London School of Economics and Political Science

Department of Economics

Psychological Factors in Welfare and Policy Design

Canishk Vasanttilak Naik

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy. London, May 2025

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgment is made. This thesis may not be reproduced without my prior written consent.

I warrant that this authorization does not, to the best of my belief, infringe on the rights of any third party.

The thesis consists of 35,585 words.

Chapter 4 of this thesis is joint work with Daniel Reck[1], and we contributed equally.

[1] Department of Economics, University of Maryland

Abstract

This thesis investigates the role of psychological factors in the design of optimal policy, focusing on mental health and the social safety net. Around 1 billion people suffer from mental disorders [WHO, 2022], and those with poor mental health are disproportionately likely to live in poverty [Lund et al., 2010]. Mental disorders cause significant disturbances in cognition, emotion regulation, and everyday functioning [Hammar and Årdal, 2009], yet their role in economic policy design remains understudied.

Chapters 1 to 3 build on my Job Market Paper, which focuses on whether social assistance effectively reaches people with poor mental health.

- Chapter 1 develops a theoretical framework showing how take-up responses to policy separately identify the marginal value of benefits (need) and the cost of barriers.
- Chapter 2 presents new empirical facts about mental health and the targeting of social assistance using Dutch administrative data.
- Chapter 3 combines the theory and empirics to show that people with poor mental health have a 2× higher need for benefits yet face a 64% higher cost from barriers. I also show that reducing barriers would be twice as effective as increasing benefits.

While Chapters 1 to 3 take a revealed preference approach, a question remains: should the planner normatively respect the observed choices of people with poor mental health? Chapter 4 (with Daniel Reck) generalises this idea, tackling the fundamental challenge of behavioural welfare economics: psychological factors can cause inconsistencies, forcing policymakers to take a stand on which choices reflect an individual's true normative preferences. We show that incorporating normative uncertainty leads to a structured welfare criterion, and explore how the resulting notion of robustness shapes optimal policy in several examples.

Throughout the thesis, I argue that understanding psychological mechanisms, and their normative consequences, is essential for designing effective policies.

Acknowledgements

So many people made this all possible. I am incredibly thankful for all the support given to me by my advisors Nava Ashraf, Daniel Reck and Johannes Spinnewijn. Throughout the PhD journey, you pushed me to find out new things about the world and myself, and enabled me to enjoy the process of doing so. Nava, thank you for emphasizing the importance of qualitative research and hearing about lived experience. Daniel, thank you for modelling a curious, calm and fun attitude towards doing research. Hannes, thank you for telling me to be my own devil's advocate and making me trust in myself that I could do it.

Whilst not official advisors, I received so much help from the brilliant community at LSE and beyond. Special thanks to Neil Thakral for generously guiding me through the process from before I even started the PhD. I also benefited greatly from insightful discussions with Gharad Bryan, Francesco Caselli, Matthias Doepke, Xavier Jaravel, Camille Landais, Joana Naritomi, Kate Smith, Alwyn Young and many more.

I also gratefully acknowledge the financial support from the ESRC and STICERD.

Thank you to my PhD colleagues and friends Arnaud Dyevre, Jack Fisher, Nilmini Herath, Amen Jalal, Michelle Rao, Hugo Reichardt, Veronica Salazar Restrepo, Pol Simpson, Patrick Schneider, Sarah Winton and the rest of the 3.12, DICE and Public Econ seminar families. I have been constantly inspired by your passion, kindness and positivity.

Gaby, Gabriel, Cat and Will–I really could not have done the PhD without you. You were so giving and you always said "yes" when I said "can I ask you something?" Thank you also for all the pub trips, bike rides, coffee breaks, dancing and easy peelers along the way.

I am grateful also to Amaara, Aoife, Becca, Callum, Georgie, Gideon, Harsh, Jaimin, Livvy, Michael, Nikhil, Rohan and Sid for all the years of unconditional friendship, supporting me through the PhD despite having no idea how long I had left and reminding me that there's more to life than research.

To all my family, your love and encouragement got me through the hardest parts. Dada, I really feel you're with me as I explore the world of the social sciences more, and Grandma, well, "where does it all come from?!" Pete & Fiona - you were always there to look after me and Cian, your level-headedness and care both encouraged and reinforced me. Mom, thank you for being where I get my passion and enthusiasm from and Pop, you have been the best teacher I've ever had.

I owe my confidence as a researcher and in myself to you, KB. Thank you for being there to chat anything through with me, sharing everything and, of course, the copy of Thinking Fast and Slow that was the beginning of this amazing journey.

To Anupriya

Contents

Declaration i

Abstract ii

List of Tables xii

List of Figures xv

Introduction 1

1 Theoretical Framework 10

1.1 Model of Social Assistance Take-up 11

1.2 Welfare 12

1.3 Identification 16

2 Empirics 21

2.1 Context and Data 22

2.2 Average Take-up Levels 29

2.3 Barrier Screening Effects 34

2.4 Benefit Take-up Effects 43

3 Calibration of Welfare Effects 53

3.1 Quantifying Sufficient Statistics 53

3.2 Quantifying Welfare Effects 58

Conclusions of the first three chapters 65

4 Intrapersonal Utility Comparisons as Interpersonal Utility Comparisons 68

4.1 Setup 74

4.2 Policy Problems and Normative Criteria 82

4.3 Comparability 90

4.4 Examples 98

4.5 Robust Optimal Policy in Applications 110

4.6 Discussion: Identifying Normative Weights. 127

4.7 Conclusion 129

Bibliography 150

A Appendix to Chapter 1 **151**

A.1 MVPF Formulae 153

A.2 Identification 155

B Appendix to Chapter 2 **158**

B.1 Context and Data 158

B.2 Average Take-up Levels: Additional Material 162

B.3 Barrier Screening Effects: Additional Material 165

B.4 Benefit Take-up Effects: Additional Material 177

C Appendix to Chapter 3 **202**

C.1 Eligibility 202

C.2 Identification 203

C.3 Relaxing Modelling Assumptions 204

C.4 Welfare Effects 208

C.5 Robustness to Bias 208

D Appendix to Chapter 4 **211**

D.1 Theory Appendix 211

D.2 Comparability 221

D.3 Perturbations 222

D.4 Examples 226

D.5 Convex Sets of Frames and Interpretation of Normative Weights . . 229

D.6 Counterfactual Normative Consumers and Identification of Normative Weights 230

List of Tables

B.1 Summary Statistics for General and Eligible Populations 159

B.2 Summary Statistics for General and Eligible Populations 160

B.3 SA Receipt vs Mental Health (different conditions) 162

B.4 SA Receipt vs Mental Health: progressively adding controls 163

B.5 SA Receipt vs Mental Health (different measures) 164

B.6 Obligations 166

B.7 Full RKD Benefit Take-up Effect Results 192

C.1 Table summarising the calibration of the key sufficient statistics . . . 203

C.2 Summary Statistics for Middle-Age Couples and Single Employees . . 206

C.3 Summary Statistics for Middle-Age Couples and Single Employees (Mental Health) 207

List of Figures

2.1 Prevalence of poor mental health in the Netherlands. 27

2.2 Equalised Access to Mental Healthcare 28

2.3 Eligibility and Receipt of SA by Mental Health 30

2.4 SA Receipt vs Mental Health (different measures) 33

2.5 SA Receipt vs Mental Health (different conditions) 34

2.6 Barrier Screening Effects by Mental Health 37

2.7 Barrier Inflow and Outflow Effects by Mental Health 38

2.8 Barrier Screening Effects: Depression vs Psychoses 40

2.9 Barrier Screening Effects by Mental Health (different measures) . . . 41

2.10 Benefits schedule as a function of Income 43

2.11 Take-up around Eligibility Threshold by Mental Health 44

2.12 Density of Income around Eligibility Threshold. 47

2.13 Take-up in Small Window around Eligibility Threshold by Mental Health 49

4.1 Illustration of the Optimal Default 113

4.2 Illustration of Optimal Reference Points 117

B.1 Take-up of SA over time 158

B.2 Coverage of the “Participation Act” over time 165

B.3 SA Receipt and Inflow over time 168

B.4 Effects of Participation Act for the Always-Eligible 169

B.5 Effects of Participation Act with/without controls 170

B.6 Expanded Sample: Including Younger Couples 171

B.7 Additional Control: Use of WMO Home Support 172

B.8 Event-Time Dynamics of Subjective Mental Health (K10) 173

B.9 Robustness to Timing of Poor Mental Health Definition 174

B.10 Always Poor Mental Health (2011–2020) 175

B.11 Buffer Sample: Income Below 90% of Threshold 176

B.12 Measurement Error of Income Concept 177

B.13 $\mathbb{E}[B|SA, Y = y]$ 181

B.14 Benefits Schedule Imputation by Mental Health 184

B.15 RKD Polynomial Simulations 187

B.16 Heterogeneous RKD Simulation 188

B.17 RKD First Stage 189

B.18 Benefit Take-up Effects 190

B.19 RKD First Stage by Mental Health 191

B.20 Benefit Take-up Effects by $\mathbb{1}\{\text{Anti-Depressants}\}$ 193

B.21 Benefit Take-up Effects by $\mathbb{1}\{\text{Anti-Psychotics}\}$ 194

B.22 Benefit Take-up Effects by Surveyed Mental Health 195

B.23 Use of Psychotropic Drugs around Eligibility Threshold 195

B.24 Density around Eligibility Threshold 196

B.25 RKD Covariate Test 197

B.26 RKD Covariate Test by Mental Health 198

B.27 RKD Permutation Test 199

B.28 RKD Bandwidth Test 200

B.29 RKD Bandwidth Test by Mental Health 201

Introduction

Poor mental health is an urgent societal issue. Almost 1 billion people live with a mental disorder [WHO, 2022]. In 2010, the economic cost of mental illness due to lost productivity and bad health was estimated to be \$2.5 trillion, and is expected to more than double by 2030 [Bloom et al., 2012]. Symptoms of mental disorders include worthlessness, confused thinking, withdrawal from support networks, fear, fatigue, guilt and, in the extreme case, suicidality [APA, 2013]. Additionally, people with poor mental health face up to three times the risk of poverty [Ridley et al., 2020]. Therefore, people struggling with mental disorders are especially vulnerable.

Modern welfare states are rooted in the principle that society should protect its most vulnerable members. Ensuring that safety net programs effectively reach those in need is essential to upholding the social contract. However, administrative and psychological costs often make it difficult to access social support, which leads to widespread non-take-up [Ko and Moffitt, 2024]. In theory, application barriers could help filter out people with lower need [Nichols and Zeckhauser, 1982], but in practice people suffering from mental illness find it more challenging to overcome take-up barriers than those with good mental health [Bell et al., 2022].

The following chapters investigate whether social assistance effectively reaches people with poor mental health. A key concern is that the very source of vulnerability is also what makes it difficult to overcome barriers to help. Nevertheless, the inefficiency arising from excluding individuals with mental disorders from assistance has remained undocumented.

Mental disorders pose important theoretical and empirical challenges in determining effective targeting. I focus on low-income welfare benefits in this study. Theoretically, eligibility for these programs is determined by people having few resources. The goal of barriers is then to target for *general* unobservable need. The challenge is that poor mental health affects the *cost* of overcoming barriers, which decreases take-up, as well as in principle the *need* for support, which increases take-up. Therefore, take-up does not distinguish between these channels, but separating them is essential for assessing effectiveness. This is because barriers target well if the needy can afford the cost and the less needy cannot.[2] Empirically, measuring mental health at scale is challenging with survey data, due to small samples and under-reporting due to stigma [Bharadwaj et al., 2017], but also with administrative data: objective outcomes are often extreme and people with poor mental health forgo care [Cronin et al., 2024].

I address these challenges in the following three chapters. First, I develop a theoretical framework to disentangle need for benefits from the cost of overcoming barriers using take-up responses to changes in benefits and barriers. Second, I empirically estimate take-up levels and responses of low-income benefits, heterogeneously by mental health, using Dutch administrative data. The data contain rich information on mental health from administrative sources and a large ($N \approx 400$k) linked survey, as well as on social assistance eligibility and take-up. Finally, I combine theory

[2] This theoretical challenge applies to the wide range of social programs where eligibility does not directly depend on mental health but the eligible population contains many people with mental disorders. The exception is disability insurance. Here, Godard et al. [2022]; Haller and Staubli [2024] emphasize that the key policy challenge with mental disorders is that they are hard for case-workers to observe.

and empirics to calculate how need and cost vary with mental health and evaluate welfare consequences of the targeting of social assistance.

The key theoretical finding of Chapter 1 is that combining differences in average take-up levels across groups with take-up responses to changes in benefits and barriers is sufficient to evaluate the marginal value of benefits (need) and the cost of barriers. To show this, I develop a theoretical framework allowing for heterogeneity in both need and cost.[3] There are three components to identification. (i) Differences in average take-up levels reflect how *average* value net of cost compares across the population. (ii) If an individual responds more to a change in benefits, either they have high need (*marginal* value) or they were at the margin of taking-up versus not (i.e. *average* value net of cost closer to 0. This can be isolated by difference in take-up levels). (iii) Once need has been separately identified through (i) and (ii), this information can be combined with take-up responses to changes in barriers to identify cost.

Identifying how the need for benefits and the cost of overcoming barriers depend on mental health is crucial for policy-making. The former is the social welfare gain from transferring €1 from someone with good mental health to someone with poor mental health. The latter reflects the welfare costs that barriers impose on individuals. Therefore, these key primitives characterise the benefits and costs of targeting social assistance using barriers. For example, need, cost and take-up responses to benefits and barriers are sufficient to calculate the welfare effects of a budget-neutral increase in barriers, where the money saved due to lower take-up is used to finance an increase in benefit level.[4]

[3] In the framework, need and cost can vary across people with the same income. Thus, need cannot be controlled for by holding income constant, creating a new identification challenge relative to past work on targeting [Finkelstein and Notowidigdo, 2019; Rafkin et al., 2023].

[4] This is an example of a policy experiment which captures the essence of how effective it is to use barriers to target [Zeckhauser, 2021; Ko and Moffitt, 2024].

My theoretical framework adopts a reduced-form revealed preference approach to identify marginal benefits and costs across groups without making strong assumptions about the underlying mechanisms. This accommodates various psychologies but assumes that perceived costs and benefits are welfare-relevant. Insofar as misperceptions exist, I conduct a robustness exercise to assess how policy conclusions change depending on the extent to which choices reflect true welfare, relying on the framework from Chapter 4.

Empirically, I study social assistance take-up and mental health using administrative data for the population of the Netherlands (17 million people). I examine the flagship Dutch social assistance program, the *algemene bijstand*,[5] a cash transfer designed for people who don't have enough money to subsist. I combine detailed information on socio-economic demographics for the years 2011 - 2020 to newly construct an accurate measure of eligibility with low measurement-error.[6] Furthermore, the data contain rich mental health information, coming from three classes of outcomes: care usage, extreme outcomes and subjective mental health from a large survey which is linked to the administrative data.[7] I combine these outcomes to reliably proxy for mental health status: this is not possible with survey or admin data alone.

Three key findings arise from my empirical analysis in Chapter 2. The first is descriptive. I find that people with poor mental health are substantially more likely to be eligible for social assistance than those with good mental health, however,

[5] Literal translation: general assistance. Information about the benefit can be found on the Dutch Government website. I will refer to this program as social assistance (SA). Social assistance is more prevalent than unemployment or disability benefits, with around 400,000 recipients every year. Eligibility is primarily determined by income being below 100% of the full-time net minimum monthly wage for couples (70% for singles).

[6] Accurately calculating eligibility is a key challenge facing the take-up literature [Ko and Moffitt, 2024]. I find that the probability of a Type-II error is small: the estimated $\mathbb{P}[SA|\text{ Ineligible}] = 1\%$.

[7] The outcomes are: care usage (mental healthcare spending, dispensations of psychotropic drugs), extreme outcomes (hospitalisations for a mental health condition, deaths by suicide) and subjective mental health from a large survey (psychological distress, loneliness and perceived control over own life).

conditional on eligibility, they take-up at the same rate. I find that one quarter of people eligible for social assistance have been diagnosed with a mental disorder, more than double the rate for the general population. However, the average take-up levels (60%) do not meaningfully differ by mental health status conditional on eligibility, income and other covariates.

Second, increases in barriers to accessing social assistance disproportionately screen out people with poor mental health. I exploit the introduction of the Participation Act [Ministerie van SZW, 2015], a policy which increased access barriers by intensifying the obligations that recipients have to satisfy and incentivising municipalities to restrict inflow [SCP, 2019], a goal they pursued through (threat of) sanctions [Ministerie van SZW, 2022]. I use a difference-in-differences design to show that the reform disproportionately discourages people with poor mental health from flowing into the program, reducing their inflow by 10% compared to those with good mental health.

Third, people with poor mental health respond twice as much to a change in benefits compared to those with good mental health. In the Netherlands, social assistance tops up income to an eligibility threshold, creating a kinked benefit schedule as a function of income (100% marginal tax rate below the threshold, 0% above). I leverage this kinked schedule as a novel instrument for benefits, made possible by my construction of the income concept used to determine eligibility. Using this source of variation in a regression kink design, I estimate elasticities of social assistance receipt with respect to benefits of 0.38 for those with poor mental health and 0.16 for those with good mental health.

Combining theory and empirical estimates in Chapter 3 yields the final key finding: people with poor mental health need benefits twice as much as those with good mental health, conditional on income, but also have a 64% higher cost of overcoming barriers. These primitives suggest that governments have an incentive to redistribute money to people with poor mental health, but that barriers are not an efficient way to do so. I estimate the marginal value of public funds [Hendren and Sprung-Keyser, 2020], capturing the direct welfare effect of the policy divided by net government cost, of a reduction in barriers as 2.34 and of an increase in benefits as 0.91. This implies reducing barriers is an effective use of government funds, 2.4× more so than increasing benefits.

Contribution to the Literature:

I contribute to the public economics literature on the targeting of government programs. There is an ongoing empirical debate on who gets screened out of assistance by barriers in terms of income and other proxies for need [Alatas et al., 2016; Deshpande and Li, 2019; Giannella et al., 2023; Homonoff and Somerville, 2021; Wu and Meyer, 2023]. Studies estimating welfare effects highlight how take-up frictions [Finkelstein and Notowidigdo, 2019] or adverse selection [Shepard and Wagner, 2022] can undermine effectiveness. The classic view from Nichols and Zeckhauser [1982] suggests that ordeal mechanisms are effective when need and the cost of ordeals are weakly negatively correlated and Rafkin et al. [2023] recently argues that self-targeting can be socially beneficial on average. However, a full cost-benefit analysis requires quantifying the trade-off between the costs of ordeals and the need for benefits—potentially extending beyond poverty—that can be redistributed to inframarginals.

My theoretical framework shows that neither take-up levels nor responses to ordeals are sufficient statistics for characterising this trade-off, precisely because need can co-vary with cost across the population. An additional moment—take-up responses to changes in benefits—is necessary to evaluate welfare implications. While this framework is broadly applicable, it also brings attention to an especially understudied dimension: mental health.

I provide one of the first quantifications of the welfare consequences of excluding people with poor mental health from assistance. Although the behavioural public policy literature has explored how mental health correlates with take-up [Arulsamy and Delaney, 2022; Bell et al., 2022; Martin et al., 2023a,b], understanding welfare effects requires assessing need. I show that individuals with mental disorders need benefits more than those with good mental health, even when conditioning on income. This highlights that vulnerability is often multi-dimensional and extends beyond poverty.

The idea that mental disorders not only increase need but also make it harder to navigate barriers mirrors the dual effects highlighted in the scarcity literature [Mullainathan et al., 2012]. Financial strain can impair cognition [Mani et al., 2013; Kaur et al., 2021], yet it can also sharpen focus and lead to better decisions [Shah et al., 2012; Fehr et al., 2022]. I propose a theoretical framework to discipline these opposing forces and implements it empirically using rich data and policy variation in benefits and barriers.

Lastly, there is a growing literature in psychology and economics studying mental disorders. Poor mental health not only imposes cognitive burden [Bierman et al., 2008; Hammar and Årdal, 2009] but also impairs emotion regulation [Gross and Muñoz, 1995], both of which hinder everyday functioning [Kessler et al., 2003; Evans et al., 2014]. In economics, studies demonstrate that mental healthcare interventions, such as psychotherapy and mindfulness, improve self-confidence, patience, risk-tolerance and reduce decision costs [Bhat et al., 2022; Shreekumar and Vautrey, 2021; Angelucci and Bennett, 2024a,b]. The literature also explores how mental healthcare affects economic outcomes [Barker et al., 2021; Baranov et al., 2020; Serena, 2024], how income impacts mental health [Christian et al., 2019; Miller et al., 2024; Schmidt et al., 2021; Silver and Zhang, 2022] and the drivers of psychotherapy demand [Abramson et al., 2024; Cronin et al., 2024; Roth et al., 2024].

I quantify the policy relevance of the cognitive and emotional burdens that mental disorders impose on individuals by empirically estimating the welfare costs of ordeals for these people. Moreover, I use a revealed-preference approach as in Deshpande and Lockwood [2022]; Haller and Staubli [2024] to show that people with poor mental health have a higher *perceived* need for welfare benefits than those without mental disorders. This new finding shows that non-take-up of assistance among people with poor mental health does not stem from under-valuation, but rather the challenges of accessing benefits.

