assumption that the importance placed by individual i on each domain is stable over the time dimension of the panel: the solution that I am looking for is therefore a single scalar for all t. The subscript t for the weightings in equation (1.1), is therefore dropped from the model:

$$U^*_{i,t} = \sum_j p_{i,j} u^*_{i,t,j} \tag{1.2}$$

Note that this decision to identify a stable weighting over time from multiple waves of reports from an individual requires the assumption of intra–personal comparability of satisfaction reports: the individual interprets the satisfaction scale in the same manner in each wave. This seems a reasonable assumption: an individual who instinctively interprets the scale as linear, rather than logarithmic for instance, may well continue to do so each time they face a question of the same type. However, level shifts may occur, if the individual's perceived 1 or 7 (worst and best cases) alter as a result of life shocks.

This assumption is made primarily for simplicity and does not deter from the primary goal of the chapter: the recovery of a single weighting per individual still avoids the inter–personal comparability assumption. Making the assumption then enables more straightforward cross–sectional descriptives of the results that demonstrate the richness of the ME output and offer some interesting insights into patterns of weightings with individuals' characteristics (see the results in Section 4). The decision to remove variability over time can be justified by the notion that satisfaction questions are designed to measure underlying states that are reasonably consistent, stable and long–lived, as opposed to more instantaneous and volatile feelings classified as happiness (Ellison, Gay, & Glass 1989).[24] Indeed, the dataset observations are mostly only one year apart, so the assumption of stability over these small timeframes is not too strong.

In Section 5 I discuss the potential of dropping this assumption, and allowing the weights for individuals to vary over time. This would enable interesting dynamic analysis: how do relative weights change in response to life shocks? The ME procedure can recover weights that vary over time for each

[24] Indeed, when the ME procedure is run over each individual–period observation separately, the average within–person variability of weightings is smaller than the between–person variability of weightings (individuals are more similar to themselves overtime than to each other). An interesting test for further consideration of this stability assumption could be to split the sample into two halves, with four years' of data in each, run the ME procedure on the two halves separately, and compare the difference between the two weights, and the weights recovered in the full–sample analysis in this chapter.

individual, however when taken together with the addition of a quantisation error (details below) the model becomes more difficult to falsify, in the sense of it becomes easier to recover weights that adhere to all the constraints when the single set of weights need to solve the weighted mean in only one year, rather than up to eight. The model would therefore need additional constraints to render its findings more meaningful and this is left as a next step for this research agenda.

Assumption 2: quantisation error in reported satisfaction

In surveys, respondents are required to choose an integer between 1 and 7 in order to record their satisfaction. This is a constraint on true satisfaction, which could be any real number: true satisfaction levels, $u^*_{i,t,j}$ and $U^*_{i,t}$, are unobserved. When responding to these questions, respondents must calibrate the lower and upper bound of the satisfaction scale to their own worst and best possible satisfaction levels. Individuals then round the satisfaction level they feel to the nearest whole number. Hence, a reported satisfaction of 1 could stem from a true satisfaction of any real number between 1 and up to 1.5, a reported satisfaction of 2 could stem from a true satisfaction of between 1.5 and up to 2.5, and so forth, up to reported 7 which could be a true satisfaction of between 6.5 and 7. I model this as introducing a simple *quantisation* error into each reported satisfaction level, which is a standard format for rounding errors:

$$U^*_{i,t} = U_{i,t} + \epsilon_{i,t}$$
$$u^*_{i,t,j} = u_{i,t,j} + \epsilon_{i,t,j}$$

The quantisation error $\epsilon_{i,t,j}$ varies according to where on the scale the report lies (1, between 2 and 6, or 7). To achieve this, let $\omega_{i,t,j}$ be some number between 0 and 1, and let $\phi_{i,t,j}$ be a number which is $+\frac{1}{2}$ if $U_{i,t}, u_{i,t,j} = 1$; $+1$ if $U_{i,t}, u_{i,t,j} \in \{2, \ldots, 6\}$; and $-\frac{1}{2}$ if $U_{i,t}, u_{i,t,j} = 7$. Then quantisation error, $\epsilon_{i,t,j}$, is defined as follows:

$$\epsilon_{i,t,j} = \phi_{i,t,j}\omega_{i,t,j} - |\phi_{i,t,j}| + \frac{1}{2}$$

Therefore, each respondent's unobserved true satisfaction levels, $u^*_{i,t,j}$ and $U^*_{i,t}$, are made up of the reported satisfaction levels, $u_{i,t,j}$ and $U_{i,t}$, together with the scaled quantisation errors:

$$\overbrace{U^*_{i,t}}^{\text{true}} = \overbrace{U_{i,t}}^{\text{reported}} + \overbrace{\phi_{i,t}\omega_{i,t} - |\phi_{i,t}| + \frac{1}{2}}^{\text{quantisation error: } \epsilon_{i,t}} \tag{1.3}$$

$$\overbrace{u^*_{i,t,j}}^{\text{true}} = \overbrace{u_{i,t,j}}^{\text{reported}} + \overbrace{\phi_{i,t,j}\omega_{i,t,j} - |\phi_{i,t,j}| + \frac{1}{2}}^{\text{quantisation error: } \epsilon_{i,t,j}} \tag{1.4}$$

Therefore, in addition to the domain weightings, in the model I also need to identify these quantisation errors. This forms part of the Generalised Maximum Entropy model laid out later in this section, in which the parameters $\omega_{i,t,j} \in [0,1]$, which underpin the quantisation errors $\epsilon_{i,t,j}$, are treated as probabilities, and as such can be recovered by the entropy method.

1.2.3 Solving the model

An ill–posed problem

To fix ideas, first consider the unquantised problem (I will then describe the generalisation to accommodate the quantisation errors). For any satisfaction dataset that is either a cross–section or a panel with $J > T + 1$, identifying the weights is an ill–posed problem. As outlined in the introduction, the domain weights can be treated as probabilities, and hence the full problem is to choose the set of weights for each individual, i, that solve this system:

$$U_{i,t} = \sum_{j=1}^{J} p_{i,j} u_{i,t,j}$$
$$\sum_{j=1}^{J} p_{i,j} = 1$$
$$p_{i,j} \geq 0 \; \forall j$$

Figure 1.1 illustrates the problem in the simplified context (i.e. still setting aside the quantisation errors). For the individual in the figure, there are just 2 time periods and 3 life domains. This simple system can be written as:

$$4 = 3 \cdot p_1 + 4 \cdot p_2 + 5 \cdot p_3$$
$$4 = 2 \cdot p_1 + 4 \cdot p_2 + 6 \cdot p_3$$
$$p_1 + p_2 + p_3 = 1$$
$$p_1, p_2, p_3 \geq 0$$

The set of weightings that could solve the system of equations is infinite.[25] To choose which single set of weightings is "the best" solution for this problem, a criterion function is needed. The one I propose is the Maximum Entropy (ME) criterion function. I first explain its application for this simplified context without the quantisation errors, and then write out the generalised formulation of the problem, which can allow for the errors which arise from the rounding procedure.

[25] In the case where $T + 1 \geq J$, there is either no solution, or there is a unique solution.

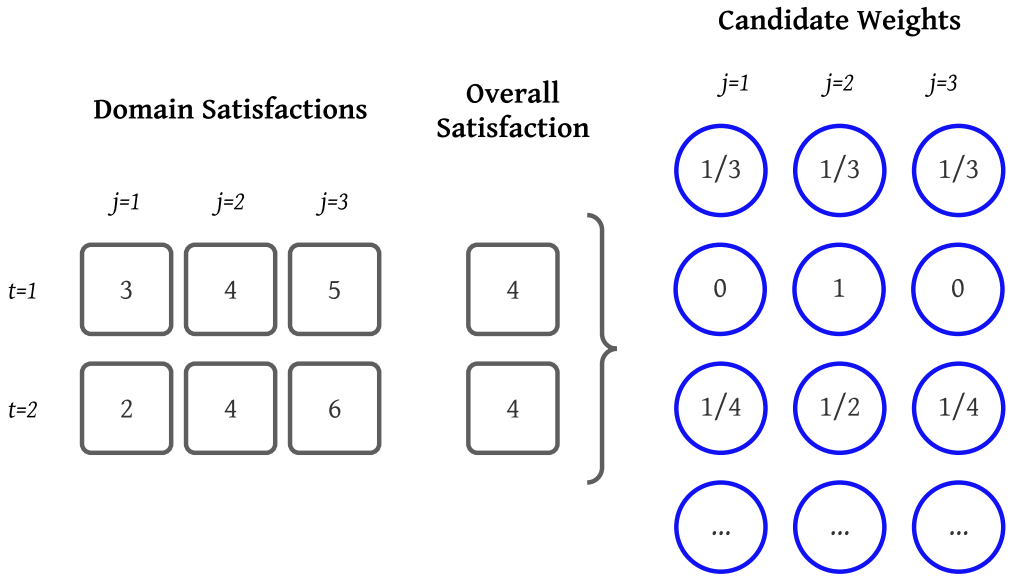


Figure 1.1: An underdetermined problem

Notes: The boxes represent satisfactions for each domain and overall; the circles are possible weightings for each domain. Three sets of candidate weights are proposed, but there are an infinite set that could solve the system of equations. The introduction of the quantisation errors has been set aside in this simplified example.

Solving ill–posed problems: Maximum Entropy

To solve partial information problems, Jaynes suggested that the unknown entities (domain weights) be expressed as probability distributions over the feasible set (1957a, 1957b). He then offered a constructive principle that enables the researcher to state that one probability distribution is "better" in a particular respect than all the others. The distribution chosen as the solution is the one that reflects the level of uncertainty that remains after having taken into account the available data (the reported overall and domain satisfactions). Jaynes' suggested approach is therefore to choose the distribution that is true to these observed data whilst avoiding making any additional informational assumptions. This approach provides a representation of the model that is consistent with what is observable, but is otherwise maximally agnostic about the parameters.

To choose a solution on the basis of its level of uncertainty requires a measure of uncertainty: this is *entropy*. Shannon provides a measure of entropy (1948):

$$h(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \tag{1.5}$$

where $P(X)$ is the probability mass function of discrete random variable X, with possible values $\{x_1, ..., x_n\}$, and $h(X)$ is the entropy of X. Shannon proposed using logarithms in base 2. The units of

entropy are *bits*. $h(X)$ is maximised when all events are equally likely hence all component probabilities are equal: maximum entropy is the "least informative" probability distribution, where there is no way of determining that any outcome is more likely than another. The entropy function is minimised where the probability distribution is "most informative": entropy has 0 bits when it is known that one particular outcome will occur with probability 1 (i.e. the distribution is degenerate on a particular outcome). Shannon's requirements of the measure of uncertainty are as follows: the measure should be continuous, symmetric, maximised when all events are equiprobable, and additively decomposable (1948).

Once uncertainty can be measured, a criterion function can then be to *maximise the uncertainty*: choose the probabilities that makes $h(X)$ as large as possible. This criterion, taken together with the constraints made available in the data, can be used as a way to select model parameter estimates. Jaynes defined the resulting ME solution as the one that is *the least biased* given the available information: he justified this approach as a way to use inductive logic to assign probabilities in a manner that is invariant to the observer.[26] The criterion can also be interpreted as an extension of LaPlace's *Principle of Insufficient Reason*: the maximally uninformative uniform distribution should be used as an estimate when we are maximally un–informed. This principle requires that the parameters are treated as probabilities, hence are non–negative and sum to one (as well as adhering to the constraints in the observed data).

Jaynes applied this criterion to a simple example now known as the Jaynes die problem. If the mean outcome after a million throws of a die is observed, the best guess for the weight on each side of the die is the one that maximises the entropy: choose the distribution of weights that does not unduly favour any particular outcome any more than the observed moment suggests. In Figure 1.2, the observed outcome is 4.5. The solution for weights on each side of the die maximises the entropy, by choosing weights as close to each other as possible: entropy is 1.61 bits for this solution (versus the maximum possible of 1.79 bits, for the case where the sides are all equally weighted).[27] The weightings selected (to 3 significant figures) keep the weights as flat as possible, lying on a smooth curve that increases from left to right, reflecting the fact that the observed moment sits to the right of the uniform expectation of 3.5. The weights also ensure that the assigned probabilities to each side of the die result in an expected outcome of 4.5. The weights sum to one, and none are negative.

The die problem illustrates that by choosing the set of weights that is as close to uniform as possible

[26] See Siendenfeld for a discussion of ME and in particular its connection with Bayesian theory (1986). For a full discussion of the ME principle, see Mittelhammer, Judge and Miller (2000).

[27] In this chapter I do not interrogate the entropy result for each individual. However, if this were an area of interest, it would be useful to normalise the entropy by the size of the support.

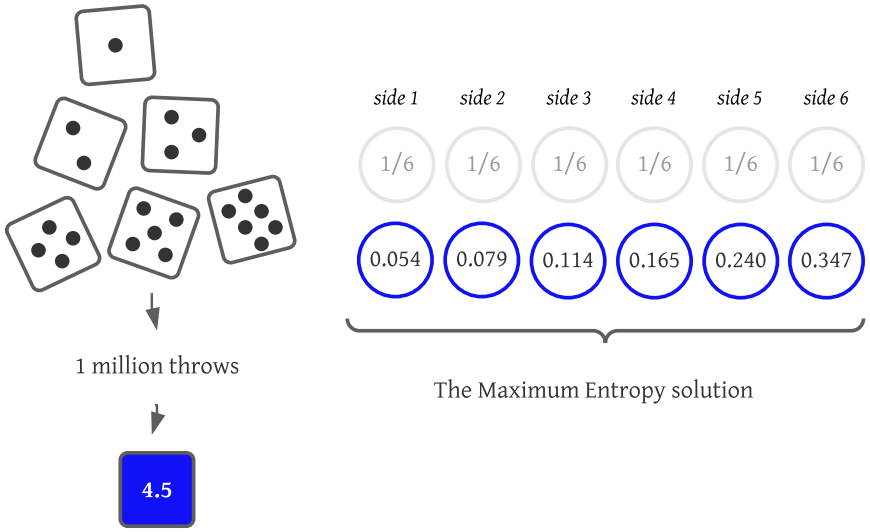


Figure 1.2: Jaynes' die problem

Notes: Suppose a die has been rolled a million times, and the average outcome is 4.5. The researcher wishes to know what weights each side of the die has. Jaynes' solution: choose the weights that are closest to the uniform distribution ($\frac{1}{6}$ on every side) whilst also matching the observed moment (the expected outcome must be 4.5). The weightings selected (to 3 significant figures) keep the weights as flat as possible, lying on a smooth curve that increases from left to right, reflecting the fact that the observed moment sits to the right of the uniform expectation of 3.5.

whilst remaining consistent with the observed moment, no side of the die is favoured more than another by any assumptions about the weights that are not supported by the available information.

Applying Maximum Entropy to the SWB context

Returning to Figure 1.1, the underdetermined satisfaction problem for an individual, the equivalence to the Jaynes die setting is as follows: the observations of domain satisfaction reports are the equivalent to the sides of a 3–sided die, and the overall life satisfaction report is the mean outcome from throwing the die. Looking for candidate weights using the ME principle in this case looks for the weights that are each as close to $\frac{1}{3}$ as possible. With 2 time periods, the set of weights must solve the arthimetic mean in both period equations. The problem can be expressed as follows:

$$\max_{\{p_1,p_2,p_3\}} h(\mathbf{p}) = -(p_1 log_2 p_1 + p_2 log_2 p_2 + p_3 log_2 p_3)$$

$$\text{s.t.:}$$

$$4 = 3 \cdot p_1 + 4 \cdot p_2 + 5 \cdot p_3$$

$$4 = 2 \cdot p_1 + 4 \cdot p_2 + 6 \cdot p_3$$

$$p_1 + p_2 + p_3 = 1$$

$$p_1, p_2, p_3 \geq 0.$$

The choice in this example is $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{3}$, reaching a maximum possible entropy of 1.10 bits. The solution stays as close as possible to the probability distribution in which an individual weights all domains equally (as per the majority of well–being indexes), unless the data tell otherwise. There are no assumptions of any domain being more or less important than another; this is like the assumption of a die being fair, until evidence says otherwise.

1.2.4 Justification of Maximum Entropy for Subjective Well–being data

The life satisfaction problem is a simple system made up of observed data, a linear model and constraints, all of which contain some information. Once the weights are re–interpreted as probabilities and the weighted sum satisfaction model is chosen, the ME principle finds a solution for this linear model which conserves the information that is available and does not add any additional information into the system. A flat distribution across the weights is the distribution that would be allocated in a simple well–being index. The ME principle for selecting the weights seems appropriate in this context. If a solution were chosen that had greater entropy — i.e. weights were selected that reflected greater uncertainty about the distribution of weights across the domains — there would not be full use of the information available in the reported satisfaction data, which do reveal something about what matters more or less to this individual when they consider their overall life satisfaction. Likewise, if a solution with less entropy is chosen — i.e. more constraints are added than those the reported satisfaction data directly impose — information is added for which there is no evidence for that individual.

The key advantage of using the approach is that it avoids the inter–personal comparability debate entirely: it can choose a solution without pooling satisfaction data across people. Thus, no matter a researcher's opinion regarding the comparability of data of this sort, the method can be acceptable as a means to extract information about well–being from satisfaction reports.

A further advantage is that ME solution to the life satisfaction problem is hugely rich, enabling a second set of analyses into the weights' relationships to individuals' characteristics, not used in the ME procedure: the weights results matrix can help determine which social, economic or environmental factors are correlated with the choice of these weights. Whilst the satisfaction scales are not unanimously agreed to be inter–personally comparable, the relative importances of one domain to another certainly are. I illustrate this potential in Section 4.

An alternative: using regression The natural alternative to using ME to solve the ill–posed problem would be to assume that the weightings do not vary across across individuals: weightings are homogeneous across people (as well as stable over time). In the simplified problem in which the quantisation errors are again put aside for the moment, regression can be used to estimate a $\hat{\rho}_j$ using Ordinary Least Squares (OLS) regression: $U_{i,t} = \sum_j \hat{\rho}_j u_{i,t,j} + \hat{\mu}_{i,t}$, where subscript i has been dropped from the weights. In taking this approach, identifying a single weighting scalar, it is possible to identify a unique solution that minimises the squared residuals.

A technical problem with the regression above is that using the original quantised responses would bias the OLS estimates and also increase the standard errors (the rounding introduces measurement error into both the true overall satisfaction and the domain satisfactions). There is also a problem with the assumption of homogeneity of the weights. Campello, Galvao and Julh show that OLS–FE (Fixed Effects) estimates, on a econometric model with homogenous slope parameters, bias the estimates if there is individual slope heterogeneity in the true model (2019).[28] Regression will not recover the average of the heterogeneous slopes.

Crucially, the key conceptual issue with this approach is that to identify $\hat{\rho}_j$, observations must be pooled across people. By doing so, inter–personal comparability of reported satisfaction must be assumed, and the primary goal of this chapter is to avoid this because of the nature of SWB data. Additionally, assuming homogeneity of weightings is not in the spirit of SWB as a *democratic* measure that acknowledges the agency of citizens and their own life–evaluations (Clark et al. 2019). Thus, if it is possible to infer from the data a set of domain weights specific to an individual, this takes us closer to the heterogeneous evaluative mechanism that each person goes through when considering their satisfaction with their lives.

Note that the ME technique recovers the weightings given a particular functional form assumption: that overall satisfaction is the weighted arthimetic mean of the domain satisfactions. This is broadly

[28] The authors' model can also be interpreted as a Random Coefficients Model, and they link this to discussion by Hsiao and Peseran (2008) and Wooldridge (2010) on biased estimates of this type of model.

equivalent to the assumption of a linear model for regression, adding a parameter restriction that summing over $\hat{\rho}_j$ must equal 1. Using regression would allow a researcher to experiment with alternative functional forms: interaction terms or restrictions across parameters. The ME technique can be similarly adapted. This could be done by formulating:

- An alternative functional form for the overall satisfaction equation, for instance allowing for interactions between particular domains, then maximising the entropy across the weightings as before (keeping the weights as close to each other as possible);

- Alternative assumptions as to the shape of the distribution across the weightings (for instance using the Kullback Leibler distance), which are then added as inequality restrictions to the weightings in the optimisation.[29]

1.2.5 Applying Maximum Entropy to the full model

Returning to the full model, which includes the quantisation errors, I apply the ME principle to the BHPS dataset in the following way: for each individual, maximise the entropy of the weightings and of the (unscaled) quantisation parameters such that (1) the weightings and (scaled) quantisation errors solve the life satisfaction equations in every period and (2) the weightings and unscaled quantisation parameters sum to one and are all non–negative. The inclusion of the quantisation errors in the model requires the use of the Generalised Maximum Entropy (GME) function, which is Golan, Judge, and Miller's generalization of Jaynes's ME criterion to include stochastic moments of the data (1996). This involves additively including the entropy of the parameters that underly the quantisation errors in the objective function: like the domain weights, the unscaled $\omega_{i,j,t}$ parameters are treated as probabilities.[30]

The full optimisation is therefore as follows. For each i,

$$\max_{\mathbf{p},\omega} h(\mathbf{p}, \omega) = -\left(\sum_{j=1}^{8} p_{i,j} log_2 p_{i,j} + \sum_{j=1}^{9}\sum_{t=1}^{T}\left(\omega_{i,j,t} log_2 \omega_{i,j,t} + (1-\omega_{i,j,t}) log_2 (1-\omega_{i,j,t})\right)\right) \quad (1.6)$$

[29] This would involve having additional inequality restrictions on the weightings (e.g. income should be the highest weighting), running the ME procedure, then comparing the two probability distributions using the Kullback Leibler distance. See Golan, Judge and Karp for an overview of this (1996).

[30] As with the domain weightings, the $\omega_{i,j,t}$ parameters can be interpreted as probabilities (they also comply with the Kolmogorov axioms). Here, for one individual in one time period, GME includes in the optimisation the unscaled quantisation parameter ω_j and $(1-\omega_j)$, hence treats these parameters as Bernouilli probabilities.

subject to the following constraints:

$$U_{i,t} + \epsilon_{i,t} = \sum_{j=1}^{8} p_{i,j}(u_{i,t,j} + \epsilon_{i,t,j}) \tag{1.7}$$

$$\sum_{j=1}^{8} p_{i,j} = 1 \tag{1.8}$$

$$\epsilon_{i,t,j} = \phi_{i,t,j}\omega_{i,j,t} - |\phi_{i,t,j}| + \frac{1}{2} \tag{1.9}$$

$$p_{i,j}, \omega_{i,j,t} \geq 0 \ \forall j, t \tag{1.10}$$

$$\omega_{i,j,t} + (1 - \omega_{i,j,t}) = 1 \ \forall j, t \tag{1.11}$$

where:

$$\omega_{i,j,t} \in [0, 1]$$

$$\phi_{i,j,t} = \begin{cases} +\frac{1}{2} & \text{for } U_{i,t}, u_{i,t,j} = 1 \\ +1 & \text{for } 2 \leq U_{i,t}, u_{i,t,j} \leq 6 \\ -\frac{1}{2} & \text{for } U_{i,t}, u_{i,t,j} = 7 \end{cases}$$

Equation (1.6) keeps $p_{i,j}$ as close to $\frac{1}{8}$ as possible for all j, and the unscaled errors $\omega_{i,j,t}$ close to $(1-\omega_{i,j,t})$, and hence close to $\frac{1}{2}$, for all j and t. Note that there is an error on the 8 domain satisfactions and on the overall satisfaction, and therefore 9 in total. The constraint in equation (1.7) states that the unquantised weighted mean domain satisfaction in each period is equal to that period's unquantised overall satisfaction. The constraint in (1.8) simply ensures the weights add up to 1. Equation (1.9) states the relationship between the parameters $\omega_{i,j,t}$ that appear in the entropy function and the scaled quantisation errors in the weighted mean constraint. Equation (1.10) requires the weights and the parameters underlying the quantisation errors to be non–negative. Equation (1.11) ensures each unscaled quantisation parameter and its Bernouilli counterpart sum to 1 in each domain and in each time period.

Once the steps are applied to each individual, GME recovers a J x n matrix of domain weightings: in the BHPS dataset there are $J = 8$ domains and $n = 22,211$ individuals, so the matrix of weightings is 8 x 22,211. The solution also contains the quantisation errors, but I treat these as nuisance parameters and do not consider them further in this chapter as they are not of direct interest. I take the weightings matrix, and incorporate individual and household characteristics that are available in the BHPS. This enables a further set of analyses to explore how the weights vary with observable data about an individual: (1) investigate heterogeneity in the distributions of what matters for well–being and (2) search for natural groupings within the whole set of individuals using clustering techniques, and use regression to try to predict these groups from observable characteristics.

1.3 | Data

In this section, I introduce the *British Household Panel Data Survey* (BHPS), and the key BHPS life and domain satisfaction questions used in this chapter. I then lay out how respondents are selected from the BHPS for my analysis and how I deal with missing domain satisfactions in the sample. Finally, I describe the socio–demographic characteristics of the individuals, and briefly compare these to the ONS data on the UK population during the time of the BHPS collection.

1.3.1 The British Household Panel Survey and sample selection

The full BHPS: The BHPS was carried out by the Economic and Social Research Council (ESRC) UK Longitudinal Studies Centre and the University of Essex.[31] This project collected survey responses on a yearly basis from 1991 to 2009 to investigate social and economic changes in Britain. 5,000 households were interviewed in the first wave, including approximately 9,000 adults. All eligible individuals within a household were invited to respond, and all were contacted for reinterview in each of the subsequent 17 waves. If a person moved out of a household, all individuals in their new household were interviewed; new members to existing households were also interviewed. The sample was designed to be nationally representative in each year: additional sub–samples were therefore added in later years to maintain a representative set of households, with just over 10,000 households having been interviewed by the end of the 18th wave. Not all individuals provided interview responses in all waves; this is therefore an unbalanced panel dataset, with an average of 4 waves of data available for each individual.

Usable BHPS Waves: Figure 1.3 illustrates this chapter's sample selection process from the full BHPS. To perform the ME procedure, I need domain–specific and overall life satisfaction reports from individuals. Questions about domain satisfactions appeared in only 8 of the 17 waves of the BHPS (1996–2000 and 2002–2004).[32] I can therefore only use responses from these 8 years for the analysis in this chapter.

Unsolvable cases: The 8 waves of the BHPS sample that collect domain satisfaction reports from survey respondents contain 25,107 individuals, with an average of 4 years of responses per individual. However, the ME method cannot find a solution for the following individuals:

(1) Individuals who did not report an overall life satisfaction level in all waves i.e. $U_{i,t}$ is missing for all t. This is the case for 4.1% of individuals.

[31] Full details regarding the BHPS are in the BHPS manual (Taylor et al. 2018).
[32] The precise wording of the satisfaction questions are as previously outlined in Section 2.

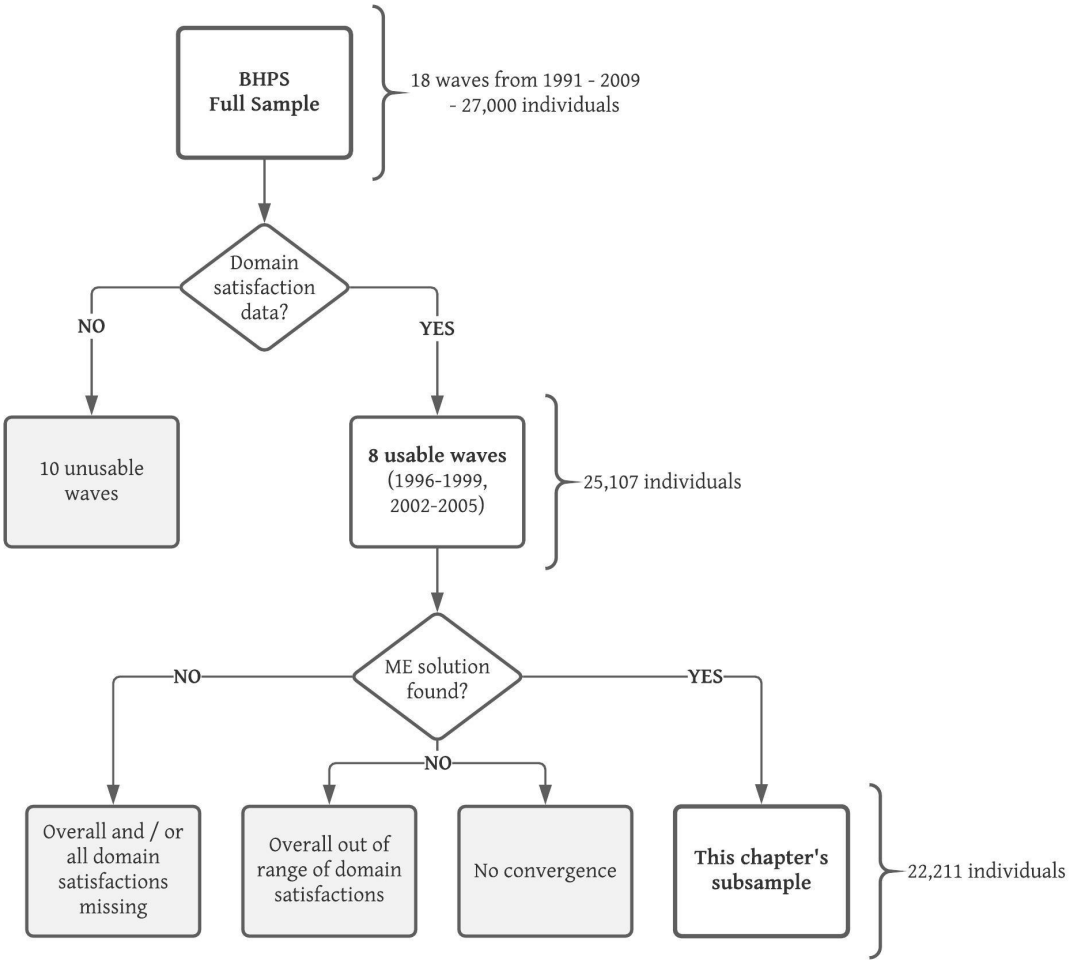


Figure 1.3: BHPS Sample

Notes: This figure demonstrates the sample selection for the paper. Only 8 of the 18 waves of the BHPS panel contain domain specific satisfaction data. The ME procedure requires (1) that the individual reports satisfaction overall and at least one domain satisfaction (i.e. does not skip all or most of the question) and (2) that the individual's overall satisfaction level is within the range of the domain satisfactions. "No convergence" occurs when the loop in R is unable to find a solution that meets all the constraints.

(2) Individuals who do not report any domain satisfactions in all waves i.e. $u_{i,t,j}$ is missing for all t and all j. This is the case for 0.1% of individuals.

(3) Individuals whose reported overall satisfaction is outside of the range of their domain satisfactions, i.e. they reported a $U_{i,t}$ that is larger than any of the $u_{i,t,j}$, or smaller than any of the $u_{i,t,j}$. These individuals could be interpreted as having a domain that was not in the survey and that this missing domain is the most highly weighted for them. Alternatively, these individuals were not considering their answers with care. Irrespective of the reason, there can be no solution in these cases. This is the case for 0.5% of individuals.

(4) Individuals for whom the ME procedure cannot find a solution: no solution is found at all, or the final solution fails to meet the constraints in one or more of the waves (the weighted sum of

domains does not equal the overall satisfaction). This is the case for 6.8% of individuals.

As a result, the final sample of respondents for whom the ME method can find domain weightings is made up of 22,211 individuals.[33]

In addition to satisfaction data, I use individual and household level characteristics in order to investigate the patterns of the domain weightings: household income, type, size and presence of children together with the individual's age, sex, employment and marital status. All of this information is available for BHPS respondents.[34] I use an associated dataset for a household income variable: the *Derived Current and Annual Net Household Income Variables, BHPS waves 1–18* (Levy et al. 2012). This dataset is a complement to the main BHPS dataset, in which a set of derived net income variables per household are calculated. The household identifiers per wave are matchable to households (and hence individuals) in the main BHPS dataset.

1.3.2 Sample descriptives

Table 1.1 presents the summary statistics for the average domain–specific and overall life satisfaction reports in the subsample. Note that these statistics are calculated using the averaged quantised satisfaction reports per person: a single average raw report per person, not the series of estimated unquantised reports each period that are used in the ME procedure. The distributions are illustrated in Figure 1.4. There is clear variation across the reports, with differences both in the means but also in the shape of the distribution for each domain. These summary statistics and distributions of satisfactions do not differ significantly from the full BHPS sample.

Missing domain satisfactions

Table 1.1 highlights that the number of individuals with a report for the Job or Partner domains is considerably smaller than for those with reports in the other domains. This is because individuals could tick "Doesn't apply to me" for these domains, and thus the satisfaction level is missing for these respondents. One would expect that the individuals who do not answer the Job satisfaction question

[33] OLS regression analysis demonstrates that an individual being dropped from the sample for any of the reasons is not significantly correlated with observable characteristics. There is a marginally significant negative correlation of health and income satisfaction with being removed from the sample due to (4): those who report lower satisfaction with their health or job are less likely to have a successful ME solution. Perhaps those feeling ill or financially stressed are less inclined to provide thought–through answers to all the domain questions that fit into the model, but this cannot be ascertained.

[34] Note that the ME procedure provides a single scalar of weights per individual; to describe differences in household or individual characteristics, I therefore collapse the panel of these characteristics into a single average value for each individual. To achieve this, I proceed as follows: (1) Average across the continuous variables, such as age, number of children or household income, to find each individual's mean level across the panel; (2) Take the modal response across the waves for the categorical characteristics, such as marital or employment status, to find the most common status for each individual across the panel. It is these mean or modal responses that are used in the weights description in the results section, and also this section's discussion on missing reports of Job or Partner satisfaction, and for comparing the sample's characteristics to UK population parameters.

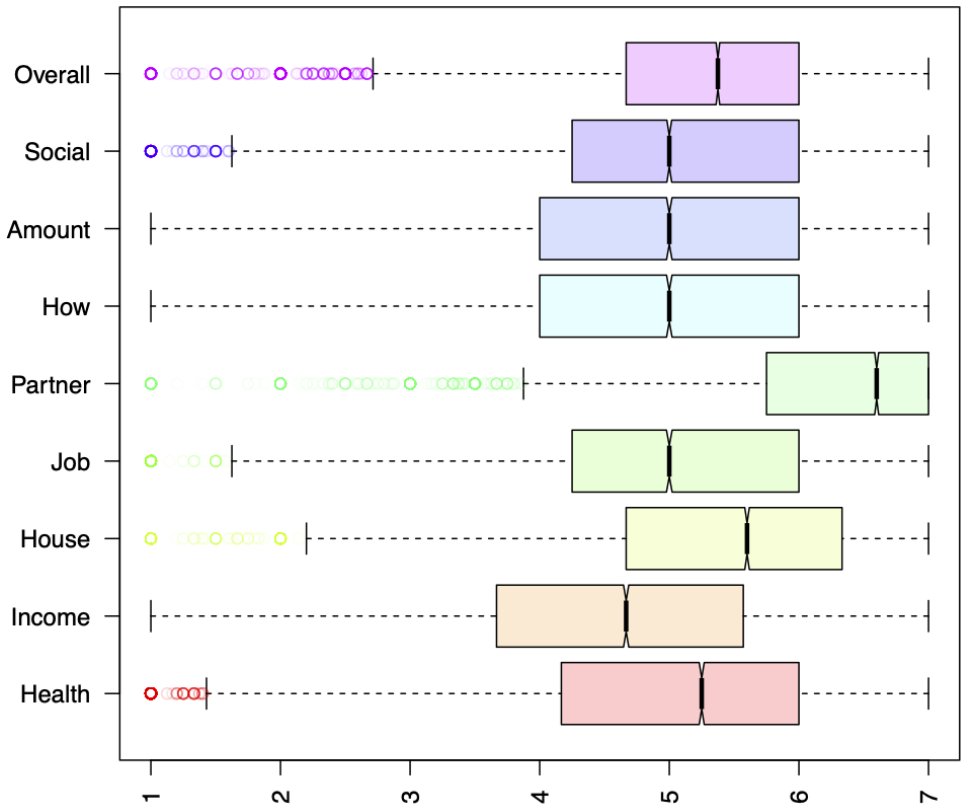


Figure 1.4: Satisfaction reports

Notes: These are box and whisker plots of the average reported satisfactions in all 8 domains and for life overall across the sample of 22,211 individuals who have a ME solution. Note that these are based on the quantised satisfactions, as reported, rather than the unquantised satisfaction. The median is marked by the vertical line inside the box. The ends of the boxes are the upper (75%) and lower (25%) quartiles. The whiskers extend the highest and lowest observations. The pale coloured circles are considered outliers.

Domain	N	Mean	St. Dev.	Min	Pctl(25)*	Pctl(75)*	Max
Health	22,208	5.03	1.39	1.00	4.17	6.00	7.00
Income	22,172	4.55	1.41	1.00	3.67	5.57	7.00
House	22,183	5.39	1.25	1.00	4.67	6.33	7.00
Partner	16,675	6.14	1.19	1.00	5.75	7.00	7.00
Job	15,702	5.00	1.27	1.00	4.25	6.00	7.00
Social	22,208	5.03	1.30	1.00	4.25	6.00	7.00
Leisure Amount	22,204	4.85	1.39	1.00	4.00	6.00	7.00
Leisure How	22,207	4.93	1.32	1.00	4.00	6.00	7.00
Overall	22,211	5.25	1.13	1.00	4.67	6.00	7.00

Table 1.1: Average satisfaction reports: summary statistics

Notes: These are the summary statistics for each individual's averaged domain–specific and overall life satisfaction reports in the BHPS subsample used for the paper's analysis. Note that *N* varies across the domains because (1) individuals are not obliged to answer every question, and therefore could opt to leave their answer blank for any domain and (2) individuals can choose "Doesn't apply to me" for the Partner and Job domains. There is further discussion on these missing domains in section 3.3. *Pctl(25) and Pctl(75) are the 25th and 75th percentiles.

are not in the workforce (retired / taking care of children) or not presently working (studying / unemployed / on sick leave). However, it is possible that some individuals falling into these categories would interpret the question as regarding their feelings about their present status in the workforce. Retirees for instance could make a report based on their satisfaction with being retired, or indeed their satisfaction with the work they did before they retired.

The proportion of individuals who do not report a Job satisfaction level are presented in Table 1.2 (breaking the sample down by employment age and gender and then non-employment status). Overall, 29.3% of the sample do not report satisfaction with their job. Table 1.2 shows that the majority of these individals are not working age (16–64); of those who are in the working age category, there are approximately 50% more women not reporting Job satisfaction than men. Table 1.2 also shows that of those not currently employed, the largest number of non–responses come from retirees.

	Job Sat. Reported (%)	Job Sat. Missing (%)	Total
Gender & Age			
Males 16–64	7,413 (85.4%)	1,272 (14.7%)	8,685
Females 16–64	7,570 (80.1%)	1,879 (19.9%)	9,449
Non–working age	719 (17.6%)	3,358 (82.3%)	4,077
Employment status			
Unemployed	739 (70.0%)	317 (30.0%)	1,056
Retired	1,019 (24.4%)	3,152 (75.6%)	4,171
Family Care	1,182 (56.5%)	912 (43.6%)	2,094
FT student / school	1,944 (65.6%)	1,019 (34.4%)	2,963
Other*	831 (44.1%)	1,053 (55.9%)	1,884

Table 1.2: Missing Job satisfaction reports

Notes: Rows 1 to 3 present the proportion of the sample who do or do not report a satisfaction in the Job domain, broken down by whether they are working age, by gender. Rows 4 to 8 present the proportion of the sample who do or do not report a satisfaction in the Job domain, broken down by their reported employment status. *Other includes: Maternity Leave, Long term sick or disabled, Government training scheme, or Something else.

The same situation occurs in the Partner domain: individuals may have chosen not to make a report because they do not have a partner, but those people may also choose to report on their satisfaction with their lack of partner or consider their satisfaction with a recent partner. Overall, 24.9% of the sample do not report a satisfaction in this domain. A variable available in the BHPS dataset that can offer some insight into the characteristics of these non-respondents is marital status. Table 1.3 reports the proportions of individuals not reporting satisfaction levels for the Partner domain, broken down by their marital status. This is naturally an imperfect correlate of having a partner, given the restrictive nature of the options made available for people to select: civil partnerships are not included, nor an option for those with partners yet legally defined as separated, divorced, widowed or never married.

This is exemplified by the fact that 54.2% of those individuals categorised as never married do report a Partner satisfaction level.

	Partner Sat. Reported (%)	Partner Sat. Missing (%)	Total
Married	9,526 (99.7%)	28 (0.3%)	9,554
Separated	1,246 (88.1%)	168 (11.9%)	1,414
Divorced	1,194 (64.7%)	651 (35.3%)	1,845
Widowed	620 (33.1%)	1,252 (66.9%)	1,872
Never married	4,060 (54.2%)	3,425 (45.8%)	7,485

Table 1.3: Partner satisfaction reports by marital status

Notes: This table presents the proportion of the sample who report or do not report a satisfaction in the Partner domain, broken down by their reported marital status.

Implications for ME: Those who have a missing domain satisfaction require an adjustment to the ME procedure. If an individual does not report a domain satisfaction, this missing domain is given a reported satisfaction of 0 and a weighting of 0; the ME weighting solutions of the 7 remaining non–missing domains are rescaled by $\frac{7}{8}$, to ensure they are comparable to weightings recovered for an individual who has all 8 domain satisfactions. For an individual who does not report a domain satisfaction in some years, but does report it in others, the ME solution looks for weightings that solve the weighted sum in every year (with the quantisation error adjustment in each year), with those years with a missing satisfaction having the missing domain satisfaction dropping out of the weighted sum (due to being allocated a satisfaction of 0).

A simple example: To fix ideas in a simple example, I consider again a situation with unquantised data. Figure 1.5 illustrates the process in a simplified set–up with just 2 time periods and 3 domains. In the first case, the individual makes a full report across the domains in both periods; the resulting ME solution is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. In the second case, the individual does not make a report in domain 2 in either period; the resulting ME solution first finds weights $(\frac{1}{2}, 0, \frac{1}{2})$ then these are rescaled to $(\frac{1}{3}, 0, \frac{1}{3})$ to render the importance of domains 1 and 3 comparable to the first case's solution. In the third case, the individual makes a full report in period 1 and misses a report in domain 2 in period 2. The ME solution finds a compromise set of weights between period 1 and re–scaled period 2, to propose $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

By choosing to include all the individuals in the sample and addressing the missing domains issue in this way, I am implicitly interpreting non–response as meaning that the person does not feel satisfaction or disatisfaction, but indifference. Hence, the domain does not feature in their overall life satisfaction contemplation at all. This seems like a reasonable assumption. The subsequent rescaling

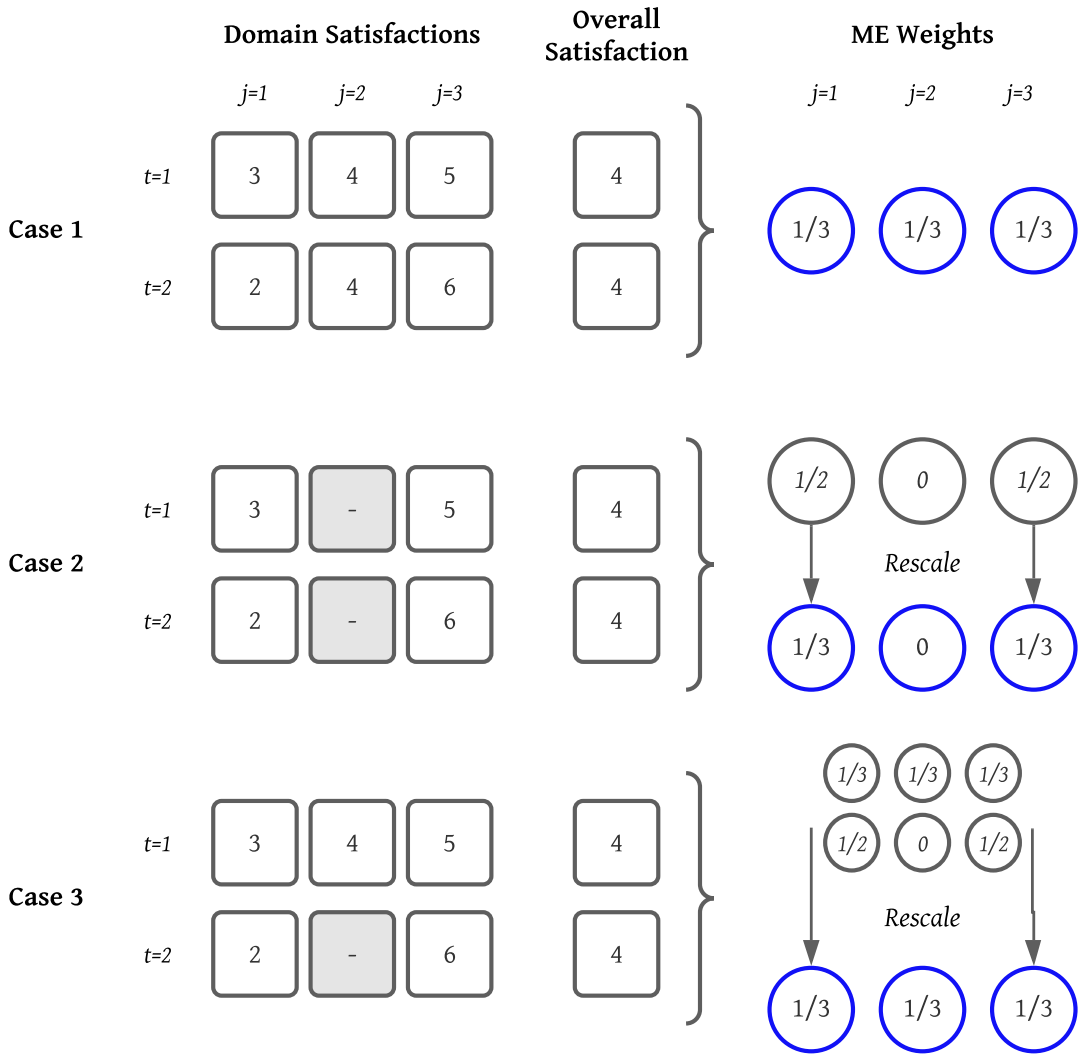


Figure 1.5: Dealing with missing domains in the ME procedure

Notes: The boxes represent the reported satisfactions for domains and overall; the circles are the ME solution for weightings for each domain. Note that the full optimisation allows for a quantisation error in each period in each domain report, hence the weights times the reported domain satisfaction, adjusting for the error, equals the overall satisfaction in each period.

ensures that I can make cardinal comparisons of weights between people with differing numbers of domain satisfaction reports, rather than only making ordinal comparisons of domain weightings. Making cardinal comparisions without rescaling would not be comparing like for like, as the higher weightings for an individual who is missing a domain would be driven by the constraint that weightings must add to one, rather than information within the reported satisfactions.

Alternative solutions to the issue of missing domain satisfactions include:

(1) Imputing a satisfaction level for the missing domain, using a predicted report based on similar

individuals' reported satisfaction in that domain, and running the ME procedure over the full 8;

(2) Dropping individuals with a missing report in every year from the sample, and dropping years in which individuals have a missing report but keeping any years in which they have the full 8; this would result in having just 12,925 individuals, which would be a loss of 58.2% of the sample.

The first of these would require the assumption that a missing domain reflects a mistake rather than an active choice to not report satisfaction in a particular domain; the fact that individuals who did not report a Partner or Job satisfaction (accounting for 99.6% of the missing domain cases) chose "Doesn't apply to me" invalidates this approach. The second option loses many data points which are of interest — those choosing not to report a satisfaction in a particular domain ought still to feature in an account of the importance of different domains for individual life satisfaction, and play a part in determining the appropriate weights for an overall well–being index.