These results support Sen's "capabilities approach" [Sen, 1999, 2008]; those facing greater daily challenges, such as disabilities, require more resources to satisfy basic needs. My analysis indicates that the same cognitive bandwidth and emotion regulation constraints that heighten the costs of overcoming barriers also appear to exacerbate everyday stressors enough to significantly raise the marginal value of additional income. These constraints plausibly affect everyone to varying degrees and have significant implications for inclusive program design, indicating that ordeals may not be effective.

Outline:

Chapter 1 sets out my theoretical framework to characterize the social welfare consequences of targeting. Chapter 2 establishes new facts about mental health and the targeting of social assistance using Dutch administrative data. Finally, Chapter 3 combines the theory and empirical estimates to calculate welfare effects.

Chapter 1

Theoretical Framework

I adapt the model from Finkelstein and Notowidigdo [2019]. I allow for heterogeneous marginal value of €1 (need), even across individuals with the same income or consumption. This generalisation is motivated by the vulnerability of people with poor mental health going beyond their risk of poverty. I propose a method for separately isolating need from the cost of overcoming barriers using take-up responses to changes in benefits and barriers.[1] The framework yields empirically-implementable formulas for the welfare effects of targeting. Proofs and extensions are in Appendix A.

[1] This distinction relates to Shepard and Wagner [2022], who show that adverse-selection can undermine ordeal-mechanisms due to the correlation between value and cost. Importantly - in their setting cost refers to cost of insurance (borne by the government), whereas I focus on the cost of ordeals (borne by the individual).

1.1 Model of Social Assistance Take-up

1.1.1 Setup

Individuals are indexed by θ.[2] Social assistance is defined by two policy parameters. B is the (monetary) benefit, Λ is the barrier that individuals have to overcome to receive B. Each θ makes one key choice: whether to receive social assistance:

$$SA = \mathbb{1}\{\text{overcome barrier } \Lambda \text{ to receive benefit } B\} \tag{1.1.1}$$

Preferences are defined as follows. Individuals derive value $v_\theta(B)$ from benefits B. There is an take-up cost $\kappa_\theta(\Lambda)$, which represents the individual-specific dis-utility from overcoming barrier Λ. I also model take-up to depend on an independent additive choice-shock $\varepsilon \sim F$ which can be thought of as decision-relevant unobservables which are unaffected by policy. Therefore, the take-up equation for each θ is:

$$SA = 1 \iff v_\theta(B) > \kappa_\theta(\Lambda) + \varepsilon \tag{1.1.2}$$

This means that behaviour follows a threshold-rule: if $\varepsilon \leq \varepsilon^*_\theta = v_\theta(B) - \kappa_\theta(\Lambda)$, $SA = 1$ and if $\varepsilon > \varepsilon^*_\theta$, $SA = 0$. Therefore, rate of receipt is given by:

$$\mathbb{P}[SA]_\theta = F\big(v_\theta(B) - \kappa_\theta(\Lambda)\big) \tag{1.1.3}$$

This model takes a stylised reduced-form revealed-preference approach, where indi-

[2] In my empirical setting, θ will represent mental health status, but the following model applies to any other dimension of heterogeneity which could influence the marginal value of €1 as well as the take-up cost.

vidual values and costs are modelled as catch-all quantities that could arise from various psychological factors and are reflected by behaviour. Given the limited evidence on welfare effects for individuals with poor mental health, simplicity is crucial. As such, I minimise structural assumptions and focus on identifying the key statistics that are sufficient for policymakers to assess targeting effectiveness.

Nevertheless, Appendix A presents a micro-foundation of $v_\theta(B)$ for completeness. Value arises from extra consumption and recovered costs of work. Income depends on take-up but is fixed otherwise: $y_\theta^{SA=1}$ refers earned-income while receiving social assistance and $y_\theta^{SA=0}$ represents earned-income when not. All income (including benefits) is taxed at marginal tax rate τ.

1.2 Welfare

1.2.1 Individual Welfare

Denote U_θ as θ's utility (which depends on take-up), and $\mathcal{U}_\theta$ expected utility. Following the setup in Section 1.1.1, I normalise utility to 0 if $SA = 0$.

$$\begin{aligned}\mathcal{U}_\theta = \mathbb{E}[U_\theta] &= \mathbb{P}[SA]_\theta \cdot \mathbb{E}[\text{Utility}|\text{ SA} = 1] + \big(1 - \mathbb{P}[SA]\big) \cdot \underbrace{\mathbb{E}[\text{Utility}|\text{ SA} = 0]}_{\text{Normalised to 0}} \\ &= \int_{-\infty}^{\varepsilon_\theta^*} [v_\theta(B) - \kappa_\theta(\Lambda) - \varepsilon] \; dF(\varepsilon)\end{aligned}$$

where $\varepsilon_\theta^* = v_\theta(B) - \kappa_\theta(\Lambda)$. Importantly, this formulation assumes rationality.[3]

[3] See Section 1.3.2 for a discussion of all key assumptions.

1.2.2 Social Welfare

Let $\mu(\theta)$ be the distribution of types, and λ_θ social welfare weights. The government's problem is given by:

$$W = \max_{\Lambda, B} \int \lambda_\theta \, \mathcal{U}_\theta \, d\mu(\theta)$$

$$\text{s.t.} \quad \underbrace{\int \tau y_\theta^{SA=0} \cdot \big(1 - \mathbb{P}[SA]_\theta\big) + \tau(y^{SA=1} + B) \cdot \mathbb{P}[SA]_\theta \, d\mu(\theta)}_{\text{Tax Revenue}} = \underbrace{\int B \cdot \mathbb{P}[SA]_\theta \, d\mu(\theta)}_{\text{Program Costs}} \tag{1.2.1}$$

In this framework, I assume eligibility criteria for benefits are fixed (though not explicitly modelled).[4] τ is also fixed. The government does not observe individuals' private types (θ, ε), making targeted policy design challenging. Instead, it must rely on blunt instruments—benefit levels (B) and barriers to access (Λ)—which do not vary by θ to indirectly target those most in need. The policy-maker's goal is to allocate benefits to individuals with a high, unobservable marginal value of benefits ($v'_\theta(B)$, i.e., need). Barriers (Λ) effectively target when neediest receive assistance, while those with lower need do not. Section 1.2.3 derives formulas for the welfare effect of an example policy experiment capturing this mechanism.[5] This is one way of characterising the effectiveness of targeting using barriers.

[4] I discuss how to explicitly model eligibility in detail in Appendix A.

[5] See, e.g. Ko and Moffitt [2024] who say that the "presence of costs induces the less needy to not apply, which saves government funds that can then be used to pay higher benefits to those in greater need, who have a higher probability of ending up as recipients."

1.2.3 Welfare Effects of a Budget-Neutral Increase in Barriers

I consider a policy experiment capturing the essence of using barriers to target social assistance: increase barriers, saving government funds due to lower take-up, in order to finance an increase in benefit level. This is a budget neutral increase in Λ (B adjusts).

Proposition 1.2.1. *The marginal welfare effect of a budget-neutral increase in ordeals financing an increase in benefits is given by:*

$$\frac{dW}{d\Lambda} = \int \lambda_\theta \; \mathbb{P}[SA]_\theta \left[\underbrace{v'_\theta(B)}_{Need} \cdot \frac{dB}{d\Lambda} - \underbrace{\kappa'_\theta(\Lambda)}_{Cost} \right] d\mu \tag{1.2.2}$$

Budget Neutrality implies:

$$\frac{dB}{d\Lambda} = \frac{-\int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} d\mu}{(1-\tau) \cdot \int \mathbb{P}[SA]_\theta \; d\mu + \int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial B} d\mu} \tag{1.2.3}$$

where:

$$FE_\theta = \tau \cdot (y_\theta^{SA=0} - y_\theta^{SA=1}) + (1-\tau) \cdot B \tag{1.2.4}$$

Equation (1.2.2) follows from an application of the Envelope Theorem. The expression shows that the overall welfare effect is large whenever take-up is high ($\mathbb{P}[SA]_\theta$ large) among the θ's with the highest need (v'_θ large) and *lowest* ordeal-costs (κ'_θ small). Analogously, $\frac{dW}{d\Lambda}$ will be negative when need and cost are strongly positively correlated.

The intuition behind Equation (1.2.3) is as follows. Budget-neutral policy changes depend on aggregate responses only. The government can increase B more if *more* people are screened out by ordeals, if people take-up *less* in response to changes in benefit level and if there are *fewer* beneficiaries at baseline. FE_θ is the fiscal externality of θ applying: there is a moral hazard fiscal externality due to labour supply response $y^{SA=0} \to y^{SA=1}$ which costs the government $\tau(y^{SA=0} - y^{SA=1})$, and a direct cost $(1-\tau)B$ paid out to θ.

The welfare effects depend on four key sufficient statistics. Increasing barriers imposes a direct cost on infra-marginal individuals: $\kappa'_\theta(\Lambda)$. However, the government saves money due to lower take-up. This depends on the strength of *barrier screening effects*, $\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}$. Increasing benefits has redistributive value for infra-marginal individuals: $v'_\theta(B)$. However, it costs the government money. This depends on the strength of *benefit take-up effects*, $\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}$.[6]

$\frac{dW}{d\Lambda}$ is my overall metric for the social welfare consequences of targeting using ordeal mechanisms. However, the units are hard to interpret. In the calibration, I use the framework of Hendren and Sprung-Keyser [2020] to aid intuition by also deriving the marginal value of public funds (MVPF) of a decrease in barriers vs an increase in benefit level. The MVPF is defined as willingness-to-pay divided by government cost (both money-metric). The formulae for the MVPF of dB and $d\Lambda$ are derived in Appendix A.1. While MVPFs have interpretable units, they do not capture θ's having different marginal values of income unless social welfare weights are included. Therefore, I also calculate MVPFs with utilitarian (rather than money-metric) social welfare functions. The comparison of utilitarian $MVPF_{dB}$ and $MVPF_{d\Lambda}$ is isomorphic to $\frac{dW}{d\Lambda}$.

[6] The policy experiment differs from Rafkin et al. [2023] in what we compare ordeal-costs $\kappa_\theta(\Lambda)$ to. They consider moving to automatic enrolment (comparing $\kappa_\theta(\Lambda)$ to 0) whereas I consider reducing barriers (comparing $\kappa_\theta(\Lambda)$ to $\kappa_\theta(\Lambda - \delta\Lambda)$). Hence, I require variation in barriers as well as benefits.

1.3 Identification

How should we empirically characterise the welfare consequences of targeting using barriers? Proposition 1.2.1 is an example showing that in order to know whether barriers target effectively, we must estimate four key "sufficient statistics": need (v'_θ), cost (κ'_θ), benefit take-up effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}\right)$ and barrier screening effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}\right)$.[7]

My goal is to quantify these statistics empirically. Therefore, it is helpful if there are as few as possible. First, I use theory to reduce the number of sufficient-statistics from 4 to 3. The key idea is that benefit take-up effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}\right)$ depend on the marginal value of benefits, i.e. need (v'_θ). Similarly, barrier screening effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}\right)$ depend on the cost of barriers (κ'_θ).

Remark 1.3.1. *Barrier screening effects are characterised by Equation (1.3.1), and benefit take-up effects by Equation (1.3.2).*

$$\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} = -\kappa'_\theta \cdot f_\varepsilon(v_\theta - \kappa_\theta) \tag{1.3.1}$$

$$\frac{\partial \mathbb{P}[SA]_\theta}{\partial B} = v'_\theta \cdot f_\varepsilon(v_\theta - \kappa_\theta) \tag{1.3.2}$$

Intuitively, Λ is a price of taking up. Therefore, responsiveness to take-up is large when consumers are price-responsive (κ' large) or just at the margin of take-up $\big(f_\varepsilon(\cdot)$ large$\big)$. Similarly, responsiveness to a change in benefit level is governed by need (v') and the probability of being marginal. This means that there are only three key primitives which determine welfare effects: need, cost and $f_\varepsilon(v_\theta - \kappa_\theta)$, the likelihood of being on the margin of take-up. The latter is an scaling factor

[7] Therefore, my model aligns with the sufficient-statistics approach to public economics [Einav and Finkelstein, 2011; Baily, 1978; Chetty, 2008].

allowing for the inference of infra-marginal costs/benefits through marginal take-up responses.

1.3.1 Three-step Identification

In this section, I present a three-step method to identify the three key statistics sufficient for evaluating welfare effects. The method takes as inputs: take-up levels $\mathbb{P}[SA]_\theta$, barrier screening effects $\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}$ and benefit take-up effects $\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}$ and uses these to identify need (v'_θ), cost (κ'_θ) and the likelihood of being marginal ($f_\varepsilon(v_\theta - \kappa_\theta)$). The intuition is as follows:

Difference in take-up levels $\mathbb{P}[SA]_\theta$ across types cannot distinguish between average value (v_θ) and cost. However, they reflect how average value *net* of cost compares across types. This, in turn, influences how $f_\varepsilon(v_\theta - \kappa_\theta)$ compares across types. Using this information, cost can be inferred from barrier screening effects $\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}$. The idea is that $\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}$ being large reflects either large κ' or average value net of cost being close to zero. The latter can be isolated using difference in take-up levels. Similarly, the contribution of $f_\varepsilon(v_\theta - \kappa_\theta)$ to benefit take-up effects $\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}$ can be isolated from need.

Step 1 (Average take-up levels): To aid intuition, suppose that we are in a special case of equalised take-up levels: $\mathbb{P}[SA]_\theta = \mathbb{P}[SA]_{\tilde{\theta}}$.[8] I.e. $F(v_\theta - \kappa_\theta) = F(v_{\tilde{\theta}} - \kappa_{\tilde{\theta}})$. Then, $f(v_\theta - \kappa_\theta) = f(v_{\tilde{\theta}} - \kappa_{\tilde{\theta}})$ because the cdf F is monotonic. More generally if $\mathbb{P}[SA]_\theta \neq \mathbb{P}[SA]_{\tilde{\theta}}$, $f(v_\theta - \kappa_\theta)$ is identified in terms of $f(v_{\tilde{\theta}} - \kappa_{\tilde{\theta}})$ using a first-order Taylor expansion of difference in average take-up levels $\mathbb{P}[SA]_\theta - \mathbb{P}[SA]_{\tilde{\theta}}$.

[8] This is motivated by the empirical application, where I find no meaningful difference in average take-up levels by mental health.

This requires additional structure, and is set out in Appendix A.2. At the end of **Step 1**, we know how $f(v_\theta - \kappa_\theta)$ compares across types.

Step 2 (Benefit take-up effects): If we know how $f(v_\theta - \kappa_\theta)$ compares across types, and estimate benefit take-up effects for each type - then we can quantify how need varies across types. This done by dividing Equation (1.3.2) *across types* to give:

$$\frac{\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}}{\frac{\partial \mathbb{P}[SA]_{\tilde{\theta}}}{\partial B}} = \frac{v'_\theta}{v'_{\tilde{\theta}}} \cdot \underbrace{\frac{f_\varepsilon(v_\theta - \kappa_\theta)}{f_\varepsilon(v_{\tilde{\theta}} - \kappa_{\tilde{\theta}})}}_{\text{Estimated in } \textbf{Step 1}} \tag{1.3.3}$$

Then, if we normalise $v'_{\theta_0} = 1$ for some θ_0 we calculate v'_θ for all other θ using Equation (1.3.3). This normalization is without loss, and effectively scales all welfare effects in terms of θ_0's WTP for €1.[9]

Step 3 (Barrier screening effects): Finally, divide barrier screening effects from Equation (1.3.1) by benefit take-up effects from Equation (1.3.2) *within type* to identify κ'_θ for all θ as follows:[10]

$$\frac{-\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}}{\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}} = \kappa'_\theta \cdot \underbrace{\frac{1}{v'_\theta}}_{\text{Estimated in } \textbf{Step 2}} \tag{1.3.4}$$

[9] v'_θ is then understood as θ's need relative to θ_0.

[10] This within-type identification method is the same as the method used in Haller and Staubli [2024], and echoes the identification of the marginal rate of substitution in Russo [2023]. The across-type identification is new. Here, the key novelty is that I can estimate take-up levels with information on eligibility and use these to inform differences in likelihood of being on the margin of take-up across types.

1.3.2 Discussion of Key Assumptions

Before presenting the empirical analysis, it is important to discuss the key assumptions underlying the identification of need and cost. In Chapter 3, I return to these assumptions and characterise how relaxing them impacts welfare effects.

I assume that ε is an additive independent shock to the take-up equation: $SA = 1 \iff v_\theta > \kappa_\theta + \varepsilon$. Independence, as assumed in random utility models [McFadden, 1981; Woodford, 2020], enables **Step 1** in the identification. The assumption can be probed by examining how including additional covariates changes the three-step identification.[11] Without independence, the model is not identified and either v_θ constant across θ or κ_θ constant across θ must be assumed. Seeing as the purpose of the framework was to separate need and cost when both could depend on mental health, neither of these cases is useful.[12] Additivity allows me to separate need/cost from the scaling factor $f_\varepsilon(v_\theta - \kappa_\theta)$ in Equations (1.3.1) and (1.3.2).

The framework assumes that the likelihood of being on the margin of take-up, $f(v_\theta - \kappa_\theta)$, is the same for benefit and barrier instruments. This comes from θ and ε being one-dimensional. The assumption is called into question by recent work arguing that the compliers to an instrument depend on the instrument itself [Kline and Walters, 2019; Mogstad et al., 2024]. This assumption allows for minimal structure on the take-up equation. Relaxing $f(v_\theta - \kappa_\theta)$ to depend on the instrument is possible under additional parametric assumptions as long as $f(v_\theta - \kappa_\theta) = f(v_{\tilde{\theta}} - \kappa_{\tilde{\theta}})$ for all $\theta, \tilde{\theta}$, i.e. as long as the difference in complier characteristics across instruments is orthogonal to mental health. Appendix C shows welfare effects in this case.

In the theory, θ is treated as an immutable type, but in practice mental health evolves

[11] Throughout my empirical analysis, including additional covariates does not meaningfully change the comparison between good and poor mental health, providing support for $\varepsilon \perp \theta$.

[12] In settings where it seems reasonable that $v_\theta \perp \theta$ or $\kappa_\theta \perp \theta$, models from Finkelstein and Notowidigdo [2019]; Rafkin et al. [2023] should be used.

over time and in response to stimuli. This assumption is made in order to set out a tractable static framework. In Section 2.4, I show that social assistance does not appear to have a strong dynamic positive effect on mental health. However, ordeals likely worsen mental health, a dynamic I cannot quantify in this project. This effect would imply that the welfare costs of increasing ordeals are a lower-bound.

Finally, I use a revealed-preference framework. Assuming rationality allows me to reveal need and cost from take-up responses, and to use the Envelope Theorem when deriving welfare effects. In Chapter 3, I use a specific case of the general framework of Chapter 4 to characterise how confident the government needs to be about bias to reverse the estimated sign of the welfare effects.

Chapter 2

Empirics

In this chapter, I examine the relationship between mental health and social assistance take-up using administrative data covering the full population of the Netherlands from 2011 to 2020. I focus on the algemene bijstand (social assistance), the Netherlands' primary cash transfer program for individuals with insufficient income. A key component of the analysis is the construction of an accurate measure of eligibility, combining detailed income, household, and demographic information to minimize measurement error. To measure mental health status, I integrate administrative records on mental healthcare usage and extreme outcomes with survey-based measures of psychological distress, loneliness, and perceived control, enabling a richer proxy than either data source alone would allow.

First, Section 2.1 explains the context and data. Then, in Section 2.2 I use these data to establish some descriptive facts about mental health and social assistance in the Netherlands. I estimate the effects of a policy which increases barriers in Section 2.3. Finally, I estimate the take-up response to exogenous variation in benefits using a regression kink design in Section 2.4.

2.1 Context and Data

2.1.1 Institutional Context

Social Assistance in the Netherlands

In the Netherlands, social assistance, or *algemene bijstand*, is a non-contributory social safety net program. It is intended for individuals who do not have enough income or assets to subsist, and who do not qualify for any other benefit. Over the period of this study, around 450,000 people claim benefits each year. This translates to around 4.5% of the adult population and is more than the number of people on disability and unemployment insurance. Figure B.1 shows the evolution of caseload from 2005 to 2021.

Eligibility: Eligibility rules are determined at the national level. The benefits are means-tested: income and assets must be below a threshold in order to be entitled. The income threshold is 100% of the full-time national minimum wage for couples, and 70% for singles. The threshold depends on household composition. Income includes not just labour income, but from capital and other benefits.

Additionally, eligibility requires being at least 18 years old and Dutch citizenship or residing lawfully in the Netherlands, not in prison or a detention center. Mental health does not directly affect eligibility.

Application: Applicants must submit information to verify eligibility (e.g. residency proof, income / bank statements etc) as well as potentially go to the municipal office for an interview. The municipality legally must make a decision within 8 weeks of application.

Receipt: If accepted, income is topped-up to the eligibility threshold - i.e. there is a 100% marginal tax rate.[1] The national minimum wage, and couples' threshold, is around €16.5k per year during the observation period. Often, people earn some income - on average, benefits paid equal around €12.7k per year. Conditional on receipt, people stay on social assistance for around 5 years - there is no time-limit to take-up. Municipalities can grant additional benefits, such as housing, health insurance and children subsidies. In this chapter, I focus on the take-up of the general welfare benefit.[2]

Obligations: Social assistance is a workfare program: conditional on take-up, recipients must comply with several obligations. These include keeping all information up-to-date and work re-integration.[3] Single parents with young children and people with full and permanent incapacity to work can apply for an exemption from these obligations. In the event of non-compliance, municipalities can impose sanctions or (temporarily) reduce benefits. Exclusion from assistance is an uncommon, extreme outcome.

[1] Basic income experiments have been trialled in some municipalities, where some treatment arms reduce/remove obligations and other treatment arms reduce the 100% claw-back of benefits [Verlaat and Zulkarnain, 2022]. Strict obligations are rationalised by wanting to incentivise activation and eventual transitioning out to paid work in the face of the 100% marginal tax rate.

[2] This is a reasonable simplification because the take-up of these additional benefits is uncorrelated with receipt of social assistance, after controlling for income and wealth [Berkhout et al., 2019]. Furthermore, these subsidies are phased-out according to different schedules to social assistance.

[3] Full list of obligations can be found in Ministerie van SZW [2015]. They include acceptance of work or voluntary activities (i.e. "participate"), wearing the correct clothing doing so, being prepared to travel a distance with a total travel time of 3 hours per day to find work, keeping all eligibility and benefit-level information up-to-date, complying with information requests and even home-visits, being willing to relocate municipality, achieving a good command of the Dutch language and acquiring and retaining knowledge and skills necessary for acquiring wealth.

Healthcare in the Netherlands

The Netherlands has a mandated and subsidised private health insurance system. GPs are the first port-of-call for mental health issues, and can prescribe medications or refer to specialized care. In the general population, around 10% of people are dispensed with psychopharmacalogical medications each year. Access to mental healthcare appears to be roughly equalised by income (see Figure 2.2 below), although quality of care may differ [Lopes et al., 2023].

Disability Insurance in the Netherlands

One potential concern about my analysis is that perhaps it's not social assistance people with poor mental health should be receiving, but disability insurance. However, disability benefits count towards eligibility for social assistance. Insofar as people receive full disability benefits (e.g. people with severe mental disorders), they have income above the social minimum, are ineligible for social assistance and do not appear in my main analysis. Moreover, disability insurance is a contributory program replacing past earnings after work-limiting health shocks. Many people receiving social assistance do not have prior work history, so are ineligible for disability benefits. In sum, those with moderately poor mental health are in the target population.

2.1.2 Data

In order to quantify the nature of selection of SA recipients with respect to mental health, I use administrative data from the population of the Netherlands ($\approx$ 17 million people) accessed via CBS, the Statistics Agency of the Netherlands. The data contain information on socio-economic demographics determining eligibility for social assistance, rich characteristics on social assistance receipt and comprehensive information about mental health.

Socio-economic information:

I create a new dataset containing eligibility for social assistance in the years 2011-2020 for all working-age individuals in the Netherlands. To do so, I extend the work of Inspectie SZW [2021] to calculate eligibility by merging detailed information on socio-economic information, including income, wealth, household composition and size, work status, education and other demographics, following the rules set out by law [Ministerie van SZW, 2015]. These data are yearly, and so the measure reflects eligibility on average each year.[4] The main analysis sample is from 2011 to 2020 because I observe all eligibility determinants in this period. I focus on working-age individuals throughout the study - age 27-65.[5] Among this population, around 10% of people are eligible for social assistance each year. Table B.1 shows summary statistics about the socio-economic demographic variables, for the general population and for the eligible.

The administrative data show receipt of social assistance (among other benefits) for each person in each month, as well as benefits received, which household-composition-dependent threshold has been applied, any income earned, exemptions and sanctions. I use these data to calculate the take-up rate of social assistance - defined as $\mathbb{P}[\text{Take-up SA}|\text{Eligible}]$. Over the study period, the take-up-rate is around 60%, in line with Inspectie SZW [2021]. I find $\mathbb{P}[\text{Take-up SA}|\text{Ineligible}] = 1\%$, suggesting

[4] I also calculate eligibility monthly for people who work - as the data contain monthly income information for employees. I use this for the regression kink design in Section 2.4.

[5] As in Inspectie SZW [2021], eligibility for students and people above pension-age is noisier and so these groups are excluded.

low measurement-error.

Mental health information:

Finally, the data contain three classes of mental health measures: take-up of mental healthcare (mental healthcare expenditures and dispensations of psychotropic medications by ATC4-code), extreme outcomes (hospitalizations with ICD-10 codes corresponding to mental health disorders, and deaths by intentional self-harm–suicides), and surveyed psychological distress (Kessler's 10), loneliness and perceived control over own life (in a linked survey for 400k people in 2012 and 2016). Table B.2 shows summary statistics about (mental) health.

Figure 2.1 shows the prevalence of poor mental health in the Netherlands, and how this varies when focusing on the general population, those eligible for social assistance and recipients. The figure shows that the eligible are at least 2.5× more likely to suffer with poor mental health than the general population. Whereas, social assistance recipients seem to have similar mental health to the eligible population.

This suggests limited self-targeting; if self-targeting were effective, the more vulnerable group would be over-represented among recipients. This implies that people with mental disorders either do not value benefits more than those with good mental health or that they do but find barriers costlier to overcome. Chapter 1 says the first step to distinguishing between these explanations is the difference in average take-up levels.

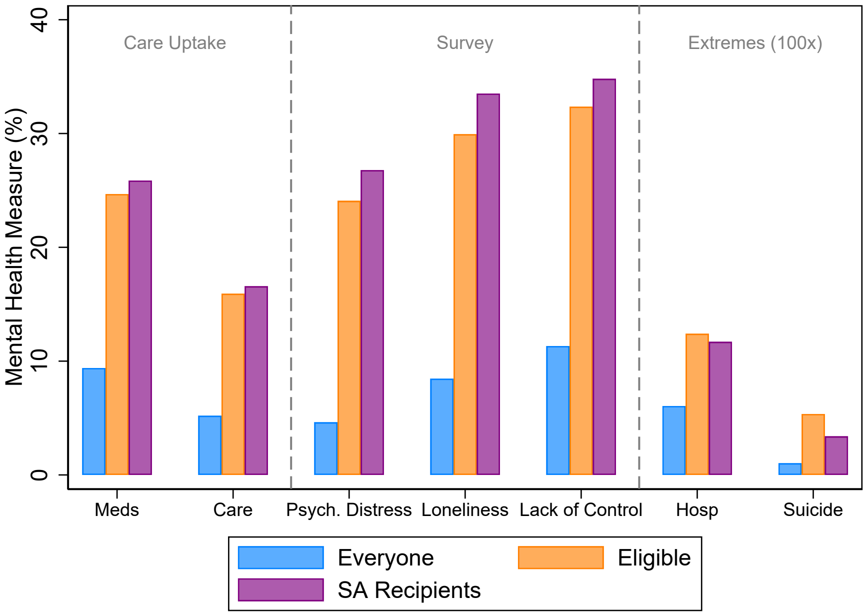


Figure 2.1: Prevalence of poor mental health in the Netherlands.

Notes: This graph shows raw means of the seven mental health measures across three populations. All measures are percentages: the probability of dispensed psychotropic medications, any mental healthcare spending, severe psychological distress, loneliness, or perceived lack of control, and (artificially inflated by 100×) hospitalization and suicide. Populations are: all individuals, those eligible for social assistance, and recipients, 2011–2020.

2.1.3 Key Analysis Variables

In the rest of the chapter, I empirically analyse the take-up of social assistance heterogeneously by mental health. Throughout, I define take-up as $SA_{it} = \mathbb{1}\{i$ receives SA in period $t\}$. For almost all results, this will refer to a stock. How should we measure poor mental health? I define Poor $\text{MH}_{it} = \mathbb{1}\{i$ dispensed psychotropic medications in year $t\}$.

Figure 2.2 shows that this is an accurate proxy for poor mental health status. In the Netherlands, usage of mental healthcare is strongly positively correlated with subjective psychological distress, and the relationship between the two does *not* depend on income. Why? Access to healthcare in the Netherlands is excellent,

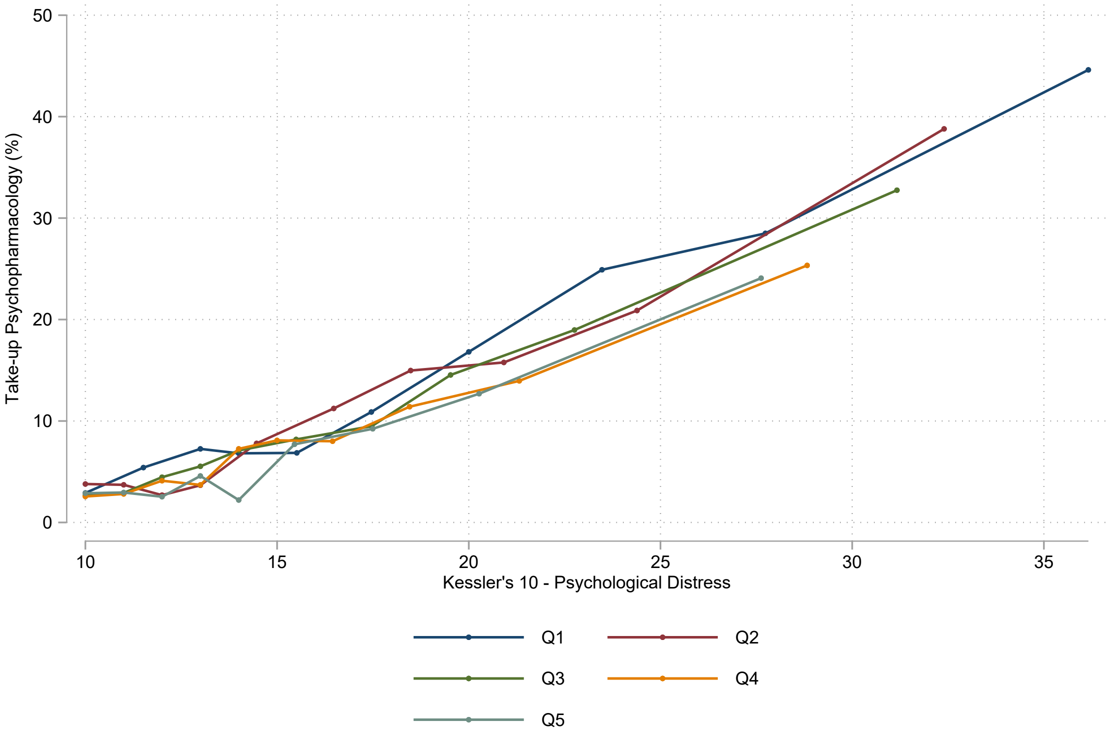


Figure 2.2: Equalised Access to Mental Healthcare

Notes: This graph shows the the relationship between two measures of mental health, split by income quintile. The y-axis displays probability of dispensed psychotropic medications, and x-axis subjective psychological distress (from the survey). The lines show bin-scatters of the mean of psychotropic drug use by income quintile and level of subjective mental health. Population: survey population.

prescriptions are done by GPs, who are the first access point to healthcare and people often still receive care even if they default on their premia [Roos et al., 2021]. Indeed, 0.4% of *poor* households report unmet medical needs in the Netherlands, relative to 8.5% of *all* households in the US [Danesh et al., 2024].

Of course, even in the Netherlands there will be some non-take-up of mental healthcare by people with truly poor mental health. Therefore, throughout the empirical analysis I verify that all findings about mental health measured by dispensations of psychotropic drugs are consistent when mental health is measured in the survey.

2.2 Average Take-up Levels

Average levels of the take-up of social assistance by mental health are useful descriptives to examine targeting and important inputs to identification of need and cost.

First, in terms of raw levels, figure Figure 2.3 shows the baseline probability of being eligible for social assistance by mental health, measured by psychotropic drug dispensation, as well as the take-up levels conditional on eligibility. People with poor mental health are three times more likely to be eligible, but conditional on eligibility seem to take-up around the same rate as those with good mental health.

2.2.1 Design

Do people with poor mental health take-up social assistance more or less than people with good mental health, conditional on eligibility and income (and other observables)? This is **Step 1** in the three-step identification from Section 1.3.1.

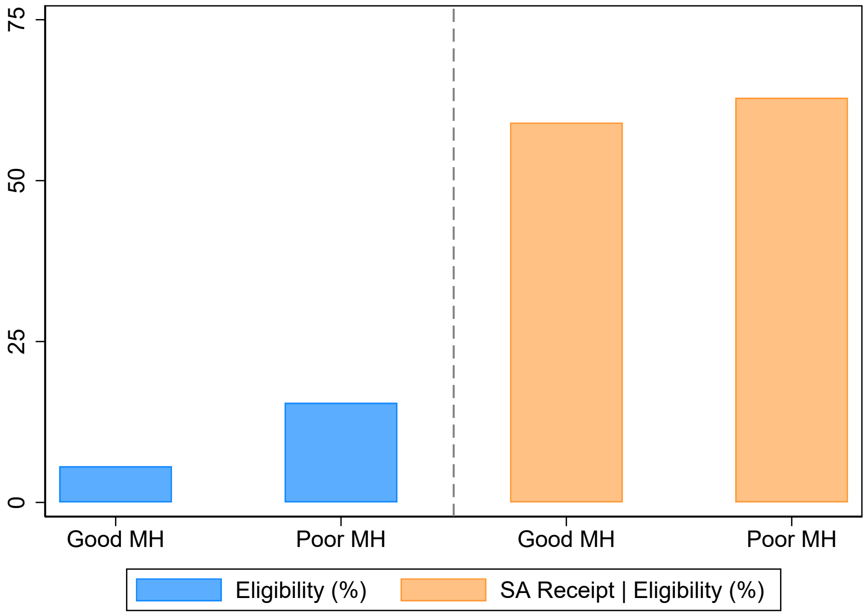


Figure 2.3: Eligibility and Receipt of SA by Mental Health

Notes: $\mathbb{P}$[Eligible] and $\mathbb{P}[SA|\text{Eligible}]$, compared for people with poor mental health (dispensed psychopharma in year previously) vs good mental health (not). Underlying population: 2011-2020 in each case.

For each individual i and year t, define $SA_{it} = \mathbb{1}\{i \text{ receives SA in year } t\}$. Poor $\text{MH}_{it} = \mathbb{1}\{i \text{ dispensed psychopharma. in year } t\}$. Equation (2.2.1) represents a *correlation test* measuring the overall extent of selection.[6]

$$SA_{it} = \beta \cdot \text{Poor MH}_{it-1} + X'_{it-1}\theta + \varepsilon_{it} \tag{2.2.1}$$

X_{it} are flexible controls of income,[7] wealth, education, hh composition, work status, work sector and year, age, gender and municipality fixed effects. ε_{it} is an idiosyncratic error term. β measures the selection of social assistance recipients with respect to mental health and is the coefficient of interest.

[6] Throughout, I use a linear-probability model, but the results are not substantially different using logit or probit.

[7] I include household standardised income percentile (moving average $t-4 \rightarrow t-2$) fixed effects.

Importantly, I estimate the correlation test on the *eligible* population. Higher overall take-up rates by a group could come from higher probability of being eligible, or more frequent receipt conditional on eligibility. I focus on the latter in this thesis because non-take-up by ineligible individuals is not attributable to the main forces of interest - need and ordeal costs.[8]

2.2.2 Results

Table B.4 shows the main results using Poor $\text{MH}_{it} = \mathbb{1}\{i$ dispensed psychopharma. in year $t\}$. Aligned with Figures 2.1 and 2.3, I find that people with poor mental health take-up social assistance only slightly more than those with good mental health. The table shows estimates of β, for seven specifications of increasing saturation.

Throughout, people with poor mental health have similar take-up rates to those with good mental health. After conditioning on year, age, and gender fixed effects, the difference in take-up between groups ranges from around -1 to +1p.p, depending on inclusion of additional controls. Estimates are statistically significant but economically small. In the full specification (Column 5), people with poor mental health have social assistance receipt rates less than 1% higher than those with good mental health, holding all else constant.

Two controls explain a large portion of the variation. Adding lagged income controls increases R^2 from 0.05 to 0.16, as people with poor mental health have different income levels, which determine benefits-level entitlement. Adding lagged work status (including past social assistance receipt) further increases R^2 from 0.16 to 0.64, with

[8] Indeed, Muilwijk-Vriend et al. [2019] show that $\hat{\beta}$ is positive for the overall population, but of course, this could be due to people with poor mental health being more likely to be eligible.

$\hat{\theta}_{SA_{t-2}=1} = 42.35$, showing strong autocorrelation in social assistance take-up. The positive $\hat{\beta}$ with individual fixed effects supports this.

The increase in the point estimate when controlling for lagged work status suggests social assistance may improve mental health, as it indicates that part of the initial negative association between assistance receipt and mental health stems from people who have consistently not used social assistance having poorer mental health. Once past work status is controlled, the analysis shows that among those with similar social assistance histories, individuals with poorer mental health are more likely to use assistance now, suggesting social assistance could help mitigate some mental health challenges.

How does $\hat{\beta}$ compare to the coefficients on other covariates, $\hat{\theta}$? The income percentile fixed effects range from 20 to -20, age fixed effects from 0 to 10 and municipality fixed effects from -15 to 5. This reinforces the idea that $\hat{\beta}$ is economically small.

Figure 2.4 presents the results of the correlation test varying the measure of poor mental health. $\hat{\beta}$ are plotted for each mental health status variable: $\mathbb{1}\{$dispensed of psychotropic meds$\}$, $\mathbb{1}\{$positive mental healthcare costs$\}$, $\mathbb{1}\{$Hospitalized for a mental health condition$\}$, and surveyed $\mathbb{1}\{$Severe psychological distress$\}$, $\mathbb{1}\{$Severe lack of control over own life$\}$, and $\mathbb{1}\{$Severe loneliness$\}$, relative to average take-up amongst those with good mental health. Qualitatively, these estimates are broadly consistent with each other and show a small positive difference in rate of receipt by people with poor mental health vs people with good mental health. Table B.5 shows the full results.

Of course, there is a broad spectrum of different mental disorders. Using psychotropic drug dispensations to measure poor mental health provides a practical approach to distinguish these differences. Figure 2.5 shows coefficients from a re-

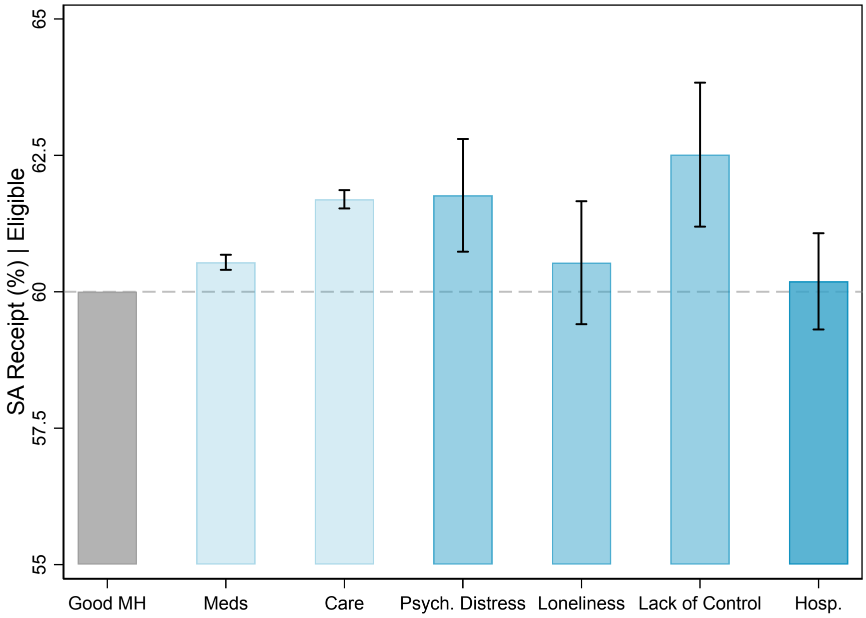


Figure 2.4: SA Receipt vs Mental Health (different measures)

Notes: Coefficients of social assistance take-up regressed on mental health status–indicators of: psychotropic drugs, mental healthcare, severe surveyed psychological distress/loneliness/lack of control over own life, or mental health hospitalisation. Point estimates added to the control mean, with 95% confidence intervals. Lagged controls include income, wealth, education, work status, household composition, municipality, year, age, sector fixed effects, physical health, and benefits schedule. Eligible population from 2011 to 2020. Standard errors clustered at the municipality level.

gression of SA receipt on dummies for *type* of psychotropic drug dispensed, and all controls. People using ADHD medication, hypnotics / sedatives and anxiolytics are no more likely to receive social assistance than those not using any psychotropic medications. Anti-depressant dispensation is associated with a higher rate-of-receipt, whereas anti-psychotic dispensation is associated with a *lower* rate-of-receipt. Table B.3 shows the full results.