As a robustness check for the strategy outlined above, alongside the main results for the full sample, I report descriptives of the weights for the sub–sample of those who had no domain satisfactions missing, across the full group and the gender and age–quartile sub–samples. Whilst of course Job and in particular Partner move up the rankings in terms of importance in the sub–samples (those given 0 for these domains have been removed), the remaining domains do not have an altered ordinal ranking relative to each other.

Individual and household characteristics

The BHPS was designed to be broadly representative of the UK's population. Tables 1.4 and 1.5 present some of the socio demographics of the subsample used for this chapter's analysis. Table 1.4 shows that there are approximately the same number of women as men in the sample (52.9% females in the sample). The mean age in the sample is 42.5 years, and the median age is 39; the ONS reports that the median age in 1994 was 37 and in 2004 reached 39, so the BHPS sample is close to this.[35]. The average household size in the sample (3.03) is slightly larger than that of the UK overall during this time period, which fell from 2.4 in 1996 to 2.3 in 2004.[36]. The mean number of children, for those households with children, is 1.76 and the median is 2, which is equivalent to the overall population between 1996 and 2004.[37]. The mean income over the whole sample is £23,388 and for those working it is £27,785.[38] The ONS reports a mean income across the years 1996–2005 of £27,872 and hence again the BHPS sample is very similar to this.

[35] ONS source: online Families and Households 2016 bulletin consulted in September 2019.
[36] ONS source: online 2011 General Lifestyle Survey results consulted in September 2019.
[37] ONS source: online Families and Households 2016 bulletin consulted in September 2019.
[38] This variable is derived from the gross income variables in the main BHPS dataset: it is the household annual net labour income. See documentation by Levy et al. for more details on this (2012).

	Mean	Range
Females (=1)	0.529	—
Spouse who works (=1)	0.666	—
Age (years)	42.5	15–97
	(19.2)	
Household size	3.03	1–13
	(1.46)	
Children*	1.76	1–9
	(0.857)	
Income (£)**	23,388	0 – 477,990
	(17,254)	
Income for those employed (£)**	27,785	0 – 399,669
	(18,313)	

Table 1.4: Individual and household characteristics in the sample

Notes: *Number of children is conditional on having at least one child. **Income is self–reported.

In Table 1.5, I present the proportion of household types in the sample for comparison with national UK statistics.[39] There are slightly fewer lone parents with dependent childen (5.9% in the sample, versus 8% in 1996 and 2004 in the UK overall), and a very similar percentage of lone parents with independent children (3.8% in the sample, versus 3% in 1996 and 2004 in the UK overall). The ONS has statistics for the number of households with 1 person aged 60 or above (16% in 1996 and 15% in 2004) and 2 people with at least one aged 60 or above (16% in both 1996 and 2004). These are not comparable directly to the BHPS categorisation of "single elderly" households; similarly comparisons between "single non–elderly" households may not be sensible either.

Household type	n (%)
Single non–elderly	1,335 (6.0%)
Single elderly	1,346 (6.1%)
Couple no children	6,078 (27.4%)
Couple: dependent children	7,072 (31.8%)
Couple: non–dependent children	2,526 (11.4%)
Lone parent: dependent children	1,299 (5.9%)
Lone parent: non–dependent children	837 (3.8%)
2+ unrelated adults	1,203 (5.4%)
Other households	515 (2.3%)

Table 1.5: Household types

[39] ONS source: online 2011 General Lifestyle Survey results consulted in September 2019.

1.4 | Results

This section outlines the results. I describe the baseline ME results: what are the weightings attached to different domains of life in the overall sample, and in subsamples broken down by sex and by age? As a robustness check, I report the differences in these results if the ME procedure is applied only to the subsample who report domain satisfaction in all 8 domains. I then describe the partial correlations of the weights with household and individual characteristics and finally use a clustering technique to see if patterns across the weightings imply natural groupings of people.

1.4.1 Weights and relative importance

Overall results

I first present the overall findings: the mean weighting of each domain across the dataset (Table 1.6) and the distribution of weights across the dataset (Figure 1.6). The table highlights that the domain most highly weighted is Health, with an average weighting of 0.146, with Social life as the next most important (average weight of 0.142). Income and Job are the least important (weights of 0.103 and 0.083 respectively).[40]

Note that the Partner domain weighting has a large standard deviation, and the largest 75th percentile (0.196); the individuals who occasionally or consistently do not report a satisfaction in this domain (driving the mean weighting down) are being balanced out by those who do report a satisfaction in this domain who weight it very highly. As discussed in the robustness check below, when those people not reporting a Partner satisfaction are removed from the sample, this domain becomes the highest weighted. In contrast, the Job domain, even when those not reporting a satisfaction are removed, keeps a low ranking (only moving from rank 8 to 6); this is clear as the 75th percentile is only 0.126. Table 9 in Appendix A demonstrates that the ordinal rankings of the other domains, when moving from the full sample to the subsample with no missings, remains reasonably constant with the exception of Income moving up one position, and Leisure amount moving down one position.

[40]The natural instinct is to consider how these results compare to those that would be found by OLS regression of overall satisfaction onto the domain weightings. As per the discussion in the Introduction and Theory sections of this chapter, using regression on the reported satisfactions (1) requires the assumption of inter–personal comparability of satisfaction reports on set scales and this is not an assumption I make and (2) even if reports were comparable, theory tells us that regression would still result in biased coefficients if the true weightings (slopes) are heterogeneous. I therefore do not present regression results. However, for readers familiar with the satisfaction literature: to enable the comparison of the BHPS domain satisfaction reports to findings that do use regression (and hence make the inter–personal comparability and homogeneous slopes assumptions), the ranking of the OLS estimated coefficient sizes is as follows: (1) Social life, (2) Health, (3) How leisure is spent, (4) Partner, (5) Job, (6) Income, (6) House and finally (8) Amount of leisure time. Note that regression drops any observation with one or more missing variables, and therefore the regressions are run only over the subsample of individuals who reported a satisfaction level in all 8 domains (n=12,975). Therefore, if a comparison is desired, this ranking of the OLS coefficient sizes should be compared not to the main ordinal results in this section, but to Table 9 in Appendix A.

	Mean (Rank)	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Health	0.146 (1)	0.157	0	0.039	0.193	1
Social	0.142 (2)	0.140	0	0.044	0.184	1
House	0.141 (3)	0.138	0	0.039	0.190	1
Leisure Way	0.135 (4)	0.133	0	0.040	0.178	1
Partner	0.134 (5)	0.154	0	0.010	0.196	0.992
Leisure Amount	0.116 (6)	0.119	0	0.025	0.161	1
Income	0.103 (7)	0.114	0	0.017	0.143	1
Job	0.083 (8)	0.109	0	0.003	0.126	0.909

Table 1.6: Mean domain weightings

Notes: These are the summary statistics of the weightings for each of the domains in the full sample of 22,211 individuals. In brackets is the ranked importance of the domain, which is the size order of mean weightings.

Weighting heterogeneity: gender and age groups

Tables 1.7 and 1.8 present the mean weights, and rankings, broken down by gender and then by age groups. The difference in ordinal rankings of the domains for men and women is not large: these have a Kendall τ coefficient of 0.79.[41] A comparison of different paired age group rankings shows larger differences: groups 2 and 4 are most dissimilar ($\tau = 0.07$), suggesting those in the early stages of adulthood (aged 26-40) are most dissimilar in terms of which areas of life are important to them to those approaching / in retirement (aged 57-97). The way leisure is spent is most highly ranked for the 4th group, Partner is ranked only 5 for this group, and How much leisure time they have and House are at 3 and 2 respectively. There is also a striking difference between the first and second groups ($\tau = 0.357$), suggesting late teenagers and those in their early 20s undergo a quick shift in what is most important to them as they transition into their late 20s and in their 30s: for the youngest group, Health and Social life are top ranked, whilst Job and Partner are ranked bottom.

In order to analyse the age and gender heterogeneities statistically, I conduct non–parametric tests of equality of distributions across groups in each domain of life. I use the Kolmogorov–Smirnov (KS), which tests a null hypothesis that the distribution of weights for different sub–samples could come from the same distribution (a two–sided test).

[41] The Kendall rank correlation coefficient τ ranges from 1 (ranks are perfectly the same) and -1 (fully different rankings). The measure was developed to measure either the discrepancy between a correctly ranked list of items and a second attempted ranking of the same items, or the difference between two alternative rankings of the same items (Kendall 1938). τ is defined as:

$$\tau = \frac{2\sum}{n(n-1)}$$

where n is the number of items in the ranking. The $\sum$ is the total accuracy of the positioning of each item relative to the others in the list. τ is a key measure of ordinal difference for the results in Chapter 3 of this thesis.

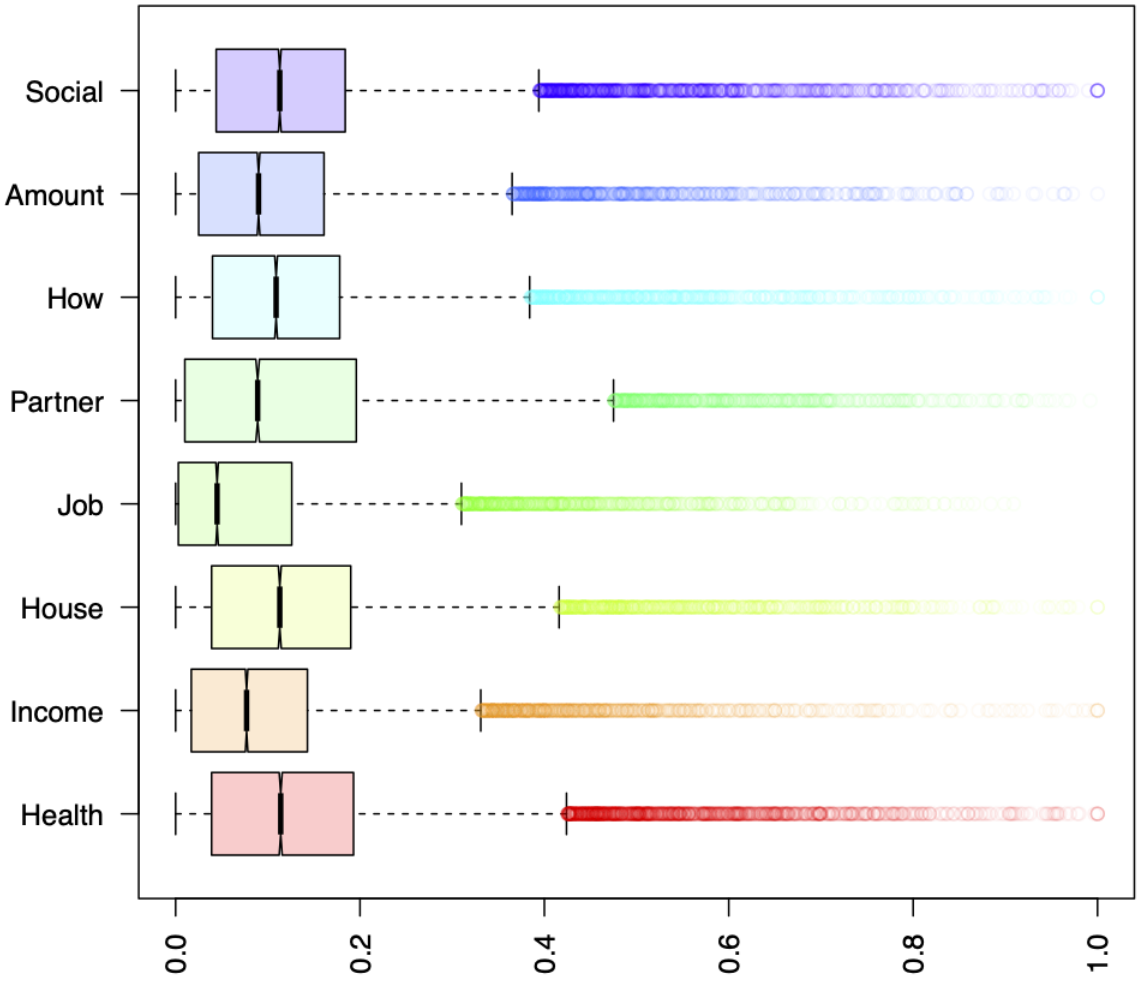


Figure 1.6: Weights

Notes: These are box and whisker plots of the Maximum Entropy weight solutions across the full sample of 22,211 individuals. The median is marked by the vertical line inside the box. The ends of the boxes are the upper (75%) and lower (25%) quartiles. The whiskers extend to indicate the highest and lowest observations. The pale coloured circles are considered outliers.

In the case of gender heterogenity, the D statistic in Table 1.7 demonstrates that there are differences in the distributions of the weightings between men and women, and this is significant at the 1% level for every domain. This is suggestive that although the average ordinal information in satisfaction reports looks similar, the cardinal information in the distributions of the weightings is significantly different for the gender subsamples. The largest difference between the distributions is for the Job domain, followed by House: across the distribution of weighting results, women place considerably less weight on the Job domain than men, and slightly more on House.

	Mean Weighting (Rank)*		D**	*p*–value
	Men	Women		
Health	0.144 (1)	0.148 (1)	0.0297	0.000
Social	0.142 (2)	0.142 (3)	0.0341	0.000
House	0.135 (4)	0.146 (2)	0.0397	0.000
Leisure Way	0.137 (3)	0.133 (5)	0.0384	0.000
Partner	0.130 (5)	0.138 (4)	0.0271	0.001
Leisure Amount	0.116 (6)	0.115 (6)	0.0259	0.001
Income	0.104 (7)	0.103 (7)	0.0291	0.000
Job	0.093 (8)	0.075 (8)	0.0947	0.000

Table 1.7: Comparison of mean domain weightings by gender

Notes: *These are the mean weightings for each of the domains, averaged for the male and female subsamples. In brackets is the ranked importance of the domain, which is simply the size order of the mean weighting. **D is the two–sample combined Kolmogorov–Smirnov (KS) statistic, with the associated *p*–value.

	Mean Weighting (Rank)			
	1: 15–26 yrs	2: 26–39 yrs	3: 40–56 yrs	4: 56–97 yrs
Health	0.172 (1)	0.157 (2)	0.136 (2)	0.119 (5)
Social	0.166 (2)	0.125 (4)	0.132 (4)	0.143 (4)
House	0.141 (3)	0.125 (3)	0.136 (3)	0.162 (2)
Leisure Way	0.129 (4)	0.113 (5)	0.131 (5)	0.165 (1)
Partner	0.093 (7)	0.177 (1)	0.153 (1)	0.116 (6)
Leisure Amount	0.112 (5)	0.091 (8)	0.104 (7)	0.155 (3)
Income	0.103 (6)	0.103 (7)	0.103 (8)	0.102 (7)
Job	0.083 (8)	0.108 (6)	0.104 (6)	0.038 (8)

Table 1.8: Comparison of mean domain weightings by age group

Notes: These are the mean weightings for each of the domains, averaged for each age quartile. In brackets is the ranked importance of the domain, which is simply the size order of the mean weighting.

Considering age heterogeneities, a KS test between the 1st and 4th age groups finds significant differences across the distribution of weightings in all domains (all with 1% significance). The KS statistics are as follows: Health 0.2190, Income 0.0988, House 0.1023, Partner 0.1056, Job 0.2851, Social 0.1216, Leisure Amount 0.1888 and Leisure How 0.1460. The largest difference is the importance placed on Job, with much more weight placed on this by those in the youngest age bracket relative to those in the oldest: those older than 57 may be for the most part retired, not have reported a satisfaction in the Job domain, and therefore have been allocated a 0. The next biggest differences in the distributions are for Health, How leisure is spent and How much leisure time is available: in the Health domain, more weight is attached by the youngest quartile, and in both Leisure–related domains, more weight is placed on these by those in the oldest age quartile.

Robustness check: Tables 9 and 10 in Appendix A present the ordinal and cardinal domain weighting

results for the subsample that reported a satisfaction level across all 8 domains; these results therefore make no assumptions about what to do about the domains reported as "do not apply to me". Table 9 demonstrates that overall, and in the gender and age specific results, there are very minimal changes in the ordinal rankings across the domains, with the exception of Partner and Job which are more highly ranked in every case. This is natural because moving from the full sample of 22,211 to a subsample of 12,925 loses all those who considered that Partner and / or Job were not applicable to them, and therefore were allocated a 0 or very low weight to this domain (depending on whether they made a report for this domain in other waves or not). There is a change at the bottom of the rankings: Income moves up a rank and Leisure amount moves down a rank. Table 10 shows a similar pattern: the main shift is Partner moving up to 1st position for all subsamples, with the exception of the 15–26 age group, for whom it moves only to 3rd position. Job moves up only a little (by one or two positions) and remains in position 8 for the 56–97 year olds. Income moves up one position for women and the 15–26 group, is unmoved for men and for the 26–39 and 56–97 groups and it falls by one position for the 40–56 age group. Overall, this check is reassuring: the assumption regarding the missing Job and Partner satisfactions does not hugely alter the overall patterns of relative weighting sizes across the other domains, although there are a few small shifts in the importance of Income in particular, once those who choose "does not apply to me" for Job and / or Partner as removed. This seems natural, as the income stream for this subsample of individuals may more uncertain, by choice or not, and hence have different importance for overall satisfaction.

Other correlates with weights

In order to investigate wider patterns of heterogenity in weightings, I also conduct OLS regressions of each weighting on a set of covariates. Appendix B1 presents the full regression tables. The covariates are grouped into household characteristics (income, household type or whether there are children at the address) and individual characteristics (age, age squared, employment status, marital status), and Tables 11, 12 and 13 in the Appendix present full results of regression specifications involving combinations of these covariates as explanatory variables for each of the domain weightings.

Overall, across the majority of the specifications for the domain weights, it is clear that currently working, being married and having children in the home are significant correlates for the importance attached to domains for overall satisfaction. *Working* (employee or self employed) is associated with less importance being attached to Health (-0.016 or -0.014), House (-0.016 or -0.013), Amount of leisure time (-0.016 or -0.018) and Social life (-0.024), and with higher importance being placed on Partner (+0.015 or +0.010) and Job (+0.066 and +0.065). All these coefficients are significant at 1%. *Married* is associated with less weight on Income (-0.011), Amount of leisure time (-0.027 or -0.010), Leisure activities (-0.028 or -0.019), Health (-0.016 or -0.010) and House (-0.024 or -0.017), and more

weight on Partner (+0.138 or +0.077). *Children* in the home also is associated with a greater weight being placed on the Partner domain (+0.062 or +0.039) and also on the Health domain (+0.021 or +0.014) — perhaps associated with the goal of being able to provide for a family — and less weight is placed on Amount of leisure time (-0.035 or -0.016), Leisure activities (-0.031 or -0.016) and Social life (-0.028 or -0.025).

Age does not appear to be a significant covariate for many of the domains, with the exception of the Job domain (+0.001 at 5% significance level) in both relevant specifications, and in the full specification for Partner (+0.001 at 5% significant level) and Social life (-0.001 at 5% significance level). This is weak evidence for a linear association of increasing importance attached to Job as people age, and on Partner, and decreasing importance on Social life. The significance of a negative coefficient on *Age squared* in the Job and Partner regressions suggests the effects grow weaker as the individual ages (at a very slow rate).

The *household income* covariate has a significant, but very small effect, on the Income, Partner and two leisure–related domains. The higher the level of household income, the less the weight attached to the two leisure–related domains: richer individuals place less weight on how much leisure time they have and how they spend it. There is a positive assocation with Partner and with Income: the richer an individual is, the more weight they place on their partner and their income for their satisfaction.

This descriptive analysis invites two considerations in particular: firstly, are there patterns across domains as to the types of poeple who place importance on them? I investigate this in the next section. The second I address in the discussion in Section 5, as an avenue for further research: what could dynamic analysis reveal about any causal relations between the weights and their covariates?

1.4.2 Weights and natural groupings of people

The per–person domain weightings invite an investigation into whether there are patterns across people as to which set of life domains matter most to them. I present here the findings of k–means clustering analysis of the weights. This is a way to see whether people could be grouped according to which areas of life matter most to them. I describe the characteristics associated with each of the groups, and then see if it might be possible to predict which group an individual falls into by observable characteristics.

k–means clustering: grouping individuals

k–means clustering is a method to determine natural groupings in a dataset by assigning observations in a sample into groups. The groups are iteratively constructed with the aim of optimising an objective function: maximising the similarity of people within each group, maximising the distance between groups or having groups as evenly sized as possible.[42] A typical algorithm to conduct k–means clustering analysis is the Hartigan–Wong algorithm whose objective is to construct groups that made up of individuals as similar to each other as possible (Hartigan and Wong 1979); "similarity" is measured by the Within Sum of Squares (WSS) which is the sum of the squared Euclidean distances between each individual's weights in that group and the group's mean weighting scalar.[43]

A key consideration is how many clusters (k) is the right number? Gordon (1999) discusses the best performing methods for estimating the optimal number of clusters, complementing a survey of possible methods to choose k by Milligan and Cooper (1987).

The most common way to choose k is to plot an "Elbow curve" of the number of clusters against an evaluative criteria: for example, if maximising similarity within groups is the evaluation criteria, one computes the total WSS for a set of values of k; the elbow method then recommends choosing the k at which the returns to adding more groups (reducing the total WSS) starts to diminish.[44] The weights data do not have a clear elbow, but the returns to k seem to start to diminish from k=6 (see Figure 10 in Appendix B2 for an illustration).

An alternative method to determine k is known as the Silhouette method: an observation's Silhouette

[42] Alternatively, the objective function could be a combination of these factors. For instance, the Dunn Index is used in machine learning to choose the clusters: it is the ratio of the minimum of within–cluster distances and the maximum of between–cluster distances. See Dunn for the initial proposition (1973).

[43] When applied to the weightings data using the *cluster* command in R, the precise algorithm steps are follows:

1. I specify k number of groups that are to be created.
2. R randomly chooses k seed scalars to be the initial group means (*centres*) which in this case is a set of 8 mean weightings, one for each domain of life. Because the final clusters are sensitive to the random start centres, the program repeats 25 times, each time with new seed centres. The chosen cluster result is the one that has the lowest within–cluster variation (individuals as similar to each other as possible).
3. The algorithm places each individual in the group that has the nearest centre to the individual's own weighting scalar.
4. New group centres are computed, using the weightings of the individuals currently assigned to that group.
5. Individuals are reassigned to the group with the newly computed group centre that is closest to their own scalar.
6. The algorithm repeats steps (3) – (5) until no individuals swap groups.

The algorithm is a *local* search that moves individuals into a different group in each iteration provided that the move improves the objective function. The approach does not necessarily guarantee that the final solution is the global optimum, however the repetition of the search 25 times with different seed centres offers some reassurance in this regard.

[44] Tibshirani, Walther and Hastie proposed an additional method known as "The Gap" as a formalised version of the Elbow method (2001): compare the WSS change as k increases with a reference null distribution.

(S) is a measure of how cohesive (similar) it is with its own cluster and how separate (dissimilar) it is to the other clusters.[45] As k increases: (1) the fit of individuals to clusters improves which improves S, but also (2) the size of each cluster falls which harms S and (3) a given individual may be close to other newly created groups which also harms S. S therefore climbs to a global maximum then falls. Figure 10 in Appendix B2 shows that the global maximum for the S value of the weights data is at k=8 and there is also a local maximum at k=6.

A final consideration for choosing the optimal k is simply looking at the number of individuals allocated to each group, and choosing a k small enough that a meaningful number of observations appear in the majority of the groups. Table 11 in Appendix B2 shows the number of individuals in each group as k increases from 2 up to 10. After k increases above 5, there are always at least 3 groups who have fewer than 9% of the sample within them, suggesting that from 6 onwards, some groups become increasingly sparsely populated.

Evidence from the Elbow method, Silhouette measure and group size analysis together suggest a k of between 5 and 8. I chose 6 and present the results for the clustering procedure in Table 1.7: the minimum, mean and maximum weight attached to each domain by the individuals allocated to each of the 6 groups.[46] Any minimum weighting of $> \frac{1}{8}$ is in bold: such an instance means that every individual in the group places a greater than proportional weight on this domain in terms of its importance for their overall well–being. Mean weightings of $> \frac{1}{8}$ are also highlighted, and in the case of groups A, B, D, E and F the mean for one domain in particular is close to or above $\frac{1}{2}$: the domain therefore has about 4 times the "fair" weighting of $> \frac{1}{8}$ attached to it.

Led by these descriptives, it is possible to define the following "types" of group: **(A)** Extremely Health focused, with a minimum of over $\frac{1}{4}$ on this domain and nearly $\frac{1}{2}$ on average; **(B)** Extremely Partner focused, with a minimum of nearly $\frac{1}{4}$ on this domain and nearly $\frac{1}{2}$ on average; **(C)** Balanced across Income, Social life, Amount of leisure time and How leisure time is spent, these people could be termed

[45] S ranges from +1 (cohesive within the group and separate to other groups) to -1 (incohesive within the group and indistinguishable from other groups), and takes the value 0 if an observation in a cluster on its own. As with the WSS, the distance (cohesion and separation) is the Euclidian Distance.

[46] In Appendix B2 I present the corresponding results for clusters formed with k=5 and k=7, as a robustness check for the choice of the size of k (Tables 12 and 13). These show that the findings are indeed robust to the choice of k, with similar proportions of individuals being allocated to the same types of groups as k increases and decreases away from 6. In particular: the robustness tables show the survival of both the Health and Social life focused groups, both with a k of 5 and 7, and with a similar proportion of the sample allocated to each group in all k cases (9%). Likewise the Partner focused group survives, with a similar proportion allocated when k is 7 (14%), and a larger group for k is 5 (18%), relative to the choice of 6 presented above. A Job focused group appears at k is 7, which does not feature in the results for 6 or 5. The Balanced group appears in for k is 5, 6 and 7, with 54% 41% and 45% of the individuals respectively. The component domains of the Balanced group vary a little: k is 5 is made up of domains Income, Social life, Amount of leisure and Way leisure is spent; k is 6 has Income, Social life, Way leisure is spent and Amount of leisure time; k is 7 has Income, House, Social life Amount of leisure and Way leisure is spent.

"Work Life Balanced", valuing income but also non–work time; (D) Extremely House focused, with a minimum of over $\frac{1}{4}$ on this domain and nearly $\frac{1}{2}$ on average; (E) Extremely Social–life focused, with a minimum of over $\frac{1}{4}$ on this domain and over $\frac{1}{2}$ on average; (F) Extremely Leisure–activities focused, with a minimum of over $\frac{1}{4}$ on this domain and nearly $\frac{1}{2}$ on average.

		Health	Income	House	Partner	Job	Social	Amount	Way
A	*min*	**0.283**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	**0.484**	0.072	0.085	0.065	0.057	0.091	0.072	0.075
2,233	*max*	**1.000**	0.556	0.493	0.426	0.532	0.484	0.539	0.413
B	*min*	0.000	0.000	0.000	**0.218**	0.000	0.000	0.000	0.000
	mean	0.104	0.071	0.105	**0.419**	0.073	0.086	0.067	0.075
3,445	*max*	0.497	0.585	0.492	**0.992**	0.720	0.449	0.585	0.477
C	*min*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	0.121	**0.141**	0.124	0.093	0.110	**0.130**	**0.154**	**0.127**
10,927	*max*	0.327	1.000	0.330	0.313	0.909	0.347	1.000	0.324
D	*min*	0.000	0.000	**0.277**	0.000	0.000	0.000	0.000	0.000
	mean	0.094	0.076	**0.467**	0.064	0.052	0.091	0.076	0.080
1,983	*max*	0.484	0.603	**1.000**	0.407	0.559	0.479	0.641	0.469
E	*min*	0.000	0.000	0.000	0.000	0.000	**0.301**	0.000	0.000
	mean	0.082	0.057	0.075	0.071	0.055	**0.510**	0.071	0.078
1,673	*max*	0.471	0.563	0.466	0.490	0.543	**1.000**	0.579	0.462
F	*min*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.270**
	mean	0.083	0.052	0.086	0.067	0.041	0.101	0.111	**0.457**
1,950	*max*	0.477	0.558	0.486	0.469	0.568	0.482	0.659	**1.000**

Figure 1.7: Cluster centres

Notes: These are the minimum, mean and maximum weights — centres — of the individuals within each cluster e.g. in Group A, the minimum weighting on Health of individuals in the cluster is 0.283, the mean is 0.484 and the maximum is 1. Darker colours illustrate a greater weight. The number of individuals in each group is in italics below the group letter, for instance group A has 2,233 individuals.

Nearly half the sample of 22,211 individuals are in (C), the work–life balance individuals, who are placing a slightly more weight on Income, Social life, Amount of leisure time and Leisure time activities than on the other 4 domains: between them, these domains play an important role in their overall satisfaction, with a total mean weighting of 0.552. The next largest group is (B), whose overall satisfaction is driven mostly by their satisfaction with Partner. The smallest group is made up of the most extremely focused individuals, whose Social life is playing centre stage in their overall satisfaction evaluations.

Describing the clusters: groupings' covariates

The next step is to understand whether these groupings could be predicted by observable characteristics of an individual or their household. I compare the characteristics of people in each group to the full sample: the mean values, and standard deviations for non-binary characteristics, for each group are presented in Table 14 in Appendix B2. In each group I conduct a two–sided Student t–test (or test of proportions in the case of binary variables) that the mean characteristic for the group is equal to the overall sample mean. There are significant patterns:

- (A) Health focused: this group has a higher proportion of women in comparison to the overall sample (56% rather than 53%), and the mean age is 4 years younger than the full sample. The individuals are more likely than average to be unemployed, a student or engaged in family care (and hence less likely to be currently working or retired). The proportion who are married is significantly lower than overall (33% rather than 43%), with individuals being more likely to be separated, divorced, widowed or never been married. Overall, the picture is that those who are younger, not working and single may be likely to place a larger weight on their health, relative to other areas of life, when they evaluate their life overall.

- (B) Partner focused: This group is the most distinct from the main sample, with significant differences across all the characteristics. As with group (A), individuals in this group are more likely to be female, but are only 1 year younger than the full sample mean. They are in households that are wealthier than the full sample (household income higher than average by £2000). These households are also likely to be larger on average, have more children and be formed of couples. They are more likely to be working or engaged in family care (not unemployed, students or retired). There is a much higher percentage of married individuals: 67% versus 43% in the main sample.

- (C) Work–life balance focused: This group, despite being made up of nearly half of the sample, still has some distinguishing features from the full sample. These individuals are less likely to have children, have a slightly higher mean income (by about £500), be made up of unrelated adults rather than a couple, and be working or students. There is a higher proportion of never married individuals. The mean age, however, does not differ significantly from the overall sample. Overall, these individuals are single, without dependents, and working or studying full–time.

- (D) House focused: There is a much higher proportion of women in this group (59% rather than 53% overall), 3 years older on average and with a lower income. They are likely to be single — elderly or non–elderly — and if a parent, a lone parent. There is a smaller proportion of individuals currently working, and a higher proportion of retirees or individuals engaged in family care. The individuals are likely to be divorced, widowed or separated. Overall, women

operating single adult households, not in the workforce, are likely to place a large weight on house for their satisfaction.

- (E) Social–life focused: These individuals are less likely to have children, less likely to be currently working (34% versus 45% in the full sample) but instead are made up of more full time students or be in school (18% versus 13% in the full sample). They are less likely to be currently married and more likely to have never married. They are more likely to be in a single adult household. The individuals who place a high weight on Social life can therefore be characterised as not currently working, without children and without a co–habiting partner.

- (F) Leisure activities focused: This group has the highest mean age (48.4 years versus 42.5 in the full sample) and the lowest mean income (over £2000 less than the full sample). They are the most likely to be a single–elderly household, and least likely to be a couple with dependent children or co-habiting non–related adults. They are most likely to be retired: 30% are retired, versus 19% in the full sample. They are likely to be married, divorced or widowed (13%), rather than never married. Therefore, those who place a high weight on their leisure activities when evaluating their life overall are likely to be older and retired.

The *k*–means clustering and subsequent descriptive analysis of the characteristics of the groups demonstrate above all the potential of the richness of the ME output to explore the relations between observable characteristics and the domains of life that are important to the well–being of different individuals. An in–depth exploration of the weightings, using the relationship between implied natural groupings within the weightings and observable characteristics could be taken forward as rich avenue for future research.

1.5 | Discussion

88% of the analyzable individuals in the BHPS dataset had domain and overall satisfaction reports that were consistent with a very simple model: their overall life satisfaction is a weighted arithmetic mean of their domain satisfactions, with a quantisation error on those reports. This model is falsifiable, and individuals for whom the model does not work can be identified. The model then allows, through the use of the ME principle, the investigation of patterns within the solvable individuals' weights without making the assumption that satisfaction reports are inter–personally comparable. This is the key contribution of this chapter.

Weightings results and the SWB literature

The main results show that on average, the largest weightings are placed on Health (0.146), Social life (0.0142) and House (0.141). The ordinal ranking of the average weightings is roughly comparable to the ranked coefficient sizes attached to different drivers of well–being estimated in the SWB literature. Dolan, Peasgood and White's systematic review of the determinants of SWB find key drivers of low SWB are poor health, separation, unemployment and lack of social contact (2008); Boarini et al. find health, personal security and freedom (2012); Clark, Layard and Senik find in the personal features that mental and physical health and family experience are important, although key external determinants also include income, work, community and governance (2012).[47] The importance attached to health, home life and social life related variables are therefore consistent across these papers and the ME findings in this chapter.

The introduction to this chapter highlighted two papers that specifically aim to estimate the relative importance of domains for overall SWB: Benjamin et al. find that happiness, life–satisfaction and health have some of the largest marginal utilities, but that other aspects measured less commonly have at least as large marginal utilities, for instance, aspects related to family, security, values, having options (2014); Van Praag, Frijters and Ferrer–i–Carbonell find that the three most important domains are found to be finances, health and work, out of a set of domains that also include leisure, housing and the environment (2003). These findings are also relatively similar to the ME findings in this chapter: health and family are found to be important, although finances and work are found to be least important in this chapter's results.

[47] Note that Dolan, Peasgood and White caution against simple comparisions of results across the determinants of SWB literature for the following reasons (2008): (1) difference in the categorisation of variables and reference categories (e.g. marital status / employment status / education); (2) inclusion of different control variables; (3) difficulty of exploring the impact of variables that have little or no variability whilst also controlling for individual unobserved effects.

The gender and age group heterogeneity results — both the ordinal rankings of the means and the cardinal weightings distributions — also offer a comparison point to the SWB literature. Boarini et al. find evidence for age heterogeneity in drivers of well–being but little evidence for differences between men and women (2012).[48] Considering the gender heterogeneity in the ME results, it is only when the distributions of weightings are examined that men and women have clear differences (recall the Kolmogorov–Smirnov test results). The comparison of the simple mean weightings and ordinal rankings of weight importance between male and females is not suggestive on its own of clear gender heterogeneity. Therefore it may be unsurprising that without the granularity of person–by–person weightings enabling comparisions of full distributions of weightings per gender, gender heterogeneity goes undetected in regression analysis.

With regard to the age heterogeneity findings, Boarini et al. (2012) highlight that (1) income and being employed have a larger effect on SWB for those working than those younger or older than working age and (2) health importance increases monotonically with age. The ME results are similar to this first finding: Job is ranked most highly by the middle two age groups: one would therefore indeed expect that satisfaction with Job would have a larger effect on SWB for those in these age groups, than in the youngest and oldest groups. However, the regression of the ME Job weighting onto the set of covariates also finds a weak negative linear trend with age (although this relationship declines with age); this could be linked to the fact that more of those in the youngest age group are working than those in the older age group. Note that, in a similar vein to this decreasing importance of Job as individuals age, the Income domain is ranked most highly by the youngest group in the ME results.

Regarding the second age group related finding — monotonically increasing importance of health for SWB in age — the ME results are not comparable: in terms of ranked mean importance, the youngest age group rank health 1st, whilst the oldest rank it 5th. The oldest age group do not see their health as a key determinant of their overall well–being. The regression results do not find a signficant correlation between age and the size of the weighting on health, with the exception of specification (3), in which there is a negative partial correlation of -0.001, significant at 10%. The k–means analysis tells a similar tale: the mean age for those in Group (A), health–focused, is 38.7 which is significantly lower than the full sample mean age of 42.5 (1% significance level). This is puzzling in relation to what intuition may suggest: that the elderly would place more importance on their health as their health degrades overtime. Perhaps the inclination to weight Social life and Leisure activities more highly reflects the fact that these are factors an individual can better control than ailments associated

[48] Other SWB work looks at heterogeneity in drivers according to groups / quartiles determined by SWB levels (Clark et al. 2005; Binder and Coad 2011), which could also be tested in the ME weightings matrix in further work.

with age, and therefore have more importance attached to them (one has to accept health deterioration with age). A microeconomic dataset with domain–level and overall satisfaction and also observed health–related variables may be able to investigate this finding in more detail, by controlling for some observable health–related variables.

The existence of some common threads between the overall patterns in the ME results and in the existing body of empirical evidence is reassuring for a number of reasons. Firstly, it suggests that the application of the ME principle to the SWB data is sensible, in that it does not up–end 40 years' worth of careful scientific study of such data. And secondly, a key difference in the ME method (applied to each individual), as compared to using regression (on pooled reports from individuals) is the avoidance of the assumption of inter–personal comparability. The symmetry in some of the average findings across the two approaches suggests that even if such data are not comparable, making this assumption does not harm the results. Indeed, Fabian's estimation of only 15% of individuals scale–norming suggests the majority of reports may be intra–personally comparable (2018). If similarly small percentages of reports are not inter–personally comparable, then estimations based on pooling individuals may not be adversely affected by the assumption.

In the light of this symmetry in the findings between this chapter and in the existing literature on the relative importance of the determinants of SWB, it is natural to consider whether the methodological contribution offered here is necessary. In their review, Dolan, Peasgood and White highlight some problems in drawing concrete conclusions on the determinants of SWB, including the possible impact on the findings of potentially unobserved variables (2008); there are challenges in exploring the impact of variables that do not vary over time but maybe correlated with unobserved individual effects. This impacts the coefficients estimated using regression analysis. The two–step approach to the investigation of individual characteristics in this chapter — first obtaining a personalised ME solution for the weightings exclusively using the SWB overall and domain reports, then secondly investigating patterns in the weightings with respect to various characteristics — avoids the unobserved individual effects issue entirely in the ME procedure.

The fact that the ME weightings are broadly comparable to existing empirical findings gives greater confidence in the second set of analysis on the patterns within the weightings themselves. The per–person weightings are matched to other individual characteristics and can then be analyzed at a more granular level. The potential of such analysis is large. The identification of different groups which can in turn be predicted from observable characteristics enables the targetting of welfare policies to particular groups within societies. For example, Group (E), who are likely to be not working, without children and without a co–habiting partner, put over half the available weight on their Social life: for

these individuals, interventions to enhance the opportunities for social gatherings and meet–ups would provide the best return on overall well–being. Those in Group (F), who are Leisure–activities focused, and who are also likely to be in single–person household and be older, would benefit from investment into community groups offering leisure–activities suitable for the elderly.

Next steps for SWB and Maximum Entropy

This chapter's identification of the relative importance of SWB drivers assumed these are stable for individuals over time: I have identified one set of weightings per person, using their SWB reports over the panel, adjusted for the quantisation error. These results could be extended in a number of ways:

1. Adding further characteristics to the dataset that also are stable over time. For instance, one growing thread of research connects SWB levels to genes and personality traits (for example, see Lykken and Tellegan 1996 and De Neve 2011). Breaking down the weightings results matrix through the use of this type of information about individuals, or indeed other static characteristics such as ethnicity, nationality or health indicators could be a fruitful next step for the research.

2. A natural extension would be to build upon the ME approach and introduce a dynamic component to the weights and their determinants. Dynamics are important in SWB research, and are crucial to understand causality within the field (required for clear policy recommendations).[49] For example, Diener notes that current SWB research uses a range of methods to look at psychological pheneomena, such as the effects on SWB of adaptation (2013). The ME procedure would need to be adapted to provide a falsifiable model that estimates weightings per–person per–period, adjusting for reporting error. k–means clustering in each period could then offer the opportunity to test whether life shocks (marriage, divorce, having a baby, losing or getting a job) move individuals into new groups. One extension would be to test the finding in Macchia and Whillans that those who value time (leisure) more than money (work) report higher SWB (at a country and individual level) and that these groups are less likely to be negatively impacted by financial instability (2019). This would involve categorising individuals by their SWB weightings patterns, and then looking for heterogeneity across the groups in the effect of different shocks on overall SWB reports in each period: which pattern in weightings across domains offers the most SWB resilience? Likewise, changes to group membership could be plotted overtime with age. An additional benefit to working with dynamic weights would be the avoidance of any kind of comparability assumption: the SWB reports would not be pooled over time for individuals, and hence intra–personal comparability would not be needed.

3. A weakness in the results, with comparison to empirical work using regression, is the fact that the ME solutions do not include standard errors and standard inference techniques cannot be

[49] This is highlighted as requiring more research in the Dolan, Peasgood and White review (2008).

applied. As a result, it could be helpful to introduce further robustness checks, for instance the use of the Kullback–Leibler distance: the solution stemming from the introduction of alternative prior distributions can be compared to the solution stemming from the standard ME flat–distribution assumption. One possibility for this could be to test the difference in the result if a prior of individuals prioritising income above all is included as a constraint.

Concluding remarks

Overall, the core contribution of this chapter is to introduce an alterative way to handle SWB data, that avoids the inter–personal comparability assumption. The symmetry of the ME results with much of the work looking at the determinants of SWB that do make this assumption speaks to a helpful output for this work: it is reassuring that the two very different methods extract similar threads within the reports, and suggests the damage of making the assumption, even if not true, is not substantial.

However, the chapter offers a richness of output from the ME procedure that is unparalled in approaches using regression: a per–person estimation of what matters to them for their overall life satisfaction. Further research into these patterns has potential in two key areas: (1) informing the construction of well–being indices whose weightings for domains are sensitive to subpopulations' characteristics and (2) proposing well–being policies for different subpopulations, that would have maximum return on investment by targetting them to the areas of life that matter most for each group. This chapter focuses on the explanation, justification and implementation of the ME procedure on SWB panel data, and then offers simple descriptives of the results; the extension of the research could include deeper analysis in order to speak to the two policy areas above.

Chapter 2 | On the motivational power of pride: a field study with participants in a road–running event*

*We acknowledge financial support from the The Department of Economics, University of Oxford (George Webb Medley Fund), ESRC (Research Training Support Grant) as well as St John's College, University of Oxford. The Pre Analysis Plan for this study is registered under the title "On the motivational power of pride: a field study with participants in a road–running event" (AEARCTR-0004154). The Pre Analysis Plan for the replication is registered under the title "Replication of "On the motivational power of pride: a field study with participants in a road–running event"" (AEARCTR-0004710). The ethics approach from the Department of Economics in Oxford was approved on 2 April 2019, with approval number ECONCIA19-20-04. We also gained ethics approval for a replication of the study (containing some updates to the study design) on 28 August 2019, with approval number ECONCIA19-20-16.

Chapter Abstract

We investigate the effectiveness of incentives that emerge from a perceived threat to one's (self or social) image, triggered by another person's assessment: an ego threat. In a field experiment, we asked study participants to tell us their goal time to complete a running race, and then included a randomised treatment message which differed in whether we expressed doubts about their ability to achieve their goal time or offered them a reward for achieving it. We wish to understand: if you receive such a threat to your pride (an ego threat), does it change your motivations and the likelihood that you will achieve your goal? And, after the event, how does this treatment affect feelings of pride or satisfaction derived from your achievement (or failure)? We find evidence that the message expressing doubt is associated with a 12 percentage point increase in success rates (meeting the time goal). We find no evidence of an effect of offering financial rewards. We consider the welfare impact of the treatment, separately for those who succeeded and those who failed to meet their goal: any pride gain from meeting their goal may be cancelled out if the runner received our challenging message before the race. We are under–powered to examine details regarding mechanisms and heterogeneity and so we also present the planned replication (this is in progress).

2.1 | Introduction

In studying the effectiveness of incentives on behaviour, economic theory and empirical work has focused on positive incentives such as cash or material objects. This work finds only small effects (if any at all) of these incentives on individuals' motivation, effort or achievements: changing behaviour with material incentives is hard.[1] But recent work also finds that individuals derive utility from beliefs they hold about themselves (self image, such as wanting to be skillful or generous), and beliefs they feel others hold about them (social image). This is known as ego–utility (see for instance Bénabou and Tirole 2006, 2011). Empirical studies show that agents may be willing to incur costs and / or forfeit rewards to protect their social image (Bursztyn et al. 2017, 2020): feelings about your image can affect your choices.[2] In this chapter we seek to understand the effectiveness of incentives that emerge from a perceived threat to one's (self or social) image triggered by another person's assessment: an ego threat. We ask how a person's motivation to achieve their goal, and their subsequent behaviour, changes if told by another: "I would be surprised if you succeed."

It is not difficult to think of contexts where the desire to prove others wrong pushes people to perform beyond others', and possibly their own, expectations. Consider Alice who was told by her dad that she should forget about applying to Ivy League schools and concentrate instead on schools "at her level." Or think about Ben whose colleagues tell him that they think he does not have what it takes to be promoted. The power of an ego threat has not yet been leveraged in the economics literature. However, understanding pride is important in psychology. Pride is understood to be a powerful motivational force which propels individuals to push harder to achieve a goal: it is a motivator and mediator for perseverance (Williams and DeSteno 2008).[3] In recent experimental psychology research, there is a distinction between two facets of the pride emotion, *authentic* and *hubristic* (Tracy and Robins 2007). Authentic pride is considered to be positive, is associated with self–esteem and is generated by external changing events; hubristic pride is considered to be negative, is associated with narcissistic self–agrandizement and is more stable and internal. There is evidence that both facets may be part of a motivational suite of psychological adaptations underpinning a human drive for dominance, in the case of hubristic pride, and / or prestige, in the case of authentic pride (Cheng, Tracy and Henrich 2010).

In this chapter we present the design and results of a field experiment with participants in a road–running race. We asked runners to tell us their goal time to complete a race, and then included

[1] For a review, see Gneezy, Meier and Rey–Biel (2011).

[2] Individuals are willing to forfeit money to signal a political identity (Bursztyn et al. 2020) or choose to forfeit a career–enhancing opportunity to avoid signaling socially undesirable traits (Bursztyn et al. 2017).

[3] Williams and DeSteno show that pride (a positive emotion) can motivate effort for gaining social status or attaining costly skills: pride drives individuals to obtain image capital that they feel will be self–enhancing (2008).

a randomised treatment message which differed in whether we expressed doubts about their ability to achieve their goal time or offered them a reward for achieving it. Our ego threat message may provoke one or the other, or both facets of pride, as we invite the runners to prove us wrong by their own achievement (prestige) and indeed show themselves as relatively better that "most runners" (dominance), who we say tend to overestimate their abilities. We wish to understand: if you receive such a threat to your pride (an ego threat), does it change your motivations and the likelihood that you will achieve your goal? And, after the event, how does this treatment affect feelings of pride or satisfaction derived from your achievement (or failure)?

Motivation and study design

A context in which the demand for personal challenges manifests itself very clearly is participation in road-running events such as park–runs or marathons. Evidence shows that amateur runners in marathons exhibit behaviours consistent with reference dependence on round number finish times, as shown by their finish times significantly bunching *just faster* than round numbers (Markle et al. 2018; Allen et al. 2017). Runners are competitive, and research shows that running itself is good not only for physical health but is also connected to good feelings of self image, such as self esteem, and social image, such as social recognition (Zach et al. 2017).