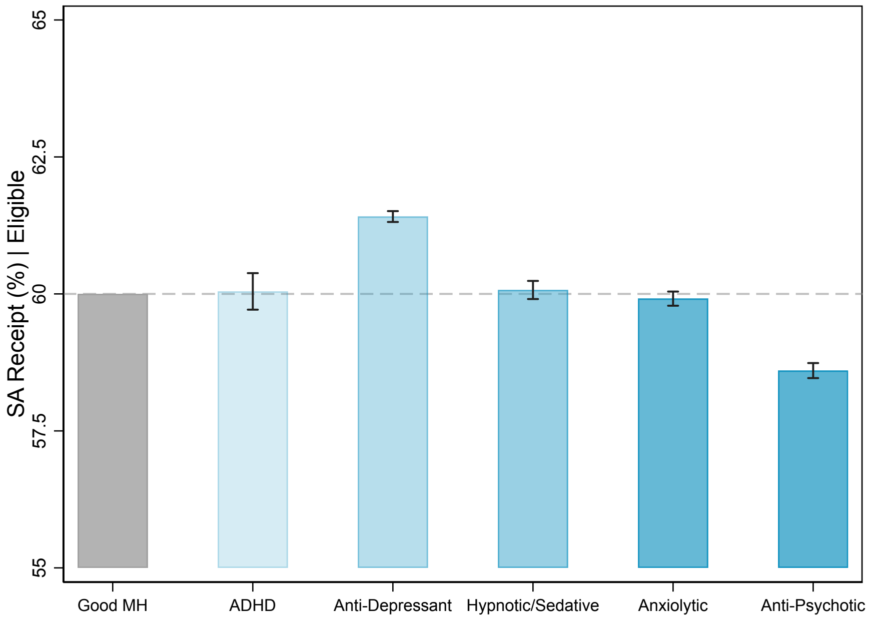


Figure 2.5: SA Receipt vs Mental Health (different conditions)

Notes: Coefficients of social assistance take-up regressed on psychopharmacology dispensation fixed effects (by type: ADHD medications, anti-depressants, hypnotics/sedatives, anti-anxiety medications and anti-psychotics). Point estimates added to the control mean, with 95% confidence intervals. Lagged controls include income, wealth, education, work status, household composition, municipality, year, age, sector fixed effects, physical health, and benefits schedule. Eligible population from 2011 to 2020. Standard errors clustered at the municipality level.

2.3 Barrier Screening Effects

Moving past average take-up levels, recall from Chapter 1 that we need to identify take-up responses to changes in barriers to assistance to reveal how costly people with poor mental health find overcoming these barriers and thus start to examine targeting effectiveness.

I examine the effects of the Participation Act, a large reform to social assistance design in the Netherlands, which increase barriers to access. The policy was announced in 2014 (with significant public discourse - see Figure B.2) and implemented in January 2015. The reform was a response to rising caseloads following the Great Financial Crisis, and it cut municipal social assistance budgets from €1.4 billion in 2010 to around €500 million by 2018 [Heekelaar, 2021]. The Participation Act intensified obligations for recipients and incentivized municipalities to restrict inflow through (threat of) sanctions [SCP, 2019; van der Veen, 2019].[9]

2.3.1 Identification

I exploit the Participation Act to estimate the heterogeneous take-up response to a change in barriers by baseline mental health. Practically, the specification, Equation (2.3.1), is a standard Difference-in-Difference design with people poor mental health as the treatment group. The interpretation of the treatment effects is the heterogeneous effect $\frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda} - \frac{\partial \mathbb{P}[SA]_H}{\partial \Lambda}$. The identification assumption is that people with poor mental health's receipt would have evolved in parallel to those with good mental health.[10]

$$SA_{it} = \alpha + \eta_i + \gamma_t + \delta_t \times \text{Poor MH}_i + X_{it}'\theta + \varepsilon_{it} \tag{2.3.1}$$

η_i and γ_t are individual and year fixed-effects respectively. X is a vector of time-varying controls including income, education and muncipality, hh composition and sector fixed effects. δ_t for $t \geq 2013$ are the coefficients of interest and represent the heterogeneous treatment effect of the policy by baseline mental health. Poor $\text{MH}_i = \mathbb{1}\{i$ dispensed psychotropic drugs at some point in the pre-period (2011 - 2014)$\}$. Throughout, I cluster standard-errors at the level of municipality of residence in 2013.

[9] For more details, see Appendix B.3.1.

[10] The formal parallel trends assumption is that the receipt of social assistance by those affected by the policy would have evolved in the same way as a (purely hypothetical) control group who did not experience the policy, for every level of baseline mental health [de Chaisemartin and D'Haultfœuille, 2023; Shahn, 2023].

I estimate Equation (2.3.1) for eligible middle-aged couples (ages 45-65), for whom the policy represents a clean exogenous increase in barriers only. I focus on couples because the eligibility threshold for single parents was cut in 2015, incentivising reclassification as single households. The take-up of social assistance pre/post 2015 by younger individuals is contaminated by inflow from a youth disability program (Wajong), where people with poor mental health are likely over-represented.[11] I focus on the eligible because the take-up responses for this group can be attributed to the change in barriers, not underlying changes in eligibility.[12] Sensitivity analyses confirm that the findings remain robust across various specifications and assumptions, as detailed in Section 2.3.5.

2.3.2 Main Results

Figure 2.6 shows the estimates $\hat{\delta}_t$ according to Equation (2.3.1). The groups are on parallel trends before the policy announcement, giving confidence to the identification assumption. [13]

The Participation Act disproportionately screens out people with poor mental health. The effect starts when the Act is announced, and then is especially pronounced in 2015. The overall difference-in-difference estimate of $\frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda} - \frac{\partial \mathbb{P}[SA]_H}{\partial \Lambda} \approx -1$p.p. This is comparable in magnitude (but opposite sign) to Finkelstein and Notowidigdo [2019] who estimate that barriers to SNAP in the US screen-out low-earnings-potential types 2 percentage points *less* than high-types.

[11] Figure B.6 shows the results when including adults aged 35-45. Although this group is more contaminated by Wajong entrants, the results are similar, suggesting the main estimates are not driven by Wajong entrants.

[12] The effect holds also for those who are 'always-eligible' (eligible throughout 2011-2020), providing confidence that the main results are not driven by eligibility churn.

[13] Note: as the policy happens in the aftermath of the GFC, I expect $M \ll 1$ in the framework of Rambachan and Roth [2023]. In this case, statistically insignificant pre-trends are meaningful.

What do we learn from this through the lens of the theory? Section 2.2 suggests that take-up levels are roughly equalised on average. Therefore people with poor mental health have similar average value of benefits net of average cost of ordeals to people with good mental health. The barrier screening effects estimated show ordeals are more costly for those with poor mental health.[14]

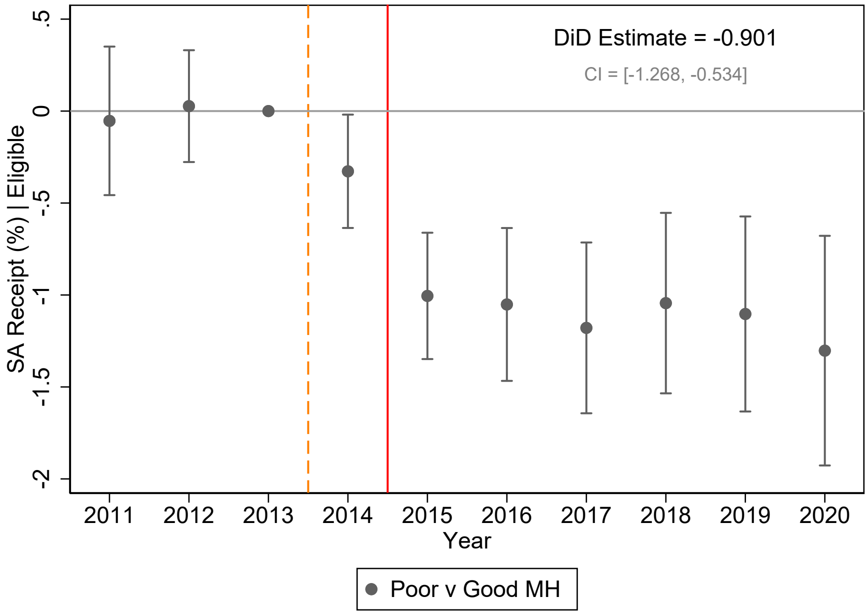


Figure 2.6: Barrier Screening Effects by Mental Health

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in pre-period. Controls include individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. The TWFE estimate $\hat{\delta}$ in the regression $SA_{it} = \alpha + \eta_i + \gamma_t + \delta \cdot \mathbb{1}\{t \geq 2013\} \times \text{Poor MH}_i + X'_{it}\theta + \varepsilon_{it}$ is also shown. Standard-errors are clustered at the level of municipality of residence in 2013.

2.3.3 Mechanism

Figure 2.7 shows the effects on inflow and outflow. Outflow is not mechanically zero, the estimates are just tiny and have tight confidence intervals. This shows

[14] The idea is that **Step 1** shows the likelihood of being marginal is similar across types. Therefore, large heterogeneous barrier screening effects are informative of differences in cost.

that the main results comes exclusively from a deterrence of inflow, aligning with Cook and East [2024] who suggest work requirements can screen-out individuals at the extensive margin. The disproportionate reduction in inflow for people with poor mental health (1p.p.) is around 10% of the baseline control mean (see Figure B.3).

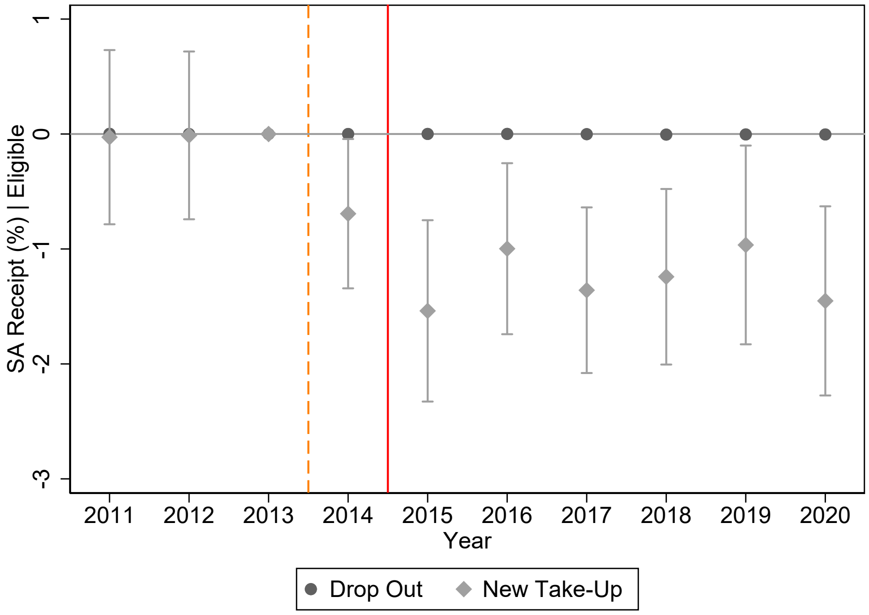


Figure 2.7: Barrier Inflow and Outflow Effects by Mental Health

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. Here, I split by inflow (receipt conditional on being ineligible last period), and drop-out (non-receipt conditional on receipt last period). The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in pre-period. Controls include individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

The reduction of inflow suggests people with poor mental health are deterred by increased obligations and also from more unpleasant interactions with the municipality, given incentives to reduce caseload. Qualitative evidence from [Ministerie van SZW, 2022] supports the latter mechanism. The authors state that beneficiaries experience a "feeling of shame" and highlight a "negative tone" from the municipality where "small event[s] can have major consequences". This creates "mutual distrust" and "fear" which creates a "barrier to applying for assistance, even when the need is great".[15] Lack of outflow corroborates SCP [2019] who find no effect of the Participation Act on transition into paid work.

2.3.4 Different Mental Health Measures

How do the results compare when mental health is measured differently or across different disorders? Figure 2.8 shows that the Participation Act screens-out those using anti-psychotics twice as much as those using anti-depressants. Figure 2.9 shows that the barrier screening effects are more pronounced when poor mental health is measured by surveyed severe psychological distress. This likely reflects the fact that the main estimates are a lower-bound since some mental disorders are not diagnosed. Taking these results together suggests that severity of mental disorders—rather than simply their presence—exacerbates the dis-utility from ordeals.

2.3.5 Robustness

The main results of Section 2.3.2 are robust to several threats to identification. First, the sample consists of couples eligible for social assistance each year, a population that changes over time due to eligibility churn from income fluctuations and the inflow of individuals with youth disabilities. This raises concerns that the main result might be driven by differential take-up rates among new entrants and exiters to eligibility. However, Figure B.4 shows that the results are consistent for the always-eligible population—couples who remain eligible throughout the sampling period, suggesting that the main results are not driven by differential selection of

[15] Translated from page 8 of Ministerie van SZW [2022]. See Appendix B.3.1 for full quote.

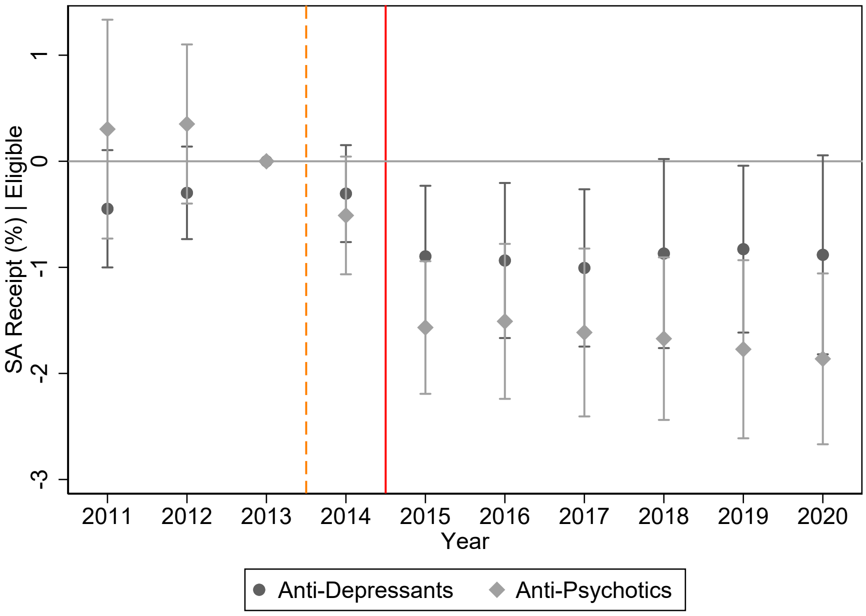


Figure 2.8: Barrier Screening Effects: Depression vs Psychoses

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. Here, Poor MH_i can now take 3 values: 0 (control), 1 (anti-depressants) or 2 (anti-psychotics). The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in pre-period. Controls include individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. I plot the estimate $\hat{\delta}_t^{\text{Dep.}}$ and $\hat{\delta}_t^{\text{Psycho.}}$. Standard-errors are clustered at the level of municipality of residence in 2013.

the newly eligible.[16] The point estimates being smaller for the always-eligible is unsurprising: these people are less likely to be on the margin of take-up.

Appendix B.3 presents a formal presentation of the sample-selection issue. Consequently Figure B.5, which shows that the estimates are virtually unchanged when removing all time-varying covariates, is further evidence that eligibility flows do not drive the results.

Secondly, there could be contemporaneous policy changes which affect the take-up of social assistance heterogeneously by mental health. One threat is a reform to the structuring of long-term care (WMO) [Kromhout et al., 2018]. The remit of home

[16] While this may seem like a stark restriction, 25% of the eligible are always-eligible.

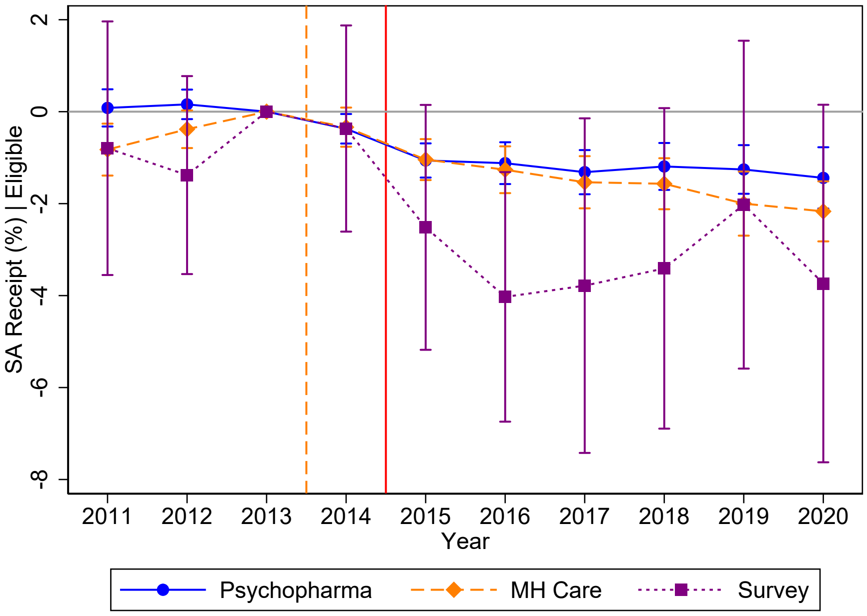


Figure 2.9: Barrier Screening Effects by Mental Health (different measures)

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. Here, Poor MH$_i$ is defined in 3 ways: dispensations of psychotropic drugs in pre-period, $>$ 0 mental healthcare costs in pre-period, surveyed severe psychological distress in 2012. The analysis population is eligible middle-age couples. Controls include individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

support for people with mental health issues was changed to be under the remit of municipalities starting in 2015. Figure B.7 shows the WMO reform does not drive the results.

The main results are not driven by inflow from Wajong (an income-support program for those experiencing disability shocks before age 18), which merged into the Participation Act in 2015. My sample restricts to individuals above age 45 to conservatively exclude those who might have transitioned from Wajong to social assistance. The only way this group could contaminate the sample is if they experienced a disability shock at 18, did not take up Wajong, survived without income support until age 45, and then opted for social assistance. Figure B.6 shows that the estimates are unchanged when including adults aged 35-45, confirming that the age restriction effectively controls for the potential contamination.

Thirdly, the Participation Act could have affected people with poor and good mental health differently due to its differential implementation, particularly the change in how eligibility was calculated based on household composition. To control for this, I include flexible controls for household size, and Figure B.5 shows that this does not drive the results.

Fourth, a concern is that the observed heterogeneous treatment effects could be due to pre-existing differences in take-up rates, rather than baseline mental health. However, when splitting the poor mental health group into subgroups—moderate (anti-depressants and anti-anxiety) and severe (anti-psychotics)—we find that both groups' take-up rates diverge from the good mental health group after 2015. Prior to 2015, the severely poor mental health group had lower receipt levels to the good mental health group, supporting the hypothesis that mental health differentially affects responses to barriers, and not simply pre-policy take-up levels.

The groups are defined based on pre-period dispensations, so we might worry that the $\hat{\delta}$'s are capturing the effect of mental health treatment on social assistance receipt. However, Figure 2.9 shows that results when defining poor mental health based on self-reported symptoms, not prescriptions, are even stronger. Additionally, Figure B.8 shows that people dispensed drugs exhibit significantly worse mental health both before and after dispensation, with scores consistently above the threshold for moderate mental illness.

A related concern might be that the results could reflect long-term positive effects on social assistance take-up following a mental health shock. However, Figure B.9 shows that even when mental health is defined after the policy, there is a noticeable drop in

$\hat{\beta}$ in 2014 that persists over time. Additionally, Figure B.10 presents similar findings when poor mental health is defined as continuous psychopharma dispensations in all years from 2011 to 2020 (compared to none in any year). This group likely suffers from chronic mental illness, reinforcing confidence that the main results are not merely capturing the effects of a one-time mental health shock.[17]

2.4 Benefit Take-up Effects

In the final empirical part of the chapter, I estimate the take-up response to exogenous variation in benefits. I leverage quasi-experimental variation in the benefit-level by exploiting the kinked benefits schedule as a function of income with a regression kink design (RKD) as in Card et al. [2015]. The statutory benefits schedule is displayed in Figure 2.10.

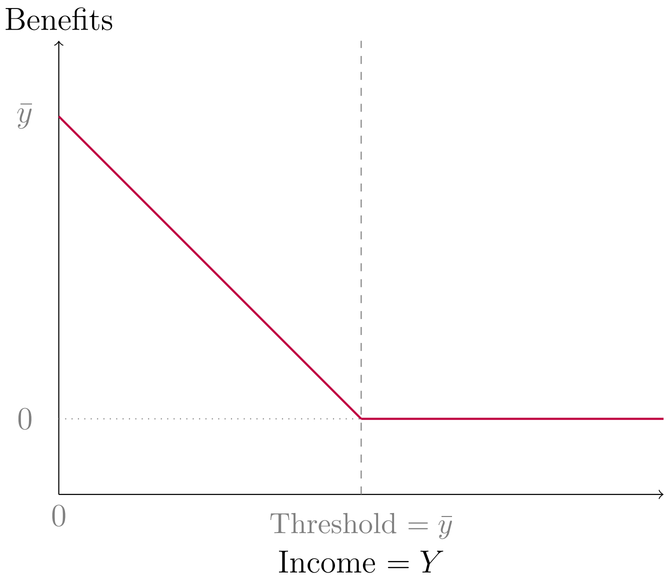


Figure 2.10: Benefits schedule as a function of Income

[17] Overall, these findings suggest that being prescribed psychopharma at any point is a consistent indicator of mental health status, which remains significantly worse than the good mental health group throughout the study period.

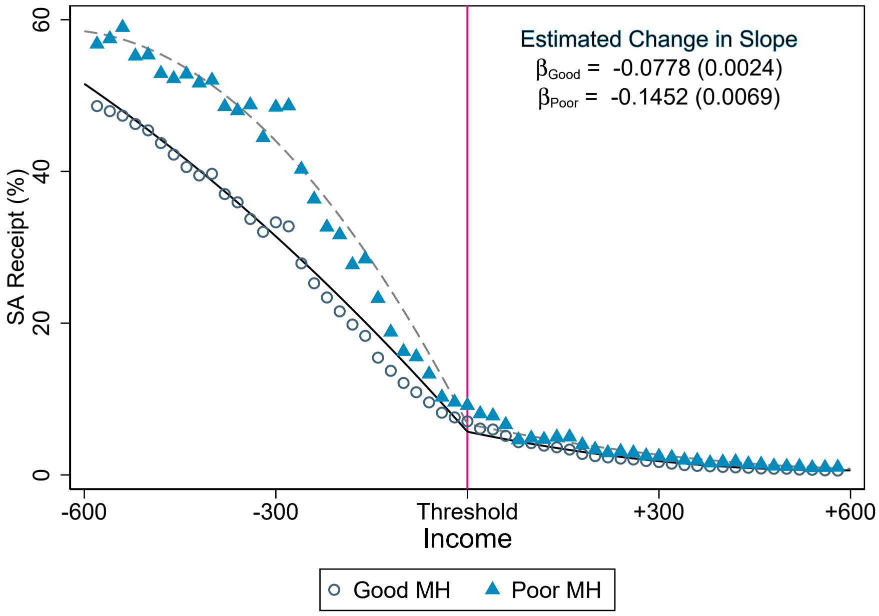


Figure 2.11: Take-up around Eligibility Threshold by Mental Health

Notes: Average rate of receipt within income slice in a large window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as receiving psychopharma in the year previously. The sample contains single employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions.

Before diving into the details of identification, Figure 2.11 shows graphical evidence of the behavioural response to a change in benefit level by poor vs good mental health. The figure indicates that people suffering from mental disorders take-up *more* in response to increasing benefit-levels than those with good mental health. I plot take-up within slices of *monthly* income.[18] The take-up functions diverge starkly at the threshold for poor vs good mental health. This is confirmed by fitting polynomials on either side of the threshold and testing for differences in their slopes. The results show that the slope change is almost twice as large for individuals with poor mental health compared to those with good mental health.

People with income above the threshold take-up primarily due to measurement error:

[18] Granular analysis is critical - hence the switch to monthly data. We can reconcile the findings in Figure 2.11 with the small difference in average take-up levels estimated in Section 2.2 by recognizing that the overall results are largely driven by the 75% of eligible individuals who do not work. The RKD, however, is a LATE capture effects locally around the eligibility threshold.

some sources of income do not count towards the eligibility threshold. Therefore, I need to calculate the income concept used to determine eligibility. This Y_{calc} differs from Y_{true} because (a) some income information (e.g. from other benefits) is only recorded yearly, yet eligibility determined monthly. Unemployment spells are imputed. (b) Y is aggregated to the family level. Families are 1 or 2 adults (+ kids) who live together and share costs. The latter is unobservable. Y_{true} is observed for the selected sample: recipients.

To minimise attenuation from measurement error, I focus on single employees.[19] The data contain monthly income information for employees–minimising error due to (a) and singles are immune from issues in (b). Figure B.12 shows a histogram of $Y_{\text{true}} - Y_{\text{calc}}$ for the analysis population. Y_{true} is negatively selected for recipients, so we expect the distribution to be left-skewed. Measurement error has significant mass around 0 and both mean and median are small (-€51, -€13 respectively).

2.4.1 Identification

Theory

Figure 2.11 measures $\frac{d\mathbb{P}[SA]}{dy}$ for income y. In order to retrieve the take-up response $\frac{d\mathbb{P}[SA]}{dB}$, we need to re-scale by $1/\frac{dB}{dy}$. The statutory benefits schedule would imply $\frac{dB}{dy} = -1$ below threshold, and 0 above.

There is a challenge: municipalities can deviate from the policy formula through income exemptions - some or all of y is ignored when calculating B. Appendix B.4.1

[19] Details of the estimation are in Appendix B.4. Around the threshold, couples are significantly mismeasured because I cannot observe which adults live together as part of a family and which don't. This does not drive the barrier screening effects. Figure B.11 shows that the results remain the same when focusing on individuals away from the threshold. Unfortunately, this does mean that internal validity concerns restrict the samples differently for barrier screening and benefit take-up effects. This does not affect results about *relative* need and cost by mental health, as discussed in Chapter 3.

sets out my theoretical approach to impute the ex-ante benefits schedule (i.e. the expected benefits a potential applicant is eligible for conditional on their income) accounting for exemptions. Figure B.14 shows the results.

The imputation process is not perfect: it measures the ex-ante benefits schedule with error. Let B^* be the imputed (mis-measured) version of B: $B^* \triangleq B + U_B$. As discussed above, Y is also measured with error: $Y^* \triangleq Y + U_Y$. Therefore, I use a fuzzy RKD specification [Card et al., 2015]. Proposition B.4.1 shows that a fuzzy RKD estimates a weighted average of marginal effects of B on $\mathbb{P}[SA]$.

Estimation

I estimate a standard fuzzy RKD specification, using local linear regression. I use a Calonico et al. [2014] robust bandwidth of €60, estimated separately for people who are (/not) dispensed psychotropic drugs in the year previously (poor/good mental health, respectively). The IV estimate $\frac{\hat{\beta}_1}{\hat{\delta}_1}$ measures $\frac{\partial \mathbb{P}[SA|Y=\bar{y}]}{\partial B}$. Standard-errors are clustered at the municipality level.[20]

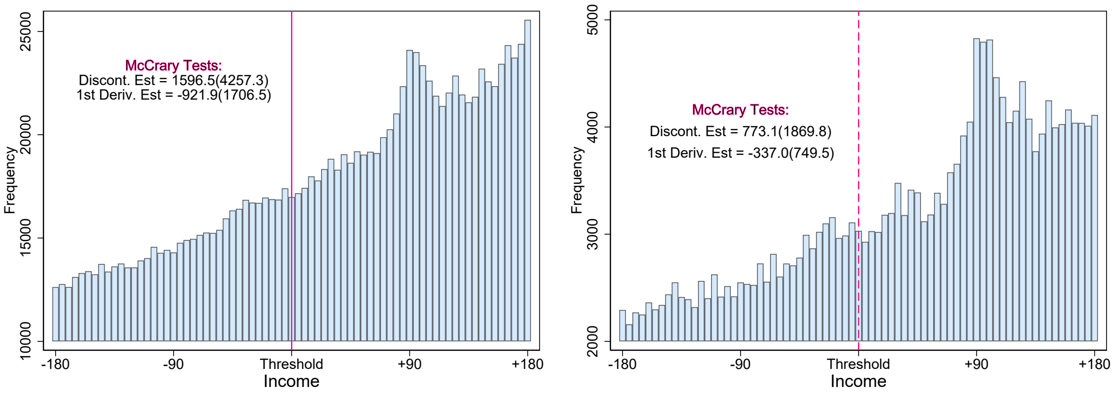
$$SA_{it} = \alpha + \beta_0 \cdot (y^*_{it} - \bar{y}_i) + \beta_1 \cdot \min\{y^*_{it} - \bar{y}_i, 0\} + \varepsilon_{it} \qquad \textbf{(Reduced Form)}$$

$$B^*_{it} = \gamma + \delta_0 \cdot (y^*_{it} - \bar{y}_i) + \delta_1 \cdot \min\{y^*_{it} - \bar{y}_i, 0\} + \varrho_{it} \qquad \textbf{(First Stage)}$$

Intuitively, the fact that the first-stage is also estimated on the mis-measured running variable y^*_{it} "accounts" for measurement-error as in Card et al. [2015].

[20] See Appendix B.4.2 for more details.

Support for Identification Assumptions: The key identification assumption is that there is no manipulation of income around the threshold. Figure 2.12 and Figure B.24 shows no evidence for strategic income targeting around the eligibility threshold: McCrary [2008] tests with seventh-order polynomials show no statistically significant bunching. Although the threshold equals the full-time monthly minimum wage, the sample works much less than full-time (around 100 hours per month) and income used for eligibility does not only come from labour. Adjustment frictions are likely a key reason for lack-of-bunching [Kleven, 2016].



(a) Good Mental Health **(b)** Poor Mental Health

Figure 2.12: Density of Income around Eligibility Threshold.

Notes: McCrary [2008] tests for discontinuity in levels and slopes around the threshold are shown. Income is monthly. Poor mental health is defined as receiving psychopharma in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions.

2.4.2 Main Results

First, I pool people with good and poor mental health together. Figure B.18 shows that employees react significantly to the quasi-experimental variation in benefit-level. I estimate $\hat{\beta}_1 = -0.0258$ which translates to take-up increasing by $\approx 2.6 p.p.$ for a €100 increase in the benefit level.

People with mental disorders have a two-times larger take-up response to a change in benefits than those with good mental health. The results are shown in Figure 2.13. I estimate $(\hat{\beta}_{1H}, \hat{\beta}_{1L}) = (-0.0218, -0.0508)$.[21] Measurement error is uncorrelated with mental health status - there is no statistically distinguishable difference in the slope above the threshold between good and poor mental health. Using the first stage in Figure B.19 to re-scale the above reduced-form and account for measurement error, we obtain IV estimates (and associated confidence intervals):

$$\text{Estimate of } \frac{\partial \mathbb{P}[SA|Y=\bar{y}]}{\partial B} = \frac{\hat{\beta}_1}{\hat{\delta}_1} = \begin{cases} \underset{[0.0080,0.0374]}{0.0227} \quad \text{p.p.} \quad \text{for Good MH} \\ \\ \underset{[0.0164,0.0842]}{0.0503} \quad \text{p.p.} \quad \text{for Poor MH} \end{cases} \tag{2.4.1}$$

Translating these to take-up elasticities with respect to changes in benefits yields 0.16, 0.38 for good and poor mental health, respectively. The elasticity for good mental health is on the lower-end of the range of previously estimated take-up elasticities of social insurance [Krueger and Meyer, 2002; McGarry, 1996], whereas for poor mental health lies within-range. The full set of reduced-form and IV estimates (with and without controls) are contained in Table B.7.

Robustness

I assess the credibility of the design with standard robustness analyses whose results are described in Appendix B.4.7. Figures B.25 and B.26 show no statistical evidence of selection along observable characteristics around the kink. Indeed, Table B.7

[21] Whilst these estimates are somewhat noisy, recall that the Calonico et al. [2014] robust bandwidth does not take into account measurement-error, nor the efficiency of estimating *heterogeneous* treatment effects across groups. Here, the zoomed-out version in Figure 2.11 gives us confidence that the take-up response is indeed twice as large for those with poor mental health. Moreover, Figures B.28 and B.29 confirm that for less extreme restrictions to the bandwidth, the estimates are very similar, but more precise.

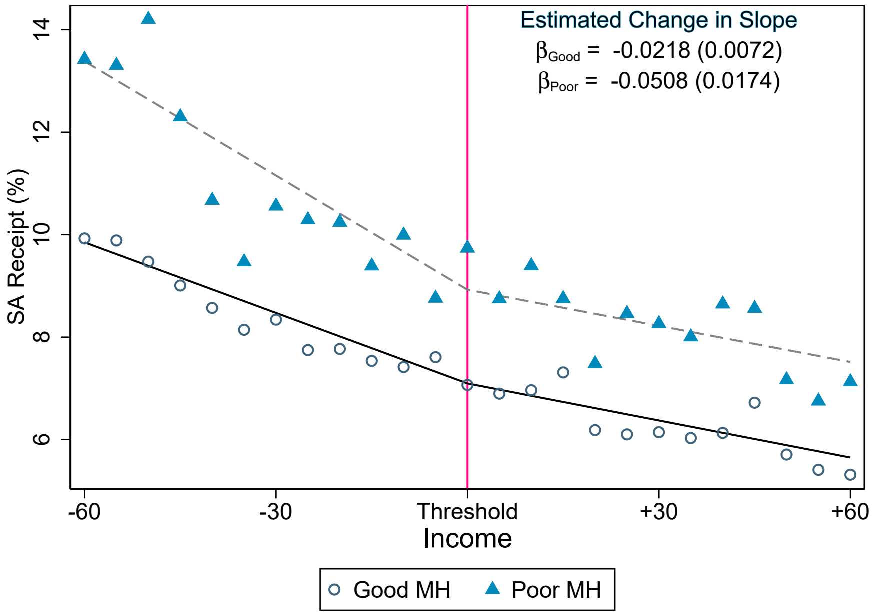


Figure 2.13: Take-up in Small Window around Eligibility Threshold by Mental Health

Notes: Average rate of receipt within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as receiving psychopharma in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as estimated change in slopes from the regression kink design. Standard-errors are clustered at the municipality level.

shows that the addition of a rich set of covariates does not meaningfully affect the results. Figure B.27 displays a permutation test [Ganong and Jäger, 2018], and shows no evidence for worrying non-linearities. Figure B.28 and Figure B.29 explore sensitivity of the results to different bandwidths. Estimates are quite robust to lower bandwidths overall, and point estimates do not vary much in the heterogeneous case despite the confidence intervals overlapping with lower bandwidths.

Different Mental Health Measures

How do the results compare when mental health is measured differently or across different disorders? Figure B.20 and Figure B.21 show the take-up response to a change in benefits estimated for those dispensed anti-depressants and anti-psychotics, respectively, in each case relative to those with good mental health. The IV estimates are 0.0704 for anti-depressants and 0.120 for anti-psychotics. This suggests that the magnitude of the take-up response to a change in benefits is increasing in severity of mental disorder.

Moreover, Figure B.22 shows that the larger take-up response for those with poor mental health is present when poor mental health is alternatively measured by severe psychological distress reported in the survey. Therefore, responsiveness of take-up to changes in benefits seems to reflect general psychological distress, rather than treatment.

2.4.3 Mechanisms

Why do people with poor mental health react more to a change in benefits, conditional on having the same income? Similarly to Section 2.3, this larger sensitivity despite similar average take-up levels reflects higher need, i.e. a larger marginal value of income from social assistance. There are two main reasons why this might be the case.

First, cash transfers improve mental health [Haushofer et al., 2020]. If people with poor mental health anticipate the protective effect of social assistance income on their mental health, it could cause them to value €1 more than people with good mental health and thus have a higher behavioural response.

However, I find no strong support for this mechanism in my setting. The reduced-form RKD induces exogenous variation in social assistance receipt, which I then regress future psychotropic drug dispensations on to estimate $\frac{\partial MH}{\partial SA}$. Figure B.23

shows a (noisy) 0. I cannot rule out social assistance improving mental health,[22] but it does not seem to be the main driver. This is perhaps not surprising - Miller and Bairoliya [2023]; Silver and Zhang [2022] also do not find strong evidence that cash improves mental health. Indeed, [Solmi et al., 2022] argue many mental illnesses start early in life.

Instead, I interpret these results through the psychology literature studying mental disorders. This literature often refers to the impairment of everyday functioning as a key mechanism in the difficulties this population face. Of course, the cognitive burden of mental illness, including effects on information processing, attention, memory and executive function can clearly hinder psycho-social functioning [Kessler et al., 2003; Evans et al., 2014]. Mental disorders can also affect everyday functioning through impaired emotion regulation - this can affect work, relationships and self-image [Gross and Muñoz, 1995].

The cognitive burden and emotion resilience tax imposed by mental disorders seemingly increases the difficulty handling common everyday stressors amongst the low-income population, thus increasing the value of support.[23] This idea aligns closely with Amartya Sen's "capabilities approach" [Sen, 1999, 2008]; people with poor mental health need more income to get by. Section 2.4.2 shows that the results are consistent when poor mental health is measured through surveyed psychological distress. Moreover, the estimate associated with anti-psychotics—typically prescribed for more severe mental health conditions—is larger than that for anti-depressants. These findings are consistent with the idea that the increased need for assistance

[22] The descriptive results of Section 2.2 where the point estimate on the correlation test increases when controlling for lagged social assistance receipt.

[23] Indeed, people with poor mental health work less than those without, and limits on earnings capacity are indicative of higher marginal utility of benefits [Deshpande and Lockwood, 2022]. An economic model of scarcity would resonate closely with this interpretation [Mullainathan and Shafir, 2013]. Given limited mental resources, people with poor mental health will have a higher value of releasing resources through additional money compared to people with good mental health with the same initial income.

stems from the common component of psychological distress inherent in poor mental health, characterized by impaired cognition and emotion regulation.

Perhaps most interestingly, this higher need is estimated through revealed-preference. Not only do people with poor mental health need benefits more, they *think* that they need benefits more. This suggests that impaired functioning seems to dominate anhedonia and other psychological mechanisms lowering the perceived value of help. If anything, this is likely an under-estimate of true need given pessimism characterises depression, one of the most common mental disorders. I return to this point in the welfare calibration.

Chapter 3

Calibration of Welfare Effects

In this chapter, I plug empirical results from Chapter 2 to the model of Chapter 1 to quantify need for benefits and cost of overcoming barriers, heterogeneously by mental health. The key primitives from Section 2.2, Section 2.3 and Section 2.4 determine the effectiveness social assistance targeting using barriers. For example, I calculate the welfare effects derived in Proposition 1.2.1. To be clear, this is not the only way of measuring effectiveness. However *any* measure will need to trade-off the differential need for benefits by people with poor mental health against differential cost of overcoming barriers.

The sufficient statistics for these welfare effects are need (v'_θ), cost (κ'_θ), benefit take-up effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}\right)$ and barrier screening effects $\left(\frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda}\right)$.

3.1 Quantifying Sufficient Statistics

For the calibration, I assume $\theta \in \{L, H\}$: mental health is either poor or good. Throughout the empirical sections, I examine take-up conditional on eligibility.

However, welfare-estimates should reflect the general population - for example because the government budget constraint should reflect the fact that the ineligible fund benefits for the recipients, and not the eligible non-takers. Appendix C.1 shows how to rescale take-up levels and responses estimated on the eligible to reflect the overall population. The tax rate $\tau \approx 37\%$

3.1.1 Identifying Need and Cost

I employ the three-step identification method set out in Section 1.3.1. Appendix C.2 shows the full set of results of this calibration. First, Section 2.2 shows no meaningful difference in average take-up levels conditional on eligibility between poor and good mental health. Therefore, I apply the special case of **Step 1**, where equalized take-up levels implies equalized likelihood of being at the margin of take-up.

$$f_\varepsilon(v_L - \kappa_L) = f_\varepsilon(v_H - \kappa_H)$$

This result reflects the fact that *average* value net of cost seems to be roughly the same across mental health states. However, this does not necessarily pin down *marginal* value (need) - nor does it separate between need and cost.

First, I normalize $v'_H = 1$. As discussed in **Step 2**, this effectively scales need by the willingness-to-pay for €1 amongst people with good mental health. Moreover, it means that the benefit take-up response for people with good mental health measures $f_\varepsilon(v_H - \kappa_H)$. To match the theory, I re-scale the response estimated in Section 2.4 by $(1-\tau)$, the net-of-tax rate, because in the theory B is understood as gross benefits, whereas the regression kink design estimates responsiveness to net benefit level. I estimate $\frac{\partial \mathbb{P}[SA]_H}{\partial B} = f_\varepsilon(v_H - \kappa_H) = \underbrace{0.63}_{1-\tau} \times \underbrace{0.000227}_{\text{Estimate from RKD}}$.

Need: I apply **Step 2** and divide the benefit take-up response for people with poor mental health by the response for good mental health. The above implies $\frac{f_\varepsilon(v_L-\kappa_L)}{f_\varepsilon(v_H-\kappa_H)} = 1$. This, combined with $v'_H = 1$ shows that need for benefits for people with poor mental health is revealed as the *relative* benefit take-up response for this group. I estimate $\frac{\partial \mathbb{P}[SA]_L}{\partial B} = 0.63 \times 0.000503$, which therefore implies $v'_L = 2.22$.[1]

Cost: Finally, I use the difference-in-differences results of Section 2.3 to calibrate $\kappa'_\theta(\Lambda)$. I use the raw descriptive drop in inflow for people with good mental health (see Figure B.3) to calibrate $\frac{\partial \mathbb{P}[SA]_H}{\partial \Lambda} = -0.014$. The main results of Section 2.3 thus imply $\frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda} = -0.023$. I then apply **Step 3** of the identification method. These results are combined with $f_\varepsilon(v_L - \kappa_L) = f_\varepsilon(v_H - \kappa_H) = 0.63 \times 0.000227$, and imply that that $\kappa'_H = 98$ and $\kappa'_L = 161$.