Road–running races also provide a suitable context for an experimental field study, in that there are frequent large scale events all over the world, providing a large set of potential study participants. Indeed, participation in such races has hugely expanded in the last two decades.[4] Race finish times are a short term outcome that can be verified without relying on participant reports: running races tend to publish results online, often with details beyond simply the finish time but also split times and performance rankings. Additionally, evidence from marathon studies suggest that the majority of runners fail to achieve their goal times: Markle et al. (2018) measure the success rate of runners across 15 US marathons in meeting their goal time, finding that an average of 25.2% of runners are successful; Krawczyk and Wilamowski (2017) present evidence from 1,000,000 marathon runners suggesting that less than a third of runners meet their goal time. Hence, there is room for an experimental treatment to increase success rates.

We therefore chose to address our question about the power of an ego threat in a road running race: we conducted a study with 430 volunteer participants in a ten–mile running race in Washington DC.[5] Participants completed a Pre–race and Post–race survey, which we merged with registration and fin-

[4] In 2018 more than 8 million people took part in road running races internationally, which is four times the number that did so in 2001. See the Run Repeat website, last consulted in May 2021.

[5] We have since begun a replication study in a set of 10–mile and half–marathon races in Scotland and England in September and October 2019. There are details on this in Section 4 of this chapter.

isher data shared with us by the race director. In the Pre–race survey, participants answered basic questions about themselves, including their past race participation, current training, and their time goal for the race. After that, they were challenged to achieve the goal they set. More specifically, we had a 2 by 2 design: runners were randomised into one of 4 treatment conditions that differed in whether (i) we expressed doubts about their ability to achieve their time goal and / or (ii) they were rewarded for achieving their goal with a gift card payment (either a $10 or a $25 gift card). In our control group, we simply wished the participants good luck. Finally, to test whether the treatment manipulation introduced doubt in the runner's mind, we elicited respondents' probabilistic beliefs about their finish time on race day. In the Post–race survey, participants were asked to reflect on their race performance and participation in the study, and we collected stated and revealed preference measures of pride.

Our choice of message for the ego threat needed to balance a set of considerations. We wanted it to be provocative but we also tried to minimize the negative wording and adopt a teasing tone. We also wanted our evaluation to be perceived as a noisy signal of future race performance i.e. as somewhat informative (otherwise our message might have been too easily discarded), but not so informative as to make participants systematically revise their expectations downwards. We justify the wording in detail in Section 2, and evaluate its success in Section 4 when we discuss the results and outline changes in our planned replication of this study.

The inclusion of the pure incentives treatment, with random variation in the value of the voucher, was primarily included to try to calibrate the monetary equivalent of the ego threat message. The vouchers offer also connects our work with a large literature on optimal rewards. Theory and empirical work point to a possible crowding out effect of intrinsic motivations when financial rewards are offered in return for particular behaviour.[6] This would predict a null effect of our voucher offer (which is indeed what we find in this study).

In our use of the survey data, we have the following primary aim: evaluate the motivational effect of doubting a runner's performance. We estimate the overall impact on a range of race performance outcomes: whether participants met their goal (extensive margin), how far ahead or behind their goal they finished (intensive margin), clustering of finish times just faster than goal times (bunching) and pacing strategies across the splits.[7] We look at possible sources of heterogeneity in the treatment effects: gender, race experience, race attachment and goal attachment. We look at evidence for a psychological

[6] Bénabou and Tirole show that performance incentives are weakly positive for performance in the short run, but negative in the long run (2003). Monetary rewards tied to performance are found to have a detrimental effect on intrinsic motivation (Gneezy and Rustichini 2000). There is a flat response of performance to monetary incentives when the task generates strong intrinsic motivations (Dessi and Rustichini 2015).

[7] Note that we pre–registered the extensive and intensive margins of goal attainment as primary outcomes for our study, but did not pre–register the bunching analysis.

channel through which the treatments could have worked, or not worked: changing participants' beliefs about their chance of success, if they believe the doubting message signal is even relevant to them. We measure both certainty (entropy) and expected finish times. We explore a possible dampening mechanism for the treatment effect: (reported) recall of the treatment message.[8]

We also seek to measure the psychological and emotional impacts of our study. Although generating ego threat may improve performance outcomes, we need to look beyond behaviour to address any emotional costs or welfare impacts of such an intervention. Does the manipulation erode the pride that runners report from their achievement? Might runners even dampen their expressions of pride, in a desire to avoid signalling that our threat affected them?

To investigate these questions, in the Post–race survey we administered the Tracy and Robins (2007) psychological pride scale, which measures both positive (authentic) pride and negative (hubristic) pride. We also offered runners who had received the voucher treatment the opportunity to make a counter signal: turn down the voucher and instead choose that we make a donation to charity on their behalf.[9] This allows us to explore whether successful participants who had both the message and the incentives chose to turn down the vouchers as a way to deny the effect of our comment: an over–justification effect may mean that runners choose to give the voucher away, to prevent the inference that they did it for the money (Benabou and Tirole 2006). However, there is evidence that working for something and feeling you won the (non–cash) reward thanks to your own efforts increases its value to an agent.[10] Our small voucher reward could become more interesting and valuable to runners who pushed harder to win it, which may mean they are unlikely to choose to donate it to charity. The charity donation signal alone is ambiguous and needs to be considered alongside reported pride: higher reported positive pride and a charity donation could suggest a boost to self–esteem and hence a lower demand for any kind of external reward (as in Bursztyn et al. 2018). Alternatively, if our ego threat causes disutility, and if a runner is then successful, they might choose to reciprocate by refusing to grant the researcher any utility from seeing that we affected their behaviour: we might expect in this case a lower reported pride (positive and / or negative) together with a charity donation.

[8] Benabou and Tirole connect a person's scepticism with respect to information from others vs. their own memories as represented by Bayes' rule, and that rational self–deception can result in biased beliefs in order to preserve self–esteem (2005). In our context, bayesian updating could be downwards (believing the message, and hence losing confidence) but it also could be upwards, if runners recognise and internalise the message's motivating effect and hence anticipate they will be more likely to meet their goal time. Or, runners might simply suppress the message as soon as they read it or later consciously forget the information that they perceived to be harmful. See Golman, Hagmann and Loewenstein (2017) for a review on information avoidance; Bénabou and Tirole (2002) and Carrillo and Mariotti (2000) on deliberate forgetting. This would dampen any behavioural response to the message. Finally, it could be that the message is simply not impactful and runners forget it for this reason, rather than *motivated* forgetting.

[9] Note that by making the donation on their behalf (and not in their own name), we eliminate the possibility of a social image benefit with regards to the charity.

[10] Schweyer, Landry and Whillans, Incentive Research Foundation 2018; Loewenstein and Issaheroff 1994; Peterson, Rowe and Anderson 2016; Buhren and Plesner 2014.

Our interest in testing both the behavioural and emotional impact of an ego threat is connected to recent work testing the connection between behaviour and self—image construction: van der Weele and von Siemens find that self—image management is important, but that it is at times a weak force, because individuals can opt to reinterpret their own poor actions or undesirable behaviour in a less negative manner, and reduce such behaviour's informational value for revising their own personal traits (2020). If there is room to avoid diagnosing behaviour potentially associated with personal moral flaws as such, an individual may take it. We note that this could have repercussions on our own findings: an individual who did not make their goal may choose to recall an easier one, choose to suppress their recall of the treatment or alter their reported pride to signal to us and to themselves that it was not an important goal to them anyway.

Results overview

Race performance results suggest that the message treatment is associated with 12 percentage point higher success rate (meeting the goal time), relative to a success rate of 44% in control. Men in particular see a intensive margin change as a result of the message: their finish times were 2.0 percentage points faster than their goal with the message vs. 1.1 percentage points slower than their goal time without the message. We find evidence for an asymmetric bunching of runners who saw the message to finish *just faster than* their goal time. Pacing evidence from split times suggests that right from the start of the race, the message treatment runners are more likely to be on track to meet their goal than the other runners. We find no effects of the financial incentives on race performance.

Our belief measures relative to actual performance suggests that our runners were not systematically overconfident, unlike existing evidence on marathons (Markle et al. 2018; Krawczyk and Wilamowski 2017). However, we do not see any difference in these beliefs between the treatments in the full sample. We do find suggestive evidence that first–timers and those more attached to the race (those who pre–ordered merchandise) are more uncertain about their chances of meeting their goal if they received our challenging message. In terms of reported recall of the treatments, we find that men are 17% more likely to recall the challenging message if they met their goal, and non first–timers are 20% more likely to recall it if they met their goal: perhaps the male and experienced runners who failed to meet their goal had suppressed the message, hence did not feel any effect. But our reported recall measures are noisy, and the direction of causality between reported recall and message effect is not clear. These reported recall findings are fairly weak and it could be the message was simply not impactful and thus forgotten (rather than deliberately not remembered) by some runners.

When we consider the welfare impact of the study, we find that runners who had the message treatment and succeeded in their goal time report lower pride than those who did not have the message and succeeded. The message may have induced some sort of negative image utility, although we may have demand effects (treated runners not wishing to report pride to the researcher). Individuals who received the message were more likely to give away their winnings: given that they also reported lower pride, this could be a signal that they did not meet their goal because of our study i.e. a counter signal. We aim to collect more welfare related measures in the replication of this study to investigate these suggestive findings in more detail, and with greater power to identify effect sizes.

Chapter overview

This chapter is organised as follows: in Section 2 we outline the experimental setup. In Section 3 we report the findings for overall race performance, and examine beliefs and reported recall data to consider possible mechanisms for the egp threat message to have had the observed behavioural effect. We also report some evidence related to the psychological impact of the treatment on our participants who met and did not meet their goals. In Section 4, we discuss the findings from the current dataset. We also highlight limitations in our understanding of the mechanisms underpinning the message effect and the gaps in our evaluation of the welfare impact of the study. In the light of these limitations, we decided to conduct a replication of the study (this is on–going).

2.2 | Experimental Setup

To investigate the motivation emerging from perceived threats to one's (self or social) image, triggered by others' assessments, we conducted a two–part study: a Pre– and Post– race survey of participants in a large 10 mile running race in Washington DC. The Pre–race survey contained cross–randomised experimental treatments. We combine the data collected in the two surveys with data collected by the organisers of the event. We partnered with the race director to recruit runners using adverts on Facebook, in Race E–News and in–person recruitment at the Race Expo. We planned for and recruited 430 runners into the Pre–race survey and had a very low attrition rate for the Post–race survey (93% completion).

2.2.1 Study design

Data sources and survey outline

For the study, we use a combination of data sources; these are summarized in Table 2.1. We collected survey data from the study participants before and after a large running race: the Pre–race survey and Post–race survey. We also have access to the Registration and Performance data from all runners in the race, and historical Registration and Performance data from the 2018, 2017 and 2016 of the same event. This section describes the content of the study surveys (the full survey questionnaires are found in Appendix A). We also outline the data available from the Registration and Performance data collected by the race organisers.

Pre–race survey	• Demographics; race preparation; prior running experience • Goal time; goal related motivations; beliefs about goal attainment • Randomisation into 1 of 4 treatments with cross–randomised incentives
Post–race survey	• Recall questions; behavioural and subjective measures of pride
Registration form	• Race distance; Socio–demographics; race experience • Additional purchases
Performance data	• Chip time and split times; rank and age–grade; relative pace information

Table 2.1: Data sources

Notes: The Pre–race survey, and Post–race survey are completed by study participants only. The Registration data and Performance data are available for all runners in the race.

In the **Pre–race survey** we collected basic socio-demographic data as well as data on past race participation and current training. We then asked participants about their goal time for the race. More precisely, participants were asked what time (hh:mm) they were aiming for on race day (*goal time*), how much it would mean for them to achieve their goal (*goal importance*), how committed they were to meet their goal (*goal commitment*), and how confident they were that they would meet their goal (*goal confidence*). This provides us the information for our primary outcome variable: did the participant meet their goal time when they ran the race? It also provides information on their certainty and dedication to their goal time prior to the experimental treatment.

We then included the main treatment for the study: we encouraged or challenged participants to achieve their declared time goal. More specifically, after submitting their time goal, participants were randomized into one of 4 treatment conditions that varied whether we (i) expressed doubts about their ability to achieve their time goal (in the Message (M) and Message and Incentive treatments (MI)); (ii) rewarded them for achieving their goal (in the pure Incentive treatment (I) and the MI treatment). The Control (C) group had neither the message nor financial incentive. Table 2.2 shows the text that was displayed in each of the treatments. Our choice of wording for the message in the M and MI treatments is discussed in detail below.

In the two incentive treatments, I and MI, we cross randomized the level of incentives to be either $10 or $25 sports clothing vouchers. We introduced this additional variation to estimate the responsiveness of race performance to financial incentives. With the MI treatment, we aimed to test whether the message manipulation has an effect even when the message is cheap talk or if putting money on the line is necessary for motivation to increase and affect performance. The level of incentives was selected based on budget and sample size considerations, with an effort to maximize the difference between the two amounts (further details below). Importantly, this was in the form of sports clothing vouchers: we may therefore underestimate any effect of financial incentives, as the willingness to pay for such a voucher (given the constraint on where and on what it could be used) may be below the purchase price of the voucher.

After the treatment manipulation, we included a question to test whether runners in the M and MI treatments would form more pessimistic beliefs about their finish time than runners in the C and I treatments.[11] To do so, we elicited respondents' probabilistic beliefs about their finish time on race day, using their goal time as the reference point.[12] We use these data to calculate the *belief entropy*

[11] Immediately before the belief elicitation question, we also included two questions about time tracking i.e., whether the participant planned to keep track of their time during the race using a watch/phone and whether they usually do so. Those questions were mainly used as buffer questions to increase distance between the treatment manipulation and the belief question in an attempt to minimize demand effects.

[12] More precisely, we asked respondents to estimate the chances that they would be N minutes faster/slower than their

Control (**C**)	Good luck with the race! Don't forget to pace yourself. Let's see how things go on Sunday!
Message (**M**)	Good luck with the race! Don't forget to pace yourself. Runners often are too optimistic about the goals they set. We have to say we would be surprised if you met the goal you chose. Although we're certain you can make it in your dreams ☺ Let's see how things go on Sunday!
Message and Incentive (**MI**)	Good luck with the race! Don't forget to pace yourself. Runners are often too optimistic about the goals they set. We have to say that we would be surprised if you met the goal you chose. Although we're certain you can make it in your dreams ☺ In fact, we are willing to bet money on your race performance: if we are wrong and you meet your goal, we will send you a $X Under Armour gift card by email after you have completed Survey 2! Let's see how things go on Sunday!
Incentive (**I**)	Good luck with the race! Don't forget to pace yourself. If you meet your goal, we will send you a $X Under Armour gift card by email after you have completed Survey 2! Let's see how things go on Sunday!

Table 2.2: Treatment messages

Notes: $ X varied to be $10 or $25. In our first Pre Analysis Plan, we abbreviated our treatments as C, P (for Pride message), M (for money) and PM (Pride message and money).

of each participant: their level of certainty or uncertainty about attaining their goal time. We also calculate participants' *expected finish*. We included this question to facilitate the understanding of a possible transmission mechanism for the treatment messages onto participants' behaviour. Do the treatments cause individuals to doubt themselves? Are some more prone than others to revise downwards their beliefs in their own capabilities? Do some internalise the motivational impact of such a message and thus actually boost their confidence that they will meet their goal and prove us wrong?

In the **Post–race survey** we collected data about (i) participants' recall of and satisfaction about their race performance, (ii) general feelings of pride, (iii) interest in race merchandise / memorabilia; (iv)

time goal of hh:mm, with time brackets {Less than one minute, 1:00–4:59 minutes, 5:00–9:59 minutes, At least 10:00 minutes}. Participants could enter any number between 0 and 100 for each time bracket with the constraint that their answers sum to 100.

recall/comments regarding the Pre–race survey. The recall questions allow us to evaluate respondents' accuracy relative to their actual goal and finish time and ascertain how well they remembered the treatment message, if at all. We use this to explore the possibility that runners might deliberately ignore and therefore not remember the treatment message as a defensive mechanism against a discouraging signal about their ability. We also offered participants in the I or MI treatments the opportunity to turn down the clothing voucher and instead choose that we make a donation to charity on their behalf. We included this option to investigate whether successful participants in the IM condition might turn down the incentives as a way to deny the effect of our message; this offered them a chance to signal "not caring" about our message. Note that by making the donation on their behalf (and not in their own name), we eliminate the possibility of a social image benefit vis–à–vis the charity.

We also have access to the **Registration** and **Performance** data from the event organisers. The performance data contains the official race time of participants, measured through a chip located at the back of their bib number, which records individual start times and finish times, as well as the time taken to cross the 5k (3.11 mi), 10k (6.21 mi), and 15k (9.32 mi) marks. We make use of these data to investigate how the treatment may affect runners' pace at different stages of the race.

Treatment message: ethical considerations

In the message treatments (M and MI), our choice of words attempted to balance several considerations. We wanted to minimise any potential harm to participants, either emotional from the message, or physical from pushing too hard as a response to the message. This had to be balanced with the desire to provoke some effect on behaviour. Also, it was important the message not be deceitful, and be perceived as somewhat informative.

Emotional Impact: Although our message was deliberately provocative, we tried to minimize the negative wording and instead adopt a teasing tone. We knew that there would be a distribution of emotional impact, and that not all participants would perceive the message as merely teasing. Therefore, participants were given two opportunities to give written feedback in the Post–race survey, and we report the findings in this text in Section 4 of this chapter, where we discuss the success of the treatment message. To minimise any long–term effects of the treatment, we also sent out a debriefing email in the autumn of 2019 to explain to participants what the study was about.

Message Credibility: We wanted our evaluation to be perceived as a noisy signal of future race performance i.e. as somewhat informative (otherwise our message might have been too easily discarded), but not so informative as to make participants systematically revise their expectations downwards.

We envisaged a potential mechanism through which the treatment would encourage a runner to push harder during the race to achieve their goal. We hypothesised that triggering too much uncertainty could prevent some runners from believing they could meet their goal at all, and therefore giving up on the goal entirely or even — worst case — the race. This would entail a runner reading the treatment as precise and accurate, becoming strongly negatively certain about the impossibility of meeting their goal, and therefore revising it downwards. This is connected to our conjecture that there may be a difference in the behavioural response to uncertainty in yourself (do I have faith in myself?) and uncertainty from another (do they believe in me?). If a message were too strong and be perceived as an undeniable truth, runners might lose faith in themselves rather than push back and wish to prove us wrong. We therefore looked for a message that would be to a degree messy in its informational value and be read as unfair by the runner, but that would be correlated with the truth and keep the goal time for the runner the same.

We therefore chose to remain ambiguous about the basis of our assessment, which was not person-specific, but instead pertained to the population of runners as a whole. The signal was based on empirical evidence showing that on average only 26.9% of marathon runners meet their goal time, computed across a sample of 283,651 runners across 15 marathons; 37.5% was the upper bound for goal attainment, for runners who reported a very high attachment to their goal time (Sackett et al. 2018). We therefore considered that our message was not deceitful: past evidence demonstrated that the majority of runners do indeed fail to meet their goal. However, to address any concern that the message might have mislead runners into thinking that we used their personal data to determine they were unlikely to meet their goal, we asked participants about their beliefs in this regard in some follow up questions, and discuss the findings of these in Section 4 of this chapter.

Physical Harm: Participants were challenged to attain the goal time that they set for themselves: there was a non-zero risk that participants might injure themselves by pushing too hard in order to meet their goal. To minimize this risk, the instructions emphasized that participants should pace themselves. In the consent form, we also emphasized that we "hold no responsibility for potential injuries or accidents that might occur during the race." Finally, although the self–set goals might be over–optimistic, we did not challenge participants to go above and beyond their personal goals.

2.2.2 Study context and set–up

The Credit Union Cherry Blossom 10–mile Run

We chose to conduct our study with participants in a large race in Washington DC, the Credit Union Cherry Blossom Ten Mile Run 2019 (CUCB). This is usually oversubscribed and race entries are

allocated through a lottery mechanism: in 2019, there were 23,223 applicants, 19,691 of whom received a place to run and 17,800 completed the 10–mile distance on the day. We chose this race for a number of reasons: (1) the number of participants offered a large pool of runners from which we could recruit, (2) the 10 mile distance being challenging for participants but not so long we felt a intervention to push runners might be physically harmful (e.g. in a marathon), (3) the level of interest of the race–director in helping us with recruitment and his willingness to share with us the highly detailed data collected in the Registration form and the Performance data and (4) the existence of data on ex ante and ex post purchase decisions commemorating race performance in the Registration data.

Recruitment and study timeline

We recruited 430 runners to take part in a Pre– and Post– race survey of less than 10 minutes each.[13] Figure 2.1 shows the data collection process.

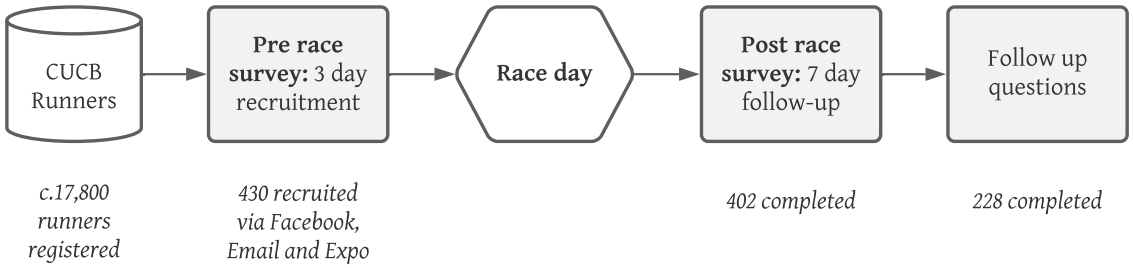


Figure 2.1: Data collection timeline

Notes: The follow–up questions were optional and were combined into a lottery–winner announcement, in which participants entered their bibnumber to find out if they won a prize or not.

Participants were invited to sign up and complete the Pre–race survey from 3 days before the race. At the beginning, they read and signed a consent form, the precise content of which is in Appendix A. The study was advertised as a "study about participation of runners in races" conducted by social science researchers at the University of Oxford in partnership with the race organisers. As an incentive for runners to take part in the study, we advertised that participants would be entered into a lottery to win prizes provided that they (1) completed both surveys and (2) completed the Post–race survey within 7 days of receiving it via email. We recruited participants through 3 channels: (1) Facebook adverts on the official CUCB race event page;[14] (2) An advert in the Pre–Race E–News from the race director, sent to all runners; (3) In–person recruitment at a stand at the Race Expo. In total, 75% of study participants were recruited via email or Facebook, with the remainder recruited at the Expo. In our analysis we

[13] The full survey questions are in Appendix A.

[14] On the date of the advert, there were 16,096 individuals who followed this page. The adverts were posted in the group by the official CUCB social media Facebook account.

control for recruitment channel, acknowledging there may be demographic differences between those recruited via the different channels and also the potential for psychological differences in study participation from having "seen the researcher" at the Expo, versus the greater anonymity of online sign up.

Table 15 in Appendix B demonstrates that the characteristics of participants in the full CUCB 10 mile race are broadly comparable to the runners who signed up for our study.[15] However, using historical data we had estimated that the CUCB race would have approximately 60% female runners: the 2019 had 61% female participants, 2018 had 59%, 2017 had 60% and 2016 had 62%. We wanted to recruit a 50:50 split between male and female runners in the study, because we anticipated that gender could be an important moderator for the treatment effect and therefore we took action to deliberately over–sample male runners. Female runners were much faster to complete the first survey than men: at the end of the first recruitment day (only the Facebook link had been shared so far) we had already reached the cap of 214 female sign–ups but had fewer than 50 male sign–ups. At this point the survey was closed to female runners and recruitment at the Race Expo was exclusively targeted at male runners. The fact that only male runners, and indeed approximately half of the male runners, were recruited in person at the Expo, confounds the gender heterogeneity analysis, as we were forced to have a different recruitment style for men and women.

We look at both first–timer and pre orders of medals as possible dimensions of heterogeneity in our treatment effects. Table 15 shows the number of CUCB first–timers is approximately the same in the study sample as in the full population, which historically is always around 50%. The study participants ex ante appear to be more enthusiastic about the race: purchasing medals or technical t–shirts via the Registration form for the race. This could be explained by more enthusiastic racers perhaps being more likely to (1) be members of a Facebook group for the CUCB event, (2) read the full Pre–Race E–News or (3) attend the Expo, and hence they would be more likely to see the advert for the study via one of the three recruitment channels. Alternatively, conditional on seeing the study advert, those who pre–purchase CUCB merchandise may simply be keener to interact with other CUCB related activities, and hence sign up to a research study advertised to CUCB runners, especially one that advertises sports gear vouchers as prizes for participants.

[15] Note however: looking at the distribution of finish times displayed in Figure 14 in Appendix B, the study participants had a slower mean time to complete the 10 mile distance, by 6 minutes. This is not surprising, as there are a number of elite athletes who compete in the race, and these runners were no doubt more unlikely to sign up to the running study than amateur running enthusiasts.

2.2.3 Sample size and balance tables

Sample size

The choice of sample size was partly dictated by budget constraints. Besides the voucher payments offered to participants in the I and MI treatments for meeting their goal time for the race, we ran a lottery with voucher prizes available to all participants. Our limited budget together with power considerations, the anticipation of at least a 10% attrition rate between the Pre– and Post–race surveys, and the expectation that not all runners would meet their target race time resulted in our decision to recruit as close to 430 runners into the study as possible.

The choice of sample size — resulting in the comparison between groups of approximately 100 to 110 survey participants in the control and each treatment respectively — dictates a minimum detectable effect size. The main outcome variable is the percentage of runners who meet their goal time. Based on prior evidence, we estimated that 25% of runners in the control group would do so.[16] With $\alpha = 0.05$, a control mean of 25%, 100 runners per condition, and 80% power, we would be able to detect a 19 percentage point difference (or larger) between treatments and control. To be able to detect smaller effect sizes such as a 10 percentage point difference, one would need over 300 runners per treatment group (thus, 1,200 overall). Note that we investigate four dimensions of heterogeneity: gender, experience with CUCB, goal importance and race attachment. Each moderator splits the sample approximately in half. We therefore also report the power analysis for half the full sample: with alpha = 0.05, a control mean of 25%, 50 runners per condition and 80% power, we would be able to detect a 27 percentage point difference (or larger) between treatments and control. Full power calculations can be found in Appendix B.

Balance tables and attrition

Table 15 in Appendix B demonstrates that the randomisation of treatments was effective across the socio–demographic characteristics of the survey participants: for instance, there is approximately a 50:50 split of males to females in each of the treatment categories and control. This is useful as gender is one of the key dimensions of heterogeneity in our empirical analysis. Likewise, there is a similar level of reported training in each of the treatments, and % of first–timers.

430 eligible runners completed the Pre–race survey, and 402 (93%) completed the Post–race survey. One concern we had was that those in the M treatment would be offended by the message and therefore

[16] Markle et al. (2018) measure the success rate of runners across 15 US marathons in meeting their goal time, finding that an average of 25.2% of runners are successful. Krawczyk and Wilamowski (2017) present evidence from 1,000,000 marathon runners suggesting that less than a third of runners meet their goal time.

would be more likely to drop out of the study, relative to the control group or those offered money. This is not the case: attrition across treatment groups did not vary substantially across treatments (see Table 16 in Appendix B), and indeed was extremely low overall. 93% of runners in C completed the Post–race survey, 92% of those in I, 93% of those in M and 96% of those in MI. Additionally, similar numbers of participants from the treatments answered the optional follow up questions, with 76% responding overall. Overall, this is a very low level of attrition, with fewer participants failing to complete the Post–race survey than we anticipated, and more participants being willing to answer the optional follow up questions; we consider this is a positive sign that runners really engaged with and enjoyed their participation in the study, justifying the wording of our treatment message.

2.3 | Results

The main focus of this section is to present evidence of any behavioural changes from our experiment: differences in race performance across our treatments. We test for differences in goal attainment (extensive margin), percentage deviation of finish time from the goal time (intensive margin), clustering of finish times just faster than goal times (bunching to the left of the goal time) and pacing strategies across 5k splits. We then delve into evidence for a potential psychological channel through which the treatments could have worked: changing participants' beliefs about their chances of success. We explore a possible dampening mechanism for the treatment effect related to reported recall of the message. Beyond understanding behavioural change, we want to understand the welfare impacts of our study, and so we also present evidence on participants' post race pride and satisfaction, and their choices to keep or donate any voucher winnings.

For all the outcomes reported in this section, we conduct separate estimates for the dimensions of heterogeneity that were pre–registered: gender (*male* vs. *female*), past experience with the CUCB race (*first–timer* vs. *not first–timer*), and attachment to the event as measured through pre-race purchase of the race medal (*medal ordered* vs. *medal not ordered*). We also look at differences according to reported *goal attachment* (=1 if above the median attachment to the goal, which is calculated as the mean of goal commitment, importance and confidence, as reported by the participants before the experimental treatment in the Pre–race survey). All 4 moderators split the sample into approximately equal groups: in the sample there are 49.7% females, 45.0% first–timers, 54.0% pre–ordered the medal, 52.4% are attached to their goal. Table 2.3 presents the correlations between these moderators.

	Female	First–timer	Medal order	Goal attachment
Female	1.0			
	(NA)			
First–timer	0.13***	1.0		
	(0.01)	(NA)		
Medal order	0.14***	0.08*	1.0	
	(0.00)	(0.09)	(NA)	
Goal attachment	0.11**	0.10**	0.19***	1.0
	(0.02)	(0.04)	(0.00)	(NA)

Table 2.3: Moderator correlations

Notes: These are all binary variables, with the reported correlation between the 1 values in each case. Reported goal attachment = 1 for those with above median goal attachment (calculated as the mean of goal commitment, importance and confidence, as reported before any treatments in the Pre–race survey). First–timer is a first time runner of the CUCB 10 mile race. The Pearson correlation is reported, together with the p–value in brackets.

The key primary outcome variable that we pre—registered is the percentage of runners who met their goal in each treatment. However, the central question of the research – of a change in behaviour within the race resulting from the experimental treatment – is tested across a set of outcomes, and for different subgroups within the main sample. This requires a consideration of the issue of multiple hypothesis testing and possible adoption of associated corrections. In the results presented here we have not conducted such corrections, and we therefore note that with such corrections we would lose some power of our statistical tests of significance on our set of additional outcome variables for race-performance, and also for the set of outcomes measured to understand the underlying mechanisms and welfare impact.[17]

2.3.1 Behavioural change: race performance

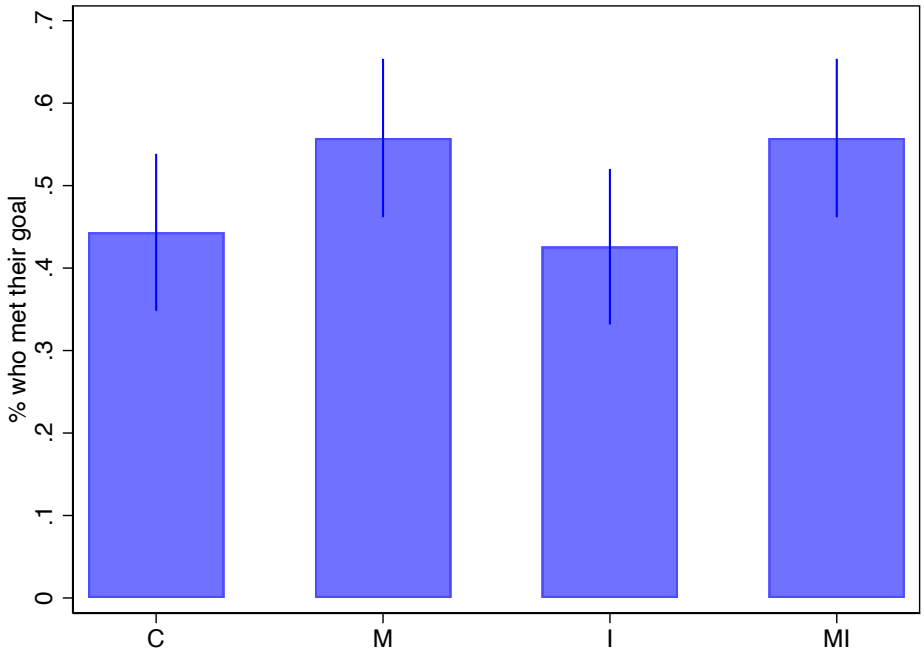


Figure 2.2: Goal attainment by message — Full sample

Notes: C: 44.3% goal attainment with 95% CI = [34.8 — 53.9]. M: 55.8% goal attainment with 95% CI = [46.2 — 65.4]. I: 42.6% goal attainment with 95% CI = [33.2 — 52.0]. MI: 55.8% goal attainment with 95% CI = [46.2 — 65.4]. This is the strict definition of goal attainment. The effect of the pride message is slightly weaker when using the lenient definition of goal attainment but the finding is robust (9.5 percentage point difference, $p<0.05$).

[17] One approach would be to combine outcomes into families and compute a single summary outcome for each family (correcting for the existence of multiple families), or indeed a single index combining all outcomes in all families (see for example Bessone et al. 2021). In our case, it is not straightforward to combine the race performance variables into a single outcome because of the heterogeneous nature of the outcomes: we have an individual level indicator variable for meeting the goal and individual percentage deviations from goals, but also aggregate bunching data and aggregate split time data. We do report an index outcome for the Tracy and Robins pride scores, collapsing the 14 measures into 2 summary measures: positive and negative pride. When reporting outcomes for each individual outcome within a family of outcomes, p-values could be adjusted to control for the family-level error rate (the chance of at least one false rejection), using the Westfall and Young procedure (1993). When we complete the replication data collection, which has more outcomes, especially in the welfare and beliefs families of outcomes, we will investigate applying this correction technique.

Goal attainment (extensive margin) Figure 2.2 presents the proportion of runners who met their goal in each treatment.[18] A clear, and surprising, initial finding is the very high success rate in the control group (44.3%). This is much larger than our hypothesis of 25%, which we based on prior evidence (Markle et al. 2018; Krawczyk and Wilamowski 2017). This surprised us, and we discuss in Section 4 why this may have been the case in our context and how it motivates the replication.

In Figure 2.2, the message increases goal attainment by about 12 percentage points (55% vs. 43%, z=-2.53, p=0.01). We also conduct regressions to estimate the effect of the treatment manipulations relative to control: $message_i$ is an indicator variable taking value 1 if participant i was in either message treatments (M or MI) and Z_i is a continuous variable for the level of incentives offered ($10 or $25 in I and MI, $0 in C and M). For goal attainment, and for the other outcome variables reported in this section, we adopt the following specifications:

$$Y_i = \beta_0 + \beta_1 message_i + \beta_2 Z_i + \mathbf{X}_i'\gamma + \alpha_s + \epsilon_i \tag{2.1}$$

$$Y_i = \beta_0 + \beta_1 message_i + \beta_2 Z_i + \beta_3 (message_i \times Z_i) + \mathbf{X}_i'\gamma + \alpha_s + \epsilon_i \tag{2.2}$$

where $\mathbf{X}_i$ is a vector of controls and α_s is a dummy for the mode of recruitment (online or in person at the Expo).[19] The regressions in Table 17 in Appendix C find a positive effect of the message, with a 10–12 percentage point increase in goal attainment robust to the addition of race and socio–demographic controls. The treatments interaction term is not significant. We did not have a strong prior regarding the effect of the message, but this evidence suggests that on average it pushed individuals to achieve their goal.

The positive effect of the message is particularly large for first–timers, with the message boosting goal attainment by 14 percentage points (45.7% vs. 59.4%, z=-1.99, p=0.06), and not significant for non first–timers. Perhaps first–timers are most susceptible behaviourally to a psychological intervention of the type we employed. For those attached to their goal, the effect of the message is even larger, with a difference of 16.7 percentage points (39.3% vs. 56.0%, z=-2.48, p=0.01), and no significant difference for those not attached to their goal. The other dimensions of heterogeneity (gender and medal ordered) have similar goal attainment rates to the full sample.

[18] After completing the study, we considered with hindsight that in our treatment message "meeting your goal time" was ambiguous as to the criteria for success. For a runner who had as a goal time 1:30, would a chip time of 1:30:59 be a success? We define "strict" goal attainment as running up to or under the goal time, i.e. 1:30:01 would be failure, and "leniant" as running up to 59 seconds slower that the goal time, i.e. 1:30:59 would be success. Our effect size is strongest for the strict measure, but the positive effect of the message is robust when we look at the leniant measure.

[19] We control for race–specific variables, socio–demographics, race performance measures and survey response variables; the choice of which sets of controls to include for regressions of the sets of outcomes was pre–registered, and full details of the controls are specified in the regression table notes.

Figure 2.2 shows no effect of the voucher incentives on performance; this is confirmed by the regressions in Table 17. This was counter to one of our main hypotheses: we expected that participants' performance would be higher when offered vouchers relative to control. We also expected a positive interaction effect of the message and incentives, but we find adding vouchers to the message has no effect on goal attainment. These null effects may be due to the fact that $10 and $25 vouchers are too small to influence behaviours. In our replication we offer a higher level of incentives to some participants.

Percentage deviation of finish time from goal time (intensive margin) We find small effects of the message treatment on this outcome. Runners without the message finish on average slower than their goal by 2 percentage points (95% CI = [0.6, 3.6]); those with the message are 0.4% faster (95% CI = [-1.1, 1.9]). This difference is not significant in the full sample and regression results in Table 18 find no evidence for a message effect. We do find that there is a significant difference for males. Pooling the message treatments, men in M and MI were faster than men in I and C (2.0 vs. -1.1, t=2.28, p=0.02). Note that the median finish time is 95 minutes, so a 2.0 percentage point deviation for the median runner represents approximately a 2 minute deviation from their goal time. There are no significant effects for women. Figure 17 in Appendix 3 illustrates the kernel density plots for those with and without the message in the full sample, and for men and females separately. We find no differences by first–timer, medal pre order or goal attachment.

Bunching of runners just faster than their goal time Figure 2.3 presents the number of runners whose finish time is within 8 time intervals relative to their goal, by message treatment. The intervals are chosen to facilitate comparison to the reported beliefs data, which are illustrated in Figure 2.5. We look for evidence of bunching to the left of the goal time (the two palest green bars), as evidence of caring about beating or just meeting the goal time. Those with the message are more likely than those without the message to beat their goal by between 1 to 5 minutes (25.0% vs. 17.3%, z= -1.94, p=0.05). Pooling runners who missed their goal by between 0 and 5 minutes (the 2 palest bars each side of the goal time) finds that those without the message are more likely than those without to miss their goal by 5 minutes or less (32.2% vs. 25.0%, z=1.64, p=0.1). We test for a difference in differences: the difference in the number of runners who are 0 to 5 minutes slower than their goal time (the two palest red bars) minus the number who are 0 to 5 minutes faster than their goal (the two palest green bars), by message treatment. For those with the message, there are 64 more runners in the 0 to 5 faster interval than the 0 to 5 slower interval, which is 30.8% of the message sub sample; for those without the message, there are 25 fewer runners in the 0 to 5 faster interval than the 0 to 5 slower interval, which is 11.7% of the non message sub sample (-30.1 vs. 11.7, t=2.32, p=0.02). The bunching of treated runners just to left of the goal time is suggestive of increased motivation for these runners

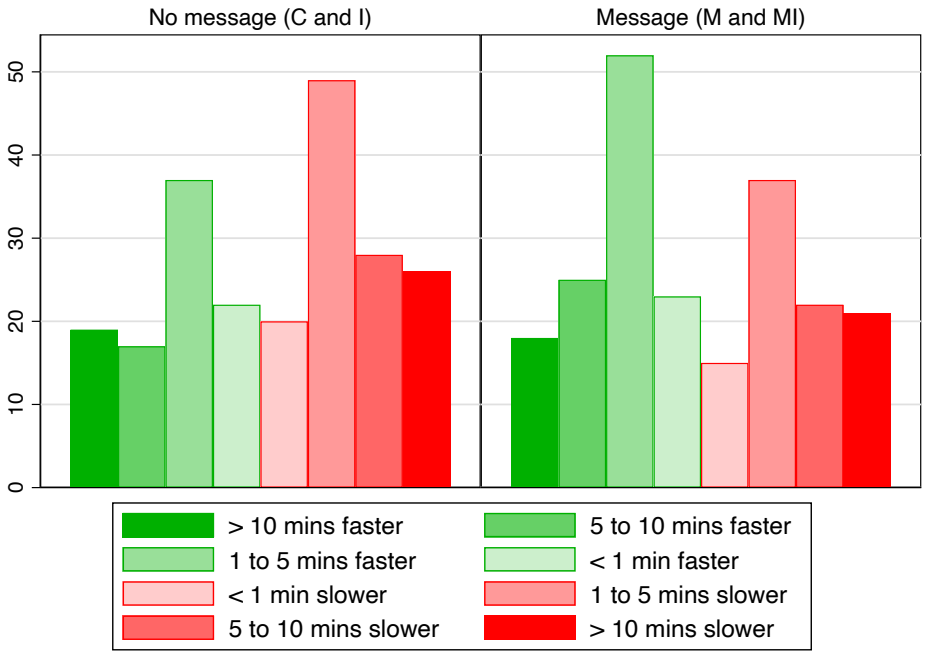


Figure 2.3: Closeness to goal time: by message

Notes: This graph presents the number of runners whose finish time is within a set of 8 intervals set by their own goal time. The intervals are chosen to facilitate comparison to the reported beliefs data, which are illustrated in Figure 2.5. We interpret bunching just to the left of the goal (the two palest green bars) as evidence of caring about beating or just meeting the goal time.

who would finish close–*ish* to their goal to push just a little more to prove us wrong.

Pacing strategy We construct variables for whether the participant was on track to meet their goal at the 5k, 10k and 15k marks. The proportion of runners on track in the message versus non message treatments are in Figure 2.4. Those in the message treatment exhibit more consistent effort across the race, with an increasing proportion of runners in this group on track to meet their goal as the race draws to a close. The majority of those without the message start too slowly and then also slow down in the final 5k. 10% more runners in the message treatment are on track at 5k (48.1% vs. 37.9%, z=-2.12, p=0.03), an insignificant 7% more runners are on track at 10k, and 12% more are on track at 15k (55.8% vs. 43.9%, z=-2.43, p=0.02).

Chip times (absolute performance) We also look at the raw finish times of the runners who were exposed to different treatments, to ascertain whether irrespective of goals, study participants ran faster or slower after they read the message / were offered financial rewards tied to their performance. We hypothesised that there could be a higher variance in the distribution of performance in the message treatments due to heterogeneity in the emotional reaction to the message. We do not find significant differences in variance nor mean finish times in the full sample, nor when we examine the sub samples

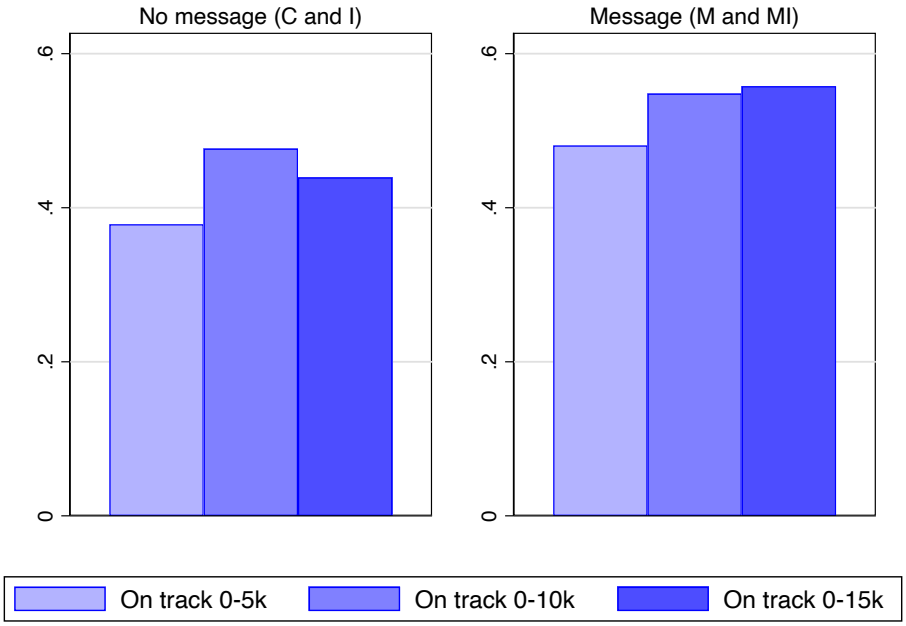


Figure 2.4: On track to meet goal, by message — Full sample

Notes: The bars represent the proportion of individuals who were on track to meet their goal at each 5k mark of the race.

by heterogeneity. This may in part be due to the large spread of finish times across a relatively small sample size.

Overall, we find evidence that challenging individuals to reach their goal results in greater success. There is no clear effect of incentives on race performance. Our results are weakened by the small effect sizes overall (in part due to the surprisingly high success rate in the C group).

2.3.2 Mechanisms

Ex ante beliefs about performance

We want to test for evidence of participants updating their beliefs about their chances of success after the treatment manipulation. In principle, the effect of the treatment on beliefs is ambiguous, because runners could internalise the treatment effect. In the Pre-race survey we asked participants to form beliefs about their chances of reaching their time goal, t^*, plus or minus N minutes, for different values of N. More precisely, participants formed beliefs over 7 possible finish time intervals: $\mathcal{I} := \{[t^-, t^*-10], (t^*-10, t^*-5], (t^*-5, t^*-1], (t^*-1, t^*+1), [t^*+1, t^*+5), [t^*+5, t^*+10), [t^*+10, t^+]\}$ where t^+ and t^- are some upper and lower bounds. The first (last) interval only specified "being at least 10:00 minutes faster (slower)" than the time goal. We use this data to compare the raw weightings on each interval, and also to construct runners' *expected finish*, *belief entropy* and a binary variable

pessimist which takes the value 1 for runners whose expected finish is slower than their goal and 0 otherwise.[20]

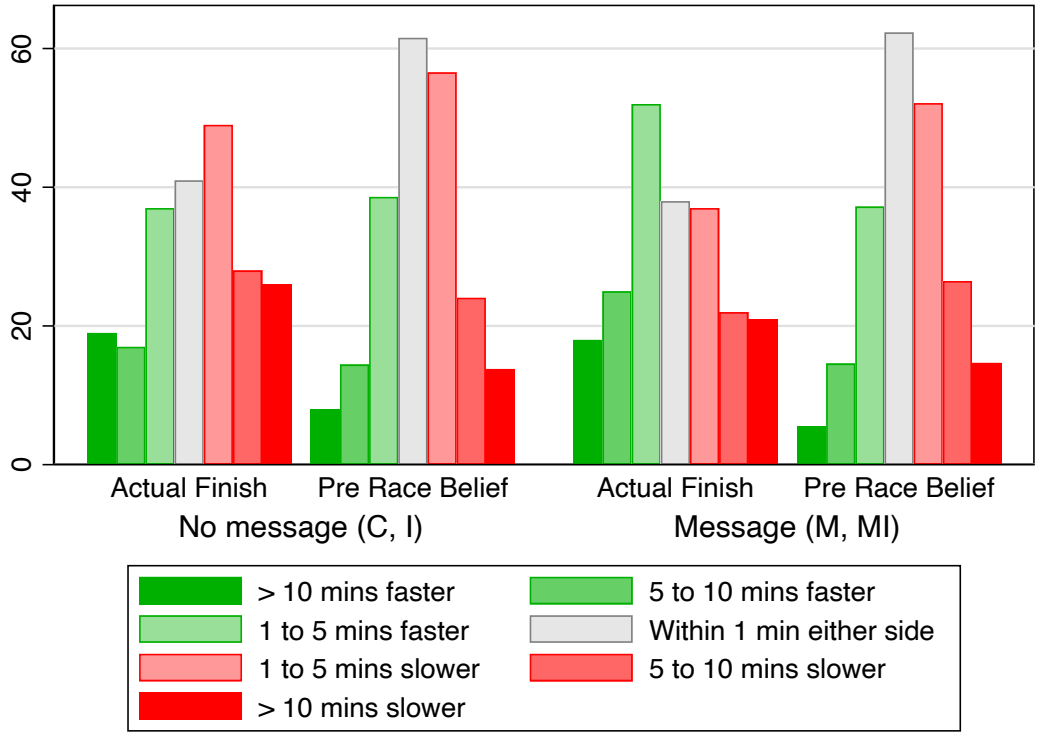


Figure 2.5: Raw beliefs vs. actual race performance, by message

Notes: In the case of Pre–Race Beliefs, each bar represents the sum of weighting attached by each runner in the full sample to that interval, relative to their goal time i.e. the aggregate weighting placed on that interval. For the Actual Finish, this is the sum of participants who finished in a time different from the goal, by that interval.

An initial observation is that runners are not overconfident: in Figure 2.5, for runners in either treatment the skew in beliefs is to the right of the goal: on average runners think their goal is an upper bound of their race performance. 59.1% of our study participants think there is more than a 50% chance of that they will finish at least 1 minute slower than their goal. This is contrary to recent findings where runners are characterised as naive about their goal and overconfident (Krawczyk and Wilamowski 2016).

Figure 2.5 shows no movements in raw beliefs for runners who had the message versus those who did not. Neither do we find a significant effect of the treatments relative to control on belief entropy,

[20]We take the midpoint of each interval, times the weighting on that interval, to compute each runner's expected time (where the top (bottom) interval uses the slowest (fastest) time as the highest (lowest) bound). This makes an assumption of uniform distributions of times within each interval, which is a strong assumption. We also compute the belief variance, but as we discussed in the Pre Analysis Plan, we focus on belief entropy because this requires fewer assumptions for its construction.

expected finish time or pessimism in the full sample.[21] Why might we have this null effect on ex ante beliefs, despite the boost in actual race performance for those with the message? One answer could lie with heterogeneity: an explanation might be that after reading the message some participants update downwards, resulting in them expecting a lower performance than the control group's beliefs, but some update upwards if they anticipate a boosting effect of the message. This would result in the net change indeed being zero for the full sample. To investigate heterogeneity, we focus our attention on belief entropy and expected finish times, both of which collapse all the belief information into a single measure.

For first–timers the belief entropy of those without the message is lower than those with the message (0.58 vs 0.66, t=-2.19, p=0.03): the message made first–timers less certain (higher entropy) about meeting their goal. A KS test shows the CDF for the first–timers with the message sits to the right of the CDF of those without the message (p=0.06, illustrated in the top left panel of Figure 2.6). However, while first–timers with the message are more uncertain about their race outcome, their expected finish does not vary significantly with those without the message: they are less certain, but do not become more pessimistic nor optimistic relative to the other runners. There are no significant differences on entropy for non first–timers who had the message or not.

We do not find significant differences between males and females (middle panels of Figure 2.6). In terms of medal orders, we find for those who ordered a medal, having the message caused more uncertainty than not having it (0.66 vs. 0.59, t=-2.06, p=0.04). A KS test, and inspection of the bottom left panel of Figure 2.6 illustrates this, with a difference of 0.178 (p=0.05). These runners also have a slower expected finish time if they got the message (104.2 minutes vs. 98.1 minutes, t=-2.18, p=0.03). This is not a particularly intuitive result, but it could be plausible that who pre–ordered medals to commemorate their participation are more vulnerable to an informational shock that tells them they are not going to achieve their goal: they are less certain about their finish time and become more pessimistic. There are no significant differences for the sub sample of runners who did not pre–order medals (bottom right panel of Figure 2.6). There are no differences by treatment on the goal attachment dimension.

Overall, although in the full sample there are no clear movements of beliefs in response to the message, the first–timer and medal order dimensions of heterogeneity offer some insight into the psychological effects of the different treatments. For first–timers, there is a small decrease in certainty about finish time for those who had the message. Those who pre ordered medals also experienced a fall in certainty when receiving the M or MI message.

[21] Note that pessimism and entropy have an insignificant correlation (-0.01, p=0.87). Being uncertain about the finish time does not correlate with expecting to fail nor expecting to succeed.

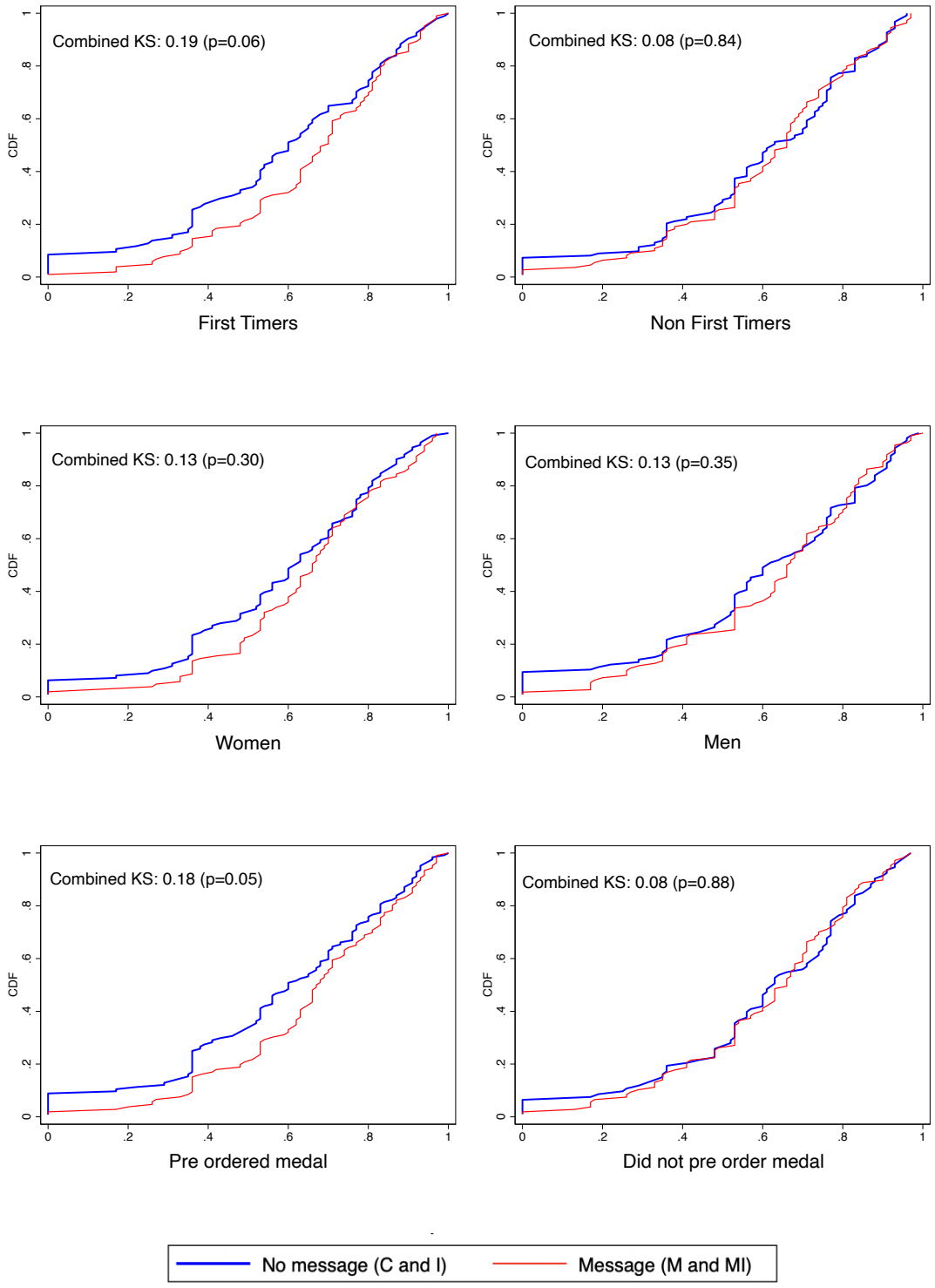


Figure 2.6: Heterogeneity in Belief Entropy, by Message

Beliefs about treatment information

A reason why the message might affect the performance of some runners but not others, and affect beliefs very little, would be the belief the participant holds about the relevance of the message to them.

We asked runners to guess what kind of information we used in the treatment messages: personal information about them, general information about runners or no information at all. We hypothesised that the treatment would have the strongest effect on those who believed that the information was personal. However, in the full sample and in the samples broken down by our dimensions of heterogeneity, we did not find significant differences in treatment effect on any of the measures of race performance between those who reported different beliefs about the treatment's informational content. In the replication, we randomise that 50% of participants are told that if they answer this question correctly they could win a £40 voucher. We do this to test whether there is any evidence of systematic misreporting, which could partly explain this null effect.

Suppression and ex post recall

Any effect of our treatment message was fairly small. There are a range of reasons why this could be the case: runners did not read it, they read it and felt it was not relevant to them and later forgot it, or they could have deliberately suppressed the message as a defensive mechanism.[22] We test for a combined effect of these mechanisms. In the Post–race survey we asked participants if they remembered seeing any information or comment regarding their time goal in the first survey and, if so, what the information was. We also asked if they had any comments for us about the study. We make use of participants' comments in both questions to construct recall–related variables as to whether the participant accurately remembered his or her treatment. Of course this is a noisy measure of recall: recalling the message is necessary to report it, but the opposite is not true. Participants might choose not to report it, despite recalling it (they found it hurtful and did not wish to give us any satisfaction through showing they recall it or they were simply in a rush to complete the survey).

In the full sample (pooling the successful and unsuccessful runners), we find that those in the M treatment are less likely to recall the ego threat than those in MI, who were also offered vouchers (26.0% vs. 48.0%, z=-3.20, p=0.00). Those not offered financial incentives may be less likely to pay attention to the message, or may be more likely to be motivated to forget it (it offers no discernible positive use for them). The proportion of runners who report remembering the message is 10% larger for those who met their goal than for those that did not, but this is not significant. These differences are significant when we examine two sub samples in particular: successful men are 17% more likely to report recalling the message than unsuccessful men (39.7% vs 22.5%, z=-1.78, p=0.08); successful non first–timers are 20% to report recalling the message than those unsuccessful (50.9% vs 31.3%, z=-2.01,

[22]This is connected to theories on information avoidance. Golman, Hagmaan and Loewenstein (2017) state two necessary criteria for information avoidance to be classed as active: (1) the individual knows the information is available, and (2) the information is freely accessible. Both of these conditions hold in the context of the treatment message. One form of information avoidance identified by the authors is *forgetting*, whereby people may "deliberately and selectively fail to rehearse negative information and therefore forget it overtime" (p.103). See also Benabou and Tirole 2002.

p=0.04). Neither the women nor the first–timers, nor either dimension of medal pre orderings or goal confidence, has a difference in reported message recall when we examine those successful or not.

Do runners ex ante suppress the message as unwelcome, and then fail to meet their goal? Or are those who do not meet their goal less likely to ex post report recalling the message (because they do not wish to record that they failed to prove us wrong)? We look patterns of recall of the message with a binary variable for whether a runner correctly or falsely recalls success or failure in meeting their goal (constructed using reports on recalled finish times and recalled goal times). We expected that participants who failed to meet their goal might self–enhance by recalling an easier goal and/or a better finish time. We suspected that selective recall by those who failed would be stronger in the message treatments. Out of the 204 unsuccessful runners, 59 (28.9%) recalled a goal time too long and / or a chip time too short, resulting in falsely recalling success. Of the 198 successful runners, only 13 (6.6%) recalled a goal time too short and / or a chip time too long, resulting in falsely recalling not meeting their goal time. However, we find no significant difference between treatments. 16.8% of those who do not recall the treatment message also falsely recall successfully meeting their goal, whilst 9.6% of those who do recall the treatment message falsely recall success. This difference is barely significant, but it does suggest the message recall variable may be a noisy measure of actual recall. This motivates an update to the survey design in the replication, in which we tell study participants that a random selection of runners will win a £40 prize if they correctly answer any of the recall or guessing questions (more on this in Section 4).

2.3.3 Welfare

We presented evidence that the message treatment may have induced more individuals to meet their goal successfully. Is this therefore a positive intervention? For those who otherwise would not have met their goal, do the psychological mechanisms inducing the behavioural change detract from the good feeling of success? For those that did not meet their goal, do they feel worse for failing given they were unable to prove us wrong? We consider two measures related to welfare — decision to donate vouchers to charity (for those who won them in I or MI), and reported pride (all groups) — to evaluate the welfare costs or gains to runners who participated in the study.

We conjectured in the Pre–analysis plan that there might be two types of individuals who feel differently about acknowledging their own pride. (1) Individuals who feel good about telling others that "I did it and you were wrong": these types might be more likely to remember the treatment manipulation, more likely to take the gift card ("I earned it"), and more likely to acknowledge their own pride. (2) Individuals who do not feel proud of their own pride and need counter-signals: they may be less

likely to remember the treatment and express their thoughts on it, more likely to donate the money they earned, and less likely to report feelings of pride. Basic correlations between the variables in the Post–race survey indicate the following patterns (illustrated as heat maps in Appendix C, Figure 18): overall, those reporting a strong sense of achievement in the race are more likely to also report pride in their race (0.689), score highly in the Tracy and Robbins positive pride score (0.182) and report recalling the pride message (0.169). Those deciding to donate any winnings to charity are less likely to score highly on positive pride (-0.178) and they are more likely to record a lower sense of achievement in the race (-0.135). Unwillingness to take a voucher is associated with less desire to report feelings of pride or achievement.

Psychological pride scores We report participants' Tracy and Robins psychological pride scores (2004, 2007).[23] For 14 different feelings and emotions, participants were asked to rate on a scale from 1 to 5 the extent to which they felt this way (1- not at all; 5- extremely). The first 7 expressions were positive (accomplished, like I am achieving, confident, fulfilled, productive, like I have self-worth, successful), while the remaining 7 terms were negative (arrogant, conceited, egotistical, pompous, smug, snobbish, stuck-up). We aggregate these ratings to construct two scores capturing the positive and negative facets of pride: (1) *positive pride score* is a respondent's mean rating on the 7 positive feelings and emotions; (2) *negative pride score* is a respondent's mean rating on the 7 negative feelings and emotions. A score close to 5 is extremely proud, and close to 0 is not at all proud.

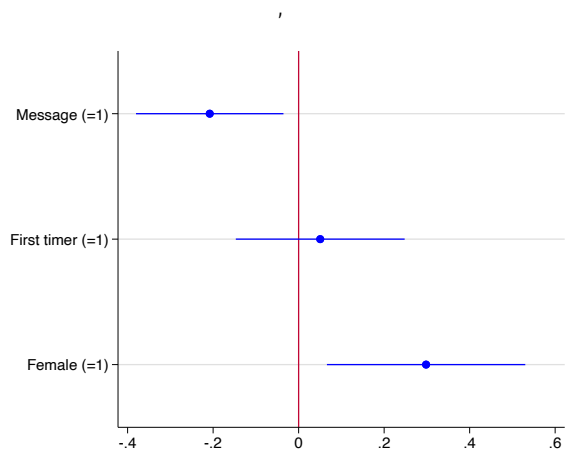


Figure 2.7: Positive pride: met goal

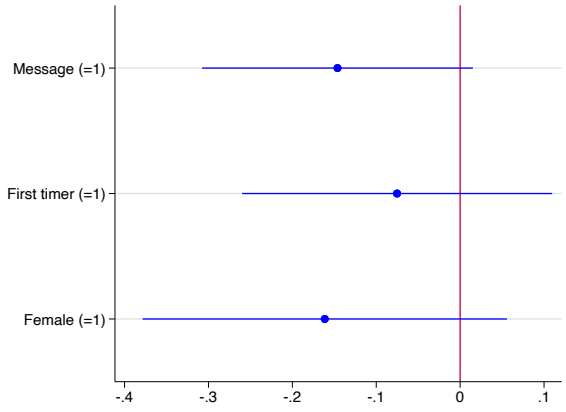


Figure 2.8: Negative pride: met goal

Notes: On the left are the coefficients from Regression column (3) in Table 20, predicting positive pride score, for the sub sample of runners who met their goal, and in which specification there are the full set of race and socio–demographic controls (coefficient estimates for these are not shown in the figure). On the right are the coefficients from Table 21. prediciting negative pride scores. The blue lines are 95% confidence bounds.

For successful runners, the treatment message is associated with a drop of approximately 0.15 positive

[23] We also measured participants' personal assessment of their race performance and self-reported feelings of pride. These variables did not have much variation in our sample.

pride points (4.20 vs. 4.04, t=1.69, p=0.09). Regressions in Table 20 in Appendix C demonstrate this effect is robust (columns (1)–(3)). Likewise, those without the message have a lower negative pride score than those without: a difference of approximately 0.16 points (1.59 vs. 1.43, t=1.86, p=0.06). Regressions in Table 21 find a similar sized difference across all 6 columns (all significant at $p<0.05$). We illustrate the coefficients from Regression (3) for those who met their goal in Figure 2.7, for positive pride score, and Figure 2.8, for negative pride score. Those who had been challenged to meet their goal (and who were successful) are less likely to express pride (positive and negative) if they were exposed to the treatment message. For the unsuccessful runners, we do not find significant differences between those who did and did not get the message in terms of positive or negative pride scores.

The unsuccessful runners overall report both lower positive pride scores than the successful runners (3.99 vs. 4.11, t=-1.89, p=0.06) and lower negative pride scores (1.41 vs. 1.51, t=-1.61, p=0.11). There is evidence that getting the message and succeeding is better than getting the message and failing. However, we cannot reject the hypothesis that all those who failed and did not get a message are as proud (positive score of 4.01) as those who met their goal but had received the message (4.04) i.e. getting the message and succeeding is not necessarily better than not getting the message and failing.

Donated to charity is an indicator for whether a successful participant in treatments I or MI chose to make a donation to charity instead of receiving the voucher payment (those who did not meet their goal would not receive the voucher, so this variable is not meaningful for these runners). Table 19 in Appendix C presents regressions of charity donation onto the treatment message and other controls. In the first specification, we find that having been in MI, as opposed to the pure I group, increases the likelihood of donating the vouchers to charity by 19 percentage points ($p<0.05$). In the second, when we control for the level of incentives offered via the interaction term, the likelihood of donating to charity increases by between 37% and 42% if a person was in the message treatment ($p<0.05$). Taken together with the fact that successful runners in the message treatment expressed less pride (positive and negative), this is suggestive that the message induces some sort of negative image utility associated with exerting effort in response to an ego threat, which runners make up for by avoiding reporting pride or accepting a prize associated with success. Individuals may have wished to signal to us that they did not meet their goal because of our study i.e. a counter signal. Note also that, unsurprisingly, the more valuable the voucher, the less likely it is to be donated.[24]

[24] Note also that in Table 19, the significance of the variable *In person recruit* to explain the charity donation, estimated to have a large and significant effect of between 32% and 34% in columns (3) and (6), could be evidence that those recruited at the Expo were more likely to feel they ought to give the money away: those who met the researcher had a stronger negative image cost associated with having responded to and wanting the winnings from our challenge. This effect is confounded by the fact that only men were recruited at the Expo, and hence we cannot disentangle a gender versus recruitment style effect. The recruitment dummy is only significant in the regressions in which we control for gender via the socio–demographic control sets.

Overall, those who met their goal have lower levels of self-reported pride if they were exposed to the message and were more likely to donate to charity. Small group comparisons are suggestive that the pride gains of achieving a goal may be cancelled out if exposed to a challenging message. The sub sample sizes are small and so effect sizes may not be robust. This is one reason for the planned replication's larger sample size. In particular, this will facilitate investigation of heterogeneity in these sub samples of success / failure and message / no message. We also include in the planned replication additional measures of welfare, to interrogate this aspect in greater depth.

2.4 | Discussion

In this section, we discuss the findings in the CUCB study. We also highlight limitations in our understanding of the psychological mechanisms underpinning the message treatment effect and gaps in our ability to evaluate precisely the welfare impact of the study. In the light of this, we decided to conduct a replication of the study. In particular we intend to (1) see if the behavioural responses replicate in other running race contexts (2) add some additional questions around beliefs and welfare measurement (3) have a larger sample size of approximately 800 runners. We give more details on the planned tools to achieve these aims and we also sketch the high level findings of the replication to date (this is on–going).[25]

2.4.1 Study findings

Behavioural change Runners in the full sample who received our ego threat message were 12 percentage points more successful in meeting their goal than those without. We also found evidence of a significant bunching of runners with the message to be just fast enough to prove us wrong, and we see this effort right from the first 5k of the race. Thus in the running race context, the ego threat appears to have had a positive effect. On the extensive margin (goal attainment), the effect was particularly strong for first–timers (13.7 percentage point increase for those with the message). This is consistent with the idea that these runners have less strong priors about their ability to meet their goal (no previous CUCB race experience) and so take more notice of the message when they read it. Indeed, we do see that first–timers also had a small but significant shift in their beliefs after the treatment message: an increase in belief entropy (i.e. they became more uncertain about meeting their goal), although they did not become more pessimistic. We also found that those with a stronger attachment to their goal saw a particularly large boost to goal attainment (16.7 percentage points). For these runners, however, we see no movement in beliefs. This might indicate an alternative psychological response for these runners: they read the message, it did not immediately change their beliefs about their capability, but they did then push harder to meet their goal to prove us wrong, because this was something they really cared about.

Appropriateness of the ego threat We presented evidence that the ego threat had an effect on runners' race performance: our choice of words induced a behavioural response. But we needed to balance the ego threat within our message with minimising harm to our participants. There is evidence that our message did not inflict grave emotional harm on our participants: (1) Only 2.5% of participants' textual feedback (10 runners) was graded as containing negative feelings[26] (2) We have a very low attrition

[25] Appendix D presents additional details on the proposed analysis and associated hypotheses for this replication.
[26] Of these runners, we reached out to apologise personally to 2 individuals who expressed particularly hurt feelings

rate, suggesting participants were engaged and motivated to complete the study, with no significant drop–out from those in the M and MI treatments, relative to C and I. Neither did we see significant drop–out for those who failed to achieve their goal relative to those who did achieve their goal. This contrasts to Markle et al., who found participants were less likely to complete a post race survey if they had not met their goal (2017). In terms of any physical harm, the evidence from the written feedback in the Post–race survey did not suggest that runners had pushed themselves more than they should to meet their goal. Some runners reported that they did not finish the race or did not meet their goal due to injury / sickness but did not report that they pushed too hard because of the study message.

Welfare effect of the ego threat In addition to avoiding immediate emotional and physical harm of our message, we wanted to understand the welfare impact on our runners from taking part in our study. We did not have a strong hypothesis about the effect of our treatment message on those who met their goal: "proving wrong" could either be satisfying or it could detract from satisfaction of reaching a goal through your own determination. In the final results section, we presented evidence that "proving us wrong" negatively impacted the reported pride boost runners would have had from achieving their goal time. We also found evidence that successful runners who had the message were more likely to give away their winnings. We cannot diagnose for certain the cause of these findings: perhaps the message induced less *actual* pride in success (proving us wrong reduced the self esteem or social recognition boost), or it induced less willingness to *report* pride (over–justification effect in the case of the vouchers i.e. "I didn't do it for the money", or reciprocating our hurtful message by refusing to let us see their pride). We found little evidence for a difference in pride for those with or without the message in the case of failing to meet one's goal. The welfare impact is an important part of a psychological manipulation of this type, and we therefore have designed additional questions in the replication Post–race survey to be able to evaluate the welfare impact, including multiple price list mechanism questions to elicit runners' willingness–to–pay for future racing or trophies and also compensation–related vignettes (see Appendix D for more details).[27] Likewise, the larger sample size in the planned replication will allow us to delve deeper into this area, by examining heterogeneity.

Baseline success rate We were surprised at the control success rate (44%), which was considerably higher than we anticipated. This could be due to a number of race–specific factors: the CUCB weather conditions were perfect for runners to achieve fast times and the distance was mis–measured and hence slightly shorter than 10 miles (9.96 miles). It is also plausible that the shorter distance (10 miles) than a marathon (26.2 miles) may have enabled our participants to perform closer to their expectations than

related to the treatment message.

[27] Note that our Post–race survey already aimed to measure feelings of pride using a set of measures: a direct self–reported pride question, the Tracy and Robbins (2017) pride scale, and self–reported race–enjoyment. We also looked at changes in choices for medal or t–shirt purchase after the race as measures of revealed pride in achievement. With the exception of the Tracy and Robbins scale, we found very little variation in these measures, and hence drop the others from the replication.

those in the marathon studies we used for our prediction: as racing time lengthens so too do possible disruptive elements such as injury / inclement weather / pacing mistakes. Markle et al. present evidence that asking experienced runners for their goal time *16 days before the race* results in faster goal times and faster finish times, relative to those not asked their goal time before the race (2017); this is consistent with a mechanism by which runners are encouraged to stick to their true goal time at this point, rather than revise it downwards in the final days / hours before the race due to nerves. As our Pre–race survey was just *3 days before the race*, we may have conducted our study in the period where runners have already revised their expectations downwards, and have hence less ambitious stated goals, which offered greater scope for all our runners to meet them once we had asked them to write it down.[28] Another consideration is whether the primary outcome measure — meeting the goal time — is the most accurate measurement tool in the running context to quantify behavioural responses to our treatment. Something more deliberate and more difficult in terms of a behavioural effort might offer a better measure of change: for instance, challenging runners to complete the race in a negative split i.e. run the second half faster than the first half, which would be difficult for most amateur runners.

Truthfulness of the signal Irrespective of the reasons for the high baseline success, the result affects our prior for success for a replication of the study. Our message stated "We would be surprised if you meet your goal" which was based on existing empirical evidence on marathon runners. For the CUCB race, to address any concern that our message might have mislead runners into thinking that we used their personal data to determine they were unlikely to meet their goal, we asked participants about their beliefs in this regard in the follow up questions: 47% of participants thought we used information about runners in general, 37% thought we used personal information and 16% thought we used no information at all. We do not therefore deduce that runners were systematically misled by the message they received. Since we found 44% of those in the CUCB control group met their goal and 55% who received any sort of pride treatment met their goal, we considered updating the phrasing of our message in the replication, to reflect the fact that we would be less surprised if runners did meet their goal. However, in the end we did not implement an changes to the message as we conjectured we may have had an unusally high success rate in the particular CUCB race context.

Small effect: recall and suppression The movements in behaviour that we find are quite small. We investigated differences in recall of the message, to try to understand whether some study participants may have deliberately ignored the message (suppressed it) and hence had no effect. However, a clear issue with our recall measures is that they were not incentivised, and it is not hard to imagine that runners may have felt inclined to misreport to save face. We did find that across all treatments, includ-

[28] In our planned replication, outlined in more detail below, we consider that further analysis on 10 mile or half marathon (13.1 mile) distances would be useful to understand this high baseline success rate further. We do not change the timing of our Pre–race survey.

ing control, individuals tended to report less ambitious goals than they had entered in the Pre–race survey. In the replication, we plan to incentivise recall–related questions for 50% of the participants (randomised), by telling them that 5 randomly selected participants will receive a £40 gift card if their answer is correct for a randomly selected question. We aim to test whether recall accuracy increases with monetary incentives and if any potential biases in the non–incentivised participants have a systematic direction.

Small effect: a lower bound We conjecture that we might have a lower bound for the effect of an ego threat: runners may not really care about proving an anonymous researcher wrong. It is likely that the person delivering this kind of message is important for the way in which it will be received: a close friend or family member expressing doubt about one's ability is likely to be more believable (they definitely have personal information about you) and possibly more motivating (it is worth it to prove them wrong).[29] A second reason why we may have a lower bound is that in our experiment there was a risk of contamination: participants may have discussed the treatments and this may have then diminished the effects. We tried to mitigate this risk in the instructions in the Pre–race survey (asking runners not to discuss the survey with others), but they may not have read this, or have been sufficiently startled by the message that they wished to discuss it with others regardless. To try to estimate the contamination, we include two extra questions in the replication Post–race survey: "Do you know others in the study?" "Did you discuss the survey with them?" Also, our participants knew that they were in a research study, which may have reduced the impact of the message, with runners realizing that they were "being played" and so possibly acting like defiers and putting in less effort. On the other hand, they may have wanted to "play the game" and we could have some experimenter demand inflating the goal attainment for the treated, relative to a natural field setting. To eliminate these confounding factors, an ideal setting for this experiment would be a natural field setting, but this is not something we have been able to achieve in the planned replication.

Null effect of incentives We did not see any behavioral changes related to the financial incentive treatments, neither pure incentives nor when mixed with the message. This finding may reflect a true invariability of performance in response to such treatment: perhaps amateur runners are simply not interested in running for prizes, and hence would not alter their behaviour in response to a

[29] We did find that recruitment in person at the Expo was associated with a higher chance of meeting the goal for those with the message, but given we exclusively recruited men at the Expo, we cannot disentangle the researcher–effect from a gender effect. We plan to investigate this in the replication Post–race survey. We ask firstly: "How would you rate the effect of our comment on your motivation to meet your goal?" And secondly: "Whether you succeeded or not, how important were each of the following factors in your efforts to meet your goal? Proving to myself that I could do it; Proving to the researchers that I could do it; The fact that the researchers would see how I did; The fact that my friends/family would see how I did; [For M and PM] The gift card I could win if I met my goal." We also recruited all runners in one of the completed replication races (Oxford) in person at the race–village, and all runners in the other races (Scotland, the South and Birmingham) online.

financial offer of any level. But it may also be that $10 and $25 gift vouchers were insufficiently large to provoke a behavioural response. The estimated mean reported household income in our sample was $136,370 and the median $125,000: the runners are much wealthier than the average US household (the median income in the US in 2019 was $68,000). $10 and $25 gift vouchers may simply be uninteresting. Secondly, our chosen financial incentives were not pure cash prizes, but instead vouchers for sportswear. A runner's willingness to pay for these vouchers, and hence motivation to win them, may be lower than the cash value of them.[30] In our replication we plan to test whether increasing the financial incentives changes this finding: we offer £10, £20 and £40 (the replication takes place in the UK).[31]

Replication sample size and progress

Our study was under-powered to detect a treatment difference of 10 percentage points at the 5% level assuming a success rate of 45% in the control group ($1 - \beta = 0.54$). In order to have 80% power to detect such an effect size, we would have needed to nearly double the sample size (N=784). Several reasons mentioned above lead us to believe that a baseline success rate of 45% in the CUCB 10 mile race is likely to be an upper bound. In addition, in the replication we recruit runners from both 10 mile distances and half marathons: we would expect the additional distance to make it slightly more difficult for runners to achieve their goal (at least after controlling for experience with the race distance). We therefore consider a 10 percentage point improvement from a 45% baseline success rate to be a conservative estimate for our replication. With a 25% (30%) baseline success rate, our calculations show that we need a sample of 658 (712) participants to have 80% power to detect a 10 percentage point difference between the message and control conditions (M and MI vs. C and I). Allowing for a 10% attrition rate between the Pre– and Post– race surveys, we therefore aim to recruit a sample of about 880 participants across all the listed races (220 per treatment).[32]

The replication data collection process started in Autumn 2019 but is still incomplete, due to the on–going Coronavirus pandemic prohibiting large running races. We started the replication of the study using four running races taking place in the UK in Autumn 2019: the Bank of Scotland Great

[30] To get an idea of the possible actual value of the gift cards, we looked at online resale sites such as cardcash.com, which buys unwanted gift cards and then resells them. For Under Armour vouchers (the sportswear vouchers we offered) the site advertises purchasing a voucher at a price of up to 85% of its value, and for very low amounts (such as $10) it may be less.

[31] We also change the voucher to be an Amazon voucher, which is the closest we could think of to cash. Almost anything can be bought on Amazon, although, as a downside, some study participants may be unwilling to use the site due to ethical objections. Also, in our updated Post–race survey we also include a question on why people run (detailed above): we specifically ask if the money offer motivated them, and we aim to compare the importance of this aspect with the importance of image–related motivations.

[32] For our continuous outcome measure *percentage deviation from goal*, a sample of 800 participants will give us 80% power to detect a treatment difference of 2 percentage points. In a subsample of 400 men, we estimate that we will have about 80% power to detect a difference of 2.8 percentage points or more, an effect size in the ballpark of our CUCB findings.

Scottish Run (September 29th), the Simplyhealth Great Birmingham Run (October 13th), the Virgin Oxford Half (October 13th) and the Simplyhealth Great South Run (October 20th). There are very few running events over the winter months, and therefore we planned to recruit runners in these Autumn 2019 races, and then conduct a second wave of data collection in the Spring 2020 to complete the dataset when the running season recommenced.

Two unforeseen factors have resulted in the data collection for the replication being incomplete:

- The Birmingham Great Run (a half marathon, so 13.1 miles) had a course alteration and shortening to just 11 miles on the morning of the race due to a potential safety threat on the planned course; this news was announced via megaphones to the runners on the start line. We asked in our Post–race survey what runners knew about the changes when they started the race: 74.9% of runners reported hearing an announcement of a course alteration, but only 8.0% of runners heard the exact distance they would be completing. As a result, the goal times submitted for the race were not relevant for the course distance and hence we cannot use the chip times to calculate the main outcome variable for our analysis. This therefore reduces our current replication sample to 439 from 622 in terms of the primary performance outcome variables.[33]
- We planned to conduct a second wave of data collection in one or several large running races in the Spring of 2020. Due to the Coronavirus pandemic there were no large events of this sort in 2020 and hence we have not had the opportunity to collect any more data.

Table 2.4 presents the numbers recruited across these 4 races, reporting the number of runners who completed all three surveys (Registration, Pre– and Post– race surveys) and the number of runners with usable chip times. 439 survey participants with usable chip times is just over half the required sample of 800 needed for the replication. We therefore will resume data collection once larger races are happening, or consider an alternative online survey of runners to focus on just some aspects of the study.

In the primary outcome measure — goal attainment — the initial findings of the partial data collection do not look as though they will straightforwardly replicate the findings from the CUCB race. Figure 2.9 shows the percentage who met their goal per race per treatment. The Oxford race appears to have a positive effect of each of M, MI and I, and in particular M, but these differences in raw means are not significant at any conventional level (the sample size is only 100, hence c. 25 in each treatment). For the Scottish runners, the I group do have a significantly positive response to the financial incentives, increasing by 22 percentage points the success rate relative to the C (p=0.01). In the South,

[33]We nevertheless honoured our financial bets to runners in the I and MI treatments who completed the shorter course at or faster than their goal pace.

	Screened In	All surveys completed	Chip time usable*
Great Scottish Run	135	131	✓
Great Birmingham Run	188	183	x
Virgin Oxford Half	100	97	✓
Great South Run	227	211	✓
Total	**650**	**622**	**439**

Table 2.4: Replication participants to date

Notes: * The Great Birmingham Run course was altered and the race distance altered significantly on the morning of the race. The final distance — which many runners were not aware of — was just over 11 miles, rather than the full 13.1 mile half marathon distance.

the pure financial incentive also appears to have had a positive effect on goal attainment, but this is not a significant difference with C. In both the South and Scotland, the M group have near zero difference with the C. These sketch results are a first look at an incomplete data collection: the full replication will involve pooling results across the set of races in the second wave of data collection, as well as pooling results with the CUCB data. This will enable investigation of treatment effects over a full pooled sample of 1200 runners and on subsamples according to the pre–registered dimensions of heterogeneity.

It is, however, already clear at this point of the data collection that we will not have simple replication of the CUCB result of improved performance upon receipt of either just the ego threat, or the ego threat with added financial incentives. Neither does it seem that we will replicate the result that financial incentives alone do not have an effect on performance. Once the data collection is complete, looking at pre–race beliefs and the post–race welfare measures for the different groups of runners may offer some insight into the channels through which the treatments do and do not work. The replication data collection will be important to derive any concrete messages regarding the effects of ego threat vs. financial incentives in our context of road running races. Additionally, careful study design and data investigation is needed to understand the complexities of the impacts and channels of such a psychological treatment, in particular in terms of heterogeneity of subject (the runners) and context (the race).

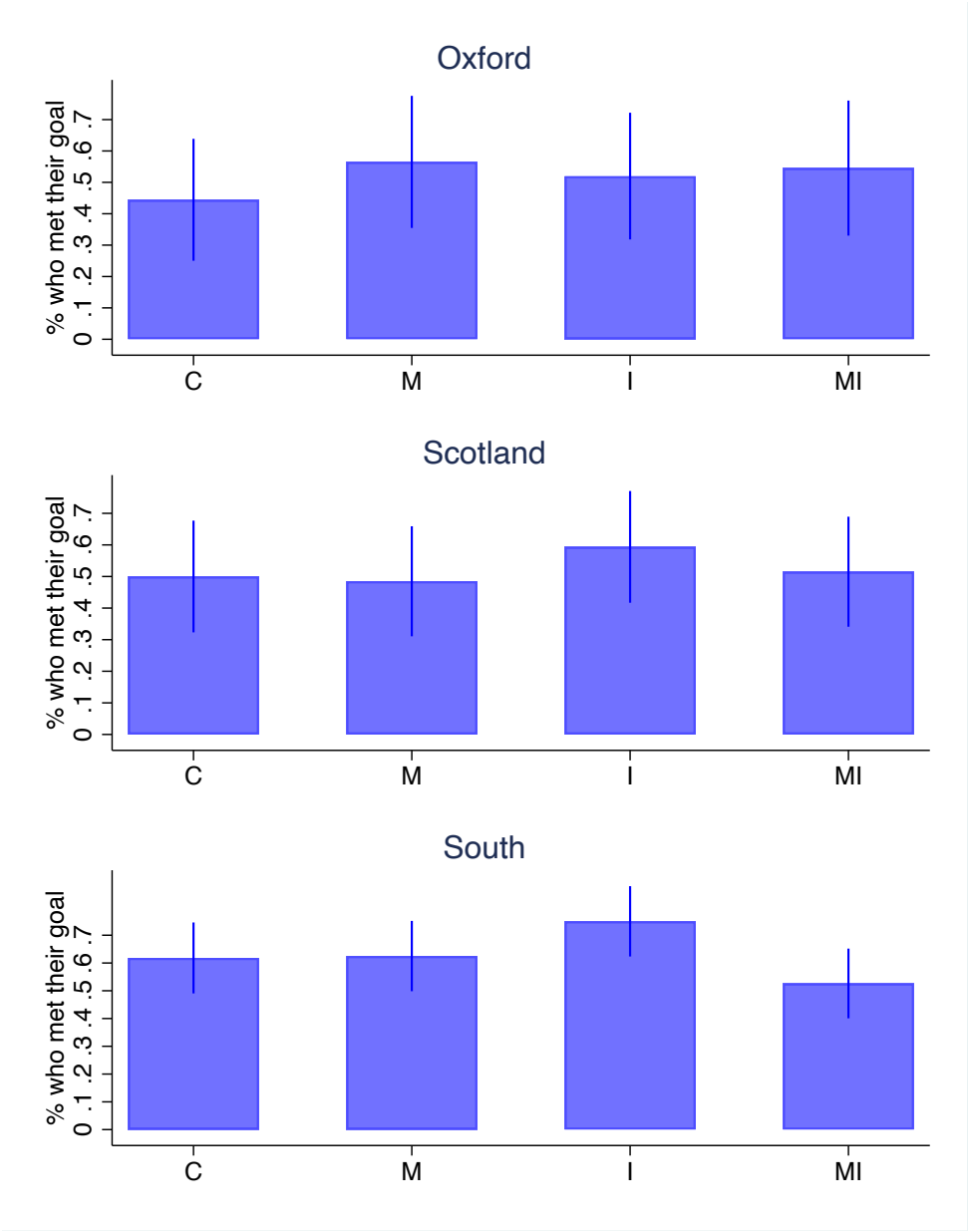


Figure 2.9: Replication Goal Attainment

Notes: The raw means of goal attainment per treatment and their 95% confidence intervals are as follows for the replication races. Oxford (n=100): C 0.44 [0.25, 0.64], M 0.57 [0.35, 0.78], I 0.52 [0.32, 0.72], MI 0.55 [0.33, 0.76]. There is no significant effect of the treatments on goal attainment for the Oxford runners, which is a very small sample, but all three treatments appear to push goal attainment higher in M, I and MI than in the C group. Scotland (n=135): C 0.50 [0.32, 0.68], M 0.48 [0.31, 0.66], I 0.59 [0.42, 0.77], MI 0.50 [0.32, 0.68]. In this case, the effect of I is significant at p=0.01, increasing goal attainment by 22% above the C group. South (n=227): C 0.62 [0.49, 0.75], M 0.62 [0.49, 0.75], I 0.75 [0.62, 0.88], MI 0.53 [0.40, 0.65]. There is no significant effect of the treatments on goal attainment for the South runners, but those who were in group I show a similar pattern as the Scottish runners of a substantial increase in % who attained their goal.

Chapter 3 | If you're happy, can you show it? Non-parametric estimation of the consistency between revealed preference rankings and a well-being measure

Chapter Abstract

It has been argued, notably by Sen (1979), that traditional welfare economics would be greatly enriched if we had access to more comprehensive and cardinal measures of well–being. Happiness, or specifically Subjective Well–Being (SWB) data have been proposed as an alternative. An important question arises: how different is our understanding of well–being if SWB data rather than traditional utility–theory based measures are used? I use consumption data across years and countries to rank years (within–country) and countries (within–year) by their Revealed Preference (RP) relations. I use SWB data (from non–incentivized self–reports of well–being) to build equivalent rankings for the same years and countries. I then conduct tests for the similarity of the RP and SWB orderings. I find evidence that within–country RP and SWB orderings are consistent over time, whilst within–year orderings are less so. I make suggestions as to why the two measures may diverge in ordinal ranking when cross–country welfare is examined. I also outline ideal datasets to which this ordinal comparison method could be applied for further research on well–being measures.

Chapter Notation

- u = utility level
- h = SWB level
- U = RP adjacency matrix
- H = SWB adjacency matrix
- $\tau \in [-1, +1]$ = Kendall's Tau rank correlation coefficient

3.1 | Introduction

—Dumbledore: "It is our choices, Harry, that show what we truly are."

J K Rowling, The Chamber of Secrets

It has been argued, notably by Sen (1979), that traditional welfare economics would be greatly enriched if we had access to comprehensive and cardinal measures of well–being. Subjective well–being (SWB) data are growing in use. A question arises: how different is our understanding of well–being if SWB data rather than traditional utility–theory based measures are used?

This chapter presents a simple method to compare and quantify the difference between ordinal rankings derived from price and quantity data — Revealed Preference (RP) orderings — and ordinal rankings derived from SWB data. To illustrate the method, I apply it to a small international timeseries dataset (comparing ordinal rankings of representative consumers)[1]: for example, the method measures how similar the following preference orderings are, where the first is the UK's derived RP ordering and the second is the UK's SWB ordering: (1) $u_{1975} \prec u_{1980} \prec u_{1985} \prec u_{1996} \prec u_{2011} \prec u_{2005}$; (2) $h_{1975} \prec h_{1985} \prec h_{1996} \prec h_{1980} \prec h_{2005} \prec h_{2011}$.[2] These rankings look reasonably similar (1975 is worst in both, and 2011 and 2015 are in the top two positions in both), but 1980 for instance is second worst in the RP ordering, but third best in the SWB ordering. In the dataset, the method finds that within–country RP and SWB orderings across different years (such as the UK's above) are similar over time, whilst within–year orderings of countries are less so. I discuss possible interpretations of these results and also make suggestions as to more suitable datasets to which the technique could be applied in future research.

This method, together with the example results from applying it to the international panel dataset, is useful because national governments are becoming increasingly interested in alternatives or complements to traditional measures of well–being like Gross Domestic Product (GDP), derived from price and quantity data. Bhutan, for instance, measures Gross National Happiness (GNP), using nine domains of life which are summarised in 4 pillars (Ura 2008). Each domain is measured using both subjective and objective indicators. The UK's Office of National Statistics also collects individual and societal well–being indicators.[3]. Understanding the differences in the information within well–being

[1] I use data in the International Comparison Program, which are the same price and consumption data underlying the Penn World Tables. These are routinely used to derive international comparisons of living standards (Feenstra, Inklaar and Timmer 2015). I merge these data with SWB data from the World Database of Happiness.

[2] For the SWB ordering, I directly interpret the SWB levels as cardinal, such that if a year has a higher SWB level, it is taken that the representative agent is strictly better off in that year.

[3] ONS source: consulted in August 2021.

measures is an important step before choosing a measure and making policy decisions on the basis of this. For example, the high–level result mentioned above points to a possible risk to comparing subjective measures across countries, which would be important to know before making policy decisions in response to differences in such measures.

Background and Motivation

The aim of this chapter is to compare ordinal rankings based on different measures of well–being, the first of which is grounded in the economic theory of *utility*. Originally used to describe the usefulness of an object, its meaning has developed in modern economics. Bentham (1843) describes utility as a usefulness property: creating some benefit, advantage, pleasure, good or happiness; Broome (1991) states that utility is an action's ability to create Bentham's usefulness properties. Varian describes the Victorian principle of utility providing a measure of a person's overall well-being (2010). Utility defined as such could be affected by a plethora of factors but it is most often understood as the relative value accorded to a bundle of quantifiable items such as goods, services or time spent doing activities (Snyder and Nicholson 2012). Understood in this way, it is determined by a person's tastes and what they consume. A utility function then assigns a real number to consumption bundles available to an agent such that a bundle giving more utility is assigned a larger number. The function thus provides a ranking across available options. The agent faces a set of prices and has an income level that determines the various bundles of goods that they can afford; given this budget constraint, their tastes and fixed prices, the agent chooses their favourite — utility maximising — bundle.

Researchers cannot directly observe utility, but can use revealed preference orderings derived from consumption observations to infer ordinal utility rankings. Such an approach does not capture all aspects of well–being: there clearly is an informational gap between well–being rankings as derived from price and quantity data and well–being as a whole. For instance, economists are well versed in the fact that resources may be priced differently by people and some may be un–priceable; a fully rounded measure therefore needs to have more inputs than just prices and quantities.[4] Indeed, the findings in Chapter 1 of this thesis suggest that on average consumption–related domains of life (household income and / or Job) are not highly weighted in overall life satisfaction. This invites the consideration of additional, or indeed, alternative measures; one suggestion is to use SWB.

Modern-day happiness philosopher Haybron (2013) defines three theoretical strands of *happiness*: (1)

[4]Consumption or expenditure data are also used to construct cardinal measures of welfare (money–metrics). These approaches are subject to similar critiques. Stiglitz, Sen and Fidoussi (2009) were commissioned to write a report on Gross Domestic Product (GDP)'s limitations for measuring economic and social well-being. The authors suggest the financial crisis surprised us either because the measurement for progress failed or nations were not measuring the correct variables. The report's outcome is that due to the informational gap between GDP and other factors affecting well–being, economies need to measure well–being directly.

Emotional state (a positive emotional condition); (2) Hedonism (pleasure); (3) Life Satisfaction (being satisfied with your life). From this triad of happiness components emerges the notion of SWB: Stiglitz, Sen and Fidoussi (2009) suggest SWB comprises both taking stock of life and considering prevalent feelings (positive and negative effects). This combines two of Haybron's strands: emotional state theory and life satisfaction theory.[5] In this chapter, I define SWB as per Stiglitz et al. and note in particular that SWB measures what is good for an agent in the context of that agent's own standards.[6] For instance, the precise question used to collect the SWB data used in this chapter is as follows: *Taken all together, how satisfied are you with your life as a whole these days? (1) Completely dissatisfied (2) Very unsatisfied (3) Unsatisfied (4) Neither satisfied nor unsatisfied (5) Satisfied (6) Very satisfied (7) Completely satisfied.* The individual then interprets the response scale in their own fashion.