3.1.2 General Policy Implications

These estimates suggest that people with poor mental health have more than twice as high a marginal value of additional income (need) versus those with good mental health. The differences are not explained by differences in income, as the regression kink design estimates $\frac{\partial \mathbb{P}[SA]_\theta}{\partial B}$ conditional on income being equal to the eligibility threshold for both groups. This implies a strong redistributive motive towards people with poor mental health. Moreover, need is revealed from behaviour, suggesting that policy instruments to transfer income to those with poor mental health can be effective in practice.

The fact that people with poor mental health need money more, but take-up at the

[1] In this section, I use the regression kink design estimates of the take-up response to a change in benefits with the Calonico et al. [2014] robust bandwidth of €60. In Appendix C I explore how welfare consequences change when alternatively using the magnitudes estimated with a wider bandwidth as in Figure 2.11.

same rate as those with good mental health suggests they are inefficiently excluded from social assistance by barriers.

The quantities also imply ordeals impose a 64% higher cost on those suffering from mental disorders relative to people with good mental health. This suggests that although governments have an incentive to redistribute money towards people with poor mental health, barriers to access are a costly way target. These statements are formalised with a specific example policy experiment in Section 3.2.

3.1.3 General Psychological Implications

Taking the results together, my findings suggest that people with poor mental health have both a higher marginal value of additional income and a higher marginal cost of administrative and psychological barriers. The observed patterns in Section 2.3 and Section 2.4 indicate that these differences reflect general psychological distress rather than being driven by specific characteristics of particular conditions or the effects of mental health treatment.

Heterogeneous take-up responses, as proxied by dispensations of psychotropic drugs, are amplified when poor mental health is measured through surveyed psychological distress and increase in magnitude with disorder severity (e.g., anti-psychotics are associated with higher take-up responses than anti-depressants). This pattern suggests that the higher need for and cost of accessing assistance reflect a common underlying component of poor mental health across disorders.

The psychology literature identifies impairments in cognitive function and emotion regulation as central mechanisms underlying this common component of poor mental health [Bierman et al., 2008; Hammar and Årdal, 2009; Hyman et al., 2006; Gross and Muñoz, 1995]. My results align with a model in which deficits in these areas not only exacerbate the challenges of navigating access barriers to social assistance but also impede everyday functioning in other domains, such as work and social life [Kessler et al., 2003; Evans et al., 2014]. These impairments likely explain why the marginal value of additional income support is particularly high for this group.

3.1.4 The Role of Bias

What if individuals are biased? In this case, take-up responses reflect *perceived* need and cost, rather than true need or cost. On the cost side, the baseline model suggests substantial out-screening by barriers reveals large costs for the infra-marginals, but in reality κ'_θ calibrated above can be thought of "as-if" costs of barriers [Goldin and Reck, 2022a] which could differ from the truth. The costs are revealed by a policy which intensified obligations and increased unpleasant and stigmatising interactions with the municipality. The policy disproportionately deters people with poor mental health from flowing-in to the benefit. Outflow does not change differentially.

These findings imply that perceived costs may overestimate actual costs. While the psychological impact of fear and shame, as discussed by Ministerie van SZW [2022], is likely non-negligible, the absence of differential outflow suggests that compliance costs from obligations may be over-stated. Thus, while barriers could actually target well, welfare outcomes hinge critically on the extent to which perceived costs diverge from true costs. I formalize this argument in Section 3.2.2 for the specific policy experiment studied, following Chapter 4.

On the need-side, pessimism is a common symptom of depression [Alloy and Ahrens, 1987], a common mental disorder. This suggests that the disproportionate perceived need for people with poor mental health is likely an *under-estimate*. Bias likely only increases the welfare effects of redistributing income to people with poor mental health.

3.2 Quantifying Welfare Effects

In this section, I calibrate the welfare effects of marginal changes in benefits and barriers as a function of the sufficient statistics. In the data, the prevalence of poor mental health conditional on eligibility is $\mu(L) = 0.25$.[2] I set $\mathbb{P}[SA]_L = \mathbb{P}[SA]_H = 0.6$. I start from the baseline case of no social welfare preference for poor mental health. This means that the heterogeneous monthly net fiscal externalities $FE_\theta = \tau(y^{SA=0} - y^{SA=1}) + (1-\tau)B$ are, on average:

$$FE_L = 0.37 \times (€512.22 - €331.27) + (1 - 0.63) \times €972.22 = €679.45 \quad (3.2.1)$$

$$FE_H = 0.37 \times (€574.29 - €390.95) + (1 - 0.63) \times €916.29 = €645.09 \quad (3.2.2)$$

The fiscal externality of inducing someone with poor mental health to apply is larger than for good mental health. People with poor mental health receive more benefits than those with good mental health because they earn less when on social assistance. Here, the fact that $y_L^{SA=0} \approx y_H^{SA=0}$ comes from restricting to the eligible population. Intuitively, the change in policy induces the eligible to change their take-up rather than the ineligible. Focussing on the general population would likely imply $FE_H \gg FE_L$ as $y_L^{SA=0} \ll y_H^{SA=0}$.

[2] This is a sizeable fraction. The prevalence among eligibles is more than double the general population, as discussed in Section 2.2. This highlights how the economic vulnerability of people with poor mental health contributes to the welfare effects, as their overrepresentation among the eligible population amplifies the importance of addressing their needs.

3.2.1 MVPFs of Ordeals and Benefits

For the calibration, I recast welfare effects in terms of the Marginal Value of Public Funds (MVPF) [Hendren and Sprung-Keyser, 2020]. This is defined as the willingness-to-pay for a policy change divided by the cost to the government's budget constraint. I estimate $MVPF$s for barrier and benefit changes.

In the baseline case, I follow Hendren and Sprung-Keyser [2020] and write the direct effects of policy changes in terms of each type's *own* willingness-to-pay. This money-metric social welfare function has the advantage of having interpretable units (€'s) for inter-personal utility comparisons. However, it does not capture any heterogeneity in marginal value of income across types - a factor which is crucial in this context. Proposition A.1.1 derives the formula for the MVPF of a change in barriers ($d\Lambda$) as follows.

$$MVPF_{d\Lambda} = \frac{\overbrace{-\int \mathbb{P}[SA]_\theta \cdot \frac{\kappa'_\theta(\Lambda)}{v'_\theta(B)}\, d\mu}^{\text{Direct Effect} <0}}{\underbrace{\int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} d\mu}_{\text{Behavioral Revenue Effect} <0}}$$

The numerator reflects the fact that $d\Lambda$ imposes dis-utility on infra-marginals and the denominator reflects the government saving money due to lower take-up. The direct costs of ordeals are rescaled by need so that they are measured in each type's WTP (in €'s). I estimate these quantities on the eligible population but extrapolate to the general population, as shown in Proposition C.1.1. For the numerator, this requires integrating against the conditional density of mental health for the eligible.

For the denominator, we rescale the eligible take-up response by a constant for each θ, representing "effective eligibility." The intuition is that we must adjust for baseline incomplete take-up and the fact that some ineligible individuals may be on the margin of take-up, as they could be just indifferent between earning slightly above the threshold or below it to qualify for social assistance.

$$MVPF_{d\Lambda} = \frac{-0.6 \times \frac{161}{2.2} \times 0.25 - 0.6 \times \frac{98}{1} \times 0.75}{679.45 \times 0.25 \times (-0.023) \times \underbrace{\frac{1}{1 - 0.6 \times 0.907}}_{\text{Effective Eligibility}_L} + 645.09 \times 0.75 \times (-0.014) \times \underbrace{\frac{1}{1 - 0.6 \times 0.954}}_{\text{Effective Eligibility}_H}}$$

$$= 2.34$$

An $MVPF_{d\Lambda}$ of 2.34 means that ordeals impose a direct cost of €2.34 on infra-marginals for every €1 saved by the government through lower take-up. $MVPF_{d\Lambda} \gg 1$ suggests that $d\Lambda$ is a costly way to raise government revenue. Notice the *money-metric* barrier costs of people with poor mental health are €72.6, whereas €98 for good mental health. However, €1 is more than twice as valuable to the person struggling with a mental disorder - which means that the monetary cost does not reflect the much greater dis-utility imposed by ordeals on individuals with mental illness.

Proposition A.1.1 derives the formula for the MVPF of a change in benefits as follows, again extrapolating from the eligible population. The numerator reflects the value of the transfer dB to infra-marginals and the denominator captures the mechanical (government must pay for the transfer) and behavioural (government must also pay for increased take-up) revenue effects.

An $MVPF_{dB} < 1$ is to be expected since social assistance is a re-distributive program. It means means that beneficiaries gain 91 cents for every €1 spent raising the benefit level. The estimated value lies in the range surveyed by Hendren and Sprung-Keyser [2020].

$$MVPF_{dB} = \frac{\overbrace{\int \mathbb{P}[SA]_\theta \; d\mu}^{\text{Direct Effect >0}}}{\underbrace{(1-\tau)\cdot \int \mathbb{P}[SA]_\theta \; d\mu}_{\text{Mechanical Revenue Effect >0}} + \underbrace{\int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial B} d\mu}_{\text{Behavioral Revenue Effect >0}}}$$

$$MVPF_{dB} =$$

$$\frac{0.6 \times 0.25 + 0.6 \times 1 \times 0.75}{0.63 \times (0.6 \times 0.25 + 0.6 \times 0.73) + 679.45 \times 0.25 \times \frac{0.63 \times 0.000503}{1-0.6\times 0.907} + 645.09 \times 0.75 \times \frac{0.63\times 0.000227}{1-0.6\times 0.954}}$$

$$\approx 0.91$$

Comparing $MVPF_{d\Lambda}$ to $MVPF_{dB}$ suggests that reducing ordeals is a 2.4× more effective policy than increasing benefits.[3] The *comparison* characterises the welfare effects derived in Proposition 1.2.1 and implies that the government is actually willing to reduce benefit levels to finance a reduction in take-up barriers.[4]

I find that people with poor mental health have twice the need than those with

[3] Note that the social marginal utility of the beneficiaries of the two policies should be taken into account when comparing the $MVPF$'s [Hendren, 2016]. In Appendix C, I show that the social marginal utility of beneficiaries of dB is 1.30 and of $d\Lambda = 1.24$

[4] These $MVPF$s do not directly capture redistribution as social welfare is money-metric. If I calculate $MVPF$'s where the numerator is written using utilitarian social welfare functions, I set social welfare-weights to correspond to the marginal value of income $\lambda_\theta = v'_\theta(B)$. This effectively writes the numerators in constant units of people with good mental health's WTP for €1. Then, $MVPF_{d\Lambda}^{\text{Utilitarian}} = 2.91$ and $MVPF_{dB}^{\text{Utilitarian}} = 1.19$. The latter being above 1 highlights the redistributive motive. This exercise assumes comparability of utility across individuals, and cannot allow for type-specific scalars multiplying utility.

good mental health, but only a 64% higher cost. Why, then, does reducing barriers dominate increasing benefits? The reason is that poor targeting through barriers actually *reduces* the effectiveness of increasing *benefits*. Social assistance is poorly targeted on average, since take-up is similar across mental health states. This makes *dB costly* as it redistributes to all infra-marginal individuals. If people with good mental health had a much lower take-up rate, $MVPF_{dB}$ would be higher.

3.2.2 The Role of Bias

Consider the case that individuals are biased, and perceive barrier costs to be larger than their true value. Namely, let a share ψ of the as-if cost revealed through take-up responses be a true cost, and $1 - \psi$ pure hassle costs (which affect behaviour and not welfare). In this model, take-up is too low relative to the private optimum. Therefore, the $MVPF_{d\Lambda}$ scales down the direct cost of barriers by a factor ψ, but also includes a negative behavioural welfare effect $\times(1 - \psi)$ since individuals are not privately-optimizing so the Envelope Theorem does not hold. $MVPF_{dB}$ now contains a new term in the numerator - an internality correction $\times(1-\psi)$ as increasing benefits helps individuals take-up closer to their private optimum. Appendix C shows the formulae.

When we calibrate these using the sufficient statistics estimated above, we find that the government must be confident that less than 35% of the perceived costs are true welfare costs in order to reverse the $MVPF_{d\Lambda} > MVPF_{dB}$. Here, I take the approach of Chapter 4: if revealed preference does not hold, the government does not know how much behaviour reflects true welfare. However, policies still need to be set. In this case, it is optimal for the government to choose policies which are *robust* to normative ambiguity. The result states that reducing barriers being more effective than increasing benefits is robust as long as more than 35% of the as-if costs are normatively relevant. Similarly, as long as pessimism is weak enough such that perceived need is not below 35% of true need, barriers are more effective than benefits.

3.2.3 Relaxing other Assumptions

I conclude with a discussion of two key identification assumptions, how to relax them through the use of additional structure and the effect this has on welfare consequences. The two key modelling assumptions are (i) Take-up depends on an additive independent choice shock, (ii) $\mathbb{P}[\text{Marginal to Barrier Change}]_\theta = \mathbb{P}[\text{Marginal to Benefits Change}]_\theta$. For full details, see Appendix C.3.

(i) Relaxing independence involves adopting models from Rafkin et al. [2023] or Finkelstein and Notowidigdo [2019] where $v'_\theta(B)$ is independent of θ conditional on income. These models do not fit my context appropriately because they would imply that people with poor mental health have an easier time overcoming barriers, and are substantially *less* pessimistic about the benefit level. Both of these results contradict psychological evidence [Martin et al., 2023b; Evans et al., 2014; Alloy and Ahrens, 1987].

(ii) For internal validity, I focus on subsamples in Section 2.3 and Section 2.4 which are different and call into question the extent to which marginal take-up responses can be compared. However, **Step 2** of the 3-step identification Section 1.3.1 can be applied separately to the two policy designs. Therefore, people with poor mental health having a relatively 2× higher need and a relatively 64% higher cost does not rely on assumption (ii). (ii) is relevant for the comparison between need and cost *within-individuals*. I show in Appendix C.3 that relaxing (ii) through additional structure on the take-up equation suggests that $\mathbb{P}[\text{Marginal to Barrier Change}]_\theta < \mathbb{P}[\text{Marginal to Benefits Change}]_\theta$. This suggests that ordeal costs are a *lower-bound* and pushes further in favour of reducing barriers over increasing benefits.

Conclusions of the first three chapters

These chapters show that people with poor mental health are high-need, yet inefficiently excluded from low-income welfare assistance due to high cost of overcoming barriers. I use a theoretical framework to show how to disentangle need for benefits and cost of barriers using take-up levels and how take-up responds to changes in benefits and barriers. Empirically, I use Dutch administrative data containing detailed information on social assistance take-up and mental health.

Descriptively, while people with poor mental health are three times more likely to be eligible for low-income benefits, conditional on eligibility, they take-up at around the same rate as those with good mental health. A policy which increases barriers disproportionately screens out those with poor mental health, while they also take-up more in response to a change in benefits. This is identified with a regression kink design on the kinked benefits schedule. Combining theory and empirics shows that reducing barriers is twice as effective as increasing benefits.

Future work:

As mentioned in Chapter 3, the policy recommendations depend on whether costs of overcoming barriers are true welfare costs, or just hassle costs which affect behaviour and not welfare. Therefore, in future work I plan to elicit behavioural biases for people with poor mental health and use this quantification to determine optimal policy.

Moreover, throughout I have assumed a static model where mental health is not directly affected by ordeals. This simplification could mean that my estimates of the welfare effect of a change in ordeals is underestimated because barriers likely worsen mental health directly [Brewer et al., 2022]. In this context, mental health is unique, for example in comparison to income or education, because of its potential to respond to aspects of the take-up environment.

Due to these issues, work in progress calibrates a dynamic structural model of evolving mental health type affecting and responding to receipt of social assistance. Through this exercise, I aim to quantify the discrepancies between welfare effects under a static model with those under a dynamic setup. For example, people with poor mental health are more likely to be screened out. If this directly worsens their mental health, there would be evidence of a psychological poverty trap [Haushofer, 2019; Ridley et al., 2020] which could decrease welfare effects.

Finally, the theoretical framework described above is designed for analysing the targeting of social assistance, however can easily be applied to study the welfare consequences of people with poor mental health being screened out of other programs. One program of particular relevance is mental healthcare itself. There is evidence of forgoing mental health treatment by people with serious mental disorders. For example, Cronin et al. [2024] develop a discrete choice model which suggests that people with poor mental health could have increased psychological cost of talk therapy, despite needing it more, which could cause them to forgo. My framework can be applied to evaluate the welfare consequences of this, and determine whether those suffering from mental disorders take-up mental healthcare at the optimum rate. Work is underway along these lines.

Chapter 4

Intrapersonal Utility Comparisons as Interpersonal Utility Comparisons

Joint work with Daniel Reck[1]

[1] For valuable discussions and comments, we thank Scott Elliott, Jacob Goldin, Louis Kaplow, Ben Lockwood, Yusufcan Masatlioglu, Emel Filiz-Ozbay, Alex Rees-Jones, Daniel Schaffa, Joel Slemrod, Luminita Stevens, and Dmitry Taubinsky.

There's always a reality in what you are doing

Sometimes it's so hard to see which one is the true one

–Gene Clark

In many settings, evidence suggests choices are inconsistent, making it difficult to know which choices reflect an individual's normative preferences. A simple example illustrates the problem: at a new job, Sam contributes 0% to her employer-sponsored pension when there is no automatic sign-up, but if her employer automatically enrolls employees at a default contribution rate of 5%, Sam is passive and saves 5% of her salary instead of 0%. So, Sam initially revealed a preference for 0% when 5% was available, but later chose 5%. Which choice reflects Sam's normative preference? Is it the choice she makes when she chooses actively, or does remaining at the default reflect a true welfare cost of opting-out?

In this paper, we analyze the problem faced by a benevolent planner who sets policy while facing uncertainty about which choices reveal normative preferences. We develop a framework for analyzing optimal policy problems in such contexts, drawing insights from behavioral welfare economics and decision theory.

One way of tackling this problem is to formulate the policy-maker's problem in terms of expected-utility maximization: the policy maker seeks to maximize welfare w defined as:

$$w = \sum_{\theta \in \Theta} \psi(\theta) u(x(\theta), \theta). \tag{4.0.1}$$

Each $\theta \in \Theta$ represents a frame: a feature of the decision-making environment that affects choice ($x(\theta)$) but not welfare. Meanwhile, $\psi(\theta)$ represents the policy-makers *belief* about the probability that θ indexes the individual's normative preferences ($u(x(\theta), \theta)$), which may disagree with preferences expressed in other frames ($u(x(\theta'), \theta')$ for $\theta' \neq \theta$). We impose behavioral incentive compatibility: the individual's choices maximize utility in the frame in which the individual chooses [Rees-Jones and Taubinsky, 2018; Danz et al., 2022].

Equation (4.0.1) has an obvious resemblance to utilitarian social welfare. With this parallel in mind, the behavioral welfare criterion of Bernheim and Rangel [2009] – a policy-maker only prefers allocation x to y if for *all* $\theta \in \Theta$, $u(x, \theta) \geq u(y, \theta)$ – resembles the Pareto criterion. In the real world, incompleteness makes it difficult to fully describe optimal policy from criteria like the Pareto criterion and that of Bernheim and Rangel. Our model provides axiomatic conditions under which we can characterize whether a policy-maker prefers x to y, even if x is not chosen over y in all frames. We do not advocate one single approach; rather we explore a few welfarist objectives, depending on how we think about normative uncertainty (risk versus ambiguity) and how much structure we can impose on comparability of welfare across frames.

The core assumptions of our framework are as follows. (i) There are *normative preferences*, which describe what an individual *should* choose in any situation. (ii) Holding the frame fixed, the individual's revealed preferences are rational, i.e. all inconsistencies are explained by framing effects. (iii) Normative preferences coincide with revealed preferences in some frame, called the *normative frame*.

These assumptions have limitations. The most important limitation, in our view, concerns the specification of the primitives, especially the set of frames, in application. One must either observe behavior in all potentially normative frames or extrapolate such behavior from observed choices using non-choice information. However, much of the behavioral public economics literature maintains substantially stronger assumptions, assuming that a social planner has full knowledge of choices in all relevant frames, *and that they know which frame is normative* [e.g. O'Donoghue and Rabin, 2006; Mullainathan et al., 2012; Allcott and Taubinsky, 2015; Allcott et al., 2019]. We focus on settings in which normative preferences exist but are unknown, which we formulate as uncertainty over the normative frame.

Beginning from these three assumptions, we present conditions under which the objective of a benevolent social planner can be represented by the maximization of a intra-personal welfare function as in Equation (4.0.1). First, we show that the welfare criterion of Bernheim and Rangel [2009], which we label BR-dominance, is formally admissible under these assumptions, and that normative preferences must be constant over the frame. The former emphasizes our relationship to social welfare theory–our assumptions restrict to welfare functions which are increasing in Pareto improvements.

Noting the similarity of BR-dominance and Pareto dominance, we adapt an argument from Kaplow and Shavell [2001]. So long as there is a good the individual always prefers in strictly larger amounts, the information in revealed preferences in each frame must be sufficient to evaluate the planner's objective, i.e. welfare must take what Kaplow and Shavell [2001] call a "welfarist" form–some function of the set of utility functions $\{u(x(\theta), \theta)\}_{\theta \in \Theta}$, and increasing in each argument.