Many welfare economists care about measuring and understanding SWB variables: "the question is not whether well-being is an intrinsically important variable for moral analysis, but whether it is uniquely so" (Sen 1985, 186). SWB could be used to evaluate the relative merits of equilibria obtained by optimising an income–based utility measure, but measures of SWB have also been proposed as substitutes for such classical measures of economic welfare or measures of capabilities: SWB is directly optimised itself, rather than being used to evaluate the outputs obtained by optimising classical inputs.

Critics suggest that utility theory may be insufficient in three regards. Firstly, utility cannot be used to compare individuals to each other (interpersonal non–comparability). Secondly, researchers cannot directly observe utility, and therefore use revealed preference orderings derived from consumption data, or monetary equivalents (money metrics). Finally, measures derived in these ways do not capture all aspects of wellbeing: there is an informational gap between an individual's well–being (as derived from price and quantity data) and true well–being. Which elements are not captured, and how to incorporate these into a well–being measure, are matters for debate.[7]

[5] See a comprehensive account of social scientific use of the term SWB in Kahneman, Diener and Schwarz (1999).

[6] A comment on subjectivity and well–being: Philosophies of *happiness* date back to the ancient world, as do conflicting definitions of what happiness actually is. Aristotle's *eudomonia* describes how a person's way of life benefits them and how filled with virtuous activity it is: this is an external consideration within a value framework (Diener 1984). In Plato's *Gorgias*, Socrates takes an Aristotelian view: a man's happiness can be judged by his deeds (see the translation of the collected dialogues of Plato, by Benjamin Jowett (1953)). He elaborates that evil men must be unhappy until they are punished. In response, his colleague Polus argues that an evil man may be happy —certainly he will likely be rendered unhappy if punished!— but that we can only know if we *ask him to tell us*: the seed of subjectivity's relation to happiness is planted. 2000 years later, a theoretical consensus on happiness has not yet been reached. Wilson (1967, 294) concludes that a happy person is a "young, healthy, well–educated, well–paid, extroverted, optimistic, worry–free, religious, married person with high self–esteem, high job morale, modest aspirations, of either sex and of a wide range of intelligence": an objective recipe! The current debate centres on a Polus–type interpretation of happiness having a subjective component.

[7] Two further details: (1) The money–metric approach is cardinal to an extent: if two households face the same price increases, their compensating or equivalent variations can be calculated. Following a price change, there is an indication of who loses more relative to another. However the actual change in levels of utility is unknown. See Blundell, Preston and Walker (2009) for more on money–metrics for welfare economics. (2) Abramovitz (1959) discusses to what extent output trends are sufficient to understand trends in welfare, in particular in the cases when social conditions change alongside output. Rothschild (1961) generalises the problem: "anybody who is aware of the many practical and conceptual difficulties in constructing historical series of national product and of the rather shaky foundations of welfare economics even in rather

The question of whether SWB addresses the critiques of a utility–based approach to welfare, and hence provides a viable and useful alternative, has been investigated by researchers, with differing outcomes. For example, Fleurbaey, Schokkaert and Decancq (2009) argue that SWB data ought to be interpreted as ordinal indications of well–being rather than levels and that large individual effects on SWB levels may mar our ability to conduct inter–personal comparisons. Krueger and Schkade (2007) find higher test–retest reliability for person–specific components of affect than situation–specific components of affect, which again suggests problematic inter–personal comparability. However, Kahneman et al. (2004) propose that SWB measures could indeed be inserted into social welfare functions and then — having captured the aspects of well–being relevant to the individuals, including any personal or social judgements not included in utility measures — use welfarism techniques as before.

Setting aside the debate of whether SWB or utility–based measures are complementary to one another, or substitutes, the two approaches to measuring well–being used in this chapter — inferring preferences from observed consumption patterns and collecting self reports of satisfaction — are clearly different in their style and conceptual understanding of well–being. The World Database of Happiness (WDH) states that a key advantage of using measures of happiness is the fact that it is the broadest output measure available: unlike health indicators it is not limited to bio–physiological functioning and unlike suicide it is not limited to extreme despair (Veenhoven 2016). This suggests why there might be a divergence of SWB and RP orderings. The RP approach is based on the domain of consumption. SWB measures are designed to be broader and include many aspects of life. This potentially huge domain difference is illustrated in Chapter 1 of this thesis.

Having acknowledged the differences in the two approaches, the question is how different is our understanding of well–being when using one approach versus the other. This paper offers a way to quantify the differences in the basic ordinal information conveyed by the two types of data. The simple question I ask is: how can we measure the similarity between ordinal rankings derived from each approach? And when applied to the dataset: how consistently does a representative agent who *reveals by their choices* a preference ordering also *report via their SWB levels* a similar ordering?

Literature and Contribution

There is a close connection of this research with the microeconomic literature that seeks to determine whether individuals' hypothetical or actual choices coincide with what they think would maximise their SWB. This work uses survey data to investigate the correlation between estimated marginal rates of

simple contexts will not be surprised to learn that the difficulties of deriving a long term index of social welfare from national income data are almost insurmountable" (p.491).

substitution derived from choice and SWB related data. Key papers in this space are Benjamin et al. (2012, 2014). Benjamin et al. (2012) find a reasonably high prediction rate between binary hypothetical choices and anticipated SWB, but that controlling for other factors such as family happiness, social status, or sense of control improves predictions of stated choices. These findings are complemented by later work in which hypotheticals become real choices: in the 2014 paper, the researchers elicit choice rankings over residency programs, as well as anticipated SWB rankings and expected features of the residencies; there are differences between the anticipated SWB rankings and choices when implied trade–offs between residency attributes are examined. However, the life satisfaction and Cantril's ladder style SWB questions are more closely correlated with the choices than affect measures. The findings are shown to be robust to measurement error and survey biases. In a similar vein, Goldsmith (2009) uses household data in parametric and semi–parametric random coefficient models to investigate SWB's relation with households' reports on what matters to them: UK household SWB does not correlate with individuals' perceptions about what is important.

Benjamin et al. (2020) review the evidence that survey respondents do not always choose options (with implied real or hypothetical trade–offs) that maximise their well–being, as measured by a single question, such as the one used to collect the SWB data used in this chapter. This connects to Chapter 1 of this thesis, in which I investigate the relative importance of the multiple dimensions of SWB: individuals are motivated by more than what is captured in their response to a single well–being question. Hence it is not surprising that in this chapter I find weaker evidence for consistency between RP orderings and SWB orderings for cross country comparisons, although in the case of within–country comparisons, these two do — perhaps surprisingly — appear to have stronger positive correlations with each other.

A difference in my own approach is firstly that I use a distance measure to compute the difference between RP and SWB ordinal information. Another difference is that the data are not hypothetical: the RP rankings are derived from observed prices and quantities, and the SWB data are reported "true" levels rather than anticipated levels. I also use entirely non–parametric methods to derive the utility rankings and to test for the existence of a rational utility ranking. When I do this country–by–country, in particular I allow for unrestricted preference heterogeneity for countries. A further difference in my approach is that in my application of the method, I use aggregated data, rather than individual data; however, the method would be well–suited to microeconomic cross–sectional or panel data.

Outline

I first derive a set of RP orderings and test these for cyclical consistency. I use the Generalised Axiom of Revealed Preferenced (GARP). In the dataset, the RP orderings are in two forms: (1) per country, a ranking for chosen bundles across different years and (2) per time period, a ranking for chosen bundles across different countries.[8]

Secondly, I compare these RP orderings to the equivalent SWB orderings. This involves summarising the full set of pair–wise preference relations between items. I then compute the Kendall τ rank correlation coefficient: this measures the discrepancy between the two orderings. I compare the outcome to a benchmark outcome to determine whether the orderings can be deemed similar or dissimilar overall. I use randomness as a benchmark, following Becker (1962). Note that the method does not require the use of econometric techniques and therefore does not require the associated functional form and statistical assumptions needed for regression analyses. I also make no money–metric calculations of utility as the RP approach provides a direct mapping to the ordinal rankings required.

To illustrate the method, I use data from the International Comparison Program (ICP): average national prices and volumes, classified across four categories, across c.200 countries over time. I add to this with data from the World Database of Happiness (WDH): life–satisfaction surveys across countries and across time. I select a country–year observation provided that there is both ICP data and WDH available. The resulting subsample is a panel made up of 27 countries over up to 6 time periods; the panel is unbalanced because not all countries have both ICP and WDH data in every year. The RP orderings are constructed from the ICP data and the SWB orderings from the WDH data. For the within–country analysis, each of the 27 countries has an RP and SWB ordering (these orderings have up to 6 years in them). For the within–year analysis, each of the 6 years has an RP and SWB ordering (these orderings have up to 27 years in them).

My use of the ICP and WDH data is subject to some caveats: (1) the data are aggregated to country–level, demanding assumptions on both RP and SWB data; (2) the SWB data in particular may be subject to a series of measurement errors; (3) the panel is small and unbalanced. I discuss these issues in Section 3. In Section 5 I discuss data that might be better suited for the method: my use of ICP and WDH datasets in this chapter is driven by the fact that these data are freely available online, they contain both consumption and SWB data for the same units of analysis and hence provide a small,

[8] A natural final step to this work would be to apply these tests of cyclical consistency across the entire pooled dataset, derive a RP ordering across chosen bundles by all countries in all time periods, and then compare this to the SWB ordering. Unfortunately, the R program used to run this analysis cannot solve this problem. It is therefore left as a proposed avenue for further research to find a computational solution.

proof–of–concept testing ground for the method.

Structure

Section 2 outlines a test of cyclical consistency in choice data — the Generalized Axiom of Revealed Preference (GARP); this test is used to establish whether the ICP dataset provides evidence for countries having stable preferences over time, and whether preferences may be homogeneous across countries. I then set out the method used to generate the ordinal comparison results: I discuss potential measures of distance between the two preference orderings, and choose the Kendall τ statistic. Section 3 describes the dataset. Section 4 reports the results. The GARP tests find evidence that each country has stable preferences over time, and likewise evidence that countries have homogenous preferences in each time period. The comparison of the RP orderings with SWB orderings per country find that these are similar; when compared to a benchmark comparison of RP orderings with randomly generated data (i.e. uncorrelated), the true comparison with SWB data gives a stronger correlation result. In contrast, in the within–year analysis I find weaker evidence against the null that the true SWB orderings are no more correlated with the RP orderings than random orderings are. Section 5 discusses potential interpretations of these results and suggests that one explanation could be that there are different interpretations of the SWB scale by different nations. I outline potential extensions for this research.

3.2 | Methodology

In this section, I briefly outline the theory that underpins the GARP tests for the RP orderings, and then discuss the range of methods that could be used to calculate the differences between the RP and SWB orderings, and choose the Kendall τ.

3.2.1 Tastes, behaviours and revealed preferences

There are three key issues regarding the relationship between a consumer's tastes and their demand behaviour: (1) when is observed choice data consistent with utility–maximising behaviour? (2) given observed choices, what are the consumer's tastes? and (3) how can the observed behaviour predict responses to prices not observed? (Varian 1978). This chapter uses a test of issue (1): are the RP orderings derived from observed price and quantity data consistent with utility–maximising behaviour? This informs the discussion of the similarity and differences with the SWB orderings.

There are two standard approaches to the three issues listed above. The first approach requires that preference relations (tastes) are taken as the primitive characteristics of the individual. The individual's choice behaviour can be derived from such preferences and a demand function is determined for goods and services at different prices (Mas–Colell, Whinston and Green 1995). This requires a hypothesis about the functional form of the individual's demand which is then fitted to the data observed. Provided the functional form was a reasonable fit, the parameters estimated from the data can be used to compute predicted demand responses to changes in prices.[9]

A second approach instead imposes consistency requirements directly on observed choices (these are now the primitives); the choices are used to back–out individuals' preferences. No hypotheses about the forms of the consumer's demand functions are required, and individuals do not need to know their own preferences. This second approach can be traced back to Samuelson, who sets out key postulations regarding consumer choice that resulted in the idea later termed "revealed preference" (1938).[10] An advantage of this approach is that it can be taken directly to economic science data (Samuelson 1938).

[9] Note that, because of the background assumptions, testing the fit is subject to the Quine–Duhem problem: we cannot use empirical evidence to accept or reject the theoretical hypothesis alone, aside from the auxiliary assumptions (we have under–determination). Key extracts for the Quine–Duhem Thesis are in Curd and Cover (1998).

[10] Firstly, facing a fixed budget and a set of prices, a consumer will always choose the same set of goods. Secondly, the consumer behaves in this way irrespective of the units of the prices. Thirdly, the consumer cannot simultaneously choose batch one over batch two, and batch two over batch one (now known as the Weak Axiom of Revealed Preference). Samuelson's shows the three postulations are deducible from conventional microeconomic analysis; the patterns in consumer choices are logically equivalent to the classic demand theory defined by Hicks and Allen (1934).

The first step for this chapter's analysis of the RP orderings is to test whether the data can be rationalized by a well–behaved utility function in two respects: (1) Are the choices that the representative consumer of each country makes consistent with them having stable preferences over time (*consistency*)?[11] (2) Are the choices the representative consumer of each country makes consistent with countries having common preferences (*homogeneity*)?[12] Building on Samuelson's RP theory, Afriat's theorem (1967) and Varian's extension (1982) provide practical tests for patterns in observed data that are consistent with utility–maximising behaviour. Afriat's theorem provides the following set of four equivalent conditions:

AFRIAT'S THEOREM *The following conditions are equivalent:*

(1) There exists a locally non–satiated utility function that rationalizes the data.

(2) The data satisfies "cyclical consistency"; that is:

$$\mathbf{p}_1\mathbf{x}_1 \geq \mathbf{p}_1\mathbf{x}_2, \quad \mathbf{p}_2\mathbf{x}_2 \geq \mathbf{p}_2\mathbf{x}_3, \quad \ldots, \quad \mathbf{p}_n\mathbf{x}_n \geq \mathbf{p}_n\mathbf{x}_1$$

implies

$$\mathbf{p}_1\mathbf{x}_1 = \mathbf{p}_1\mathbf{x}_2, \quad \mathbf{p}_2\mathbf{x}_2 = \mathbf{p}_2\mathbf{x}_3, \quad \ldots, \quad \mathbf{p}_n\mathbf{x}_n = \mathbf{p}_n\mathbf{x}_1.$$

(3) There exist numbers $(U_a, \lambda_a > 0)$ for $a = 1, \ldots, n$ such that

$$U_a \leq U_b + \lambda_b \mathbf{p}_b(\mathbf{x}_a - \mathbf{x}_b) \text{ for all } a, b = 1, \ldots, n.$$

(4) There exists a nonsatiated, continuous, concave, monotonic utility function that rationalizes the data.

The condition of "cyclical consistency" in (2) states that if the first bundle is at least as good as the second, and the final bundle is at least as good as the first, it must be that the bundles are all the same. This is the key result for this chapter: choice data can be tested for cyclical consistency. Then by (2)'s equivalence to (1) and (4) the choices must have been made by a consumer who was maximising their utility, and whose preferences can be represented by a single utility function. Testing for cyclical consistency is equivalent to testing whether the data satisfy the Generalized Axiom of Revealed Preference. Varian (1982) establishes the following key definitions:

DEFINITION Given an observation $\mathbf{x}_a$ and a bundle $\mathbf{x}$:

(1) $\mathbf{x}_a$ is *directly* revealed preferred to $\mathbf{x}$, written $\mathbf{x}_a R^0 \mathbf{x}$, if $\mathbf{p}_a\mathbf{x}_a \geq \mathbf{p}_a\mathbf{x}$.

(2) $\mathbf{x}_a$ is *strictly directly* revealed preferred to $\mathbf{x}$, written $\mathbf{x}_a P^0 \mathbf{x}$, if $\mathbf{p}_a\mathbf{x}_a > \mathbf{p}_a\mathbf{x}$.

[11] The fact that I am working with a representative agent implies strong conditions on preferences, and these are discussed in Section 3.3, on the data and treatment.

[12] For full details on the definitions included in this subsection, and further information on Afriat's Theorum, see Varian (1982). Specifically, I look for a utility function that is *non–degenerate*, nonsatiated, continuous, concave and monotonic. I also thus do not allow the simple solution of $u(x) = 1$ for all x.

(3) $\mathbf{x}_a$ is (indirectly) revealed preferred to $\mathbf{x}$, written $\mathbf{x}_a R \mathbf{x}$, if $\mathbf{p}_a\mathbf{x}_a \geq \mathbf{p}_a\mathbf{x}_b$, $\mathbf{p}_b\mathbf{x}_b \geq \mathbf{p}_b\mathbf{x}_c$, ..., $\mathbf{p}_g\mathbf{x}_g \geq \mathbf{p}_g\mathbf{x}$ for some sequence of observations $(\mathbf{x}_a, \mathbf{x}_b, \ldots, \mathbf{x}_g)$. In this case, the relation R is the *transitive closure* of the relation R^0.

(4) $\mathbf{x}_a$ is *strictly* (indirectly) revealed preferred to $\mathbf{x}$, written $\mathbf{x}_a P \mathbf{x}$, if there exist observations $\mathbf{x}_b$ and $\mathbf{x}_c$ such that $\mathbf{x}_a R \mathbf{x}_b$, $\mathbf{x}_b P^0 \mathbf{x}_c$, $\mathbf{x}_c R \mathbf{x}$.

This highlights the difference between a direct and indirect preference relation. The latter is important for establishing the preference rankings of national consumption and price data: not all years and countries have directly comparable pairs because there are changes in income levels which may render previous bundles unaffordable. (3) allows the transitive closure of revealed preference relations, from R^0 to R. Using the above, GARP is defined as follows:

DEFINITION A set of data satisfies the *Generalised Axiom of Revealed Preference* (GARP) if $\mathbf{x}_a R \mathbf{x}_b$ implies not $\mathbf{x}_b P^0 \mathbf{x}_a$.

If bundle $\mathbf{x}_a$ is indirectly revealed preferred to $\mathbf{x}_b$ then it cannot also be that $\mathbf{x}_b$ is strictly directly revealed preferred to $\mathbf{x}_a$. If data do not satisfy cyclical consistency, there is a violation of GARP. $\mathbf{x}_a R \mathbf{x}_b$ circles around all the data and returns to its starting–point: preferences will contradict themselves because of the strict inequality in $\mathbf{x}_b P^0 \mathbf{x}_a$.

R^0 (directly revealed preference relations) can be represented in matrix form: If bundle a is directly revealed preferred to b, a 1 is put into element m_{ab} of matrix M, which is known as an adjacency matrix:

$$m_{ab} = \begin{cases} 1 & if \mathbf{x}_a R^0 \mathbf{x}_b \\ 0 & \text{otherwise.} \end{cases}$$

Varian (1982) outlines a simple algorithm, Warshall's Algorithm, that computes the transitive closure of the revealed preference relations. N is the matrix which is transitively closed, where a 1 in element n_{ab} indicates an indirectly revealed preference:

$$n_{ab} = \begin{cases} 1 & if \mathbf{x}^a R \mathbf{x}_b, \\ 0 & \text{otherwise.} \end{cases}$$

In the transitively closed matrix, if $n_{ab} = 1$ and $\mathbf{p}^b\mathbf{x}^b > \mathbf{p}^b\mathbf{x}^a$ the choices violate GARP.

A related question is, if a set of choices do not satisfy GARP, how large is the violation of the inequalities? There are different ways to quantify the gap between the actual choices and the optimal

ones. Afriat (1967) proposed an index specifically applicable to the consumer choice setting: Afriat's Efficiency Index. The idea is calculate how much a consumer wastes on a bundle chosen, when there was one cheaper that was also revealed preferred. The index measures the smallest amount of this wasted budget, as a fraction of the full budget. It can be thought of as how much less the consumer would have had to have spent on $\mathbf{x}_b$ before it would have been thought of as a worse choice than $\mathbf{x}_a$, which is the one that is observed:

$$\mathbf{x}_a R_e^0 \mathbf{x}_b \text{ if and only if } e_b \mathbf{p}_b \mathbf{x}_b \geq \mathbf{p}_b \mathbf{x}_a.$$

e_b is the Afriat Efficiency of observation b. When $e_b = 1$ for all b, the data perfectly satisfies the GARP conditions. When $e_b = 0$ for all b, the data have no revealed preferences. The Afriat Efficiency Index is a single e for all observations. This is the index I will report to determine how close each RP ranking is to fulfilling the GARP consistency requirements.[13]

3.2.2 Measurement of ordinal differences: UK example

Once I have the RP orderings and have tested these for cyclical consistency, I turn to the core question: how similar are these orderings to SWB orderings? For example, in the WDH data, the UK has the following SWB ordering (where h denotes the average SWB level in the UK): $h_{1975} \prec h_{1985} \prec h_{1996} \prec h_{1980} \prec h_{2005} \prec h_{2011}$. Derived from the ICP data, the UK has the following RP ordering (where u denotes the representative agent's utility level): $u_{1975} \prec u_{1980} \prec u_{1985} \prec u_{1996} \prec u_{2011} \prec u_{2005}$. In some instances the RP ordering may be only partial, but the SWB ordering will be full. These ordered lists could be compared directly.

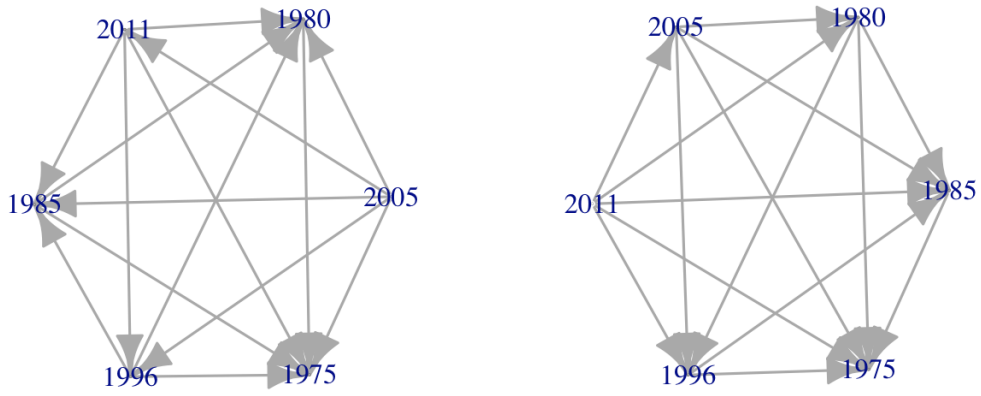
Alternatively, the orderings can be represented in adjacency matrix form as per the theory above, where 1 indicates the row time period is at least as good as the column time period, and a time period is assumed to be as least as good as itself. The SWB and RP orderings look as follows:

$$H = \textit{from}: \begin{array}{c} \textit{to}: \\ 1975 \\ 1980 \\ 1985 \\ 1996 \\ 2005 \\ 2011 \end{array} \begin{array}{c} \begin{array}{cccccc} 1975 & 1980 & 1985 & 1996 & 2005 & 2011 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

[13] See Varian (1990) for more details on this.

$$U = \textit{from}: \begin{array}{c} \textit{to}: \\ 1975 \\ 1980 \\ 1985 \\ 1996 \\ 2005 \\ 2011 \end{array} \begin{array}{c} \begin{array}{cccccc} 1975 & 1980 & 1985 & 1996 & 2005 & 2011 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

Once transitively closed, the adjacency matrices can be represented in graph form. For each 1 in the matrix, which denotes the fact that time period a is at least as good as b, draw an arrow from a to b. This is known as a directed graph. Figure 3.1 illustrates these for the UK's RP and SWB orderings, where each 1 in the UK's H and U adjacency matrices is represented as an arrow connecting the row time period to the column. One could then look for a way to measure the difference between the two directed graphs / the two adjacency matrices, instead of the direct comparison of the two ordered lists.



(a) RP ordering (b) SWB ordering

Figure 3.1: UK directed graphs

3.2.3 Measurement options

For this chapter's analysis, I considered a range of techniques to quantify the differences between either the preference orderings directly or the difference between vectors summarizing the information containted in the adjacency matrices (i.e. stacking the rows of 1s and 0s in the matrices). I looked at the standard Kendall τ rank correlation measure versus a range of alternative string matching techniques. A "string" in this chapter could be either the ordered lists directly, or the summary vectors. String matching techniques can be divided into three categories (van der Loo 2014):

(1) **Edit based distances**: how many operations are needed to transform one string into another?

Operations include actions such as: deletions, insertions, swaps, reversals. The fewer operation types are allowed, the more costly the transformation from one string to another. I looked at the Levenshtein and Hamming distances. The Levenshtein distance allows three forms of single character edits: deletions, insertions or substitutions. The Hamming distance allows only substitutions, and hence is weakly larger than the Levenshtein. Both of these distances require a normalisation, to prevent longer orderings accessing larger measures of difference simply by virtue of the fact that they summarize more preference orderings.

(2) ***q*–gram based distances**: how many character sequences of length q appear in both strings? The choice of the length of the sequence sets the degree of difficulty of determining strings to be similar to one another (there are fewer counts of a run of characters common to both strings as the run lengthens). I looked in particular at the Jaccard distance.

(3) **Heuristic distances**: these are designed to measure typed errors in human–generated texts, so I did not consider any of these types of measurements for this chapter.

In the main text, I only give the details for computing Kendall τ, as this is the measure I chose to use in this chapter. However there are details and examples of how the Levenshtein, Hamming and Jaccard distances are computed in Appendix 3A. I include below a summary table and brief outline of why I choose the Kendall measure as opposed to the other measures, but give further detail on the evaluations I made in Appendix 3A.

Kendall τ

The Kendall τ measure of rank correlation was developed to measure either the discrepancy between a correctly ranked list of items and a second attempted ranking of the same items, or the difference between two alternative rankings of the same items (Kendall 1938). τ is defined as:

$$\tau = \frac{2\sum}{n(n-1)}$$

where n is the number of items in the ranking. The $\sum$ is the total accuracy of the positioning of each item relative to the others in the list. Kendall highlights its key use is in psychological work. The measurement method mimics the logical method by which an individual would create a ranking: taking an item, and comparing it with the remaining items one by one to find its optimal place in the ordering (1938).

Kendall's paper gives the following example: A set of 10 items are ordered in the natural sequence $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ (O^N). Another ordering of the same items is $4, 7, 2, 10, 3, 6, 8, 1, 5, 9$ (O^A). The difference between O^A and O^N can be measured by taking each item in O^A and noting its relative

positioning to the other items in O^A. For instance, 4 appears before 7. Then consider whether this relative position is the same as in the ordering O^N. In O^N, 4 also appears before 7. Add 1 for the item being in the correct placement relative to another item; subtract 1 for the item being in the wrong place. This is repeated for each item, up to the penultimate item, and then the totals are summed.[14]

The maximum score, corresponding to a situation in which all the items in O^A are in precisely the same order as in O^N, forms the denominator of the τ coefficient. Thus, τ will be $+1$ if the rankings are perfectly the same. It will be -1 if they are exactly inverted.[15]

The Kendall τ coefficient is similar in concept and construction to a simple normalized count, in which 1 is counted for every inconsistency in the adjacency matrices, and then the total is divided by the dimension of the matrices (36, in the UK case above). Indeed, the normalisation of dividing by the total preference pair relations is the same as the method of dividing by the maximum possible Kendall τ score, $\frac{n(n-1)}{2}$. The only difference is that while the normalised count will penalise every incorrect pair–wise ordering by adding 1, Kendall's τ *adds* 1 for a correct pair ordering and *subtracts* 1 for an incorrect ordering.

3.2.4 Choosing a measure

To determine the strengths of the different measures, I considered two key criteria: (1) the appropriateness of the measure for application to either ordered lists or the summary vectors of the adjacency matrices and (2) the appropriateness of the set of transformation rules in the context of RP and SWB orderings. A summary of the discussion is in Table 3.1.

	Summary Vectors?	Ordered Lists?	Transformation Rule?
Hamming	✓	✗	✓
Levenshtein	✓	✗	✗
***q*–distance**	✗	✓	✗
Kendall τ	✗	✓	✓

Table 3.1: Choosing a distance measure

The discussion in Appendix 3A highlights that in the context of RP and SWB orderings for countries and years, only the Hamming and Kendall τ have suitable transformation rules to measure the distance

[14] The first two set of workings: $+1-1+1-1+1+1-1+1+1=+3$ (4 is correctly positioned relative to 7, 10, 6, 8, 5 and 9 but not to 2, 3 and 1). Likewise, 7 is scored: $-1+1-1-1+1-1-1+1$, totalling -2 (7 is correctly positioned relative to 10, 8 and 9 but not to 2, 3, 6, 1 and 5). Once completed for all items, the totals are summed: $+3,-2,+5,-6,+3,0,-1,+2,+1 \implies \sum = +5$.

[15] Kendall's paper shows that τ has good statistical properties. If all possible rankings occur equally frequently, its standard error has a known form; for large values of n it may be taken to be normally distributed.

between the orderings. The normalised Hamming distance is suitable when applied to the summary vectors for the adjacency matrices. The Kendall τ coefficient is suitable when applied to the ordered lists. Of these two measures, in Section 4 I report the Kendall τ for two reasons:

- τ highlights instantly whether a ranking can be considered "mostly similar" or "mostly dissimilar" by virtue of the fact that $\tau \in [-1, 1]$;
- τ may also have the advantage of familiarity to some readers, as it is a commonly used measure of difference in rankings.

3.2.5 Benchmarking and inference

In order to determine whether the data reject the premise that SWB and RP are similar, a benchmark is required. How *close to 1* does the Kendall score need to be to conclude that RP and SWB orderings could be described as similar?[16] Following Becker (1962), I take random behaviour as a benchmark. For uncorrelated data — simulated SWB orderings and true RP orderings — $E(\tau)$ is 0.[17] Therefore, the mean Kendall measure of the true SWB and RP distance can be compared to 0: a positive τ implies that the SWB data perform better than random behaviour, in terms of their consistency with the RP ordering. For each section — the over–time comparisons and then the cross–country comparisons — I conduct the following two tests:

- A t–test.[18] Is the mean Kendall result across the group different from 0?[19] The null hypothesis is that there is no correlation between RP and SWB orderings.
- A non–parametric test, the Kolmogorov–Smirnov test, of the distribution of Kendall results against the distribution of Kendall results if RP orderings are compared to simulated SWB orderings. The null hypothesis is that there is no evidence that the distribution of Kendall results is different from the distribution of Kendall results should each country's SWB ordering be drawn at random (i.e. RP and SWB orderings are no more similar to each other than random orderings).

[16] In order to interpret these tests, I define the population about which I am making inferences as the infinite set of representative agents for all countries in all time periods.

[17] See Kendall (1938) for further details on τ's distribution.

[18] Note that the t–test is of the mean Kendall result over the set of ordering comparisons i.e. the mean within–country τ and the mean within–time τ. As τ is already a sample estimate, I bootstrap to estimate the mean τ in the group and its standard error. Given the small n (25 countries in the within–country estimation and 6 time periods in the within–time estimation), I sample with replacement.

[19] Kendall (1938) shows that the Kendall coefficient's distribution is normal as n gets reasonably large. I have a small sample here (degrees of freedom is 24). Bonett and Wright (2000) note that sample size formulae to determine the precise sample size needed to validate hypothesis tests are not available for the Kendall τ statistic (the authors do provide a table (p.26) for approximate sample sizes needed for varying widths of Fischer confidence intervals). However the small size of the p–value that I obtain in the results are reassuring that the Kendall true mean is statistically different from 0.

3.3 | Data

3.3.1 Data requirements and sample selection

The RP orderings require observed consumption data and the SWB orderings require reported satisfactions. To achieve this, I amalgamate data from the International Comparison Program (ICP) and the World Database of Happiness (WDH), from 1975 to 2011. The key variables are grouped consumption volumes and prices together with reported SWB levels. The unit of analysis is therefore the representative agent per country, per year. An overview of the two component datasets is as follows:

- The ICP is a global initiative to gather comparable economic data from countries over time, to enable comparisons of living standards. Every 6 years, national statistics agencies measure prices and volumes.[20] The dataset contains an average price per country per year for four groups of consumption.[21] Section 3.2 gives more details about the ICP dataset.

- The WDH *Happiness in Nations* and *Trends in Nations* datasets have yearly average SWB data from countries around the world, including average happiness, inequality of happiness (standard deviation of responses to the surveys), happy life years (normalised mean happiness multiplied by average life expectancy), inequality–adjusted happiness (an index giving equal weight to the level and standard deviation of happiness) and life expectancy. To avoid assumptions regarding equivalence across these different scales, I select only observations gathered using the 7–step verbal Life Satisfaction measure.[22] This restricts the usuable data to 27 countries. Section 3.3 gives more details on the WDH dataset.

Combining the SWB data available from the WDH together with the six–yearly ICP database results in an unbalanced panel of data across 27 countries over 6 time periods. The spread of observations over the panel is illustrated in Table 3.2. The total number of observations is 97. More countries join the dataset as time goes on — although 4 then drop out in 2011 — as countries begin to collect SWB data, and in particular in the form of the 7–step verbal Life Satisfaction data measure. Note that Colombia and Ecuador each have only 1 observation which results in these countries not being usable in the within–country analysis (they each rank a list containing only 1 item). However, I keep them in the sample, as they add information to the 2005 within–year analysis.

[20] The 1975, 1980 and 1985 ICP data collections originate from the World Bank database "STARS"; 1996 comes from the Penn World Table Project; 2005 and 2011 come from the World Bank ICP.

[21] These groups are constructed from data published by the US US Bureau of Labor Statistics (BLS): (1) Food and beverages; (2) Clothing and footwear; (3) Housing; (4) Other (For 1975, 1980 and 1985, this classification is a weighted average of the groups Transportation, Entertainment, Medical care and Other goods and services (BLS–determined weighting). For 1996, 2005 and 2011, this classification is a weighted average of the groups Transportation, Recreation, Education and communication, Medical care and Other goods and services (BLS–determined weighting).).

[22] See Diener et al. (1985) for details on the development and properties of the Satisfaction with Life Scale.

	1975	1980	1985	1996	2005	2011	country *n*
Austria (AUT)				✓	✓	✓	3
Belgium (BEL)	✓	✓	✓	✓	✓	✓	6
Bulgaria (BGR)					✓	✓	2
Columbia (COL)					✓		1
Germany (DEU)				✓	✓	✓	3
Denmark (DNK)	✓	✓	✓	✓	✓	✓	6
Ecuador (ECU)					✓		1
Spain (ESP)			✓	✓	✓	✓	4
Estonia (EST)					✓	✓	2
Finland (FIN)				✓	✓	✓	3
France (FRA)	✓	✓	✓	✓	✓	✓	6
UK (GBR)	✓	✓	✓	✓	✓	✓	6
Greece (GRC)			✓	✓	✓	✓	4
Hungary (HUN)					✓	✓	2
Ireland (IRL)	✓	✓	✓	✓	✓	✓	6
Japan (JPN)	✓	✓	✓	✓	✓		5
Lithuania (LTU)					✓	✓	2
Luxembourg (LUX)	✓	✓	✓	✓	✓	✓	6
Latvia (LVA)					✓	✓	2
Mexico (MEX)	✓			✓	✓		3
Netherlands (NLD)	✓	✓	✓	✓	✓	✓	6
Poland (POL)					✓	✓	2
Portugal (PRT)			✓	✓	✓	✓	4
Romania (ROM)					✓	✓	2
Slovakia (SVK)					✓	✓	2
Slovenia (SVN)					✓	✓	2
USA (USA)	✓	✓	✓	✓	✓	✓	6
yearly *n*	10	9	12	16	27	23	97

Table 3.2: The full sample

Notes: In order to be in the full sample, the country–year observation must have both ICP consumption data, and also 7 step Life Satisfaction data from the WDH.

A caveat: aggregated data Both the ICP and WDH datasets are constructed from individual–level data, aggregated per nation and per year: in using these data, my approach is therefore a representative consumer model which is very commonly used in macroeconomic papers but that does require some key assumptions. With respect to the ICP data, theory states exact aggregation is possible if and only if individual preferences have the Gorman polar form (individuals have expenditure functions that are linear in utility, with a common slope), and that exact aggregation implies that a representative agent exists.[23] There are theoretical arguments and empirical evidence that this assumption may not hold.[24] However, Cherchye et al. (2011) characterise a necessary and sufficient revealed preference test for the existence of a representative consumer, and then apply this to Spanish household data, finding that the observed choice patterns do not reject the existence of a representative consumer.

Putting the existence of representative agents aside, Blundell, Browning and Crawford (2003) note two other problems associated with the use of RP theory in particular on aggregated data: (1) should movements of the budget lines be large and price changes small, budget lines may not cross, so that RP conditions cannot be rejected; (2) if RP conditions can be rejected, this could be either due to individual level failure or aggregation failure (individuals do not have the same homothetic preferences). Turning to the SWB data, the use of nationally averaged data requires the assumption of inter–personal comparability of individual SWB reports to hold. This is an assumption that I do not make in Chapter 1 of this thesis, although the discussion of the results of Chapter 1 highlights the fact that the harm of making this assumption may be minimal.

A further caveat: SWB measurement It is important to acknowledge the controversy over using SWB data, in addition to the concern about inter–personal comparability. SWB measures are collected through questionnaires, and both the format of the responses and the subjectivity of the content both are sources of concerns about measurement.[25] Steps have been taken to test the accuracy of SWB

[23] The original paper is Gorman (1961). It can also be shown that if indirect utility functions of individual consumers are in this form, the preferences of the representative consumer are independent of the social welfare function used (Mas–Colell, Whinston and Green (1995))

[24] For example, Banks, Blundell & Lewbel (1997) present evidence that Engel curves are curved (require quadratic terms in log expenditure), and that failure to account for this results in distortions in the estimation of welfare in a UK household dataset. Kirman (1992) presents four theoretical reasons why — putting aside the fact that such models cannot address coordination problems between individuals — the representative agent models cannot be used to explain macroeconomic scenarios: (1) we cannot justify the assumption that an aggregate of individuals will behave like an individual utility maximiser; (2) the reaction of the representative to a change in model parameters may not be the same as the aggregate reaction of the individuals; (3) preferences of the representative may not align with the individuals, such that the judgement of an economic situation as better or worse is not aligned; (4) the sum of plausibly behaved individuals may not lead to the representative behaving plausibly, and as such we cannot be sure of correctly rejecting a behavioural hypothesis or simply rejecting the hypothesis of a plausible representative. Kirman gives some examples from empirical literature to illustrate each point. The key issue for this chapter is point (4).

[25] I group the critiques into three main categories: *(1) Response constraints:* Typically, SWB responses are in ladder form. Acemoglu (2008) highlights that Easterlin's satiation point is therefore an artifice of the scale: as income grows – unconstrained by an upper bound – happiness growth must be curtailed. Also, respondents have a more precise consideration of their happiness than the discrete scale allows, as per the discussion in Chapter 1 on the quantisation reporting error. In microeconomic theory terms, respondents have thick indifference curves. *(2) Environmental factors:* Respondents may

measurement. For example, an objective test for the truthfulness of the reports is useful. Stevenson and Wolfers (2008) find SWB responses correlate to physical signs of affect such as brain activity or smiling. Andrews and Withey (1976) find SWB responses correlate with other self–reported measures and external variables such as marriage or standards of living. Larsen, Diener and Emmons (1985) find a correlation of SWB responses with other well–being measures. Krueger and Schkade (2007) interrogate the test–retest reliability of life satisfaction and affect measures, with interesting conclusions: whilst the former is more strongly correlated with income, the latter has a slightly higher net affect over two weeks. Evidently, SWB is challenging to measure accurately and using a laddered scale constrains the information captured: these concerns may confound any certainty in interpretation of the differences between RP and SWB orderings, and are therefore important to have in mind.

Sample size and alternative datasets: The subsample used in this chapter is small, and some of the preference orderings for countries are only made up of 2 or 3 observations. It would therefore be useful to apply the method to a larger macroeconomic sample, with more countries, and more years. Additionally, applying the method to a microeconomic panel dataset with both price and consumption data, and SWB data, would be interesting (this would remove the need for the assumptions regarding aggregation outlined above, giving greater clarity to the interpretation of the results). Finally, it would be a natural extension using the current dataset to run the RP and ordinal comparison analysis on the entire pool of countries and years. This would result in a summary τ statistic for the whole dataset. This would require more efficient computational methods.

3.3.2 International Comparison Program data

The ICP partners with national and international agencies to collect national price and expenditure data worldwide using price surveys.[26] The average prices enable the calculation of price indexes for GDP and consumption for each country. Removing price level differences in countries then enables international price comparisons: indexes for each country are expressed as Purchasing Power Parities (PPPs).[27] PPPs are expressed in US dollars, such that the US's PPP is 1.0 in all years. The US Consumer Price Index (CPI–data) from the US Bureau of Labor Statistics (BLS) is used to enable

benchmark themselves against a global standard; a country maintaining the same position in an international ranking may therefore grade themselves at the same happiness level, even if global levels have increased (Deaton 2008). This would be consistent with the Easterlin paradox result, and such an effect could also occur as a peer–effect. Additionally, beliefs or aspirations vary with standards of living. Deaton (2008) does, however, find a strong correlation of wealth and life satisfaction. Likewise, agents could adapt to unhappier situations, and thus lose their perspective (Deaton 2008; Sen 1985). *(3) Consistency of responses:* Mood has a large influence on survey responses (Schwarz and Clore 1983). Likewise, external factors such as seasons may affect the SWB reports. However, Diener (1984) reports evidence for long–term reliability of SWB reports. Also, people may respond in an "off–the–cuff" manner or not understand the question (of course, the same critique can be applied to any survey: respondents may quickly fire–off a poorly considered answer).

[26] See Chapter 1, on "Sources of Price and Expenditure Data", in World Bank (2013).

[27] The ICP provide the following definition of a PPP: "the number of currency units required to purchase the amounts of goods and services equivalent to what can be bought with one unit of the currency unit of the base or reference or numeraire country" (The World Bank 2013, 19). This can vary, depending on the commodity that is being considered.

time–trend analyses.

Countries' national accounts collect data on volumes of consumption, which when combined with the PPPs, results in a directly comparable, real national expenditure level. Per capita figures can be used to compare economic welfare of representative agents across countries: the 2011 ICP dataset, which is the most recent, has comparable data from 199 countries. This ICP data can be used directly to derive the RP adjacency matrix, and then transitively close this matrix to construct the RP ordering, and test the preference ordering for GARP. This RP ordering can then also be compared to the SWB ordering.

In order to describe the ICP dataset in more detail, and enable some trend descriptives with the SWB data (Section 3.4), I calculate GDP per capita for each country in each year. I choose the USA in 1975 as the base country–year (I set the USA's four prices and four consumption bundles to 1.0 in this year). I use the GEKS and CCD indices so that prices in the ICP dataset can be compared simultaneously across countries and over time.[28]

3.3.3 World Database of Happiness data

The World Database of Happiness describes itself as *an archive of research findings on subjective enjoyment of life*; it compiles findings and data from many different sources. Across the collections, there are data from 175 nations in total, comprising survey responses from 1348 different measures of happiness.[29] The database contains information on measures of happiness, levels of happiness (national averages), inequality of happiness, correlates of happiness, changes over time and cross–cultural differences. In order to avoid making assumptions regarding equivalence of levels of SWB collected by different scales, I choose in this chapter to isolate only data collected by the same question (the 7 step Life Satisfaction). The question used is as follows: *Taking all together, how satisfied are you with your life as a whole these days? (1) Completely dissatisfied (2) Very unsatisfied (3) Unsatisfied (4) Neither satisfied no unsatisfied (5) Satisfied (6) Very satisfied (7) Completely satisfied.* In the WDH dataset, the responses have already been transformed by the database owners onto a scale from

[28] *GEKS* is the Gini–Elteto–Koves–Szulc index. The GEKS takes an average of the indices for all pair–wise combinations of the country–years items in the dataset; the index is a geometric average of the bilateral Fisher* index. The method renders the comparisons across countries transitive, such that the results obtained are not dependent on the choice of the USA as a base, or star, country. See Balk (2009), Rao (2009) and de Haan, Willenborg and Chessa (2016) for more details on the calculation. *CCD* is the Caves, Christensen and Diewert index (1982). When the Fischer is substituted by the Tornqvist* index for price comparisons, this is the CCD index. Neary (2004) discusses whether the GEKS and CCD are appropriate for making welfare comparisons; the conclusion is that they are indeed appropriate under the assumptions of (1) homogeneous and (2) homothetic preferences. **The Fischer and Tornqvist* indices are price indexes computed using the price and quantity data from two countries only (i.e. enabling bilateral comparisons). The Fischer is the geometric average of the Laspeyres and Paasche indexes. The Tornqvist is the weighted geometric average of the price relatives for each commodity. See World Bank (2013) for more details.

[29] Veenhoven, WDH website (last accessed on 08/05/2020).

0 – 10, hence the observations I have are on this 0 – 10 scale, rather than on the original 1 – 7 scale.

The resulting sample provides a panel of countries' average SWB levels over time. To then derive the SWB orderings, I do not need to treat the SWB data at all: these are already comparable within time and within countries. Considering whether there might be a difference in scale interpretation culturally — resulting in a difference in average SWB levels across countries — would invite a treatment of the SWB variable to derive a *change in SWB* variable, relative to a base–year. I discuss this in the conclusion of this chapter, but given the small sample size of the combined ICP–WDH dataset, I choose not to lose an observation per country by conducting analysis in changes in the results presented here.

3.3.4 Data descriptives

ICP dataset

The following tables provide descriptive statistics of the per capita real GDP data: by year (Table 3.3) and by country (Table 3.4). These are computed according to the GEKS method, indexed so that the USA in 2011 is 1.0. Figure 3.2 plots the indexed GDP for all countries over time.

Year	Mean	Std. Dev.	Min	Max	*n*
1975	0.368	0.105	0.192	0.568	10
1980	0.419	0.093	0.237	0.570	9
1985	0.371	0.128	0.168	0.646	12
1996	0.489	0.159	0.178	0.796	16
2005	0.494	0.239	0.120	1.009	27
2011	0.572	0.213	0.245	1.043	23

Table 3.3: Real income per capita — by year

Notes: Real GDP per capita is calculated from the ICP dataset using the GEKS method. The US in 2011 is the base country–year.

Table 3.3 shows a steady increase in mean indexed real GDP per capita across the 6 years, with a dip in the trend in 1985. There is a general tendency for increasing standard deviations (between country inequality), with slight dips, across the six observations. Crucially, as with the SWB data, note that the panel is expanding from the first set of observations (1975); any apparent trends therefore could be explained by the introduction of particular countries as time moves on. The panel is constrained by the availability of SWB survey results: the set of included countries therefore increases in time as more nations start to collect SWB data. In Table 3.4, it is useful to note which countries have fewer observations (Columbia and Ecuador with just one observations each), and also those countries that join the dataset later in the panel (for example, Romania, Bulgaria and Latvia).

	Mean	Std. Dev.	Min	Max	*n*
Austria (AUT)	0.669	0.121	0.529	0.748	3
Belgium (BEL)	0.520	0.129	0.395	0.693	6
Bulgaria (BGR)	0.214	0.044	0.183	0.245	2
Columbia (COL)	0.123	—	0.123	0.123	1
Germany (DEU)	0.639	0.083	0.551	0.716	3
Denmark (DNK)	0.549	0.125	0.412	0.697	6
Ecuador (ECU)	0.120	—	0.120	0.120	1
Spain (ESP)	0.449	0.155	0.286	0.581	4
Estonia (EST)	0.341	0.081	0.328	0.354	2
Finland (FIN)	0.580	0.130	0.441	0.700	3
France (FRA)	0.532	0.127	0.411	0.694	6
UK (GBR)	0.502	0.167	0.330	0.728	6
Greece (GRC)	0.419	0.170	0.234	0.564	4
Hungary (HUN)	0.354	0.007	0.349	0.359	2
Ireland (IRL)	0.382	0.184	0.224	0.624	6
Japan (JPN)	0.441	0.121	0.325	0.589	5
Lithuania (LTU)	0.349	0.058	0.308	0.390	2
Luxembourg (LUX)	0.669	0.269	0.420	1.043	6
Latvia (LVA)	0.294	0.045	0.262	0.326	2
Mexico (MEX)	0.209	0.042	0.178	0.256	3
Netherlands (NLD)	0.512	0.143	0.378	0.690	6
Poland (POL)	0.356	0.079	0.301	0.412	2
Portugal (PRT)	0.365	0.150	0.168	0.491	4
Romania (ROM)	0.224	0.047	0.190	0.258	2
Slovakia (SVK)	0.362	0.073	0.311	0.414	2
Slovenia (SVN)	0.486	0.033	0.462	0.510	2
USA (USA)	0.765	0.203	0.570	1.009	6

Table 3.4: Real income per capita — by country

Notes: Real GDP per capita is calculated from the ICP dataset using the GEKS method. The US in 2011 is the base country–year.



Figure 3.2: Countries' Real Indexed GDP per capita Over Time

Notes: GDP is the real GDP per capita for each country in each time period, indexed for comparison to the base country–year, which is the USA in 2011 (1.0 on the index). The GEKS method was used. **Abbreviations:** AUT: Austria; BEL: Belgium; BGR: Bulgaria; COL: Colombia; DEU: Germany; DNK: Denmark; ECU: Ecuador; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GRC: Greece; HUN: Hungary; IRL: Ireland; JPN: Japan; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MEX: Mexico; NLD: Netherlands; POL: Poland; PRT: Portugal; ROM: Romania; SVK: Slovakia; USA: USA.

Table 3.4 reports each countries' mean, standard deviation and range of indexed real GDP per capita over time. There is a broad range of average wealth, with Latin American countries towards the poorer end, and the USA and Western Europe towards the richer end. The USA's maximum of 1.009 reflects the slight drop in GDP per capita in 2011 from 2005 (pre–crisis) levels.

Figure 3.3 shows a graph of the log indexed real GDP per capita values for each country over time; again, this is computed according to the GEKS method. Therefore the USA in 2011 is now 0.0. There is an overall trend in positive growth over the years for most countries, with a slackening off of growth rates between 2005 and 2011 (the global financial crisis).

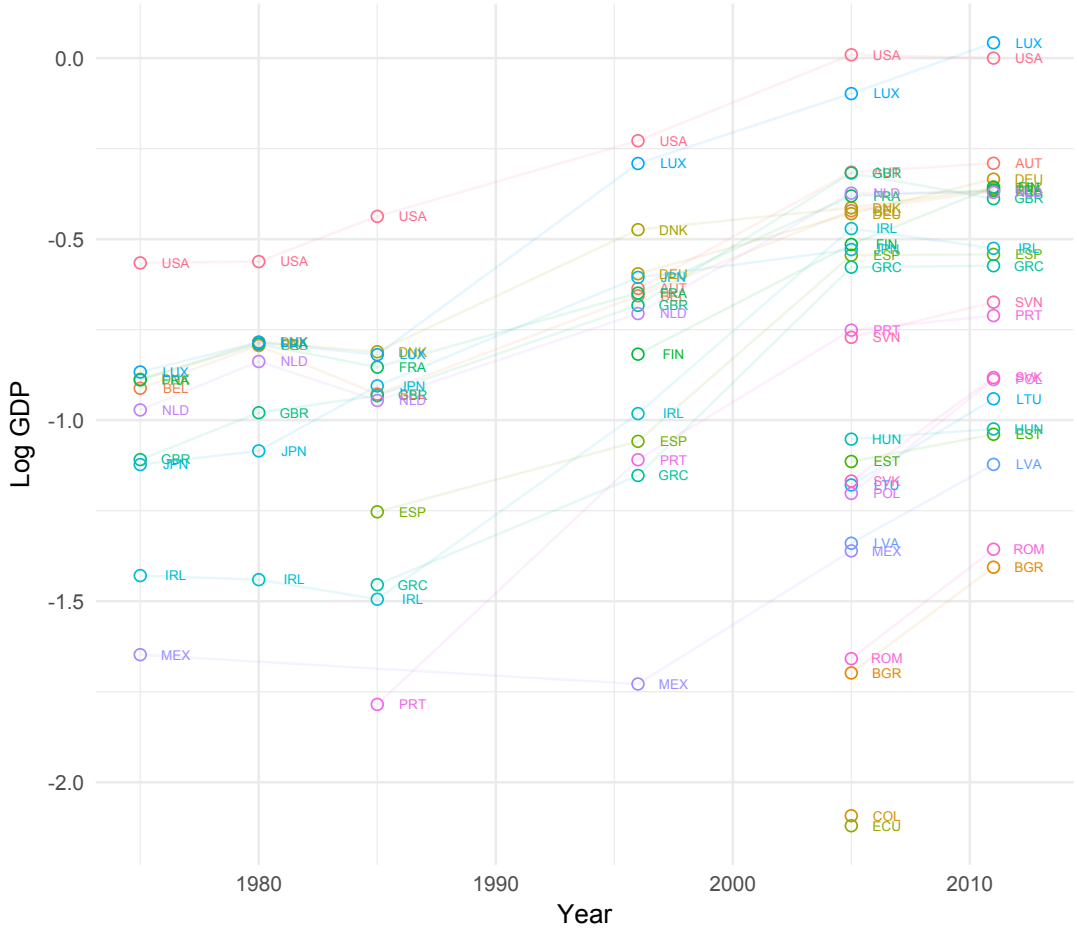


Figure 3.3: Countries' Indexed Log GDP Over Time

Notes: Log GDP is the log real GDP per capita for each country in each time period, indexed for comparison to the base country–year, which is the USA in 2011 (0.0 on the log index). The GEKS method was used. **Abbreviations:** AUT: Austria; BEL: Belgium; BGR: Bulgaria; COL: Colombia; DEU: Germany; DNK: Denmark; ECU: Ecuador; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GRC: Greece; HUN: Hungary; IRL: Ireland; JPN: Japan; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MEX: Mexico; NLD: Netherlands; POL: Poland; PRT: Portugal; ROM: Romania; SVK: Slovakia; USA: USA.

SWB dataset

The following tables provide descriptive statistics of the WDH data, by year (Table 3.5) and by country (Table 3.6). Figure 3.4 plots the SWB levels in the dataset. Looking at Table 3.5, overall there is a gradual decrease in average SWB over the 6 years of observations and a gradual increase in standard deviation; this is also clear from the illustration in Figure 3.4.

The dataset is greatly expanding from the first set of observations: an apparent downward trend in the averages for each year therefore may be explained by the arrival of particularly unhappy countries as time moves on. This may in fact be the case: Bulgaria, Hungary and Romania each join the panel in 2005, and in Table 3.6 these countries in the set of the lowest mean levels of SWB. Denmark is highlighted as the happiest country in the set, with Mexico, the Netherlands and Columbia in

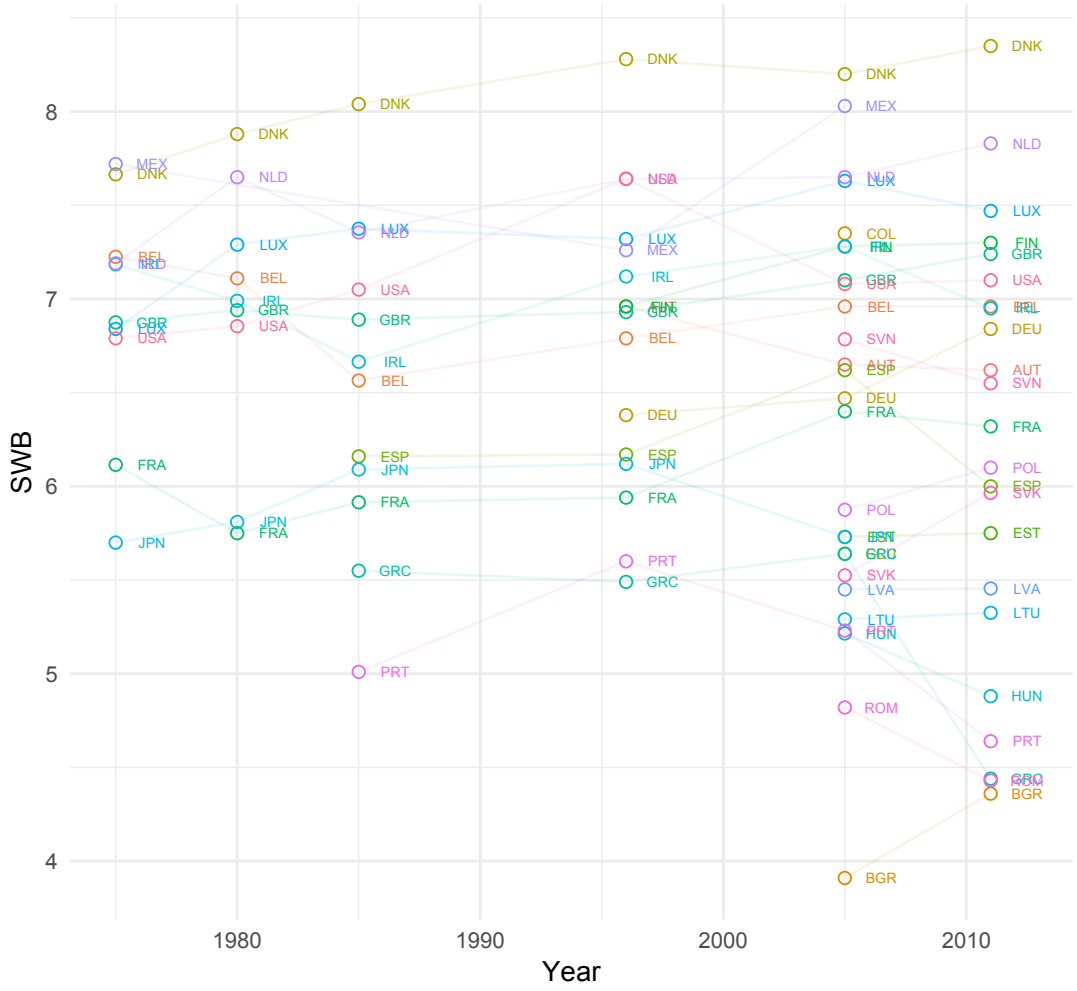


Figure 3.4: Countries' SWB Over Time

Notes: SWB is the average SWB report per country, per year, from the 7–step Life Satisfaction questionnaire. Answers are rescaled to 0 — 10. **Abbreviations:** AUT: Austria; BEL: Belgium; BGR: Bulgaria; COL: Colombia; DEU: Germany; DNK: Denmark; ECU: Ecuador; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GRC: Greece; HUN: Hungary; IRL: Ireland; JPN: Japan; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MEX: Mexico; NLD: Netherlands; POL: Poland; PRT: Portugal; ROM: Romania; SVK: Slovakia; USA: USA.

second, third and forth place. Mexico and Columbia are two of the three poorest countries in Table 3.4. Greece then Portugal report the highest standard deviations. N.B. Ecuador and Columbia do not report standard deviations as they have only 1 observation each (they may still be used as data points for between–country analysis in 2005).

WDH and ICP combined dataset

Finally, I present the correlation between the WDH and ICP data. To make the comparison as clear as possible, I transform the SWB reports into a GEKS–like index like the consumption data, with the

	Mean	Std. Dev.	Min	Max	*n*
1975	6.931	0.632	5.70	7.72	10
1980	6.919	0.728	5.75	7.88	9
1985	6.555	0.856	5.01	8.04	12
1996	6.788	0.783	5.49	8.28	16
2005	6.353	1.067	3.91	8.20	27
2011	6.212	1.154	4.36	8.35	23

Table 3.5: SWB — by year

Notes: 7 step Life Satisfaction data are transformed onto a 0–10 scale (0 is least satisfied, 10 is most satisfied)

	Mean	Std. Dev.	Min	Max	*n*
Austria (AUT)	6.743	0.188	6.62	6.96	3
Belgium (BEL)	6.935	0.234	6.57	7.23	6
Bulgaria (BGR)	4.135	0.318	3.91	4.36	2
Columbia (COL)	7.350	—	7.35	7.35	1
Germany (DEU)	6.563	0.244	6.38	6.84	3
Denmark (DNK)	8.069	0.261	7.67	8.35	6
Ecuador (ECU)	5.640	—	5.64	5.64	1
Spain (ESP)	6.238	0.266	6.00	6.62	4
Estonia (EST)	5.740	0.014	5.73	5.75	2
Finland (FIN)	7.180	0.191	6.96	7.30	3
France (FRA)	6.073	0.252	5.75	6.40	6
UK (GBR)	6.996	0.144	6.88	7.24	6
Greece (GRC)	5.280	0.563	4.44	5.64	4
Hungary (HUN)	5.048	0.237	4.88	5.22	2
Ireland (IRL)	7.032	0.217	6.67	7.28	6
Japan (JPN)	5.890	0.201	5.70	6.12	5
Lithuania (LTU)	5.308	0.025	5.29	5.33	2
Luxembourg (LUX)	7.321	0.266	6.84	7.63	6
Latvia (LVA)	5.453	0.004	5.45	5.46	2
Mexico (MEX)	7.670	0.387	7.26	8.03	3
Netherlands (NLD)	7.553	0.234	7.19	7.83	6
Poland (POL)	5.988	0.159	5.88	6.10	2
Portugal (PRT)	5.120	0.402	4.64	5.60	4
Romania (ROM)	4.625	0.276	4.43	4.82	2
Slovakia (SVK)	5.745	0.311	5.53	5.97	2
Slovenia (SVN)	6.668	0.166	6.55	6.79	2
USA (USA)	7.086	0.301	6.79	7.64	6

Table 3.6: SWB — by country

Notes: 7 step Life Satisfaction data are transformed onto a 0–10 scale (0 is least satisfied, 10 is most satisfied)

USA in 2011 again as the base country–year.[30] I then plot the log indexed SWB data against the log indexed GDP per capita data in Figure 3.5. These log variables have a positive correlation of 0.413. Appendix 3B reports the results of two regressions (indexed log real GDP per capita on log SWB and also indexed change in log real GDP on log SWB). These can offer a point of comparison for this dataset against other datasets used in regressions in the literature. However, most regressions in the literature use a set of covariates, which are not included in this dataset, and therefore are not controlled for in these regressions.

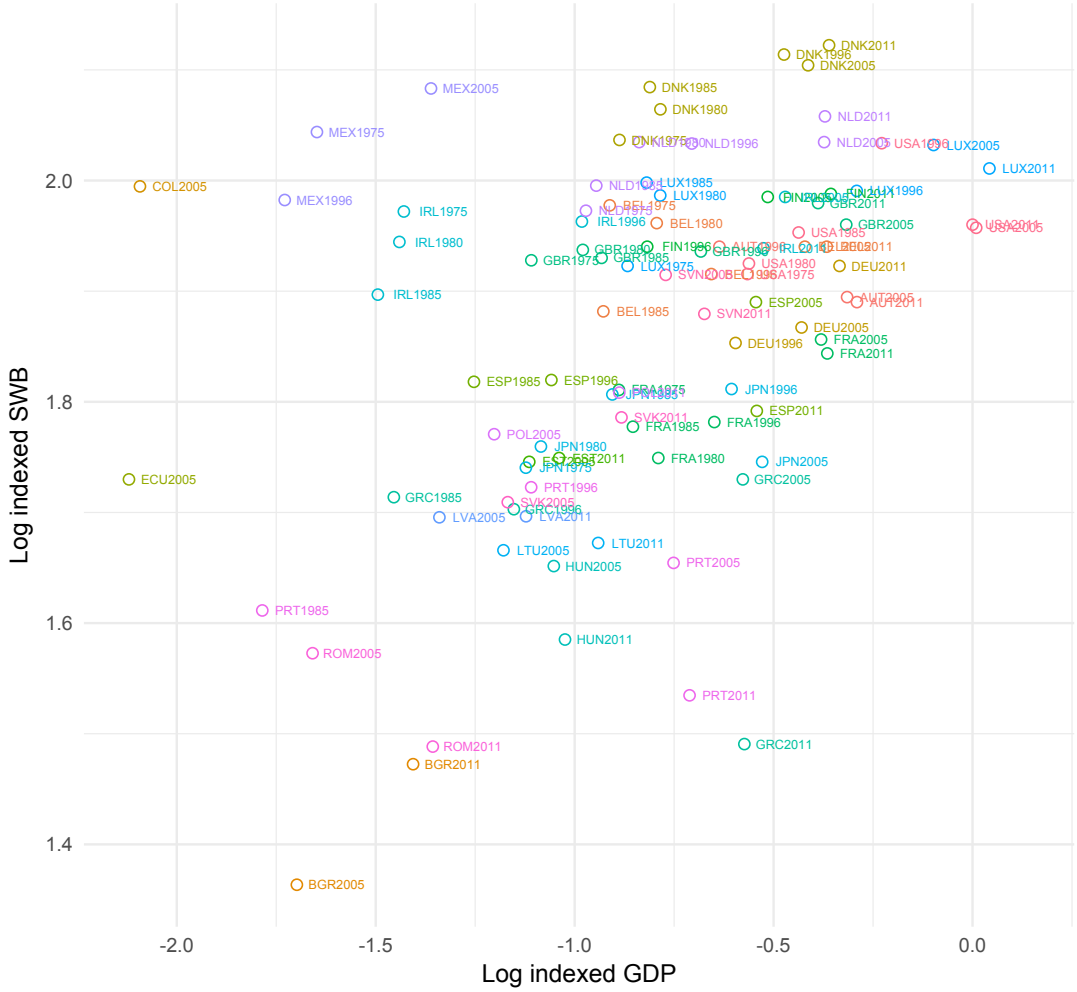


Figure 3.5: Countries' Indexed Log GDP and Log SWB

Notes: Log GDP is the log real GDP per capita for each country in each time period, indexed for comparison to the base country–year, which is the USA in 2011 (0.0 on the log index). The GEKS method was used. Indexed Log SWB the comparable SWB measure for each country–year. **Abbreviations:** AUT: Austria; BEL: Belgium; BGR: Bulgaria; COL: Colombia; DEU: Germany; DNK: Denmark; ECU: Ecuador; ESP: Spain; EST: Estonia; FIN: Finland; FRA: France; GBR: United Kingdom; GRC: Greece; HUN: Hungary; IRL: Ireland; JPN: Japan; LTU: Lithuania; LUX: Luxembourg; LVA: Latvia; MEX: Mexico; NLD: Netherlands; POL: Poland; PRT: Portugal; ROM: Romania; SVK: Slovakia; USA: USA.

[30] See Footnote 28 for some details on GEKS.

3.4 | Results

This section reports the results of the following analyses:

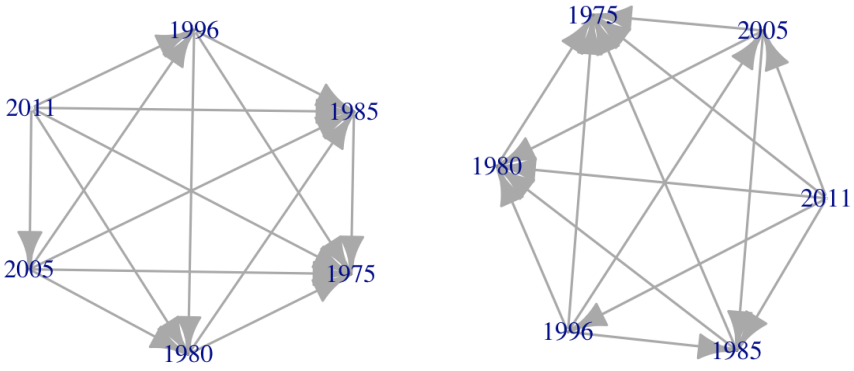
- Over–time comparisons within country, determining first if countries have stable preferences over time, and then comparing the RP orderings to SWB for each country;
- Cross–sectional comparisons per time period, determining first if countries have homogeneous preferences within a time period, and then comparing the RP orderings to SWB for each time period.

I find evidence that within–country, the RP orderings are stable over time and on average there is consistency between RP and SWB orderings. Within–year, the countries pass the GARP test in all years, thus not rejecting the possibility of homogeneous preferences, but the RP and SWB orderings are somewhat less consistent with one another.

3.4.1 Within–country analysis

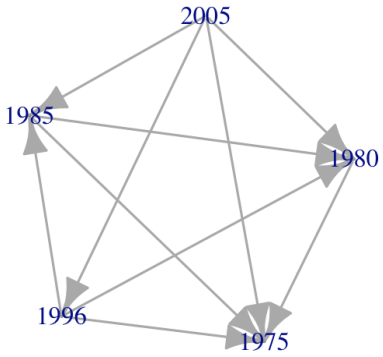
RP orderings and GARP: I use the ICP data to derive each country's representative agent's RP orderings over time. I obtain a set of RP orderings for each of the 27 countries. The left hand column of Figure 3.6 (graphs a, c, e and g) shows four example orderings as directed graphs, where an arrow from one period to another indicates that the first period is revealed preferred to the next. Each representative agent's consumption choices over time are then tested for cyclical consistency; this establishes whether the ICP data could have been generated by agents with stable preferences over time; I find that all countries pass this test (an Afriat Efficiency of 1). For those countries for whom $n > 1$ (all countries with the exception of Ecuador and Columbia) this is evidence that the representative agent's preferences are stable over time, and could be represented by a single utility function. Figure 3.6 demonstrates that the four example countries do indeed pass GARP: it is not possible to return to the starting point in any of the directed graphs.

Differences in RP and SWB orderings: I next obtain the SWB orderings from national average self–reports. The right hand column of Figure 3.6 (graphs b, d, f and h) presents four example SWB directed graphs for the same countries as the RP ordering depictions. Again, an arrow between two periods indicates that the first period has a higher SWB than the second. Figure 3.7 then depicts the differences between these four country's directed graphs; these give an indication of the size of the discrepancy between the country's RP and SWB preference orderings. For example, before calculating the Kendall τ results it is clear that Denmark (3.7a) has fewer inconsistencies than France (3.7c).

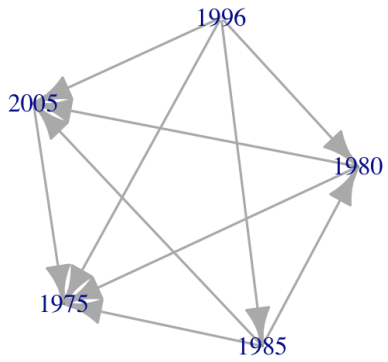


(a) Denmark — RP

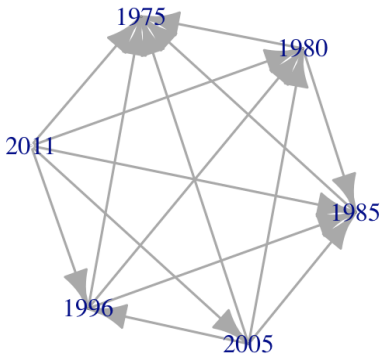
(b) Denmark — SWB



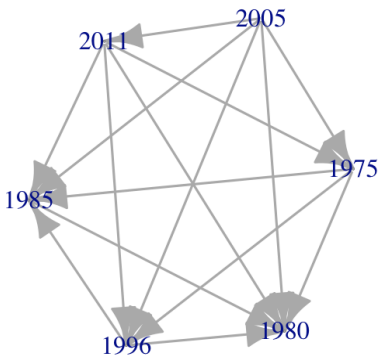
(c) Japan — RP



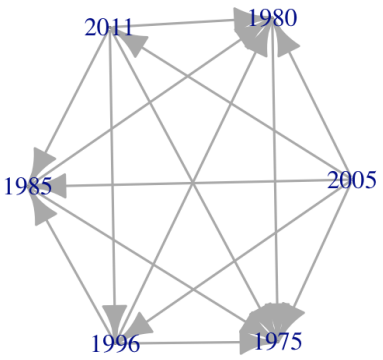
(d) Japan — SWB



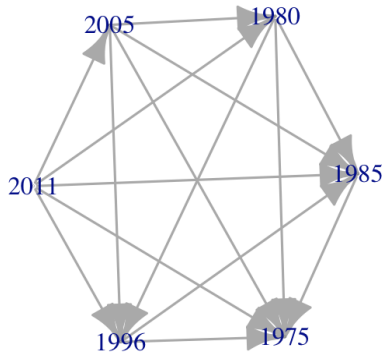
(e) France — RP



(f) France — SWB



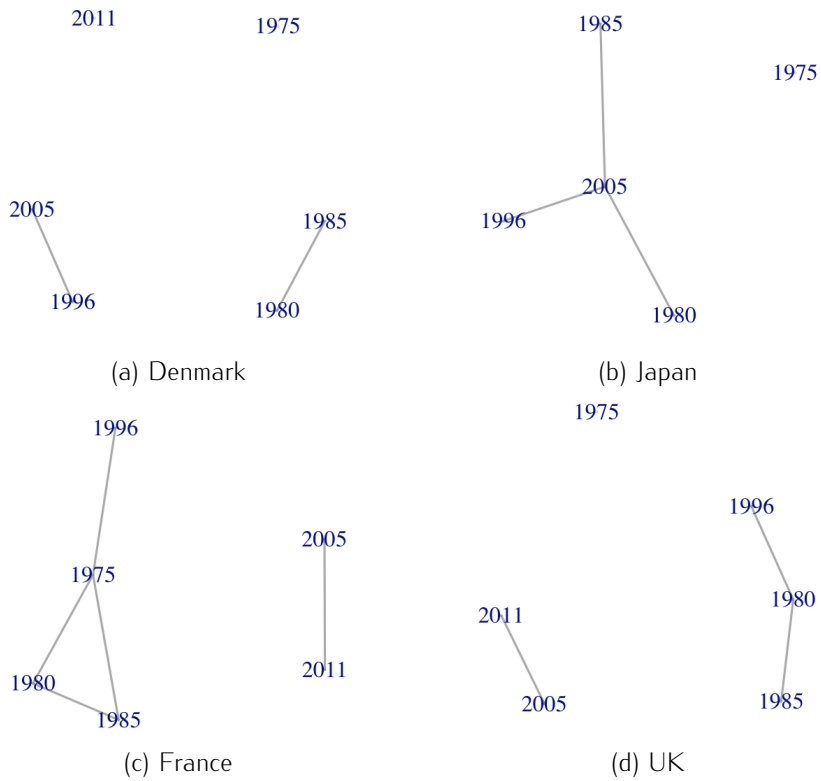
(g) UK — RP



(h) UK — SWB

Figure 3.6: Within country RP orderings

Notes: Directed graphs showing RP and SWB orderings for four example countries, derived from the ICP dataset. An arrow from one period to another indicates that the first period is revealed preferred (in the RP case) or has a greater SWB (in the SWB case) to the second.



(a) Denmark (b) Japan

(c) France (d) UK

Figure 3.7: Within country ordinal differences

Notes: Ordinal differences, showing discrepancies in RP orderings and SWB orderings. A line connecting two periods indicates that these periods were reversed in the two orderings. The more lines, the smaller the Kendall τ for that country.

I next compute the Kendall τ coefficient to measure how different the RP and SWB orderings are. The results are displayed in Table 3.7. These results can be compared to the difference graphs illustrating Denmark, the UK, France and Japan. Denmark scores the highest level of consistency out of all the countries: τ_{DNK} is 0.733. As implied by the graphs, comparatively to Denmark, France scores a more inconsistent τ_{FRA} of 0.333. The comparison of Japan and the UK shows that Japan has greater inconsistency: τ_{JPN} is 0.4 compared to a τ_{GBR} of 0.6.

There are 9 cases of perfect GARP and Kendall scores: Bulgaria, Germany, Estonia, Finland, Lithuania, Latvia, Mexico, Poland, Slovakia. This implies that not only for each of these countries can the consumption data be rationalised by a single set of preferences, but also that the orderings implied by the consumption choices are the same as those stated in these countries' SWB surveys: RP and SWB preference orderings are perfectly consistent with each other. This result is most compelling for Germany, Finland and Mexico who have three observations (the other countries, with two observations each, had only a single correctly ordered pair to obtain the perfect score). I find a τ of -1 for Slovenia, Romania, Hungary and Austria. Austria ($n = 3$) therefore has a complete reversal of all three preference relations. Despite the consumption choices being associated with a stable RP ordering over time, the SWB levels of these representative agents state exactly contrary preferences. The spread of

	Kendall	n
AUT	−1	3
BEL	0	6
BGR	1	2
COL	NA	1
DEU	1	3
DNK	0.733	6
ECU	NA	1
ESP	0	4
EST	1	2
FIN	1	3
FRA	0.333	6
GBR	0.600	6
GRC	−0.333	4
HUN	−1	2
IRL	0.467	6
JPN	0.400	5
LTU	1	2
LUX	0.600	6
LVA	1	2
MEX	1	3
NLD	0.690	6
POL	1	2
PRT	−0.333	4
ROM	−1	2
SVK	1	2
SVN	−1	2
USA	0.600	6

Table 3.7: Within country RP and SWB distances

Notes: Kendall τ scores for each country. $\tau = 1$ indicates perfect consistency between the RP and SWB orderings; $\tau = -1$ indicates perfectly inverted orderings. Note that all the countries had an Afriat Efficiency score of 1.

results from perfectly inconsistent RP and SWB preference orderings to perfectly coinciding orderings makes a overall result difficult to reach without using statistical inference.

Inference: I consider whether the Kendall results in Table 3.7 differ significantly or not from uncorrelated rankings. For uncorrelated data $E(\tau)$ is 0.[31] Therefore I simply compare the Kendall results' difference from 0: a positive τ implies that the SWB data perform better than a randomly generated ordering in terms of their consistency with the RP ordering. The (bootstrapped[32]) mean τ is 0.347, with a standard error of 0.123, and so the null of 0 can be rejected: the 95% confidence interval (0.297 — 0.398) does not include 0, and the 99% does not either (0.278 — 0.416). Hence I reject the null hypothesis that the RP and SWB orderings are no more similar to each other than if the SWB ordering had been generated at random.

I also consider the difference between the distribution of the Kendall results across the countries against the distribution of Kendall results computed from the comparison of the true RP orderings with randomly generated SWB orderings; this random SWB behaviour is simulated. I apply a non–parametric Kolmogorov–Smirnov (KS) test to the Cumulative Distribution Functions (CDFs). I first test the null hypothesis of whether the CDF of the Kendall results could come from the same distribution as the benchmark (simulated) data. This two–sided K–S test has a strong rejection of the null at a 1% significance level: there is a difference of 0.68, and $p < 0.01$. I also test whether the Kendall results are centred around a mean that is *larger* than that of the benchmark distribution: the one–sided test "less" tests the same null against the alternative that the Kendall distribution is *not smaller than* the simulation distribution. Again, this null hypothesis is strongly rejected ($p < 0.01$).

Thus far, all countries have RP utility orderings that display stability over time (Afriat Efficiency Index of 1). When compared to the SWB orderings, on average the RP orderings are more consistent with the true SWB orderings than with simulated randomised rankings. The 99% confidence interval of the bootstrapped mean τ and non–parametric KS tests support the result. I therefore conclude this section maintaining the hypothesis that RP and SWB orderings contain the same ordinal information for representative agents over time.

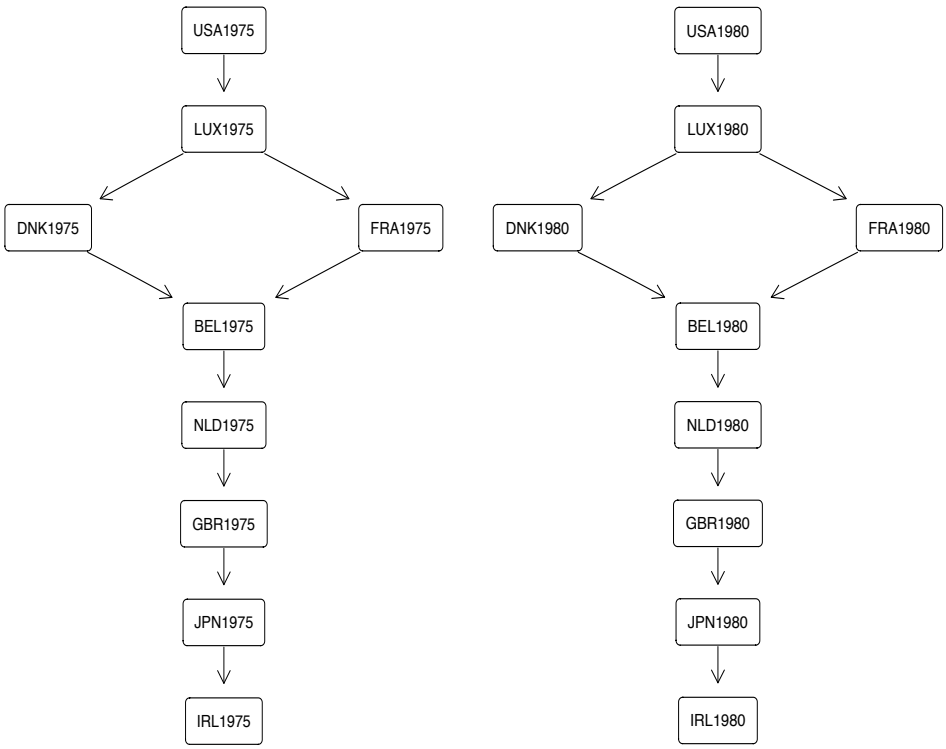
3.4.2 Within–year analysis

RP orderings and GARP: I first derive the RP orderings across countries for each time period in the ICP data. As before, these orderings could be presented in directed graphs but with such a large

[31] See Kendall (1938) for further details on τ's distribution.

[32] The ordinal τ results per country are sample estimates. The mean τ and standard error are therefore calculated by bootstrapping: I sample with replacement to obtain the empirical bootstrap distribution.

n these are no longer helpful representations. Instead, I simplify the RP orderings into Hasse diagrams: in a Hasse diagram, a country is revealed preferred to any country below it on the diagram (a downwards arrow indicates that the above country is revealed preferred to the one below). Figure 3.8 shows example Hasse diagrams for 1975 and 1980. Each is only partially ordered, as RP orderings need not be complete should two choices not be comparable. For example, France and Denmark were not directly comparable in the ICP data. Each were comparable to Luxembourg and Belgium and therefore still able to appear in the partial ordering. Ireland is ranked bottom in both 1975 and 1980, with the USA at the top of both years. Indeed, the ordering is consistent across both years.



(a) 1975 (b) 1980

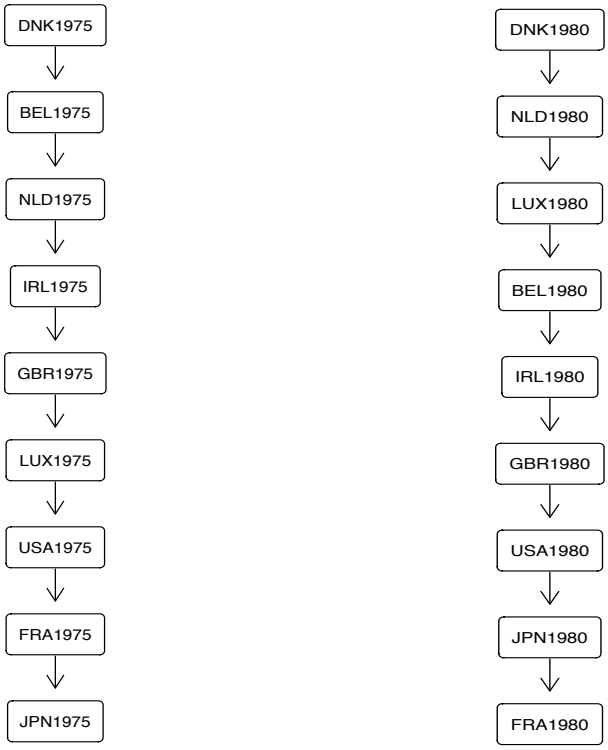
Figure 3.8: Within–year RP orderings

Notes: Example Hasse diagrams of the RP orderings for two years in the ICP data. A country is revealed preferred to any country below it on the diagram. Partial orderings occur when choices in two states are incomparable.

I next test whether the RP orderings are cyclically consistent across representative agents and find that all years pass the GARP test with Afriat Efficiency Index scores of 1. This result — perhaps surprisingly — is consistent with the hypothesis that the preferences derived from national price and consumption data are homogeneous across countries: in each time period, all the representative agents' preferences could be represented with one utility curve.

Differences in RP and SWB orderings: I derive the SWB orderings from the WDH data; these

can also be represented in Hasse diagrams; those for 1975 and 1980 are presented in Figure 3.9. The most satisfied countries are at the top of the diagrams. Comparing these to Figure 3.8, note that Denmark, not the USA, is the most satisfied country in both years. Ireland sits in the middle of the group of countries, with Japan and France in the bottom two places for each year. This immediately illustrates that there is some inconsistency across the two orderings.



(a) 1975 (b) 1980

Figure 3.9: Within–year SWB orderings

Notes: Example Hasse diagrams of the SWB orderings for two years in the WDH data. A country is reported as more satisfied to any country below it on the diagram.

I compute and report the Kendall results (Table 3.8). As with the within–countries results, there is a correlation between the two measures, with all but one year having a positive τ. However, the (bootstrapped) mean τ is now 0.316 (standard error 0.139), compared to the within–countries mean τ of 0.347, suggesting there is a little more inconsistency between the two orderings now that between–representative agent comparability is considered.

Inference: There is evidence of a greater level of inconsistency than with the within–country orderings: how much is sufficient to judge that the SWB orderings when applied within–years, contain different information to RP orderings? As before, to answer this, I consider how far these results differ from a comparison of the RP orderings to uncorrelated SWB orderings. The mean (bootstrapped)

	Kendall	n
1975	−0.180	10
1980	0.085	9
1985	0.412	12
1996	0.269	16
2005	0.398	27
2011	0.618	23

Table 3.8: Within–year Kendall results

Notes: Kendall τ scores for each year. $\tau = 1$ indicates perfect consistency between the RP and SWB orderings; $\tau = -1$ indicates perfectly inverted orderings.

Kendall is 0.316 and the 95% confidence interval [0.170 — 0.461] does not contain 0, and neither does the 99% [0.087 — 0.544]. As with the within–country results, the null hypothesis that the correlation between the RP orderings and the true SWB orderings is no stronger than the RP's correlation with randomly generated SWB data is rejected. However, the second interval is much closer to 0 than it was for the within–country results.

To compare the distribution of Kendall results across the years, I conduct KS tests as before. The KS results (two–sided and one–sided "less") both reject at a 5% significance level the hypothesis of the Kendall results being drawn from the same distribution as the random benchmark, but fail to reject at a 1% significance: the difference is 0.833 ($p < 0.05$ for the one and two sided tests).

Overall, when compared to the within–country analysis, the within–year Kendall results are somewhat less convincing that there is strong correlation between the RP and SWB orderings: although a t–test rejected the hypothesis of no more correlation than with random orderings, the KS test did not reject at a 1% significance level the test of the true and simulated results coming from the same distribution.

3.5 | Discussion

The key contribution of this chapter is the presentation of a method for interrogating ordinal differences in well–being measures. I derived RP orderings from consumption data and compared these to SWB orderings: measuring the difference between ranked years (within–country) and countries (within–year) using the Kendall τ correlation coefficient. The results when applied to the ICP and WDH dataset are mixed: Result (1) Within–country, SWB and RP orderings are consistent with each other, and the RP orderings alone pass a test for consistency for all countries; Result (2) Within–year, the SWB and RP orderings are less consistent with each other, but the RP orderings alone pass a test for homogeneity across countries. Why might this be the case?

Result (1), when considered alongside the fact that all years passed the GARP cyclical consistency test, suggest that measurement error or ill–application of theory to the consumption data cannot explain the inconsistency in within–year ordinal rankings in Result (2). Behavioural arguments related to the SWB data might suggest that people make mistakes as they report SWB (connected to the critiques of SWB, outlined earlier in this chapter), hence resulting in a discrepancy between SWB orderings and RP orderings. However, I use macro datasets, and therefore any SWB reporting mistakes would be ironed out in the aggregation procedure, provided they are not systematic within the population. And, as above, Result (1) provides evidence that the SWB data are consistent with the consumption data for the first set of within–country comparisons, and therefore less likely to contain systematic mistakes.

The multi–domain aspect of SWB could provide a hypothesis to explain ordinal inconsistencies. A finding in Chapter 1 of this thesis, that satisfaction with income is not highly weighted in individuals' reports of overall SWB, could explain divergence of RP and SWB orderings because SWB orderings take into account more than just expenditure–related well–being. If an alternative domain dominates in the SWB report, the ordinal ranking of one state over another may be altered from the RP ranking derived only using consumption data. However, the within–country orderings are consistent with each other: Result (1) suggests we must look further than domain–differences to understand Result (2).

A viable explanation could be related to differences in the cultural interpretation of the SWB scale: although each country's representative agent interprets the satisfaction scale consistently overtime (and this matches their consumption choices, derived from consistent preferences over–time), the SWB report contains level differences across countries, rendering SWB reports between countries non–comparable (despite the evidence from the consumption data of homogenous preferences). The results

of this chapter cannot prove this as an explanation, but are not inconsistent with it. To test this, SWB data could be de–meaned for each country, and then the changes used to construct orderings. This would require more observations.

Two key issues with the application of the method to the dataset in this chapter are the sample size and the aggregation of consumption and SWB to a representative agent. The purpose of the application to the data used is to demonstrate how the method works. To strengthen the findings in future research, application of the method to other datasets would be informative:

- A longer macroeconomic panel containing both consumption data and SWB data would be helpful, in order to have greater power to reject both consistency of RP orderings alone and consistency of the RP orderings with SWB orderings. For instance, yearly consumption data would be helpful (large shifts in budget sets over a 6–yearly period can result in the nesting of countries' budget sets within each other). This might also enable analysis of changes in SWB, rather than levels, to interrogate the hypothesis that cultural SWB reporting differs between countries. The Eurobarometer, World Values Survey or Gallup World Poll could offer alternative sources of SWB data that expand the number of observations for countries for whom the WDH had limited numbers.

- A macroeconomic panel with alternative measures of well–being, such as observable measures (e.g. cortisol levels, measuring stress, or the Human Development Index) would provide additional rankings for comparison against the RP and SWB rankings compared in this chapter.

- A microeconomic panel would be a natural dataset in which to apply this method to consumption and SWB data. This would avoid requiring any assumptions regarding individual preferences being in the Gorman polar form (the existence of the representative consumer), strengthen the power of RP tests (smaller budget set changes relative to price changes) and the assumption of SWB inter–personal comparability.

In addition to the use of alternative datasets, it would also be of interest to interrogate the size of the ordinal discrepancy: one could consider running regressions of country or time period characteristics on the Kendall τ. This could enable analysis such as how within–country inequality or national income could affect the level of mismatch between RP and SWB orderings. This could complement the results in Chapter 1, regarding the changing level of importance attached to satisfaction with income in overall well–being as income and other life circumstances change.

Note that Result (1) is inconsistent with the Easterlin paradox's finding of no association of income and happiness within–countries over–time, instead it falls in line with Stevenson & Wolfers (2008).

Result (2) is, however, consistent with the Easterlin finding: no / weak association between happiness and income cross–country. Clark, Frijters and Shields (2007) use econometric techniques to respond to the controversy in macro–economics between Easterlin defendants and combatants by including *relative income* in the utility function. In a similar vein, I propose two behavioural extensions to the method in this chapter, listed below:

(1) Happiness habits (adjusting happiness levels for effects resulting from lagged happiness levels). Clark et al. (2008) employ regression techniques to test for lags and leads in German SWB data and find significant evidence for this. Using non–parametric tests for evidence of habituation in both consumption and SWB data, and then measuring the differences between adjusted orderings, could provide an interesting avenue for future research.[33]

(2) Peer effects on SWB (adjusting SWB levels as the surrounding SWB changes). This would involve designing and conducting RP tests for dynamic reference points. One could then analyse differences in the adjusted orderings. This would connect to some of the experimental aspects in Chapter 2 of this thesis (and wider behavioural literature on endogeneous beliefs and evaluations about the self): we investigate and quantify the extent to which what other people say affects our actions and also our beliefs about ourselves and try to quantify the welfare effect of this.

[33]RP tests using linear programs for the existence of rational habit formation are provided in: Crawford (2010), which develops the revealed preference conditions for the Short Memory Habits model and in Demuynck and Verriest (2013), which presents a RP characterization of the Habits as Durables model.

Thesis Conclusion

The core contribution of Chapter 1 is to introduce an alternative way to handle SWB data, that avoids the inter–personal comparability assumption. The symmetry of the ME results with much of the work looking at the determinants of SWB that does make this assumption speaks to a helpful output for this work: it is reassuring that the two very different methods (ME and regression) extract similar information from the overall and domain satisfaction reports. It suggests the damage of making the assumption, even if not true, is not substantial. This finding has significance for the data used in Chapter 3: I use SWB indicators that are averages across people.

Chapter 1 finds that the weightings attached to different domains of life vary across groups of people: I find evidence for 6 natural groupings, or clusters, matched according to common patterns in which domains matter most for overall satisfaction. This speaks to one interpretation of the differences between RP and SWB orderings for countries in Chapter 3: cultural differences in which domains of life are highly weighted for individuals in each nation may result in national differences in what the overall SWB levels represent, and hence incomparability of SWB levels across countries. This would effect consistency with RP orderings.

Additionally, while the price and quantity inputs used in Chapter 3 to derive the RP orderings have been rendered internationally comparable, they are restricted to the domain of consumption. It might be interesting to interrogate the size of the RP and SWB ordinal discrepancy measure, Kendall τ: one could consider running regressions of country or time period characteristics on the τ statistic, to understand how national differences in other domains (for instance, health, housing, education) could affect the size of the mismatch between RP and SWB orderings. This would complement the comparisons of domain weightings by individual and household characteristics in Chapter 1.

Chapter 2, in investigating the welfare impacts of the challenging treatment (both for those successful and unsuccessful in meeting their goals), uses both behavioural measures of pride, such as donation to charity, and self–evaluated reports of pride. Looking at the correlations between these revealed and reported measures connects the analysis to Chapter 3: how similar are RP and SWB orderings in

terms of representing well–being. In Chapter 2's analysis, although we find correlations, these survey tools may be measuring slightly different feelings, and each will be subject to some measurement error. This motivates the inclusion of additional welfare measurement survey tools in the planned replication. And likewise, alternative well–being related data orderings for Chapter 3 could be of interest: a macroeconomic panel with alternative measures of well–being, such as observable measures (e.g. cortisol levels, measuring stress, or the Human Development Index) would provide additional rankings for comparison against the RP and SWB rankings.

There are two immediate next steps motivated by this thesis' findings. Firstly, Chapter 1 offers a richness of output from the ME procedure that is unparalleled in approaches using regression: a per–person estimation of what matters to them for their overall life satisfaction. This thesis chapter focuses on the explanation, justification and implementation of the ME procedure on SWB panel data, and then offers simple descriptives of the results. The clear extension of the research would include deeper analysis of the weightings in order to speak to policy recommendations of what types of welfare interventions would be most appreciated (and offer the best return–to–well–being) by which groups in society. Secondly, the findings in Chapter 2 are similarly preliminary: the scope to investigate heterogeneity in both behavioural outcomes (goal achievement) and psychological mechanisms is huge, and motivates our proposed replication of the study with a larger sample size.

References

Abramovitz, Moses (1959). "The welfare interpretation of secular trends in national income and product." In: *The allocation of economic resources: essays in honor of Bernard Francis Haley* pp.1–22.

Acemoglu, Daron as quoted in: Stevenson, Betsy, and Justin Wolfers (2008). "Economic Growth and Subjective Well–Being: Reassessing the Easterlin Paradox." In: *Brooking Papers on Economic Activity* 39.1(Spring), pp.1–102.