This first result does not impose structure on how the planner trades off welfare across potentially normative frames, which requires a stronger notion of cardinal comparability of welfare across frames [Debreu, 1959; Wakker, 1984; Sen, 1986]. For example, if policy A is optimal in one potentially normative frame but policy B is optimal in another, then the planner faces a tradeoff in choosing between A and B. Evaluating this tradeoff requires comparing welfare across frames.

We propose more structured approaches to such tradeoffs, drawing on the theory of normative decision-making under uncertainty. In one approach, we impose that the planner can compare welfare accross frames and that their preferences satisfy "tradeoff consistency" [Wakker, 1984; Wakker and Zank, 1999]. This assumption implies the planner's objective takes a subjective expected utility(/utilitarian) form like Equation (4.0.1). In the context of the introductory example, this means that the planner maximizes Sam's expected utility across the uncertainty about whether the "as-if" cost of opting-out of the default is a true cost or just a mistake. With a subjective expected utility model, we require the planner to have probabilistic beliefs about this normative question, which might be unrealistic.

The approach generalizes to the case of *ambiguity aversion* where the planner does not have a unique prior about which frame is normative but there is a set of priors they find plausible [Knight, 1921; Ellsberg, 1961]. Here the planner adopts a max-min expected welfare criterion over plausible priors under suitable conditions [Gilboa and Schmeidler, 1989].

Our assumptions require that the social planner knows how to compare welfare across frames, but how should we as modelers specify comparable utility functions in applied settings? How do we ensure that the utility function that represents Sam's preferences when she chooses actively is comparable to her utility function in the case of a 5% default? We do not claim to solve the comparability problem, but argue that fundamentally, deciding how to compare welfare across frames is very similar to deciding how to compare welfare across interpersonal types [as discussed in Harsanyi, 1955; Sen, 1976; Weymark, 1991; Fleurbaey and Maniquet, 2011, and many others]. Concretely, we examine the conditions under which we can use money-metric equivalent variation to represent ordinal preferences within frames, and stronger conditions under which we can compare money-metric equivalent variation (in the level) across frames. How much traction on comparability we gain from the money-metric utility concept depends on the plausibility of these assumptions, in a way that recalls prior thinking on comparing the value of money across individuals.

We illustrate the application of our framework to a few behavioral policy problems, including defaults, the manipulation of reference points, corrective taxation with unknown internalities, and "nudge" interventions. Many such problems turn on whether observed behavior reflects a bias or a true preference [e.g. Goldin and Reck, 2022b; Reck and Seibold, 2023; Lockwood et al., 2023]. We show that setting policies at the intrinsic optimum–the choice an individual would make absent behavioral frictions–is often robust, as it avoids undue influence from opt-out costs or gain-loss utility. In contrast, extreme defaults or reference points lack robustness and are generally only optimal if frictions are seen as biases. For corrective taxes, uncertainty about internalities and a preference for robustness leads to shading the optimal tax toward the worst-case marginal internality [Mullainathan et al., 2012; Allcott and Taubinsky, 2015]. When the planner additionally has access to a nudge, they trade-off the benefit of reducing the variance of bias with the worst-possible psychic cost.

Our paper is related to prior work in social choice theory, normative decision theory, and behavioral economics, especially behavioral welfare economics. We discuss this relationship throughout the exposition below. We seek to develop no new decision theory in this paper; rather, our objective is to understand how existing ideas from normative decision theory can generate normative objectives for behavioral policy problems. Relative to prior work in behavioral welfare economics [reviewed in Bernheim and Taubinsky, 2018], the novelty of our approach primarily lies in the development of robust criteria for selecting policy in the presence of uncertainty or ambiguity about normative preferences. These criteria resolve the incompleteness of the welfare criteria proposed by Bernheim and Rangel [2009] using principled reasoning. They are practically useful because incompleteness makes BR-dominance alone uninformative for a wide range of behavioral policy problems [Benkert and

Netzer, 2016].

Outline. In Section 4.1 we state the key assumptions of our framework and discuss their relation to prior work. Section 4.2 then builds from the core assumptions to characterize the planner's problem the normative criteria from axiomatic foundations. Section 4.3 discusses comparability, including money-metric welfare. Section 4.4 introduces some examples from prior literature and maps them into our framework. We develop characterizations of optimal policy that apply our robustness concepts in Section 4.5. We discuss approaches to resolving the normative uncertainty that is primitive in our theory in Section 4.6. The Online Appendix contains additional theoretical analysis, including the development of a perturbation approach to welfare analysis in our framework, a generalization to a continuous set of frames, a review of the relationship of our framework to the "counterfactual normative consumer" approach to welfare analysis, and proofs of all results.

4.1 Setup

In this section, we lay out the fundamental assumptions of our framework and discuss their relation to previous literature.

4.1.1 Core Assumptions on Individual Choices and Welfare

Primitives. A choice situation is defined by (X, θ^D), where $X \subseteq \mathcal{X}$ is a set of options within convex superset $\mathcal{X} \subseteq \mathbb{R}^N$, and θ^D is a decision-making frame drawn from a finite set Θ.

Options are denoted x, x', y, y' and the components of an option are denoted $x =$

$(x_1, \ldots, x_N)$. The individual's choice function is $x : 2^{\mathcal{X}} \times \Theta \to \mathcal{X}$, mapping a choice situation to a selection $x(X, \theta) \in X$.[2] When this selection is not unique, $x(X, \theta)$ should be understood to be any one of the options the individual chooses from the set X in the frame θ. We use standard notation for preference relations.

Assumption 1. ***Core Assumptions.***

Assumption 1.1. ***Normative Preferences Exist.*** *There is a complete and transitive preference relation $\succeq_*$ defined on $\mathcal{X} \times \Theta$, such that the individual's normative choices are those that maximize $\succeq_*$.*

Assumption 1.2. ***Frame-Dependent Rational Preferences.*** *For each $\theta \in \Theta$, there is a complete and transitive preference relation $\succeq_\theta$ on $\mathcal{X}$. Individual choices maximizes these preferences: for any $x \in X(\sigma)$, $x(\sigma, \theta) \succeq_\theta x$.*

Assumption 1.3. ***Revealed Preference Coincidence.*** *There exists $\theta^* \in \Theta$ such that for any $x, x' \in \mathcal{X}$ and any $\theta \in \Theta$,*

$$x \succeq_{\theta^*} x' \iff (x, \theta) \succeq_* (x', \theta).$$

Assumption 1.1 ensures the existence of our normative objective. Assumption 1.2 requires that all choice inconsistencies are captured by framing effects: holding the frame fixed, choices are rational. Assumption 1.3 enables normative revealed preference analysis by requiring that in some frame θ^*, choices reveal normative preferences. We label such a θ^* the *normative frame.* These core assumptions have a nuanced relationship to prior literature that we discuss below.

[2] We note that assuming choices are defined on the domain $2^{\mathcal{X}} \times \Theta$ imposes a condition Bernheim and Rangel [2009] call *rectangularity.* We discuss rectangularity further in Example 2 on intertemporal choice.

Implications of Core Assumptions. Our core assumptions have two immediate and useful implications. First, they formally define what it means for a feature of the choice environment to be a *frame*. While choices may vary across frames, Assumption 1.3 (Revealed Preference Coincidence) requires that normative preferences/choices are *frame-invariant*:

Observation 1. ***Frame Invariance.*** *Under Assumption 1, for any* $x, x' \in \mathcal{X}$ *and any* $\theta, \theta' \in \Theta$*,*

$$(x, \theta) \succeq_* (x', \theta) \implies (x, \theta') \succeq_* (x', \theta').$$

This observation makes precise the defining property of frames: they may influence *choices*, but they do not affect *welfare*.[3]

A second implication of our framework is that it admits the welfare criterion of Bernheim and Rangel [2009]. Specifically, under our core assumptions, if x is revealed preferred to x' in every situations where both are available – which is equivalent to preferring x to x' in every frame under Assumption 1.2 – then the normative preference also favors x over x'.

Observation 2. ***BR-Dominance.*** *Under Assumption 1, for any* $x, x' \in \mathcal{X}$*,*

$$\forall \theta \in \Theta,\ x \succeq_\theta x' \implies x \succeq_* x'. \tag{4.1.1}$$

Core Assumptions in the Running Example. How do these assumptions apply to our motivating defaults example? Our introduction suggests two frames: a

[3] We note that we could have imposed frame invariance directly by defining $\succeq_*$ on $\mathcal{X}$ instead of $\mathcal{X} \times \Theta$. We find our setup, in which frame invariance is a consequence of other assumptions, easier to apply in settings like Section 4.4.3, where the planner is uncertain whether an observed feature of a choice environment is truly a frame.

naturally occurring frame[4], in which there is a default option, and an active choice frame, in which Sam always chooses actively. We discuss other possibilities for the set of frames in this example below. Assumption 1.1 states Sam has normative preference that govern what savings rate she *should* choose given the default. Assumption 1.2 says Sam's revealed preferences are rational in the naturally occurring frame and the active choice frame, and Assumption 1.3 implies that one of these two frames is normative.

Technical Assumptions. We make three more assumptions that could be relaxed in principle, but which simplify our analysis. We introduce a standard continuity assumption on frame-dependent preferences, to give us an ordinal utility function denoted $u(x, \theta)$, which represents frame-dependent preferences $\succeq_\theta$ for given θ.

Assumption 2. ***Continuity.*** *For any $x_0 \in \mathcal{X}$ and any $\theta \in \Theta$, the sets $\{x \in \mathcal{X} : x \succeq_\theta x_0\}$ and $\{x \in \mathcal{X} : x \preceq_\theta x_0\}$ are closed.*

We also assume there is one good that the individual prefers to consume in strictly increasing amounts regardless of the frame. This allows us to adapt a proof from Kaplow and Shavell [2001] below.

Assumption 3. ***One Dimension of Strict Monotonicity.*** *There is some good x_n such that for every θ, $\succeq_\theta$ is strictly monotonic in x_n.*

Finally, we assume that the grand set of options is sufficiently rich that we can induce arbitrary variation in utility in each frame θ by varying options. The following assumption endows the option space with the richness of the full action space in classic subjective expected utility theory:

[4] We borrow the term "naturally occurring frame" from Bernheim et al. [2015].

Assumption 4. ***Rich Options.*** *For any two options $x, y \in \mathcal{X}$ and any frame $\theta_0 \in \Theta$, there is another option $z \in \mathcal{X}$ such that $x \sim_{\theta_0} z$ and for any frame $\theta \neq \theta_0$, $y \sim_\theta z$.*

4.1.2 Discussion of Setup and Relationship to Prior Literature

Information We interpret Assumption 1 as an assumption about the information of a social planner. We assume the planner knows the individual's preferences in each frame, and we wish to examine how we map this information to welfare. The model has a less behavioral interpretation in which the individual has different information in each frame and normative preferences correspond to the choices the individual would make under full information.

Observed Choices versus Known Preferences Our model does not necessarily require that the planner directly observes behavior in every frame. In many settings – e.g. one in which we observe choices under varying defaults – the planner might not believe that any of the observed choices come from a normative frame, so assuming that all frames are observed would threaten Assumption 1.3. When some choices are unobserved, Assumption 1 implicitly requires that the planner uses non-choice information to *infer* unobserved preferences from observed ones.[5] Such inferences might rely on knowledge of the structure of preferences or of the preferences of other individuals, as in the *counterfactual normative consumer* approach (see Section 4.6 and Appendix D.6). This approach relies on the untestable normative assumption that

[5] Most empirical applications of standard welfarist models impose practical assumptions on preferences in order to infer unobserved preferences from observed choices. In behavioral welfarist models, the main difference is that one extrapolates across frames.

experts' decisions reveal normative preferences. Incorporating normative ambiguity addresses this limitation to an extent.

The Set of Frames in Principle and in Practice Assumption 1 requires that the set of frames includes all policy-relevant decision-making frames and all the frames that might be normative. While the planner may not know the normative frame, we rule out the possibility the normative frame is something they never envisioned (a "black swan").

The planner knows Θ, but how do we specify a set of frames when applying our model in practice? A limitation of our approach is that the modeler must make normative assumptions about what are the potentially normative frames and hence what type of ambiguity to allow. That being said, there is an implicit convention in prior work to construct Θ: we begin with traditional preference forms, e.g. satisfying time consistency, vNM independence, no default-effects etc, and then introduce additional frames to capture empirically observed deviations from traditional forms. Assumption 1 says that the individual's normative preferences correspond either to the traditional form, or to the form(s) capturing deviations. Such an approach seems applicable to most behavioral policy problems.

Our discussion of two potentially normative frames in the running defaults example above follows this convention. One can think of other potentially normative frames in that example. Under costly opt-out, the default itself cannot be a frame because this would contradict Observation 1. But if we disregard normatively relevant opt-out costs, the default itself could be regarded as a frame. So a planner who thinks it is ambiguous whether opt-out costs are normative might also think it is ambiguous whether the default itself is a frame. We flesh this out further below, but we note that the difficulty we wrestle with here reflects the limits of our framework. One can always add frames and bring more normative ambiguity into the model, but the modeler must restrict normative ambiguity somehow.

Formal Definition of Frames and Normative Preferences Our model explicitly defines frames and normative preferences, addressing a core ambiguity in earlier work. While Bernheim and Rangel [2009] verbally resist explicitly introducing normative preferences as we do in Assumption 1.1, they implicitly rely on a condition resembling Frame Invariance by defining frames as conditions that affect choices without directly influencing welfare. Formally distinguishing frames from welfare-relevant factors requires a concept of normative preferences. Without this assumption, the concept of frames becomes ambiguous or circular, and it is not clear how to define features which "have no direct bearing on well-being, but instead impact biases" [Bernheim and Taubinsky, 2018].[6]

Limited Attention and Within-Frame Rationality Assumption 1.2 restricts the framework of Bernheim and Rangel [2009] by requiring that all inconsistencies in choice arise solely from framing effects. Without this assumption, our BR-dominance criterion (Observation 2) differs from theirs. In particular, Assumption 1.2 excludes models of limited attention: if individuals fail to consistently consider all available options there can be choice inconsistencies within a frame. Masatlioglu et al. [2012] show that in models with limited attention, Bernheim and Rangel's dominance criterion may fail because individuals could consistently select the best option they attend to, rather than the option they truly prefer. Limited attention thus threatens both within-frame rationality (Assumption 1.2) and RP-coincidence (Assumption 1.3).

[6] Although psychological theories describe preferences as constructed at the moment of choice [e.g., Lichtenstein and Slovic, 2006], making stable normative preferences controversial [Bernheim, 2016; Bernheim and Taubinsky, 2018], our assumption that the normative frame *exists* but is *unknown* allows us to accommodate these critiques and approach normative ambiguity using principled reasoning.

We could extend our approach to cover some forms of limited attention. One simple extension would be to require that Assumption 1.2 holds for a subset of frames in which the individual attends to all their options; then Assumption 1.3 would require that one such attentive frame is the normative frame.[7] However, this approach would not capture the possibility that variation in attention has normative importance because attention is a scarce resource, in which case a key question would seem to be be whether attention is allocated optimally by the individual [e.g. Bronchetti et al., 2023]. This echoes the type of question we take up in Example 1 below [see discussion of costly attention and defaults in Goldin and Reck, 2022b], but we defer a fuller treatment of welfare under costly attention to future work.

Role of RP-Coincidence in Prior Literature We make explicit the "RP-Coincidence" assumption (that a normative frame exists) that has been implicit in all behavioral welfare economics literature using revealed preferences. For example, Chetty et al. [2009] assume normative preferences are revealed when taxes are completely salient, while other behavioral public economics literature typically selects a single bias as the main behavioral distortion, requiring that decisions purged of this bias reflect normative preferences.

We explicitly allow for uncertainty about which frame is normative, providing a structured approach to determining optimal policy under such normative ambiguity. Whereas Bernheim and Rangel [2009]'s dominance criterion alone is often incomplete for policy decisions [Benkert and Netzer, 2016], our method delivers complete, robust characterizations of optimal policy.

While accommodating normative ambiguity, we remain fundamentally welfarist—defining welfare through revealed preferences and optimizing accordingly—unlike

[7] This is very similar to the concept of a "welfare-relevant domain" in Bernheim and Rangel [2009].

non-welfarist approaches [e.g. Sugden, 2004].[8]

4.2 Policy Problems and Normative Criteria

In this section we build on our core assumptions to characterize the planner's problem. We rely entirely on machinery from prior work in decision theory and/or social welfare.

4.2.1 Behavioral Optimal Policy Problems

Notation. Let us introduce some notation that helps us think about policy variation and welfare. Policies are real-valued vectors denoted $P \in \mathcal{P}$; we assume the set $\mathcal{P}$ is closed and convex. For a given policy P, the option set the individual chooses from is $X(P)$ and the frame in which the individual makes their choice is $\theta^D(P)$. To economize on notation, we express choices under given policies using $x(P, \theta^D(P))$ in place of of $x(X(P), \theta^D(P))$, and where it is irrelevant we also suppress the dependence of θ^D on P.

Known Normative Frame. We begin with the case where the normative frame is known to be some $\theta^* \in \Theta$, i.e. there is no normative uncertainty/ambiguity in the model. A benevolent planner's objective is to choose the policy $P \in \mathcal{P}$ that the individual would choose for themselves according to $\succeq_*$. That is, we should characterize the policy that is optimal subject to the constraint that the individual

[8] The opportunity criterion of Sugden [2004] is subject to normative ambiguity in a way that we find intriguing; does changing the default or introducing a default modify one's opportunity set? The answer seems to depend on whether or not opting out of the default requires expending real resources that reflect lost opportunity. We defer a fuller exploration of this idea, and the relationship of opportunity-based normative criteria and our approach to normative ambiguity, to future work.

will choose $x = x(P, \theta^D)$ – the Behavioral Incentive Compatibility (BIC) constraint [Rees-Jones and Taubinsky, 2018; Danz et al., 2022].[9]

RP-Coincidence implies that a benevolent planner should adopt as their the welfare function the utility function that represents $\succeq_{\theta^*}$. The planner's problem under known θ^* is therefore

$$\begin{aligned} &\max_{P \in \mathcal{P}} u(x, \theta^*) \\ &\text{subject to } x = x(P, \theta^D(P)). \quad \text{(BIC)} \end{aligned} \tag{4.2.1}$$

Many policy problems in the literature on behavioral public economics take the form above, where θ^* is implicitly or explicitly assumed to be known. From this work, we have reduced-form characterizations of the welfare effects of policy variation for known θ^* [e.g. Mullainathan et al., 2012; Allcott and Taubinsky, 2015], and structural characterizations of optimal policies for a variety of more specific structural models [e.g. O'Donoghue and Rabin, 2006].

Unknown Normative Frame Now we turn to characterizing a benevolent planner's objective when the normative preference is unknown. In this case we denote the planner's objective by the function $w(x)$. The policy problem becomes

$$\begin{aligned} &\max_{P \in \mathcal{P}} w(x) \\ &\text{subject to } x = x(P, \theta^D(P)). \quad \text{(BIC)} \end{aligned} \tag{4.2.2}$$

One possibility we will consider, for instance, is that the planner has some beliefs about the likelihood that each frame is normative, denoted $\psi(\theta)$, and maximizes

[9] Incorporating an additional constraint like the government budget constraint is straightforward – one can think of this as imposing structure on the set of feasible policies $\mathcal{P}$.

classical expected utility. In this case, w takes an expected utility form like

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta).$$

and the utility function $u(x, \theta)$ is fully comparable across frames (and represents $\succeq_\theta$). Our focus going forward is on formalize such potential forms of $w(x)$.

4.2.2 Formalizing Behavioral Welfarist Criteria

The planner's preferences over which option the individual consumes are denoted by a relation $\succeq_w$ on $\mathcal{X}$.[10] In writing (4.2.2), we are already imposing that $\succeq_w$ has a representation $w : \mathbb{R}^N \to \mathbb{R}$. We should ensure this representation exists. Traditionally, we say that a planner is *benevolent* if given any x, x',

$$x \succeq_* x' \implies w(x) \geq w(x'). \tag{4.2.3}$$

This is insufficient to fully characterize $w(x)$ when the normative frame is unknown. However, under Assumptions Assumptions 1.1 and 1.3, property (4.2.3) does imply that a benevolent planner should respect BR-dominance. We therefore begin with the following structure on $\succsim_w$:

Assumption 5. ***Basic Structure on Planner's Preferences.***

Assumption 5.1. ***Rationality.*** *$\succeq_w$ is complete and transitive.*

Assumption 5.2. ***Continuity.*** *For any $x \in \mathcal{X}$, the sets $\{x' \in \mathcal{X} : x' \succeq_w x\}$ and $\{x' \in \mathcal{X} : x' \preceq_w x\}$ are closed.*

[10] Note we are using the implication of Observation 1 in positing that $w(x)/\succeq_w$ is independent of the frame.

Assumption 5.3. ***Weak BR-dominance.*** *For any $x, x' \in \mathcal{X}$, if $x \succeq_\theta x'$ for every θ, then $x \succeq_w x'$.*

To begin with, we observe that our assumption so far impose that w must depend on information in frame-dependent preferences and it cannot depend on any other information about options.