Afriat, S. N. (1967). "The Construction of Utility Functions from Expenditure Data." In: *International Economic Review* 8.1, pp.67–77.

Allen, Eric J. , Patricia M. Dechow, Devin G. Pope, George Wu (2017). "Reference-Dependent Preferences: Evidence from Marathon Runners." In: *Management Science*, 63.6, pp.1657–1672.

Andrews, F. M., and S. B. Withey (1976). "Social indicators of well–being: The development and measurement of perceptual indicators." *New York: Plenum*.

Balk, B. M. (2009). "Aggregation Methods in International Comparisons: An Evaluation." In: *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, ed. D. S. Prasada Rao, 59–95. Cheltenham, UK: Edward Elgar.

Banks, James, Richard Blundell and Arthur Lewbel (1997). "Quadratic Engel Curves and Consumer Demand." In: *The Review of Economics and Statistics* 79.4, pp.527–539.

Becker, Gary S. (1962), "Irrational Behavior and Economic Theory." In: *Journal of Political Economy* 70.1, pp.1–13.

Bénabou, R. and J. Tirole (2002). "Self-confidence and personal motivation." In: *Quarterly Journal of Economics*, 117.3, pp.871–915.

——— (2003). "Intrinsic and Extrinsic Motivation." In: *The Review of Economic Studies*, 70.3, pp.489–520.

——— (2005). "Self-Confidence and Personal Motivation." In book: *Psychology, Rationality and Economic Behaviour*, pp.19–57.

——— (2006). "Incentives and Prosocial Behavior." In: *American Economic Review*, 96.5, pp.1652—678.

——— (2011). "Identity, Morals, and Taboos: Beliefs as Assets." In: *Quarterly Journal of Economics* 126.2, pp.805–855.

Benjamin, D., O. Heffetz, M. Kimball, and Alex Rees–Jones (2012). "What do you think would make you happier, What do you think you would choose." In: *American Economic Review* 102(5), pp.2083–2110.

——— (2014). "Can Marginal Rates of Substitution be Inferred from Happiness Data? Evidence from Residency Choices." In: *American Economic Review* 104.11, pp.3498–3528.

Benjamin, D., O. Heffetz, M. Kimball, and N. Szembrot (2014). "Beyond Happiness and Satisfaction: Toward Well–Being Indices Based On Stated Preference." In: *American Economic Review* 104.9, pp.2698–2735.

Benjamin, Daniel J., K, Cooper, O, Heffetz, and M. S. Kimball (2020). "Self-reported wellbeing indicators are a valuable complement to traditional economic indicators but aren't yet ready to compete with them." In: *Behavioural Public Policy* 4.2, pp.198–209.

Bentham, Jeremy (1843). "The works of Jeremy Bentham / published under the superintendence of his executor, John Bowring." *Edinburgh: W. Tait.*.

Bessone, Pedro, Gautam Rao, Frank Schilbach, Heather Schofield and Mattie Toma (2021). "The Economic Consequences of Increasing Sleep Among the Urban Poor." In: *The Quarterly Journal of Economics*, 136.3, pp.1887–1941.

Binder, Martin and Alex Coad (2011). "From Average Joe's happiness to Miserable Jane and Cheerful John: using quantile regressions to analyze the full subjective well–being distribution." In: *Journal of Economic Behavior & Organization* 79, pp.275–290.

Blundell, Richard W., Martin Browning and Ian A. Crawford (2003). "Nonparametric Engel Curves and Revealed Preference." In: *Econometrica* 71.1, pp.205–240.

Blundell, R., Ian Preston and Ian Walker (2009). "The Measurement of Household Welfare." *Cambridge University Press*.

Boarini, Romina, Margherita Comola, Conal Smith, Robert Manchin and Femke de Keulenaer (2012). "What Makes for a Better Life? The Determinants of Subjective Well–Being in OECD Countries — Evidence from the Gallup World Poll". In: *OECD Statistics Working Papers*, 2012/03, OECD Publishing.

Bond, Timothy and Kevin Lang (2019). "The Sad Truth about Happiness Scales." In: *The Journal of Political Economy* 127.4, pp.1629–1640.

Bonett, Douglas G. and Thomas A. Wright (2000). "Sample size requirements for estimating Pearson, Kendall and Spearman correlations." In: *Psychometrika* 65.1, pp.23–28.

Broome, John (1991). "Utility." In: *Economics and Philosophy* 7.1, pp.1–12.

Bursztyn, L., M. Callen, B. Ferman, S. Gulzar, A. Hasanain, and N. Yuchtman (2020). "Political Identity: Experimental Evidence on Anti–Americanism in Pakistan." In: *Journal of the European Economic Association*, Forthcoming.

Bursztyn, L., B. Ferman, S. Fiorin, M. Kanz, and G. Rao (2018). "Status Goods: Experimental Evidence from Platinum Credit Cards." In: *Quarterly Journal of Economics*, 133.3, pp.1561–1595.

Bursztyn, L. and R. Jensen (2017). "Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure." In: *Annual Review of Economics*, 9, pp.131–153.

Campello, Murillo, A. F. Galvao and T. Juhl (2019). "Testing for Slope Heterogeneity Bias in Panel Data Models." In *Journal of Business and Economic Statistics* 37.4, pp.749–760.

Carrillo, Juan and Thomas Mariotti (2000). "Strategic Ignorance as a Self-Disciplining Device." In: *Review of Economic Studies*, 67.3, pp.529–544.

Caves, Douglas W., Laurits R. Christensen, and W. Erwin Diewert (1982). "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity." In: *Econometrica* 50.6, pp.1393–1414.

Cheng, Joey, J. Tracy and J. Henrich (2010). "Pride, Personality, and the Evolutionary Foundations of Human Social Status." In: *Evolution and Human Behavior*, 31.

Cherchye, Laurens, I. Crawford, B. Dr Rock, and Frederic Vermeulen (2011). "Aggregation without the aggravation? Nonparametric analysis of the representative consumer." In: *IZA Discussion Paper* No. 6266.

Clark, A. E., Diener, E., Georgellis, Y., and Lucas, R. E. (2008). "Lags And Leads in Life Satisfaction: a Test of the Baseline Hypothesis." In: *Economic Journal* 118.529, pp.222–243.

Clark, A.E., Fabrice Etilé, Fabien Postel–Vinay, Claudia Senik and Karine Van der Straeten (2005). "Heterogeneity in Reported Well–Being: Evidence from Twelve European Countries." In: *The Economic Journal* 115.502, pp.C118–C132.

Clark, A.E., Paul Frijters, Christian Krekel and Richard Layard (2019). "A Happy Choice: Well–Being as the Goal of Government." Centre for Economic Performance, Discussion Papers dp1658.

Clark, A.E., P. Frijters and M. Shields (2007). "Relative income, happiness and utility: an explanation for the Easterlin paradox and other puzzles." In: *Journal of Economic Literature* 46.1.

Clark, A.E., Richard Layard and Claudia Senik (2012). "The causes of happiness and misery." Post-Print, HAL. *Available at: https://EconPapers.repec.org/RePEc:hal:journl:halshs-00846583*.

Crawford, Ian (2010). "Habits revealed." In: *Review of Economic Studies*, 77.4, pp.1382–1402.

Cummins, Robert A. (1996). "The Domains of Life Satisfaction: An Attempt to Order Chaos." In: *Social Indicators Research* 38, pp.303–328.

Curd, Martin and Jan Cover (1998). "Philosophy of Science: The Central Issues." *Norton*.

Deaton, Angus (2008). "Income, Health, and Well–Being around the World: Evidence from the Gallup World Poll." In: *The Journal of Economic Perspectives* 22.2, pp.53–72.

Demuynck, T. and Ewout Verriest (2013). "I'll Never Forget My First Cigarette: A Revealed Preference Analysis Of The "Habits As Durables" Model." In: *International Economic Review* 54.2, pp.717–738.

De Neve, J.–E. (2011). "Functional polymorphism (5–HTTLPR) in the serotonin transporter gene is associated with subjective well–being: Evidence from a U.S. nationally representative sample." In: *Journal of Human Genetics*, 56, pp.456–459.

Dessi, Roberta and Aldo Rustichini (2015). "Strong intrinsic motivation." in: *IDEI Working Papers* 847.

Diener, E. (1984). "Subjective Well–Being." In: *Psychological bulletin* 95.3, pp.524–75.

——— (2013). "Remarkable Changes in the Science of Subjective Well–Being." In: *Perspectives on Psychological Science* 8.6, pp.663–666.

Diener, E., Robert A. Emmons, Randy J. Larsen and Sharon Griffen (1985). "The Satisfaction with Life Scale." In: *Journal of Personality Assessment* 49, pp.71–75.

Diener, E., Samantha Heintzelman, Kostadin Kushlev, Lesley Lutes, Shigehiro Oishi, Louis Tay and

Derrick Wirts (2017). "Findings All Psychologists Should Know From the New Science of Subjective Well–Being." In: *Canadian Pyschology* 58.2, pp.87–104.

Diener, E., R. Inglehart and L. Tay (2013). "Theory and validity of Life Satisfaction Scales." In: *Social Indicators Research* 112.3, pp.497–527.

Dolan, Paul and Robert Metcalfe (2010). "Oops...I did it again': Repeated focusing effects in reports of happiness." In: *Journal of Economic Psychology* 31, pp.732–737.

Dolan, Paul, Tessa Peasgood and Mathew White (2008). "Do we really know what makes us happy A review of the economic literature on the factors associated with subjective well-being." In: *Journal of Economic Psychology* 29.1, pp.94–122.

Dolan, Paul, and Mathew White (2007). "How Can Measures of Subjective Well–Being Be Used to Inform Public Policy?" In: *Perspectives in Psychological Science* 2.1, pp.71–85.

Dunn, J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters." In: *Journal of Cybernetics*. 3.3, pp.32–57.

Easterlin, Richard A. and Onnicha Sawangfa (2007). "Happiness and Domain Satisfaction: Theory and Evidence." IZA Discussion Paper no. 2584.

Ellison, C. G., D. A. Gay and T. A. Glass (1989). "Does religious commitment contribute to individual life satisfaction?" In: *Social Forces*, 68, pp.100-123.

Fabian, Mark (2018). "Scale-Norming Is a Concern for Life Satisfaction Research: Evidence From a New Metric." *Available at: https://ssrn.com/abstract=3245058*.

Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer (2015). "The Next Generation of the Penn World Table." In: *The American Economic Review*, 105.10, pp.3150-182.

Ferrer–i–Carbonell, A. and P. Frijters (2004). "How Important is Methodology for the estimates of the determinants of Happiness?" In: *The Economic Journal*, 114, pp.641-659.

Fleurbaey, M., E. Schokkaert and K. Decancq (2009). "What good is happiness?" In: *ECORE Discussion Paper*, 17.

Frederick, S. and G. Loewenstein (1999). "Hedonic Adaptation." In: *Well-Being: The Foundations of Hedonic Psychology*, ed. by D. Kahneman, N. Schwartz and E. Diener, New York, NY: Russell Sage.

Gneezy, U., S. Meier, and P. Rey-Biel (2011). "When and Why Incentives (Don't) Work to Modify Behavior." In: *Journal of Economic Perspectives*, 25.4, pp.191–210.

Gneezy, Uri and Aldo Rustichini (2000). "Pay Enough or Don't Pay at All." In: *The Quarterly Journal of Economics*, 115.3, pp. 791–810.

Golan, Amos, (2006). "Information and Entropy Econometrics — A Review and Synthesis." in *Econometrics*, 2.1–2, pp.1–145.

Golan, Amos, G. Judge and L. Karp (1996). "A maximum entropy approach to estimation and inference in dynamic models *or* Counting fish in the sea using maximum entropy." In: *Journal of Economics Dynamics and Control* 20.4, pp.559–582.

Golan, Amos, G. Judge and D. Miller (1996). "Maximum Entropy Econometrics: Robust Estimation with Limited Data." *Wiley*.

Goldsmith, Glenn (2009). "Essays in the economics of subjective well–being." *Doctoral Thesis at the University of Oxford*.

Golman, Russell, David Hagmann, and George Loewenstein (2017). "Information Avoidance." In: *Journal of Economic Literature*, 55.1, pp.96–135.

Gordon, A. D. (1999). "Classification, 2nd ed." New York: *Chapman & Hall*.

Gorman, W. M. (1961). "On a class of preference fields." In: *Metroeconomica*, 13, pp.53–56.

de Haan, Jan, Leon Willenborg and Antonio G. Chessa (2016). "An Overview of Price Index Methods for Scanner Data."

Hamming, Richard (1950). "Error detecting and error correcting codes." In: *The Bell system technical journal* 29, pp.147–160.

Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A k-Means Clustering Algorithm." In: *Journal of the Royal Statistical Society* Series C. 28 (1), pp.100–108.

Haybron, Daniel M. (2013). "Happiness: A Very Short Introduction." *Oxford University Press*.

Heady, B., R. Veenhoven and A. Wearing (1991). "Top-down versus bottom-up theories of subjective well-being." In: *Social Indicators Research*, 24, pp.81-100.

Hicks, J. R. and R. G. D. Allen (1934). "A Reconsideration of the Theory of Value." Parts I and II. In: *Economica* 1.1 pp.52–76, 1.2 pp.196–219.

Hills, John (2001). "Indicators of progress: A discussion of approaches to monitor the Government's strategy to tackle poverty and social exclusion." In: *Open Access publications from London School of Economics and Political Science*.

Hsiao, C. and M. H. Peseran (2008). "Random Coefficient Models." In *Advanced Studies in Theoretical and Applied Econometrics, Vol. 46, The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice,* ed. by L Matyas and P. Sevestre. Springer–Verlag, Berlin.

Jaynes, Edwin Thompson (1957a). "Information Theory and Statistical Mechanics." In: *Physical Review* 106–620.

——— (1957b). "Information Theory and Statistical Mechanics II." In: *Physical Review* 108–171.

Kahneman, D., E. Diener, E. and N. Schwarz (1999). "Well-being: The foundations of hedonic psychology." *Russell Sage Foundation, New York*.

Kahneman, D., A. B. Kruegar, D. Schkade, N. Schwarz and A. Stone (2004). "Towards national well–being accounts." In: *American Economic Review* 94.2, pp.429–434.

Kahneman, D. and A. Tversky (2000). "Choices, Values and Frames." *Cambridge: Cambridge University Press*.

Kaiser, Caspar, and Maarten C. M. Vendrik (2019). "How Threatening Are Transformations of Happiness Scales to Subjective Wellbeing Research?" *Available at: https://doi.org/10.31235/osf.io/gzt7a*.

Kendall, M. G. (1938). "A New Measure of Rank Correlation." In: *Biometrika* 30.1, pp.81–93.

Kirman, Alan P. (1992). "Whom or What Does the Representative Individual Represent?" In: *Journal of Economic Perspectives* 6.2 pp.117–136.

Krawczyk, M. and Wilamowski W. (2017). "Are we all overconfident in the Long Run? Evidence from One Million Marathon Participants." In: *Journal of Behavioral Decision Making*, 30, pp.719–730.

Krueger, Alan B. and David A. Schkade (2007). "The Reliability of Subjective Well–Being Measures." In: *Princeton University Working Papers* 64.

Larsen, Randy, Ed Diener and Robert Emmons (1985). "An evaluation of subjective well–being measures." In: *Social Indicators Research* 17.1, pp.1–17.

Levenshtein, V. I. (1966). "Binary Codes capable of correcting deletions, insertions and reversals." In: *Soviet Physics Doklady* 10, pp.707–710.

Levy, H., E. Bardasi, S. Jenkins, F. Zantomio and H. Sutherland (2012) *British Household Panel Survey Derived Current and Annual Net Household Income Variables, Waves 1–18, 1991–2009*. [data collection]. 9th *Edition*. University of Essex, Institute for Social and Economic Research, [original data producer(s).] University of Essex, Institute for Social and Economic Research. SN: 3909.

Loewenstein, G. and S. Issacharoff (1994). "Source dependence in the valuation of objects." In: *Journal of Behavioural Decision Making*, 7, pp.157–168.

van der Loo, Mark P. J. (2014). "The stringdist Package for Approximate String Matching." In: *The R Journal* 6.1, pp.111–122.

Lykken, D., & Tellegen, A. (1996). "Happiness is a stochastic phonomenon." In: *Psychological Science* 7, pp.186–189.

Mas–Colell, Andreu, M. D. Whinston, J. R. Green (1995). "Microeconomic Theory." *Oxford University Press*.

Macchia, Lucia and Ashley V. Whillans (2019). "Leisure beliefs and the subjective well–being of nations." In: *The Journal of Positive Psychology* pp.1–9.

Markle, A., G. Wu, R. White and A. Sackett (2018). "Goals as reference points in marathon running: A novel test of reference dependence." In: *Journal of Risk and Uncertainty*, 56, pp.19–50.

McNeish, Daniel (2015). "Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences." Multivariate Behavioral Research. In Press.

Micklewright, John (2001). "Should the UK government measure poverty and social exclusion with a composite index?" Manuscript, UNICEF Innocenti Research Centre, Florence.

Milligan, G. and M. Cooper (1987). "Methodology Review: Clustering Methods." In: *Applied Psychological Measurement* 11.4, pp.329–354.

Mittelhammer, Ron C., George G. Judge, and Douglas Miller (2000). *Econometric foundations, Volume 1*, Cambridge University Press.

Neary, Peter (2004). "Rationalizing the Penn World Table: True multilateral indices for international comparisons of real income." In: *American Economic Review* 94.5, pp.1411–1428 .

Petersen, Michael J. Casey M. Rowe and M. Andersen (2016). "What Factors Affect the Valuation of Reward Objects?" In: *Management Accounting Section Meeting*.

Plato. (1953) "The dialogues of Plato." *Oxford: Oxford University Press*. Translated and edited by Benjamin Jowett.

Rao, D. S. Prasada (2009). "Generalised Elteto–Koves–Szulc (EKS) and Country–Product–Dummy

(CPD) Methods for International Comparisons." In: *Purchasing Power Parities of Currencies: Recent Advances in Methods and Applications*, ed. D. S. Prasado Rao, 86–120. Cheltenham, UK: Edward Elgar.

Rothschild, K. W. (1961). "Review: The Allocation of Resources: Essays in Honor of Bernard Francis Haley." In: *Econometrica* 29.3, pp.490–492.

Samuelson, P. A. (1938). "A Note on the Pure Theory of Consumer's Behaviour." In: *Economica* 5.17, pp.61–71.

Schwarz, Norbert, and Gerald L. Clore (1983). "Mood, Misattribution, and Judgements of Well–Being: Informative and Directive Functions of Affective States." In: *Journal of Personality and Social Psychology* 45.3, pp.513–523.

Schweyer, Allan, A. Landry and A. Whillans (2018) at the Incentive Research Foundation. "Non Cash Rewards." In: *IRF: Award Value Program and Evidence*.

Sen, Amartya (1979). "Personal Utilities and Public Judgements: Or What's Wrong with Welfare Economics?" In: *Economic Journal* 89.355, pp.537–58.

——— (1985). "Well–Being, Agency and Freedom: The Dewey Lectures 1984" In: *The Journal of Philosophy* 84.4, pp.169–221.

Senik, C. (2005). "Income distribution and well-being: what can we learn from subjective data?" In: *Journal of Economic Surveys*, 19, pp.43–63.

Shannon, Claude E. (1948). "A Mathematical Theory of Communication." In: *Bell System Technical Journal* 27, pp.379–423.

Seidenfeld, Teddy (1986). "Entropy and Uncertainty." In: *Philosophy of Science*, 53.467.

Snyder, Christopher, and Walter Nicholson (2012). "Microeconomic theory: basic principles and extensions." *South–Western Cengage Learning*.

Stevenson, Betsy, and Justin Wolfers (2008). "Economic Growth and Subjective Well–Being: Reassessing the Easterlin Paradox." In: *Brooking Papers on Economic Activity* 39.1, pp.1–102.

Stiglitz, J.E., A. Sen, J–P. Fitoussi (2009). "Report by the Commission on the Measurement of Economic Performance and Social Progress." In: *Institut national de la statistique et des etudes economiques* 291pp.

Taylor, Marcia Freed (ed). with John Brice, Nick Buck and Elaine Prentice-Lane (2018). "British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices." Colchester: University of Essex.

Tibshirani, Robert, G. Walther and T. Hastie (2001). "Estimating the number of clusters in a dataset via the gap statistic." In: *J. R. Statistical Society B*, 63.2, pp.411–423.

Tracy, J. L., and R. W. Robins (2007). "The psychological structure of pride: A tale of two facets." In: *Journal of Personality and Social Psychology*, 92.3, pp.506–525.

University of Essex, Institute for Social and Economic Research. (2019). *Harmonised BHPS: Waves 1-18, 1991-2009*. [data collection]. *12th Edition*. UK Data Service. SN: 6614.

Ura, Karma (2008). "Understanding the Development Philosophy of Gross National Happiness." *Interview with Bhutan Broadcasting Service*.

van der Weele, Joël J, and Ferdinand A. Von Siemens (2020). "Bracelets of Pride and Guilt? An Experimental Test of Self-signaling." In: *Journal of Economic Behavior & Organization*, 172, pp.280–91.

van Praag, B.M.S., P. Frijters and A. Ferrer–i–Carbonell (2003). "The anatomy of subjective well–being." In: *Journal of Economic Behaviour and Organisation* 51, pp.29–49.

Varian, Hal (1978). "Microeconomic Analysis." *W W Norton & Company.*

——— (1982). "The Nonparametric Approach to Demand Analysis." In: *Econometrica* 50.4, pp.945–73.

——— (2010). "Intermediate microeconomics: a modern approach." *W W Norton & Company.*

Veenhoven R. (2016). "World Database of Happiness." *http://worlddatabaseofhappiness.eur.nl*

Westfall, Peter H. and S. Stanley Young (1993). "Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment." *New York: John Wiley & Sons.*

Williams, L., and D. DeSteno, (2008) "Pride and Perseverance: The Motivational Role of Pride." In: *Journal of Personality and Social Psychology* 94, pp.1007–1017.

Wilson, W. (1967). "Correlates of Avowed Happiness." In: *Psychological Bulletin* 67, pp.294–306.

Winkelmann, L. and R. Winkelmann, (1998). "Why Are the Unemployed So Unhappy? Evidence from Panel Data." In: *Economica*, 65, pp.1–15.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data, 2nd edition.* The MIT Press, Cambridge, Maryland.

World Bank (2013). "Measuring the Real Size of the World Economy: The Framework, Methodology, and Results of the International Comparison Program — ICP." *Washington, DC: World Bank.*

Zach, S., Y. Xia, A. Zeeva, M. Arnona, N. Choresha, and G. Tenenbaumb (2017). "Motivation dimensions for running a marathon: A new model emerging from the Motivation of Marathon Scale (MOMS)." In: *Journal of Sport and Health Science*, 6.3, pp.302–310.

Jakina Debnam's Website. (no date) *http://www.jakinadebnam.com/research-2*. Viewed March 2021.

No Author. (2016) *http://siteresources.worldbank.org/ICPEXT/Resources/ICP_2011.html*. Viewed August 2020.

No Author. (2016) *http://www.levenshtein.net/*. Viewed August 2020.

No Author. (2016) *https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2016* (ONS Online Families and Households 2016 bulletin). Viewed September 2019.

No Author. (2019) *https://webarchive.nationalarchives.gov.uk/20160105214120/http://www.ons.gov.uk/ons/rel/ghs/general-lifestyle-survey/2011/rpt-chapter-3.html* (ONS Online 2011 General Lifestyle Survey results). Viewed September 2019.

No Author. (2019) *https://www.oecd-ilibrary.org/sites/9870c393-en/index.html?itemId=/content/publication/9870c393-en* (OECD How's Life? 2020: Measuring Well–being).

Viewed September 2019.

No Author. (2019) *https://www.oecd-ilibrary.org/sites/9870c393-en/index.html?itemId=/content/publication/9870c393-en* (OECD How's Life? 2020: Measuring Well–being). Viewed September 2019.

No Author. (2019) *https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing* (ONS Online Well–being page). Viewed August 2021.

Jens Jakob Andersen. (2021) *https://runrepeat.com/state-of-running* (The State of Running 2019). Viewed March 2021.

Appendices

Chapter 1, Appendix A: Robustness to assumption for missing domains

Click here to return to Chapter 1: Main Results

	Mean (Rank)	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Health	0.146 (2)	0.149	0	0.037	0.197	1
Income	0.100 (7)	0.115	0	0.015	0.139	1
House	0.129 (4)	0.134	0	0.029	0.176	1
Partner	0.168 (1)	0.163	0	0.039	0.244	0.992
Job	0.103 (6)	0.118	0	0.0100	0.100	1
Social	0.134 (3)	0.140	0	0.036	0.174	1
Leisure Amount	0.097 (8)	0.109	0	0.015	0.137	0.965
Leisure Way	0.123 (5)	0.130	0	0.030	0.164	1

Table 9: Mean domain weightings — no missings

Notes: These are the summary statistics of the weightings, excluding individuals who had a domain satisfaction missing (n=12,925). In brackets is the ranked importance of the domain, which is simply the size order of the mean weighting.

	Mean Weighting (Rank)					
	(1) Men	(2) Women	(3) 15–26 yrs	(4) 26–39 yrs	(5) 40–56 yrs	(6) 56–97 yrs
Health	0.144 (2)	0.149 (2)	0.167 (1)	0.156 (2)	0.137 (2)	0.126 (5)
Income	0.103 (7)	0.098 (6)	0.096 (8)	0.102 (7)	0.105 (7)	0.099 (7)
House	0.127 (4)	0.131 (4)	0.132 (4)	0.119 (4)	0.126 (3)	0.140 (2)
Partner	0.157 (1)	0.178 (1)	0.132 (3)	0.199 (1)	0.177 (1)	0.162 (1)
Job	0.111 (6)	0.095 (7)	0.096 (6)	0.113 (5)	0.115 (6)	0.088 (8)
Social	0.133 (3)	0.135 (3)	0.157 (2)	0.121 (3)	0.122 (5)	0.135 (4)
Amount	0.099 (8)	0.094 (8)	0.096 (7)	0.084 (8)	0.095 (8)	0.112 (6)
Way	0.126 (5)	0.119 (5)	0.123 (5)	0.106 (6)	0.124 (4)	0.138 (3)

Table 10: Comparison of mean domain weightings by gender and age quartile (no missings)

Notes: These are the mean weightings for each of the domains, averaged for male and females (columns (1) and (2)), and for each age quartile (columns (3) to (6)), for the subsample of individuals who had no missing domain satisfaction (n=12,925). In brackets is the ranked importance of the domain, which is simply the size order of the mean weighting.

Chapter 1, Appendix B: Investigating the weights — additional tables

(1.B1) OLS regressions of weights on covariates

Click here to return to Chapter 1: Results on weighting correlates

	Health (1)	Health (2)	Health (3)	House (1)	House (2)	House (3)	Job (1)	Job (2)	Job (3)
Female (=1)	0.002		0.000	0.008***		0.007***	-0.008***		-0.008***
	(1.07)		(0.12)	(4.30)		(3.40)	(-6.16)		(-5.27)
Age (years)	-0.000		-0.001*	0.000		-0.000	0.001**		0.001**
	(-0.72)		(-1.96)	(0.49)		(-0.08)	(3.19)		(3.09)
Age Squared	-0.000*		-0.000	0.000		0.000	-0.000***		-0.000***
	(-2.25)		(-1.10)	(1.80)		(1.81)	(-4.33)		(-4.14)
Working (=1)	-0.016***		-0.014***	-0.016***		-0.013***	0.066***		0.065***
	(-7.03)		(-5.61)	(-7.12)		(-5.56)	(41.11)		(36.45)
Retired (=1)	-0.005		-0.001	-0.004		-0.001	-0.020***		-0.021***
	(-1.18)		(-0.23)	(-1.14)		(-0.16)	(-7.14)		(-7.10)
Married (=1)	-0.016***		-0.010*	-0.024***		-0.017***	-0.007**		-0.005
	(-5.20)		(-2.33)	(-7.91)		(-4.31)	(-3.24)		(-1.82)
Prev. married (=1)	0.004		-0.002	-0.001		-0.002	-0.003		-0.003
	(1.29)		(-0.49)	(-0.31)		(-0.64)	(-1.25)		(-1.23)
Children (=1)		0.021***	0.014***		-0.011***	0.002		0.025***	0.001
		(8.33)	(4.90)		(-4.59)	(0.82)		(13.96)	(0.32)
Single hh (=1)		-0.001	0.026***		0.031***	0.015**		-0.019***	0.004
		(-0.23)	(4.70)		(6.22)	(2.86)		(-4.77)	(1.15)
Couple (=1)		-0.029***	-0.002		0.003	0.002		-0.010**	0.001
		(-5.68)	(-0.41)		(0.69)	(0.49)		(-2.66)	(0.15)
Hh income (£)		-0.000**	-0.000		-0.000**	-0.000		0.000***	-0.000
		(-3.12)	(-1.88)		(-2.70)	(-0.43)		(9.88)	(-0.78)
Constant	0.187***	0.166***	0.193***	0.135***	0.137***	0.132***	0.056***	0.077***	0.054***
	(29.63)	(33.44)	(23.56)	(22.83)	(29.45)	(17.17)	(12.91)	(20.96)	(9.48)
N	22170	18953	18927	22170	18953	18927	22170	18953	18927

t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: Predicting Health, House and Job weights

Notes: *Working* is actively working as employee or self employed. *Prev. Married* includes divorced, widowed or separated. *Children* is a binary variable, where 1 is 1 or more children living in the house. *Single hh* includes single elderly, single non–elderly, single parent with or without dependent children in the house. *Couple* includes couples without children, with dependent children in the house or with independent children.

Click here to return to Chapter 1: Results on weighting correlates

	Amount (1)	Amount (2)	Amount (3)	Way (1)	Way (2)	Way (3)	Social (1)	Social (2)	Social (3)
Female (=1)	-0.005**		-0.005**	-0.007***		-0.006**	-0.003		-0.004*
	(-2.99)		(-2.82)	(-3.71)		(-2.83)	(-1.38)		(-2.02)
Age (years)	-0.000		-0.000	0.001*		0.001	-0.001*		-0.001**
	(-0.27)		(-0.74)	(1.96)		(1.64)	(-2.49)		(-2.58)
Age Squared	0.000***		0.000**	0.000		0.000	0.000*		0.000
	(4.06)		(3.24)	(1.25)		(0.67)	(2.19)		(1.35)
Working (=1)	-0.018***		-0.016***	-0.011***		-0.008***	-0.024***		-0.024***
	(-9.88)		(-7.99)	(-5.03)		(-3.42)	(-11.04)		(-9.88)
Retired (=1)	0.015***		0.016***	0.008*		0.009*	-0.002		-0.004
	(4.60)		(4.63)	(2.28)		(2.34)	(-0.60)		(-0.90)
Married (=1)	-0.027***		-0.010**	-0.028***		-0.019***	-0.025***		-0.005
	(-10.56)		(-2.93)	(-9.80)		(-5.04)	(-8.30)		(-1.22)
Prev. married (=1)	-0.016***		-0.007*	-0.010**		-0.003	-0.012***		-0.001
	(-5.91)		(-2.31)	(-3.26)		(-0.85)	(-3.69)		(-0.31)
Children (=1)		-0.035***	-0.016***		-0.031***	-0.016***		-0.028***	-0.025***
		(-17.93)	(-7.18)		(-14.09)	(-6.25)		(-12.03)	(-9.27)
Single hh (=1)		0.011**	-0.010*		0.026***	0.004		-0.009	0.001
		(2.64)	(-2.33)		(5.41)	(0.81)		(-1.88)	(0.27)
Couple (=1)		-0.012**	-0.023***		0.010*	0.003		-0.026***	-0.013*
		(-2.89)	(-5.37)		(2.14)	(0.69)		(-5.41)	(-2.56)
Hh income (£)		-0.000***	-0.000***		-0.000***	-0.000***		-0.000**	-0.000
		(-8.20)	(-4.61)		(-5.23)	(-3.59)		(-2.88)	(-0.13)
Constant	0.117***	0.140***	0.144***	0.122***	0.138***	0.126***	0.187***	0.174***	0.202***
	(23.25)	(35.09)	(22.02)	(21.38)	(30.48)	(16.91)	(30.97)	(36.67)	(25.75)
N	22170	18953	18927	22170	18953	18927	22170	18953	18927

t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12: Predicting Social, Leisure amount and Leisure activities weights

Notes: *Working* is actively working as employee or self employed. *Prev. Married* includes divorced, widowed or separated. *Children* is a binary variable, where 1 is 1 or more children living in the house. *Single hh* includes single elderly, single non–elderly, single parent with or without dependent children in the house. *Couple* includes couples without children, with dependent children in the house or with independent children.

Click here to return to Chapter 1: Results on weighting correlates

	Income (1)	Income (2)	Income (3)	Partner (1)	Partner (2)	Partner (3)
Female (=1)	-0.000		-0.001	0.013***		0.016***
	(-0.25)		(-0.36)	(6.56)		(7.72)
Age (years)	-0.000		-0.000	-0.000		0.001**
	(-0.92)		(-1.36)	(-0.42)		(2.65)
Age Squared	0.000*		0.000**	-0.000***		-0.000***
	(2.14)		(2.62)	(-4.70)		(-4.48)
Working (=1)	0.004*		0.001	0.015***		0.010***
	(2.32)		(0.63)	(6.57)		(4.14)
Retired (=1)	-0.009**		-0.013***	0.017***		0.014***
	(-2.95)		(-3.68)	(4.43)		(3.39)
Married (=1)	-0.011***		-0.011***	0.138***		0.077***
	(-4.60)		(-3.48)	(44.68)		(19.31)
Prev. married (=1)	-0.007*		-0.007*	0.043***		0.025***
	(-2.56)		(-2.35)	(13.37)		(7.01)
Children (=1)		-0.002	0.002		0.062***	0.039***
		(-1.20)	(0.85)		(26.15)	(14.38)
Single hh (=1)		0.007	0.006		-0.046***	-0.047***
		(1.77)	(1.41)		(-9.00)	(-8.80)
Couple (=1)		0.000	0.004		0.063***	0.028***
		(0.01)	(0.85)		(12.94)	(5.51)
Hh income (£)		0.000***	0.000***		0.000***	0.000***
		(7.50)	(7.55)		(6.81)	(4.14)
Constant	0.107***	0.093***	0.098***	0.089***	0.074***	0.050***
	(21.68)	(23.97)	(15.16)	(14.36)	(15.27)	(6.41)
N	22170	18953	18927	22170	18953	18927

t statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: Predicting Income and Partner weights

Notes: *Working* is actively working as employee or self employed. *Prev. Married* includes divorced, widowed or separated. *Children* is a binary variable, where 1 is 1 or more children living in the house. *Single hh* includes single elderly, single non–elderly, single parent with or without dependent children in the house. *Couple* includes couples without children, with dependent children in the house or with independent children.

(1.B2) k–means clustering — further tables

Click here to return to Chapter 1: Results of k–means clustering

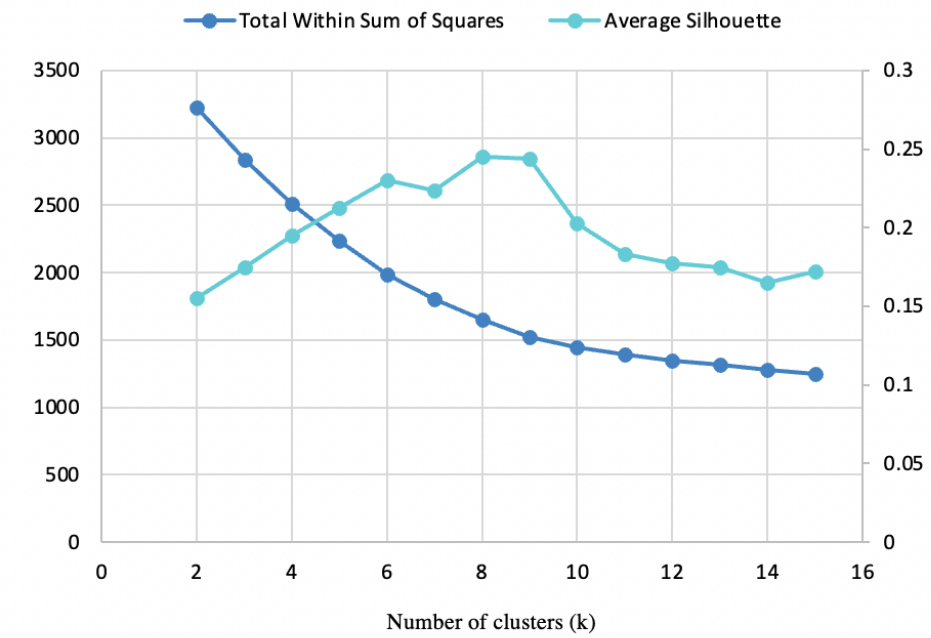


Figure 10: Determining the optimal number of clusters

Notes: Total Within Sum of Squares (TWSS): for each number of groups, k, in each group there is a deviation of each individual from the group centre of 8 domain weightings. These are squared and then summed across all individuals, creating a total squared deviation of observations from group centres per k. As k increases, TWSS falls — it is easier for individuals to be close to the group centre as the number of groups increases. The return to increasing k diminishes, and in this case this seems to begin at approximately $k = 6$ The Average Silhouette (AS) is the average of the ratio of cohesiveness to separation of individuals from their own and other groups; the higher the value (maximum of +1, minimum of -1), the more appropriate the group size.

	2 groups	3 groups	4 groups	5 group	6 groups	7 groups	8 groups	9 groups	10 groups
1	3,526	2,552	13,400	2,204	2,233	3,115	*1,392*	*1,546*	*1,418*
2	18,685	4,343	2,732	2,421	3,445	2,143	2,416	*1,723*	*1,459*
3		15,316	2,110	12,024	10,983	2,230	2,035	*1,592*	*1,284*
4			3,969	*1,820*	*1,983*	*1,585*	8,551	*1,483*	5,218
5				3,742	*1,673*	9,532	*1,584*	8,433	*1,246*
6					*1,950*	*1,741*	*1,741*	*1,372*	*1,230*
7						*1,865*	*1,519*	*1,367*	*1,498*
8							2,973	*1,964*	*1,927*
9								2,731	*1,800*
10									5,131

Figure 11: Cluster sizes per k

Notes: This table shows the size of each group as the number of groups grows, with larger groups being coloured darker than smaller groups. Groups with fewer than 2,000 (fewer than 9% of the sample) are in italics. After k is 6 at least 3 groups have fewer than 2,000 observations.

		Health	Income	House	Partner	Job	Social	Amount	Way
A	*min*	0.000	0.000	**0.255**	0.000	0.000	0.000	0.000	0.000
	mean	0.094	0.080	**0.451**	0.062	0.054	0.092	0.077	0.090
2,204	*max*	0.484	0.619	**1.000**	0.407	0.665	0.479	0.641	0.571
B	*min*	**0.269**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	**0.471**	0.074	0.084	0.065	0.059	0.091	0.073	0.083
2,421	*max*	**1.000**	0.655	0.493	0.444	0.636	0.484	0.549	0.584
C	*min*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	0.114	**0.129**	0.117	0.086	0.100	**0.125**	**0.152**	**0.177**
12,024	*max*	0.337	1.000	0.330	0.317	0.909	0.344	1.000	1.000
D	*min*	0.000	0.000	0.000	0.000	0.000	**0.287**	0.000	0.000
	mean	0.082	0.059	0.074	0.070	0.056	**0.497**	0.072	0.091
1,820	*max*	0.471	0.563	0.466	0.490	0.612	**1.000**	0.579	0.587
E	*min*	0.000	0.000	0.000	**0.194**	0.000	0.000	0.000	0.000
	mean	0.103	0.074	0.104	**0.407**	0.078	0.086	0.069	0.079
3,742	*max*	0.497	0.689	0.492	**0.992**	0.720	0.449	0.585	0.587

Figure 12: Minimum, Mean and Maximum weights for $k = 5$

Click here to return to Chapter 1: Results of k–means clustering

		Health	Income	House	Partner	Job	Social	Amount	Way
A	*min*	0.000	0.000	0.000	**0.224**	0.000	0.000	0.000	0.000
	mean	0.105	0.073	0.106	**0.433**	0.053	0.086	0.069	0.076
3,115	*max*	0.497	0.585	0.492	**0.992**	0.441	0.449	0.585	0.477
B	*min*	**0.283**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	**0.490**	0.072	0.086	0.066	0.050	0.091	0.070	0.075
2,143	*max*	**1.000**	0.556	0.493	0.426	0.449	0.484	0.539	0.413
C	*min*	0.000	0.000	0.000	0.000	**0.137**	0.000	0.000	0.000
	mean	0.113	0.096	0.098	0.116	**0.320**	0.101	0.070	0.087
2,230	*max*	0.443	0.683	0.420	0.477	**0.909**	0.462	0.592	0.423
D	*min*	0.000	0.000	0.000	0.000	0.000	**0.301**	0.000	0.000
	mean	0.083	0.057	0.076	0.072	0.048	**0.519**	0.068	0.078
1,585	*max*	0.471	0.563	0.466	0.490	0.429	**1.000**	0.579	0.458
E	*min*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	mean	0.123	**0.144**	**0.130**	0.092	0.067	**0.136**	**0.171**	**0.136**
9,532	*max*	0.349	1.000	0.352	0.324	0.294	0.379	1.000	0.355
F	*min*	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.281**
	mean	0.083	0.052	0.084	0.068	0.038	0.099	0.101	**0.475**
1,741	*max*	0.477	0.558	0.486	0.469	0.473	0.482	0.620	**1.000**
G	*min*	0.000	0.000	**0.268**	0.000	0.000	0.000	0.000	0.000
	mean	0.095	0.076	**0.476**	0.065	0.043	0.091	0.073	0.081
1,865	*max*	0.484	0.603	**1.000**	0.407	0.454	0.479	0.538	0.469

Figure 13: Minimum, Mean and Maximum weights for $k = 7$

	(A) Health	(B) Partner	(C) Balance	(D) House	(E) Social	(F) Leisure Way	Overall
Female (=1)	0.560***	0.560***	0.500***	0.593***	0.550*	0.521	0.529
Age (years)	38.7***	41.5***	42.3	45.0***	41.7*	48.4***	42.5
	(17.5)	(15.1)	(19.8)	(20.2)	(19.6)	(20.5)	(19.2)
HH income (£)	21,478***	25,639***	23,952***	21,235***	22,594	21,426***	23,387
	(14,492)	(18,630)	(17,729)	(14,938)	(19,120)	(14,670)	(17,253)
HH Size	3.07	3.19***	3.03	2.96	3.03	2.74***	3.03
	(1.45)	(1.30)	(1.49)	(1.51)	(1.50)	(1.46)	(1.46)
No. children	0.550***	0.824***	0.376***	0.449	0.390***	0.347***	0.468
	(0.930)	(1.054)	(0.822)	(0.894)	(0.832)	(0.810)	(0.894)
Household type (=1)							
Single non–elderly	0.08***	0.01***	0.06	0.08***	0.08***	0.08***	0.06
Single elderly	0.05	0.00***	0.06	0.09***	0.06	0.12***	0.06
Couple no child	0.19***	0.38***	0.27	0.22***	0.22***	0.29*	0.27
Couple: dep. child	0.34**	0.43***	0.29***	0.29**	0.32	0.27***	0.32
Couple: non–dep. child	0.11	0.10**	0.12	0.12	0.12	0.11	0.11
Lone parent: dep. child	0.12***	0.01***	0.05**	0.09***	0.08***	0.05*	0.06
Lone parent: non–dep child	0.04*	0.01***	0.04*	0.05***	0.05**	0.05*	0.04
2+ unrelated adults	0.05*	0.02***	0.07***	0.04***	0.05	0.03***	0.05
Employment status (=1)							
Currently working	0.40***	0.56***	0.48***	0.36***	0.34***	0.34***	0.45
Unemployed	0.06***	0.04**	0.05	0.04*	0.05	0.04	0.05
Retired	0.14***	0.13***	0.18	0.23***	0.20	0.30***	0.19
Family Care	0.14***	0.14***	0.07***	0.12***	0.10	0.09	0.09
FT student / school	0.17***	0.06***	0.14**	0.15	0.18***	0.12	0.13
Marital status(=1)							
Married	0.33***	0.67***	0.40***	0.36***	0.37***	0.40***	0.43
Separated	0.08***	0.07**	0.05***	0.08***	0.07	0.07	0.06
Divorced	0.10***	0.06***	0.08*	0.11***	0.09	0.10***	0.08
Widowed	0.10*	0.02***	0.08	0.13***	0.10**	0.13***	0.08
Never married	0.40***	0.17***	0.38***	0.33	0.37***	0.30***	0.34
n	*2,233*	*3,445*	*10,983*	*1,983*	*1,673*	*1,950*	*22,211*

Table 14: Cluster descriptives

Notes: For each covariate, the mean is reported and for non–binary variables the standard deviation is reported in brackets. The results of 2 sided t–tests or pr–tests of equality of each group's mean to the mean in the sample are reported: *** is $p<0.01$, ** is $p<0.05$ and * is $p<0.1$.

Chapter 2, Appendix A: Surveys text

Click here to return to Chapter 2: Experimental Setup

Pre–race survey: Introduction to the Cherry Blossom Study[34]

This is a study conducted by researchers at the University of Oxford in collaboration with the organizers of the Credit Union Cherry Blossom Ten Mile Run & 5K Run/Walk.

STUDY OVERVIEW

The purpose of this research is to study participation and performance in road running events such as the Cherry Blossom 10 mile run. If you agree to participate, you will be asked to **complete two 10-minute surveys**:

- Survey 1 will be completed today.
- Survey 2 will be completed after the race. We will send you a survey link by email and you will have at most 7 days to complete it.

During the surveys, we will ask you basic questions about your socio-economic background, your running experience, participation and performance in the Cherry Blossom 10 mile run, and you will complete a few psychology scales and decision-making tasks.

If you complete this study, you will be entered in a lottery draw for the chance to win prizes such as an **Under Armour gift card** of $10, $25 or $150, or a **guaranteed free entry in the 2020 Cherry Blossom 10 mile run**.

PRIVACY AND CONFIDENTIALITY

In order to participate in the study, we will ask you to **provide your bib number and a valid email address**. Your bib number will be used to merge responses from the two surveys and to link them to the registration and race results databases. We will use your email address to send you the second survey and, should you win one, email you an Under Armour electronic gift card.

[34] These surveys were built and survey responses gathered in Qualtrics. This Appendix reports the text but does not portray the visual layout and interactive clicking components of the Qualtrics version.

The confidentiality of your research records will be strictly maintained. At the conclusion of the study, we will delete all personally identifiable information from our dataset and replace it with a random participant number. Data from this study will be used in presentations, short reports, and academic papers. The general results of the study will also be shared with the race organizers.

To maintain the confidentiality and credibility of the study, we ask you that **you do not discuss the content of the surveys or your answers with other fellow runners** before the conclusion of the study.

ELIGIBILITY AND RIGHTS

In order to participate in the study, you must be:

1. At least 18 years of age
2. A registered participant in the 2019 Cherry Blossom **10 mile** run.

Participation in this study is entirely voluntary. You can withdraw from the study at any time. There is no penalty for choosing not to participate or withdrawing. However, note that failure to give complete and accurate answers in both surveys may make you ineligible to receive a prize.

There are no risks beyond those usually involved when participating in road running events and online activities and there are also no personal benefits. Your participation in this study may increase scientific knowledge and improve the organization of future road races. Note that we hold no responsibility for potential injuries or accidents that might occur during the race.

CONTACT INFORMATION

If you any questions, comments or complaints, do not hesitate to contact us at cucbresearch@gmail.com. Please save this email address for your records.

CONSENT

If you are willing to participate, please answer the following questions:

I am 18 years or older: YES/NO

I have read and understand the above information: YES/NO

I want to participate in this research and continue with the study: YES/NO

[NEXT button - go to new page]

Thank you for taking part in our study! As researchers (and fellow runners), we are excited to learn about your running :)

Before we start, enter your bib number in the box below. Please make sure you enter the correct information: [TEXT BOX]

[NEXT button - go to new page]

PART 1: Basic socio-demographics

What is your **gender**? Male/Female

What **year** were you born in?

What is the **highest level of school** you have completed or the highest degree you have received?

- Less than high school degree
- High school graduate
- Some college but no degree
- Associate degree in college (2-year)
- Bachelor's degree in college (4-year)
- Master's degree
- Doctoral degree
- Professional degree (JD, MD)

How much total combined money did all members of your household earn last year? Please indicate the answer that includes your entire household income in 2018 before taxes:

- Less than $20,000
- $20,000 to $29,999
- $30,000 to $39,999
- $40,000 to $49,999
- $50,000 to $59,999
- $60,000 to $69,999
- $70,000 to $79,999
- $80,000 to $89,999
- $90,000 to $99,999
- $100,000 to $149,999
- $150,000 to $199,999
- $200,000 or more
- Prefer not to answer

PART 2: Your participation in the Cherry Blossom 10 mile run

How many times have you completed the Cherry Blossom 10 mile run? [dropdown menu with max number of participations corresponding to starting year of the race]

[if previously participated]
What time did you do in your **most recent** participation? [dropdown with increments of 1 minute] *(record response time)*

[if previously participated]
What is your **personal best** in the Cherry Blossom 10 mile run? [dropdown with increments of 1 minute] *(record response time)*

How well have you trained for this year's race? (1= did not train at all, 7 = trained extremely well)

What **time** are you aiming for on Sunday? [dropdown with increments of 1 minute]

How **much would it mean to you** to achieve this goal? (1= Nothing at all, 7 = A great deal)

How **committed** are you to this goal? (1= Not committed at all, 7 = extremely committed)

How **confident** are you that you will meet your goal? (1= not confident at all, 7 = certain)

[NEXT PAGE - RANDOM ASSIGNMENT TO TREATMENT C, M, I or MI]

Control (**C**)	Good luck with the race! Don't forget to pace yourself. Let's see how things go on Sunday!

Pride Message (**M**)	Good luck with the race! Don't forget to pace yourself. Runners often are too optimistic about the goals they set. We have to say we would be surprised if you met the goal you chose. Although we're certain you can make it in your dreams ☺ Let's see how things go on Sunday!

Pride Message and Incentive (**MI**)	Good luck with the race! Don't forget to pace yourself. Runners are often too optimistic about the goals they set. We have to say that we would be surprised if you met the goal you chose. Although we're certain you can make it in your dreams ☺ In fact, we are willing to bet money on your race performance: if we are wrong and you meet your goal, we will send you a $X Under Armour gift card by email after you have completed Survey 2! Let's see how things go on Sunday!

Vary $ X to be $10 or $25.

Incentive (**I**)	Good luck with the race! Don't forget to pace yourself. If you meet your goal, we will send you a $X Under Armour gift card by email after you have completed Survey 2! Let's see how things go on Sunday!

Vary $ X to be $10 or $25.

Before ending the survey, we have a few more questions for you:

Do you plan to **keep track of your time on Sunday** with a watch/phone? [YES/NO]

Do you usually **keep track of your time** with a watch/phone when you race?

- Always
- Most of the time
- Sometimes
- Rarely
- Never
- It is my first time doing an official race.

For the next question, please enter a percentage in each box. You can enter any number between 0 and 100 in each box but your answers should sum to 100.

What do you think are the chances that you will be:

At least 10:00 minutes faster than your time goal of hh:mm	[ ]
5:00-9:59 minutes faster than your time goal of hh:mm	[ ]
1:00-4:59 minutes faster than your time goal of hh:mm	[ ]
Less than one minute above or below your time goal of hh:mm	[ ]

1:00-4:59 minutes slower than your time goal of hh:mm [ ]
5:00-9:59 minutes slower than your time goal of hh:mm [ ]
At least 10:00 minutes slower than your time goal of hh:mm [ ]

Total [100]

Thank you so much for your answers! This is the end of Survey 1.

Please enter your email address so we can send you the second survey after the race. [TEXT BOX]

Post–race survey

Welcome to Survey 2! This is the final survey. We expect that it will take you less than 10 minutes.

Before we start, please supply the following information:

Enter your **bib number** in the box below. Please make sure you enter the correct information: [TEXT BOX]

Enter your **email address**, so we can send you any Under Armour gift card you might have won. [TEXT BOX]

Please indicate how you accessed **Survey 1**:

- At the Expo
- Via the email link
- Via the facebook group

Thank you! Let's get started:

record response time

What was your **goal time** in the 2019 Cherry Blossom 10 mile run?

- [dropdown] hh:mm
- I do not know/I cannot remember.

record response time

How **long did it take you to complete** the 2019 Cherry Blossom 10 mile run?

- [dropdown] hh:mm
- I do not know/I cannot remember.

record response time

During Survey 1, do you remember having seen any **information or comment regarding your time goal** for the race? If so, do you recall what the information was?

Please provide as much detail as possible: [TEXT BOX - force text entry of at least one sentence]

NEW PAGE

If you proved us wrong by meeting your goal [If T2] / If you met your goal [if T3], you will be entitled to receive a $X gift card.

If you'd rather not receive the gift card, **we can make a donation on your behalf to a charity of your choice**.

Please select the option you prefer:

- I prefer to receive the gift card.
- I prefer to donate the amount to charity (you will be able to select your charity on the next page).

[If chose gift card] Thank you. We will verify your official time and email you a $X gift card if you met your goal. Please allow for 14 business days.

[If chose donation] Thank you. We will verify your official time and donate to your chosen charity if you met your goal. Please allow for 14 business days. Please enter the charity you would like us to make a donation to on your behalf:

- Children's Miracle Network Hospitals
- World Wildlife Fund
- DAV (Disabled American Veterans) Charitable Service Trust
- American Youth Soccer Organisation
- Other: [TEXT BOX]

How proud do you feel of your participation in the 2019 Cherry Blossom 10 mile run?

1. Nothing to be proud of
2. Somewhat proud
3. Moderately proud
4. Very proud
5. Extremely proud

How would you rate your race performance? [grade scale on Qualtrics]

On a scale of 0 to 10, with 0 being "Completely unwilling to take risks" and 10 being "Completely willing to take risks", how do you see yourself: are you a person who is generally fully prepared to take risks, or do you try to avoid risks? [dropdown] 0–Completely unwilling to take risks, 1, ..., 9, 10–Completely unwilling to take risks

How many friends or relatives also participated in the 10 mile run on Sunday? [0 to 50+]

How many friends or relatives also participated in the 5K run-walk on Sunday? [0 to 50+]

Please rate the extent to which you agree with the following statements (1 = strongly disagree; 7 = strongly agree)
I participate in road running events:

- To motivate myself to run regularly
- To push myself to run faster
- For the medal

- To feel badass
- For the t-shirt
- To have fun

Did you purchase, or do you plan to purchase, any of the race photos? YES, I HAVE ALREADY PURCHASED MY RACE PHOTOS/YES, I INTEND TO PURCHASE MY RACE PHOTOS/NO

Did you purchase the race medal when you registered? Yes/No

[If No] You still have the opportunity to **purchase the medal**. If you are interested, please select your preferred option below. Your information will be transmitted to the race organizers so they can order the right number of medals.

- Regular medal ($20)
- Enhanced medal with official time ($25)
- No medal

How many medals do you own from races you have completed?

- None.
- Between 1 and 5 medals.
- Between 6 and 10 medals.
- Between 11 and 20 medals.
- More than 20 medals.

record response time, randomize the order of the two ideas, track clicks

Not sure what to do of your medals? Feel free to click on any of the links below if you want suggestions:

- Medal hangers
  - Like these
  - Or these
- Donations:
  - Sports Medal Recycling

– Medals 4 Mettle

Below are a number of words and phrases that describe different feelings and emotions. Read each item and then indicate **the extent to which you feel this way**. [1–not at all, 2, 3, 4, 5–extremely]

- accomplished
- like I am achieving
- confident
- fulfilled
- productive
- like I have self worth
- successful
- arrogant
- conceited
- egotistical
- pompous
- smug
- snobbish
- stuck–up

Thank you for your answers! Before we end the survey, we have a message for you. Click on the **Next** button to see what it is...

Hello badass runner! You might have what it takes to create a "Race" Commitment by signing up for your next race and/or committing to a new time goal for an upcoming race.

T1 You probably won't bother, but if you think you've got what it takes, now it's time to prove it track clicks

T2 If you think you've got what it takes, now it's time to prove it. track clicks

Using StickK.com is free and very easy. You can set whatever goal you want and even set penalties on yourself if you don't make it... To make sure you stay on track, we suggest you nominate us as a referee. Let's bring on the challenge... There are just 4 easy steps to reach your goal:

1. Create your account (it's free!)
2. Make a commitment
3. Add our email as your referee: cucbresearch@gmail.com
4. Start working!

Let's see what you decide :) *P.S. We're not related to StickK in any way. We just think it is a cool tool.*

Finally, do you have any comments for us about the study? [large textbox]

Thank you for taking part in this study! We will be in touch regarding your race results and prize draw. Feel free to email us at cucbresearch@gmail.com if you have any questions in the meantime! Press the final Next button to save and submit your responses.

Follow–up questions

We have a few more questions about the study itself. After submitting your goal time in the Pre Race Survey, we made the following comment:

[INSERT CORRESPONDING MESSAGE]

Which of the following answers best matches what you thought when you read our comment? In order to form your judgement, I thought you used:

1. Personal information such as my racing history, current training, or goal time.
2. No personal information. Only information about runners in general.
3. No information at all.

How confident are you of your answer? [1: no clue, 7: certain]

Click here to return to Chapter 2: Experimental Setup

Chapter 2, Appendix B: Experiment design

Appendix B1: Summary statistics and covariates balance

Click here to return to Chapter 2: Experimental Setup

	All (1)	Study (2)	*p*–value (3)	C (4)	M (5)	I (6)	MI (7)	*p*–value (8)
Socio–Demographics								
Female (=1)*	0.61	0.50	0.00	0.50	0.50	0.47	0.52	0.88
Age on race day (years)	36	37	0.00	39	40	41	41	0.44
Locals (DC, VA or MD) (=1)	—	0.68	—	0.69	0.71	0.69	0.64	0.76
Education (years)	—	5.2	—	5.5	5.3	5.0	5.0	0.26
HH Income: $0 – 69,999 (=1)	—	0.14	—	0.11	0.14	0.14	0.10	0.11
HH Income: $70,000 – 99,999 (=1)	—	0.16	—	0.19	0.25	0.09	0.12	0.01
HH Income: $100,000 – 149,000 (=1)	—	0.24	—	0.28	0.13	0.28	0.26	0.03
HH Income: $150,000 – 199,999 (=1)	—	0.14	—	0.15	0.13	0.10	0.17	0.59
HH Income: > $200,000 (=1)	—	0.24	—	0.21	0.22	0.27	0.25	0.72
Race Experience								
CUCB first timer (=1)	0.52	0.54	0.71	0.48	0.45	0.45	0.39	0.27
Reported training (1–7)	—	4.51	—	4.5	4.4	4.4	4.7	0.23
Race Engagement								
Ex ante medal (=1)	0.38	0.54	0.00	0.58	0.52	0.56	0.47	0.38
Ex ante enhanced t–shirt (=1)	0.26	0.35	0.00	0.38	0.36	0.32	0.34	0.23
Number of observations	17,000	430	—	108	106	107	106	—

Table 15: Selection into the study and covariates balance across treatments

Notes: Column (1) indicates the mean of the co–variates in the full population of runners, where these are available from the Registration or Finisher datasets; Column (2) indicates the mean of the runners who completed the Pre–race survey; Column (3) gives the *p*–value for the difference in means across (1) and (2). Columns (4)–(7) give the mean for each treatment group, annd (8) reports the *p*–value for difference in means across the treatments (F–test). *Ex ante medal* is whether or not the runner ordered and paid for a medal as part of the registration (before the race); *Enhanced t–shirt* is whether or not the runner ordered and paid for a technical t–shirt as part of the registration (before the race). *Males were deliberately over–sampled.

Our main outcome measures focus on race performance in relation to participants' goals. For performance, we use the official chip times recorded by the CUCB organisers. As a benchmark, the median finish time was 93.5 minutes (87.5 for men and 97.5 for women), very close to the mean of 89.7 minutes (85.6 for men and 92.4 for women). Figure 14 presents histograms of the distribution of finish times, comparing our study sample (430 runners) to the entire population of finishers (17,800). The spike of fast runners in the full sample reflects the fact that the race attracts professional and semi–professional runners, as well as the amateur participants.

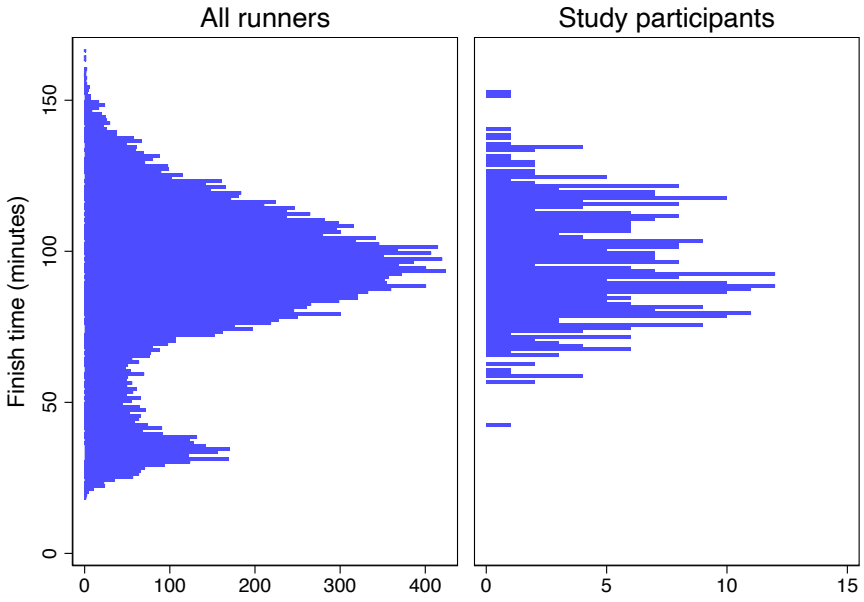


Figure 14: Histograms: finish times

Notes: Finish times are the official chip times in minutes recorded by the race organisers. There were 17,800 runners who completed the 10 mile race.

Appendix B2: Power calculations

Click here to return to Chapter 2: Experimental Setup

Below we present power calculations for differences in proportions between the control and any of treatment groups (focusing on our main outcome, *goal attained*, which is binary). Figure 15 shows power for different treatment differences, fixing the total sample size at 200 (100 in each group) and setting α at 5%. The control group mean is set at 25% as we estimated that only 1 in 4 runners would meet their time goal. To obtain 80% power, the treatment difference between two groups must be at least 19%. Figure 16 shows how the minimum detectable effect size varies with the size of the sample, holding power at 80% and α at 5%. To detect a treatment difference of 10 percentage points, we would

need at least 300 participants in each group.

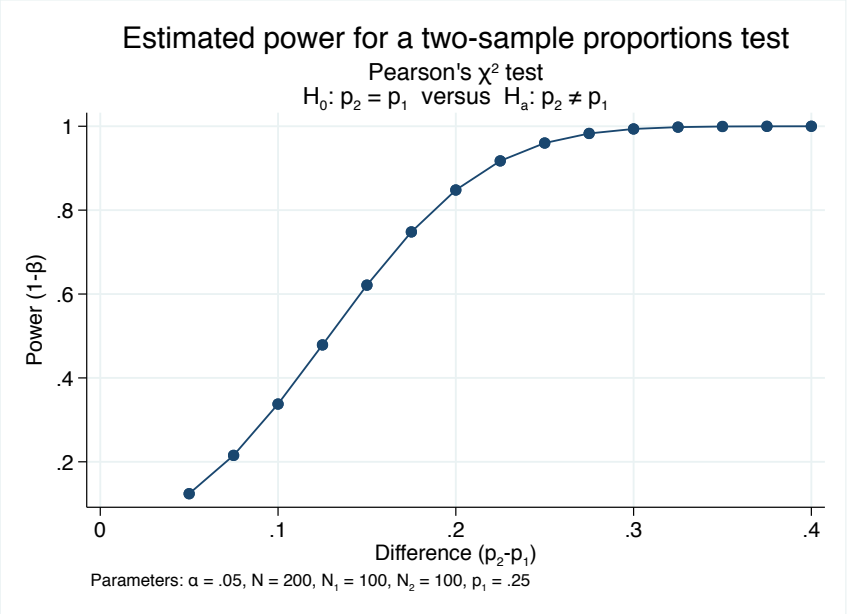


Figure 15: Power curve: 100 runners in each group

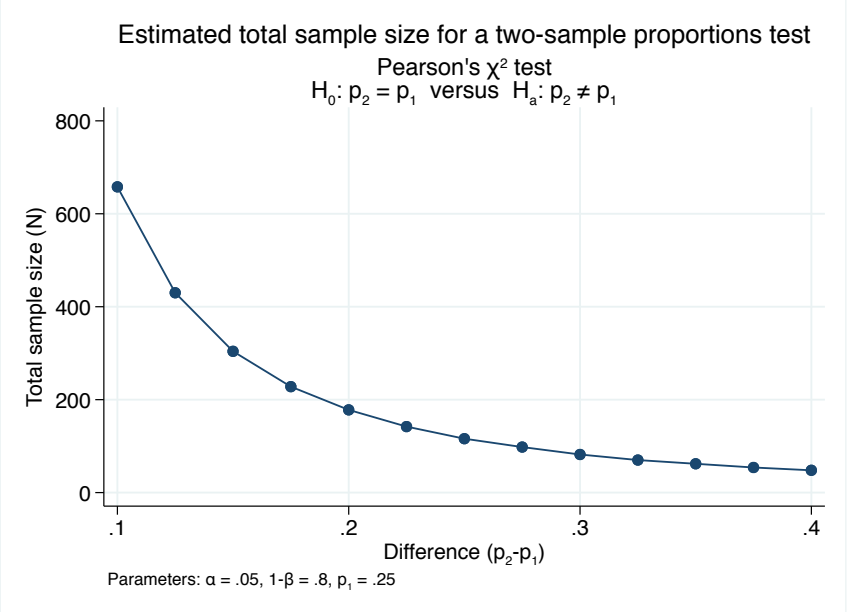


Figure 16: Sample size and effect size

Appendix B3: Study attrition

Click here to return to Chapter 2: Experimental Setup

	C	I	M	MI
Pre–race survey	108	106	107	106
Post–race survey	102 (94%)	98 (92%)	100 (93%)	102 (96%)
Exit questions*	—	79 (75%)	74 (69%)	73 (69%)

Table 16: Attrition by treatment

Notes: * The Exit questions were in an optional survey with only 2 questions, within the prize announcement link. The questions were only asked of those in the I, M or MI treatments, not C.

Chapter 2, Appendix C: Regression tables and further graphs

Click here to return to Chapter 2: Results

	(1) Goal Attained	(2) Goal Attained	(3) Goal Attained	(4) Goal Attained	(5) Goal Attained	(6) Goal Attained
Message (=1)	0.121**	0.112**	0.104**	0.111*	0.106*	0.109*
	(2.50)	(2.32)	(2.17)	(1.73)	(1.68)	(1.73)
Dollars	-0.001	-0.000	-0.001	-0.002	-0.001	-0.000
	(-0.38)	(-0.10)	(-0.29)	(-0.45)	(-0.16)	(-0.12)
In person recruit (=1)	0.053	0.064	-0.049	0.054	0.064	-0.049
	(0.87)	(1.02)	(-0.68)	(0.88)	(1.03)	(-0.68)
Goal (mins)		0.004***	0.006***		0.004***	0.006***
		(2.77)	(4.18)		(2.77)	(4.17)
First timer (=1)		0.050	0.062		0.050	0.063
		(1.03)	(1.21)		(1.02)	(1.22)
Female (=1)			-0.196***			-0.196***
			(-3.21)			(-3.20)
Age (years)			-0.007***			-0.007***
			(-2.75)			(-2.74)
Message × Dollars				0.001	0.001	-0.001
				(0.26)	(0.13)	(-0.12)
Constant	0.433***	-0.084	-0.056	0.438***	-0.083	-0.057
	(10.46)	(-0.44)	(-0.23)	(9.50)	(-0.43)	(-0.23)
Other Race Controls		✓	✓		✓	✓
Other Soc–Dem Controls			✓			✓
N	422	422	422	422	422	422

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 17: Regression: Strict Goal Attainment, Full Sample

Notes: Other race controls: reported training (0–7)) goal confidence (0–7), goal committment (0–7), goal importance (0–7). Other Socio–Demographic controls: local (=1), education (years), income group.

Click here to return to Chapter 2: Results

	(1) % Deviation	(2) % Deviation	(3) % Deviation	(4) % Deviation	(5) % Deviation	(6) % Deviation
Message (=1)	-1.657	-1.461	-1.561	-1.081	-0.912	-1.494
	(-1.53)	(-1.40)	(-1.53)	(-0.76)	(-0.67)	(-1.11)
Dollars	-0.019	-0.040	-0.025	0.013	-0.009	-0.021
	(-0.36)	(-0.79)	(-0.49)	(0.18)	(-0.12)	(-0.30)
In person recruit (=1)	-0.234	-1.079	2.557*	-0.262	-1.124	2.549
	(-0.17)	(-0.79)	(1.66)	(-0.19)	(-0.83)	(1.64)
Goal (mins)		-0.166***	-0.241***		-0.167***	-0.241***
		(-5.66)	(-7.59)		(-5.67)	(-7.58)
First–timers (=1)		-0.153	0.032		-0.110	0.037
		(-0.15)	(0.03)		(-0.10)	(0.03)
Female (=1)			6.033***			6.029***
			(4.62)			(4.61)
Age (years)			0.213***			0.213***
			(4.04)			(4.02)
Message × Dollars				-0.066	-0.063	-0.008
				(-0.62)	(-0.62)	(-0.08)
Constant	2.286**	22.249***	18.808***	2.007*	22.108***	18.787***
	(2.48)	(5.33)	(3.60)	(1.96)	(5.29)	(3.59)
Other Race Controls		✓	✓		✓	✓
Other Soc–Dem Controls			✓			✓
N	422	422	422	422	422	422

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Regression: Percentage Chip Goal Deviation, Full Sample

Notes: Other race controls: reported training (0–7)) goal confidence (0–7), goal committment (0–7), goal importance (0–7). Other Socio–Demographic controls: local (=1), education (years), income group.

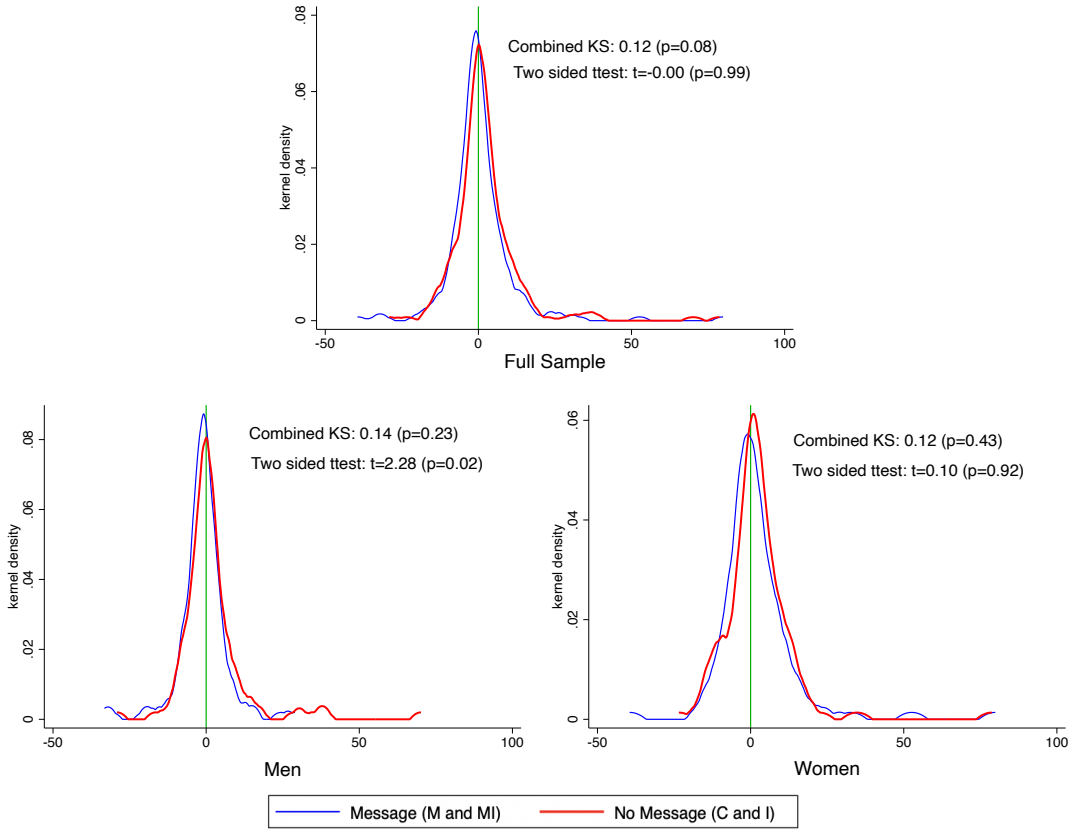


Figure 17: Percentage deviation from goal time, by message

Notes: Percentage deviation from goal time is calculated as (chip-goal)/goal, hence negative values represent runners who completed the race in a time faster than their goal time (success), and positive values represent slower times. The green vertical line is at 0, where goal time and chip time coincide.

Click here to return to Chapter 2: Results

Click here to return to Chapter 2: Results

Full sample

	Pride in race	Charity donation	Positive pride	Negative pride	Recall message	Recall money	Achievement
Charity donation	-0.217						
Positive pride	0.223	-0.178					
Negative pride	-0.022	-0.092	-0.167				
Remembered message	0.026	-0.034	-0.137	0.051			
Remembered money	0.004	0.050	-0.107	-0.022	0.345		
Achievement rating	0.689	-0.135	0.182	-0.035	0.169	0.069	
Study feelings	-0.058	0.230	0.007	0.008	-0.042	0.034	0.117

Achieved goal

	Pride in race	Charity donation	Positive pride	Negative pride	Recall message	Recall money	Achievement
Charity donation	-0.107						
Positive pride	0.236	-0.102					
Negative pride	-0.064	-0.183	-0.080				
Remembered message	0.096	-0.154	-0.024	0.098			
Remembered money	-0.057	-0.041	0.042	-0.069	0.355		
Achievement rating	0.672	-0.029	0.240	-0.098	0.196	0.034	
Study feelings	-0.160	0.293	-0.080	0.158	-0.036	-0.012	0.180

Did not achieve goal

	Pride in race	Charity donation	Positive pride	Negative pride	Recall message	Recall money	Achievement
Charity donation	-0.398						
Positive pride	0.217	-0.272					
Negative pride	0.027	0.032	-0.257				
Remembered message	-0.077	0.131	-0.273	-0.009			
Remembered money	0.042	0.151	-0.271	0.039	0.321		
Achievement rating	0.736	-0.384	0.184	0.025	0.131	0.019	
Study feelings	0.087	0.151	0.116	-0.203	-0.042	0.118	0.143

Figure 18: Correlation heat maps of post race pride and charity decision

Click here to return to Chapter 2: Results

	(1) Charity	(2) Charity	(3) Charity	(4) Charity	(5) Charity	(6) Charity
Message	0.150	0.191**	0.150	0.416**	0.416**	0.369**
	(1.65)	(2.08)	(1.59)	(2.51)	(2.52)	(2.06)
In person recruit	0.160	0.179	0.346**	0.130	0.154	0.321**
	(1.44)	(1.41)	(2.27)	(1.17)	(1.22)	(2.11)
Goal time		0.002	-0.001		0.002	-0.001
		(0.72)	(-0.35)		(0.77)	(-0.24)
First timer (=1)		-0.187**	-0.099		-0.166*	-0.085
		(-2.04)	(-0.97)		(-1.81)	(-0.83)
Female (=1)			0.240			0.221
			(1.66)			(1.53)
Age (years)			0.011**			0.010*
			(2.18)			(1.87)
Message × Dollars				-0.015*	-0.013	-0.012
				(-1.91)	(-1.63)	(-1.43)
Constant	0.175**	0.123	-0.558	0.181**	0.104	-0.592
	(2.48)	(0.32)	(-1.20)	(2.59)	(0.28)	(-1.28)
Other Race Controls		✓	✓		✓	✓
Other Soc–Dem Controls			✓			✓
N	99	99	99	99	99	99

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 19: Regression: Charity donation for goal attainers

Notes: Other race controls: reported training (0–7)) goal confidence (0–7), goal committment (0–7), goal importance (0–7). Other Socio–Demographic controls: local (=1), education (years), income group.

Click here to return to Chapter 2: Results

	(1) Positive pride	(2) Positive pride	(3) Positive pride	(4) Positive pride	(5) Positive pride	(6) Positive pride
Message (=1)	-0.153*	-0.187**	-0.208**	-0.082	-0.103	-0.106
	(-1.69)	(-2.13)	(-2.38)	(-0.79)	(-1.03)	(-1.06)
In person recruit	0.005	0.114	0.279**	0.001	0.110	0.283**
	(0.04)	(1.00)	(2.11)	(0.01)	(0.97)	(2.16)
Goal (mins)		-0.000	-0.003		-0.001	-0.003
		(-0.15)	(-1.07)		(-0.24)	(-1.17)
Female (=1)			0.298**			0.311***
			(2.54)			(2.66)
Age (years)			0.002			0.001
			(0.44)			(0.27)
Message × Dollars				-0.008	-0.010*	-0.012**
				(-1.39)	(-1.72)	(-2.01)
Constant	4.196***	3.307***	3.893***	4.197***	3.318***	3.896***
	(59.91)	(9.71)	(8.93)	(60.06)	(9.79)	(9.01)
Other Race Controls		✓	✓		✓	✓
Other Soc–Dem Controls			✓			✓
N	198	198	198	198	198	198

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 20: Regression: Positive pride score for goal attainers

Notes: Positive pride score is from 0 (none) to 7 (extremely). Other race controls: reported training (0–7)) goal confidence (0–7), goal committment (0–7), goal importance (0–7). Other Socio–Demographic controls: local (=1), education (years), income group. Also controlling for date completed Post–race survey.

Click here to return to Chapter 2: Results

	(1) Negative pride	(2) Negative pride	(3) Negative pride	(4) Negative pride	(5) Negative pride	(6) Negative pride
Message (=1)	-0.173**	-0.141*	-0.146*	-0.207**	-0.182**	-0.194**
	(-2.10)	(-1.77)	(-1.79)	(-2.19)	(-1.99)	(-2.05)
In person recruit	0.343***	0.234**	0.153	0.345***	0.237**	0.151
	(3.34)	(2.26)	(1.23)	(3.36)	(2.28)	(1.22)
Goal (mins)		-0.005**	-0.002		-0.005**	-0.002
		(-2.33)	(-0.93)		(-2.29)	(-0.88)
Female (=1)			-0.161			-0.167
			(-1.47)			(-1.52)
Age (years)			-0.003			-0.003
			(-0.67)			(-0.58)
Message × Dollars				0.004	0.005	0.006
				(0.73)	(0.92)	(1.00)
Constant	1.532***	2.788***	2.617***	1.532***	2.783***	2.615***
	(24.11)	(8.99)	(6.41)	(24.08)	(8.97)	(6.41)
Other Race Controls		✓	✓		✓	✓
Other Soc–Dem Controls			✓			✓
N	198	198	198	198	198	198

t statistics in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 21: Regression: Negative pride for goal attainers

Notes: Negative pride score is from 0 (none) to 7 (extremely). Other race controls: reported training (0–7)) goal confidence (0–7), goal committment (0–7), goal importance (0–7). Other Socio–Demographic controls: local (=1), education (years), income group. Also controlling for date completed Post–race survey.

Chapter 2, Appendix D: Further details on the replication

Click here to return to Chapter 2: Discussion

Additional details on replication changes

In an effort to identify a clearer response to the message, in a new Recruitment survey we include a screening question asking runners if they have a goal time for the race or just want to finish the distance before the cut–off time; we then screen out all those who just want to finish. Markle et al. (2017) argue that for novice marathon runners the primary objective is finishing the race and so time goals are less salient (novices in their study rated the time goal as less important than the experienced runners). We therefore hypothesise that we may see a more robust effect of the message on performance, due to this screening procedure: runners without a specific goal time may care less about the actual time and therefore be less sensitive to the message. However, we also consider that the intensive margin effect of the pride message may be smaller, as finish times may be more clustered around the goal for runners who care more about a precise time.

We make a small change to the treatment message in the replication. To ensure full clarity to the runners, we reiterate the goal time that the individual had reported within the treatment message so that it is clear that the goal is running *under* the stated goal time. For example, for a runner who had stated a goal time of of 1 hour 30 minutes, the sentence for a treatment message of type M or MI would be "We would not be surprised if you fail to meet your goal of running under 01:30." We also reiterate the goal in the pure incentive group, I. We choose to do this so that the conditions for success are more clearly stated that in the CUCB study, in which we consider with hindsight that "meeting your goal time" was ambiguous as to whether a finish time of 1 hour 30 minutes and 20 seconds could be considered as success.

We did not find clear evidence for belief–related mechanisms through which the treatment may have had an effect when we examined differences in the raw pre race beliefs, and this is an area which we would like to understand more clearly. In the replication Pre–race survey, we use the same bins as in the CUCB study, but we add a clarification question: if they finish "within one minute of their goal time", what are the chances they will be less than one minute faster and likewise the chances they will be less than one minute slower. This allows us to compute more precise posterior beliefs about success. We also elicit the participants' prior beliefs about the chances of meeting their time goal (i.e. before the treatment manipulation) to enable the calculation of a change in certainty following the treatment; we did not do this in the CUCB study.

The Post–race survey aimed to measure feelings of pride using a set of measures: a direct self–reported pride question, the Tracy and Robbins (2017) pride scale, and self–reported race–enjoyment. We also looked at changes in choices for medal or t–shirt purchase after the race as measures of revealed pride in achievement. With the exception of the Tracy and Robbins scale, we found very little variation in these measures. We therefore have designed additional questions in the replication Post–race survey to be able to evaluate the welfare impact:

- *Willingness to participate in next year's race*: We ask participants how likely they are to take part in the race in the future from 0% to 100% chance and elicit their willingness to pay for participating in the 2020 race through a multiple price mechanism. The goal is to test whether the treatment has an adverse effect on long–term motivation to race.
- *Willingness to pay for a trophy*: We use a multiple price list mechanism to elicit runners' valuation of a trophy that commemorates their participation and performance in the race.
- *Compensation*: We ask runners to state how much they think participants (in a future study of the same sort) should be compensated for taking part, varying whether the participant in the vignette is a woman or man, and whether they met or did not meet their goal. The aim is to obtain a proxy for the perceived welfare effects of the different treatments, including the potential emotional harm.
- *Reported motivation*: We directly ask runners how motivating or demotivating they found the treatment message to be, and ask them to choose what motivated them to try to meet their goal time.

In addition to using the survey tools listed above to measure the treatment effect, we also collect more data on participants' direct perceptions and feelings about the treatment messages themselves. We therefore first ask participants who got the pride message (M or MI) what beliefs they had about the information we used to form our judgement and their confidence about their guess (these were both optional components in the CUCB Exit questions). We then ask participants to guess what percentage of runners who saw that message met their goal, and then show them an alternative treatment message and ask them to guess what percentage of runners who saw this message met their goal.

Replication empirical strategy and hypotheses

The main empirical strategy for the replication is very similar to the CUCB study: presenting regressions of the performance outcomes and other pride and belief related outcomes, with a primary specification with the message and incentive level entering separately and a secondary specification

with these treatments interacting. We plan to do this for the US and UK sample separately, and pooled across them. We will use the same set of controls in the replication study (race controls and socio–demographic controls), and also present the raw treatment effects without controls. We will concentrate our analysis on two dimensions of heterogeneity: gender and running pace, with the primary regression specification being run within each subgroup. For running pace, we will split the runners into a slow and fast group according to their chosen goal.

We have a set of hypotheses for the replication study:

1. A positive but moderate effect of monetary incentives on performance but that, at higher incentive levels, runners may derive less pride in their achievement (if they feel they were paid to perform, rather than achieving it from their own determination)

2. A more robust effect of the pride message on performance, due to the screening procedure (the exclusion of "I just want to finish" runners, who may care less about the actual time and therefore be less sensitive to the message) but that the intensive margin effect may be smaller, as finish times may be more clustered around the goal for runners who care more about a precise time.

3. In the Pre–race survey, belief updating pre– and post– the treatment message will be minimal because of two effects: (1) runners think the message is not relevant to them and (2) they may internalize the treatment effect and know that their motivation has increased.

4. In the Post–race survey, participants will predict a motivating effect of both the pride messages and the incentives on performance of others.

Click here to return to Chapter 2: Discussion

Chapter 3, Appendix A: String distance measures

Click here to return to Chapter 3: Theory and Method

Levenshtein edit based distance: applicable to lists or binary strings

The Levenshtein distance is the minimum number of single–character edits required to transform one string into another; these edits can be deletions, insertions or reversals (Levenshtein 1966). Each edit operation costs 1. The lower the cost of the transformation, the more similar the strings are to each other. There may be multiple transformations with the same minimised cost.

The authors of www.levenshtein.net provide an example by measuring the Levenshtein distance between two words, *levenshtein* and *meilenstien* (2016). There are a total of 4 edits, giving a Levenshtein distance of 4:

l	e		v	e	n	s	h	t	e	i	n
o	=	+	*o*	=	=	=	−	=	=	=	=
m	e	i	l	e	n	s		t	e	i	n

[= Match; o Substitution; + Insertion; − Deletion]

The Levenshtein distance is usually solved using dynamic programming. The dynamic programming method can be illustrated by setting one string as the row of a matrix and the other string as the column, as in the table below. A step adds 0 if the row and column label coincide or 1 if an edit is required. Each cell locally minimises the cost. There may be multiple paths with the same minimised cost. The bottom right hand diagonal cell holds the total cost. The yellow boxes highlight the edit path taken: substitute an *l* for an *m*; insert a gap; substitute a *v* for an *l*; delete an *h*.

		m	**e**	**i**	**l**	**e**	**n**	**s**	**t**	**e**	**i**	**n**
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	**4**

There is an important distinction between the *Levenshtein* distance, and the *Damerau–Levenshtein* distance; the latter allows for multiple edits on the same string (van der Loo, 2014). The Levenshtein distance will always be weakly larger than the Damerau–Levenshtein because it has a more restrictive set of transformations available to it.[35] For instance, to get from ba to acb there are two possible routes. The first costs two (the Damerau–Levenshtein measure):

$$\text{ba} \xrightarrow[+1]{\text{swap b,a}} \text{ab} \xrightarrow[+1]{\text{ins. c}} \text{acb}.$$

If editing the same substring twice is prohibited, the cost is three (the Levenshtein measure):

$$\text{ba} \xrightarrow[+1]{\text{del. b}} \text{a} \xrightarrow[+1]{\text{ins. c}} \text{ac} \xrightarrow[+1]{\text{ins. b}} \text{acb}.$$

For binary strings and ordered lists, the Levenshtien and Damerau–Levenshtein return the same measure in UK data example used in the main text. They also return the same measure for all the data in the ICP and WDH data. Both allow substitutions in their transformation rules: simple substitutions of items in the ordered lists ensure the two measures will be the same. I hypothesise that the same measure will be returned for any comparison between two binary strings of the same length: with only two options (1 or 0) for each character, a discrepancy can be fully resolved by a substitution (being able to edit the same substring more than once does not offer an cost advantage). A simulation does not reject this hypothesis: after generating two randomly ordered binary strings 10,000 times, the same result for the Levenshtein and the Damerau–Levenshtein measures is returned.

Hamming distance: applicable to lists or binary strings

The Hamming distance counts how many character substitutions are needed to turn a string, s, into another string, t (Hamming 1950). Formally,

$$d_{hamming}(s,t) = \sum_{i=1}^{|s|}[1-\delta(s_i,t_i)] \text{ if } |s| = |t| \text{ and } \infty \text{ otherwise.}$$

The Hamming distance is a transformation of a simple count–based method applied directly to the difference matrices or vectors: calculate the fraction of conflicting preference directions out of the total number of preference relations i.e. (1) count the number of 1s in the adjacency matrix D; (2) normalise the total by dividing by the dimensions of D.:

$$\frac{2\sum D}{n(n-1)}.$$

[35] Further details on the mathematical implications of this difference between the two Levenshtein distances can be found on pp.116–117 of van der Loo, 2014. The key difference is that Damerau–Levenshtein distance does not obey the triangle inequality, while the Hamming and Levenshtein distance do.

This normalised count is identical to the Hamming distance divided by the dimension of the compared matrix, minus the diagonal.

Given the more restrictive set of transformation rules (no insertion or deletion), the Hamming distance will always be weakly larger than the Levenshtein measure. For example, the Hamming distance between the two words *levenshtein* and *meilenstien* is 6:

l	e	v	e	n	s	h	t	e	i	n
x		x	x	x	x	x				
m	e	i	l	e	n	s	t	e	i	n

q–gram distance: applicable to lists or binary strings

The Jaccard distance counts how many unique substrings of q consecutive characters there are in two different strings. It then compares how many of these substrings appear in both strings; this distance lies between 0 and 1. 0 corresponds to identical strings and 1 corresponds to no common substrings (van der Loo 2014).

$$d_{jaccard}(s, t; q) = 1 - \frac{Q(s;q) \cap Q(t;q)}{Q(s;q) \cup Q(t;q)}.$$

The q–gram distance between *levenshtein* and *meilenstien* is 0.54 for $q = 2$ and $0.9\dot{3}$ for $q = 4$. The distance is 1 for $q \geq 5$: there is no substring of five consecutive letters that appears in both *levenshtein* and *meilenstien* (as the substring length gets longer, the distance approaches 1 because there are fewer common substrings).

Choosing a measure

To determine the strengths of the different measures, I considered two key criteria: (1) the appropriateness of the measure for application to either ordered lists or vectors summarising the adjacency matrices and (2) the appropriateness of the set of transformation rules in the context of RP and SWB orderings. A summary of the discussion is below:

	Summary vectors?	Ordered lists?	Transformation Rule?
Hamming	✓	✗	✓
Levenshtein	✓	✗	✗
q–distance	✗	✓	✗
Kendall τ	✗	✓	✓

Table 22: Choosing a distance measure

(1) Application to binary strings vs. letter orderings: The Hamming distance is not appropriate for comparing the ordered lists. In allowing two items to swap positions within the list — irrespective

of the distance between them — the Hamming distance generates the same measurement outcome between what we could consider to be relatively minor versus major discrepancies: in the context of the RP and SWB orderings, this means a country with a relatively minor discrepancy between the placement 1980 and 1975 has the same Hamming distance as the more extreme discrepancy in which 1980 and 2011 swap positions:

$$h_{2011} \succ h_{2005} \succ h_{1996} \succ h_{1985} \succ h_{1980} \succ h_{1975}$$
$$u_{2011} \succ u_{2005} \succ u_{1996} \succ u_{1985} \succ u_{1975} \succ u_{1980},$$

$$h_{2011} \succ h_{2005} \succ h_{1996} \succ h_{1985} \succ h_{1980} \succ h_{1975}$$
$$u_{1975} \succ u_{2005} \succ u_{1996} \succ u_{1985} \succ u_{1980} \succ u_{2011}.$$

The Levenshtein distance does register such differences in the discrepancy levels when applied to ordered lists but this is still not as discriminatory as when binary strings are compared. This is because when applied to lists, the distance does not account for the distance between discrepancies and hence the discrepancies in *every* pair–wise relation. The q–distance does identify the differences between the size of the distances when we compare ordered lists and therefore would indeed suitable for application to ordered lists. In contrast, the q–distance measure compares how frequently the same substring appears *anywhere* in each substring and therefore with binary strings, the measure could be misleading because a preference pattern with substring (110) representing an ordering over years 1975, 1980 and 1985 would be considered a match to substring (110) representing an ordering over years 1996, 2005 and 2011. The Kendall τ coefficient was specifically designed to evaluate the differences between lists: τ can pick up variations in the level of discrepancy in ordered lists.

Suitability of transformation rules: The Levenshtein distance transformation set is too unconstrained: it may minimise the cost of transforming a string to match another in a manner that renders the comparison between the two preference orderings misleading. The Levenshtein can delete or insert characters, thus shifting the remainder of the string either to the left or to the right. Entering a gap in an $\overline{H}$ string moves all subsequent items one (or more) spaces right. This could then match $\overline{H}$ and $\overline{U}$, but the states now being compared have shifted relative to each other (as if rows or columns were fully or partially exchanged in the adjacency matrix). Conversely, the q–distance may keep the count of discrepancies too high because it cannot sidestep an anomaly and look at just pair–wise relations. For instance, with $h_{1995} \succ h_{2011} \succ h_{2005}$ and $u_{1995} \succ u_{2005} \succ u_{2011}$, 1995 is correctly placed relative to both 2011 and 2005. Yet, a q–distance with $q = 2$ or $q = 3$ does not see any similarity between the two strings at all.

Click here to return to Chapter 3: Theory and Method

Chapter 3, Appendix B: SWB and real income

Click here to return to Chapter 3: The combined dataset

As noted in the introduction to Chapter 3, the literature on well–being uses regression to analyse the relations between real GDP per capita and SWB. For comparison with these papers, I report here the results of two linear regressions (non–linear regressions were not found to have significance).

The linear regressions of log real GDP per capita and log SWB is reported in Table 23. There is evidence at the 1% significance level of a positive relationship between log SWB and log real GDP per capita. There is a low R^2, and the confidence interval is tightly wrapped around the regression line. Comparing with Figure 3.5 in the main text, note that poorer countries tend to have a wider range of log SWB levels than the richer countries and that 2011 observations in the main appear below the regression line, suggesting log SWB may have fallen considerably from its previous levels following the global financial crisis, relatively more so than log real GDP.

	log SWB
log real GDP per capita	0.146***
(standard error)	(0.033)
Constant	1.986***
(standard error)	(0.031)
Observations	97
R^2	0.171
Adjusted R^2	0.162
Residual Std. Error	0.145 (df = 95)
F Statistic	19.588*** (df = 1; 95)
Notes:	*p<0.1; **p<0.05; ***p<0.01

Table 23: Regression results — log real GDP per capita and log SWB

Table 24 reports the regression of the change in log real GDP per capita and log happiness. As before, the confidence interval is tightly wrapped around the regression line. There is again a statistically significant (at the 1% level) gradient coefficient, with a similarly low R^2.

The reasonably low R^2 results for both regressions, together with the small size of the gradient coefficients, provide additional motivation for this chapter' non–parametric investigation into the in–

formational similarities and differences between RP and SWB orderings. However, note that in these regressions I have not controlled for any other covariates (not available in these datasets).

	Change in log SWB
Change in log real GDP per capita	0.138***
(standard error)	(0.035)
Constant	−0.021**
(standard error)	(0.009)
Observations	70
R^2	0.185
Adjusted R^2	0.173
Residual Std. Error	0.052 (df = 68)
F Statistic	15.412*** (df = 1; 68)
Notes:	*p<0.1; **p<0.05; ***p<0.01

Table 24: Regression results — Change in log real GDP per capita and change in log SWB

Click here to return to Chapter 3: The combined dataset