Proposition 4.2.1. *Maintain Assumptions 1.2, 2 and 3. Assumption 5 holds if and only if for any representation of ordinal preferences $u(x,\theta)$, there is a function $\mathcal{W} : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planner's preferences are represented by*

$$w(x) = \mathcal{W}\left(\{u(x,\theta)\}_{\theta\in\Theta}\right), \tag{4.2.4}$$

and $\mathcal{W}$ is continuous and weakly increasing in every argument.

Discussion of Proposition 4.2.1. As with Pareto dominance, respecting BR-dominance requires that the information in frame-specific preferences $\succeq_\theta$ must be sufficient to evaluate the planner's objective. The proof is an adaptation of an argument in Kaplow and Shavell [2001], which demonstrates that if any other information about the options is used to evaluate the planner's objective, we can find violations of Pareto/BR-dominance – the proof uses continuity and the good x_n from Assumption 3 to construct such violations.

Normative Risk

Proposition 4.2.1 imposes too little structure on the planner's objective to be of much practical use. Consider, for instance, two options x and x' such that $x \succ_\theta x'$ for some frame θ but $x \prec_{\theta'} x'$ for some other frame θ'. In this case, the planner faces a tradeoff between welfare under θ and welfare under θ'. The assumptions of Proposition 4.2.1 requires that the planner must find some way to evaluate such tradeoffs, but we do not know how. Our next approach assumes the planner views approaches these tradeoffs like a subjective expected utility maximizer.

Let us introduce a little more notation to articulate the next assumption, following Köbberling and Wakker [2003] and other work by Peter Wakker. Given a utility representation $u(x, \theta)$, a frame θ, an option x, and a real number α such that for some option x', $u(x', \theta) = \alpha$, let $\alpha_\theta x$ denote the option in $\mathcal{X}$ such that $u(\alpha_\theta x, \theta) = \alpha$ and for any $\theta' \neq \theta$, $u(\alpha_\theta x, \theta') = u(x, \theta')$. That such options always exist is ensured by Assumption 4. We denote other options constructed in this fashion using $\beta_\theta x'$, $\gamma_\theta y$, $\delta_\theta y'$, etc. We say that a frame θ is *null* if $\alpha_\theta x \sim_w x$ for every α and every x. Otherwise the frame is non-null. When the planner's preferences have a subjective expected utility representation, null frames are those with zero probability.

Assumption 6. ***Tradeoff Consistency.*** *There exists a utility function $u(x, \theta)$ such that $u(x, \theta)$ represents $\succeq_\theta$ for every θ, and for any two non-null frames $\theta_0, \theta_1 \in \Theta$, any four options $x, x', y, y' \in \mathcal{X}$, and any four real numbers $\alpha, \beta, \gamma, \delta$,*

$$\begin{array}{ccc} \alpha_{\theta_0} x \sim_w \beta_{\theta_0} x', & & \alpha_{\theta_1} y \sim_w \beta_{\theta_1} y', \\ \text{and } \gamma_{\theta_0} x \sim \delta_{\theta_0} x' & \text{imply} & \gamma_{\theta_1} y \sim \delta_{\theta_1} y'. \end{array}$$

Assumption 6 structures how the planner trades off welfare across frames. To unpack the assumption, it is instructive to think of the utility function whose existence is assumed here as an amount of money that makes the individual in the given frame θ indifferent between option x and being given that amount of money (starting from some baseline situation). In this case, the option $\alpha_{\theta_0} x$ gives α dollars in frame θ_0 and the same payoff as x in other frames. The two indifference conditions on the left-hand side of the equation imply that the difference between α and β dollars in frame

θ_0 is the same as the difference between γ and δ dollars: both pairs exactly offset whatever is the difference between x and x' in other frames. The third condition says that in some other frame θ_1 that the difference between α and β dollars again offsets whatever is the difference between two options y and y' in other frames. So, if the planner is consistently trading off dollar-valued payoffs in different frames, the fourth indifference condition must follow.

Proposition 4.2.2. *Maintain Assumptions 1, 2 3, and 4. Then Assumptions 5 and Assumption 6 hold if and only if there is a probability distribution function $\psi(\theta)$ and a utility function $u: \mathcal{X} \times \Theta \to \mathbb{R}$ such that $u(x, \theta)$ represents $\succeq_\theta$ for every θ, and planner's preferences $\succeq_w$ are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta). \tag{4.2.5}$$

Moreover, u is unique up to positive affine transformation if there are at least two non-null frames.[11]

Discussion of Proposition 4.2.2. Relative to the previous proposition, this result adds that assuming the planner trades off welfare consistently across (non-null) frames, they must be maximizing a subjective expected utility function. In Appendix D.1.3, we present an alternative approach to deriving an expected utility criterion based on Von Neumann and Morgenstern [1953]. This approach requires introducing the concept of a normative lottery (the equivalent of vNM lotteries with an unknown normative frame) and then imposing the analogue of the vNM independence axiom for preferences over normative lotteries. Following our intuitive discussion of monetary values above, yet another approach would be to directly

[11] When there is only one non-null frame, we are in the case where the planner knows the normative frame and the utility representation will be unique up to monotonic transformation. When all frames are null, the planner is indifferent to all options, which is a case we generally disregard.

impose that we can compare the outcomes of choices in different frames in terms of different amounts of money, and then impose additive separability; this would lead us to an adaptation of De Finetti [1937]. A unifying analysis of these and other approaches to expected utility by Wakker and Zank [1999] shows that the key structure imposed by all of them is indeed tradeoff consistency. We prefer the approach we take here because it imposes that the planner finds a way to compare and trade off welfare across frames directly, instead of e.g. imposing comparability more indirectly via preferences over lotteries.

Remark on Comparability. The assumption that the planner can compare welfare across frames may be controversial. Given the motivation discussed above, we suspect that any approach that allows us to overcome the limitations of the welfarist criterion in Proposition 4.2.1 will require some type of comparability; in our view, the more controversial aspects of comparability involve how researchers approach these comparisons in practice, using, e.g. money-metric utility. We discuss comparability further and flesh out the idea of comparisons rooted in money-metric utility in Section 4.3. More subtly, the utility function whose existence is ensured by tradeoff consistency (Assumption 6) is not necessarily identical to the fully comparable utility function whose existence is assured by Proposition 4.2.2. However, these two functions must exhibit ordinal level comparability across both options and frames, i.e. they must be monotonic transformations of one another.

Normative Ambiguity

A potential objection to the approach to welfare analysis implied Proposition 4.2.2 is that the planner may not know what is the probability that each frame is normative; for decision theory under uncertainty generally this critique is due to Knight [1921] and Ellsberg [1961].

To address this concern, we adapt the theory of ambiguity aversion due to Gilboa and Schmeidler [1989]. In this theory, the planner does not have a unique prior distribution $\psi \in \Delta(\Theta)$ about the likelihood each frame is normative, but rather there is a closed and convex set of distributions $\Psi^* \subseteq \Delta(\Theta)$ that they find plausible.[12] The planner then maximizes a welfare crietrion of the form

$$w(x) = \min_{\psi \in \Psi^*} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x, \theta^*)) \right\}. \tag{4.2.6}$$

Our formal derivation of this objective follows that of Gilboa and Schmeidler [1989] very closely, so we focus on explaining the assumptions intuitively and present formal statements and results in Appendix D.1.5.

To formalize the ambiguity averse criterion, we begin with the same setup as the approach using vNM independence discussed above, i.e. we use the concept of normative lotteries. We require the planner's preferences over normative lotteries are rational and continuous and satisfy weak BR-dominance (similarly to Assumption 5). Comparability of u across frames is imposed directly as above. The two new assumptions require 1) that when the planner evaluates known risks, they trade off risk in a similar fashion to vNM independence (certainty independence), but 2) when faced with ambiguity, the planner prefers to *hedge* (uncertainty aversion).[13]

Forms of Ambiguity. We present two ways to operationalize normative ambiguity in planner preferences. The first is a global approach, where the planner specifies a subset of welfare-relevant frames $\Theta^* \subseteq \Theta$, and assumes the true normative dis-

[12] We endow the simplex $\Delta(\Theta)$ with a metric suitable for probability distributions e.g. the Wasserstein metric.

[13] We also require a non-degeneracy assumption that prevents the planner from being indifferent to all options/lotteries.

tribution lies somewhere in $\Delta(\Theta^*)$. This form suggests the planner fundamentally lacks a rational for privileging one perspective over another, and adopting it leads to a global max-min objective reminiscent of a Rawlsian social welfare function. The second is a more local approach inspired by robust control [Hansen and Sargent, 2001], where the planner begins with a best-guess distribution ψ but accounts for ambiguity by evaluating policies against all distributions in a neighborhood $B(\psi, \kappa)$. This form seems more appropriate when ψ is identified via a structural model (as in the counterfactual normative consumer approach), but the planner wants to allow for some robustness because they are concerned about model mis-specification.

All the objectives coincide in the limit under extreme ambiguity or extreme paternalistic risk aversion. Specifically, the planner's objective reduces to the global max-min criterion in three cases: (i) when the tolerance parameter κ is large enough to make $B(\psi, \kappa)$ cover the entire simplex, (ii) when the welfare-relevant subset of frames Θ^* spans all of Θ, and (iii) when the planner is utilitarian but extremely risk averse over a given welfare metric. We state and prove these convergence results formally in Appendix D.1.5.

4.3 Comparability

A key assumption in Section 4.2 was that the planner can compare welfare across frames. In this section, we discuss comparability in greater depth. We briefly review of debates in prior literature on interpersonal social welfare functions, which provides a useful parallel to our context. We then present a more formal analysis of the comparison of money-metric equivalent variation across frames.

4.3.1 Parallel to Interpersonal Comparisons

We constructed our proposed welfarist criteria by applying decision theory without reference to the theory of social welfare. Fundamentally, however, we view the question of how to compare utility across frames as very similar to the controversy about how to compare utility across individuals in analogous interpersonal problems. An older literature on social welfare derives forms of social welfare function using decision-theoretical axioms, similarly to our work in the previous section.[14] We find axioms in all such prior work that require that the planner can compare utility across individuals, as in our tradeoff consistency assumption. However, there is little consensus on the best way to approach comparability in applied interpersonal problems. Analogously, we do not expect to propose an approach to comparability that will prove universally acceptable to all readers and applicable in all settings.

Instead, we formalize an approach that parallels the typical approach to these problems in contemporary public economics [see e.g. Saez and Stantcheva, 2016; Hendren and Sprung-Keyser, 2020; Sher, 2023]. With this approach, one converts all options into their money-metric equivalent values, and then treats the value of money to a given individual (and by extension the value of equity) as unknown. In our setting, such an approach views both comparisons of welfare across frames and the probability that each frame is normative as judgments on the part of the planner, over which we as modelers remain agnostic. However, this approach requires normative assumptions, whose analogue an intrapersonal setting we explore next.

[14] Harsanyi [1955] axiomatizes the utilitarian social welfare function, while Hammond [1976] axiomatizes the Rawlesian (maxmin) social welfare function [see also Sen, 1970, 1976; d'Aspremont and Gevers, 1977; Maskin, 1978]. . See Sen [1986] for a technical review and see Sen [1997] for a more philosophical perspective. Local maxmin utilitarian objectives are less commonly used in the literature on social welfare, but more recent related papers here include Alon and Gayer [2016] and Mongin and Pivato [2021].

4.3.2 Money Metric Welfare

We begin with some notation and assumptions that discipline equivalent variation. We introduce a new feature of the choice environment denoted $Z \in \mathbb{R}$; the menu is now expressed as a function of policy and Z: $X(P, Z)$. As above, we economize on notation, writing $x(P, Z, \theta^D)$ instead of $x(X(P, Z), \theta^D(P))$.

Assumption 7. ***Ordinal Equivalent Variation Admissibility.***

Assumption 7.1. ***Strict Monotonicity in Money.*** *For any two Z, Z' any two frames θ^D, θ, and any policy P,*

$$Z > Z' \iff x(P, Z, \theta^D) \succ_\theta x(P, Z', \theta^D).$$

Assumption 7.2. ***Continuity over Money.*** *For any policy P, any Z, and any two frames θ^D, θ, the sets $\{Z' : x(P, Z', \theta^D) \succsim_\theta x(P, Z, \theta^D)\}$ and $\{Z' : x(P, Z', \theta^D) \precsim_\theta x(P, Z, \theta^D)\}$ are closed.*

Assumption 7.3. ***Equalizability.*** *For any option x, any policy P and any two frames θ^D, θ, the sets $\{Z' : x(P, Z', \theta^D) \succ_\theta x\}$ and $\{Z' : x(P, Z', \theta^D) \prec_\theta x\}$ are non-empty.*

Discussion of Assumption 7. Combined with RP-Coincidence, Assumption 7.1 ensures that giving the individual more Z always improves welfare in the normative frame; imposing strict monotocity over money allows us to relax strict monotonicity over some good (Assumption 3). One reason Assumption 7.1 might fail is if, for example, in an "addicted" frame θ^D, the individual spends all their money on an addictive substance that is not a "good" but a "bad" from the perspective of some other potentially normative frame.[15] Assumption 7.2 implies that welfare is contin-

[15] If we are willing to assume the "addicted" frame cannot be normative, is straightforward to relax this assumption to accommodate this possibility.

uous in Z in every frame. Assumption 7.3 ensures that all changes to welfare driven by variation in choices can be fully offset by money, which is obviously key for the existence of equivalent variation. All this is assumed regardless of which frame is normative.

Together, these assumptions discipline *Equivalent Variation* (EV) in any (potentially normative) frame θ, which is defined as $\zeta \in \mathbb{R}$ such that for a given *baseline* (P_0, Z_0, θ_0^D),

$$x \sim_\theta x(P_0, Z_0 + \zeta, \theta_0^D). \tag{4.3.1}$$

Lemma 4.3.1. ***Existence and uniqueness of EV****. Under Assumptions 1.2, 2 and 7, for any option x any frame θ and any (P_0, Z_0, θ_0^D), equivalent variation ζ exists and is unique. Moreover, $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ represents ordinal preferences $\succeq_\theta$ for every θ.*

Lemma 4.3.1 implies that for a given baseline, $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ is a unique representation of revealed preferences under θ, $u(x, \theta)$. Combining the representation result in Lemma 4.3.1 with the idea in Proposition 4.2.1, we find that introducing Assumption 7 yields the following:

Proposition 4.3.1. ***Planner's Preferences and Equivalent Variation.*** *Under Assumptions 1.2, 2, 5, and 7, for any baseline P_0, Z_0, θ_0^D, there is a function $\mathcal{W}_\zeta : \mathbb{R}^{|\Theta|} \to \mathbb{R}$ such that the planners preferences are represented by $w(x) = \mathcal{W}_\zeta\left(\{\zeta(x; \theta^*, P_0, Z_0, \theta_0^D)\}_{\theta^* \in \Theta}\right)$.*

Perturbation Approach to Welfare Analysis. Proposition 4.3.1 suggests that provided that equivalent variation is well-behaved per Assumption 7, we may use equivalent variation welfare metrics to describe the welfare effects of local policy perturbations. Under the assumptions of the proposition and additionally assuming differentiability of all relevant quantities, the welfare effect of a marginal policy reform (a perturbation) dP that does not modify θ^D will be

$$dw = \sum_{\theta} \frac{\partial \mathcal{W}_z}{\partial \zeta_\theta} d\zeta(x(P, Z, \theta^D), \theta; P_0, \theta_0^D, Z_0). \tag{4.3.2}$$

where derivatives are evaluated at the status quo (P, Z, θ^D); $d\zeta(., \theta)$ is the money-metric welfare effect of the marginal reform in frame θ.[16] With the addition of tradeoff consistency, the welfare weight term in the above expression becomes

$$\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta} = \psi(\theta) \frac{\partial u(x(P, Z, \theta^D), \theta)}{\partial Z}. \tag{4.3.3}$$

In words, the welfare effect of the perturbation is the expected value of its equivalent variation across frames (according to probabilities ψ), weighted by the relative value of a dollar in frame θ. We can therefore think of this welfare weight as involving two normative judgments: the probability that each frame is normative $\psi(\theta)$, and the value of money in frame θ, $\frac{du^\theta}{dz}$, which in turn is government by the planner's judgments about how to compare monetary payoffs.

The tendency of reduced-form work in public economics is to leave the value of a dollar under a given type/frame θ, $\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta}$, unspecified [e.g. Hendren and Sprung-Keyser, 2020]. We can take the same route in our model under the assumptions laid out so far. With a stronger assumption, we find some intuitive structure on the value of money across frames, which at the very least could help to choose a baseline situation from which to construct equivalent variation.

When will the planner's preferences satisfy tradeoff consistency over money-metric

[16] We do not consider policies that perturb the frame because we assume the set of frames is finite. We view this mainly as a technical limitation and expect the approach to generalize readily to policies that perturb frames. Note that the status quo situation here refers to the starting point from which we perturb policy. In general, this differs from the baseline situation from which we construct equivalent variation.

utility itself? This requires ordinal level comparability of equivalent variation and the planner's utility function, i.e. for any two options x, x' and any two frames θ, θ', $u(x,\theta) \geq u(x'.\theta') \iff \zeta(x,\theta) \geq \zeta(x'.\theta')$. Our next result characterizes when ordinal level comparability obtains for some baseline situation. To economize on notation we express the option the individual chooses in a given baseline as $x_0 \equiv x(P_0, Z_0, \theta_0^D)$.

Assumption 8. ***Cardinal Equivalent Variation Admissibility.*** *Let $u(x,\theta)$ be a fully comparable utility function from the representation of $w(x)$ in Proposition 4.2.2. There is a baseline situation (P_0, Z_0, θ_0^D) under which the following conditions hold:*

Assumption 8.1. ***Baseline Indifference.*** *For any θ, θ',*

$$u(x_0, \theta) = u(x_0, \theta').$$

Assumption 8.2. ***Comparable Value of Money At Baseline.*** *For any θ, θ' and any $\zeta, \zeta' \in \mathbb{R}$,*

$$u(x(P_0, Z_0 + \zeta, \theta_0^D), \theta) - u(x_0, \theta) \geq u(x(P_0, Z_0 + \zeta', \theta_0^D), \theta') - u(x_0, \theta') \iff \zeta \geq \zeta'.$$

Assumption 8.1 requires that the level of utility is the same across frames in the baseline situation. Assumption 8.2 requires that starting from the baseline and giving the individual some amount of money has the same effect on utility regardless of the frame. One can think of these as selection criteria for the baseline situation.

Lemma 4.3.2. ***Ordinal Level Comparability of Equivalent Variation.*** *Maintain Assumptions 1, 2, 4, 5, 6, and 7. Let $u(x,\theta)$ be a cardinal utility function from the representation in Proposition 4.2.2. Assumption 8 holds if and only if there is some baseline (P_0, Z_0, θ_0^D) such that $u(x,\theta)$ and $\zeta(x, \theta; P_0, Z_0, \theta_0^D)$ exhibit ordinal level comparability.*

Proposition 4.3.2. *Under Assumptions 1, 2, 4, 5, 6, 7, and 8, there is a probability distribution $\psi(\theta)$, a function $u_\zeta : \mathbb{R} \to \mathbb{R}$ and a baseline situation (P_0, Z_0, θ_0^D) such that the planner's preferences are represented by*

$$w(x) = \sum_{\theta \in \Theta} \psi(\theta) u_\zeta(\zeta(x, \theta; P_0, Z_0, \theta_0^D)). \tag{4.3.4}$$

Moreover, u_z is strictly monotonic and unique up to positive affine transformation when more than two frames are non-null.

Discussion of Proposition 4.3.2. Under these assumptions, the planner can directly compare money-metric welfare across frames and the only remaining unstructured component of the expression is the value of money itself, u_ζ. The result extends straightforwardly to the representation of the planner's preferences in the ambiguity case from equation (4.2.6).

Under equation (4.3.4), our welfare weight becomes

$$\frac{\partial \mathcal{W}_z}{\partial \zeta_\theta} = \psi(\theta) \frac{du_\zeta}{d\zeta}, \tag{4.3.5}$$

where the derivative is again evaluated at the status quo. The second term now only depends on θ through the level of ζ: the value of money is not frame-specific.

Paternalistic Risk Aversion One payoff to this additional structure is that now we can think of paternalistic risk aversion in simple terms. Specifically, we can say the planner's preferences exhibit *paternalistic risk aversion over money* if u_ζ is concave. Paternalistic risk aversion implies a preference not to have too much disagreement across frames. Intuitively, reforms that amplify disagreements across frames are riskier.

We show this formally and illustrate a perturbation approach to welfare analysis within our framework in Appendix D.3. We show that under paternalistic risk aversion, the first-order welfare effect of a marginal policy reform is the expected welfare effect plus an adjustment for the change in the variance of welfare across frames. The planner prefers to reduce this variance to an extent that is proportional to the curvature of u_ζ, i.e. the extent of paternalistic risk aversion. Paternalistic risk aversion implies a similar but less extreme preference for robustness than the one we find under ambiguity aversion.

Further Possibilities. At this point, we could leverage the parallel to prior thinking on the value of money across individuals to go even further. For instance, we could imagine an "impartial observer" with the same risk preferences over money as the individual evaluating the normative risk the planner confronts, and then construct u_ζ from the way the individual trades off risk [Harsanyi, 1955]. One problem with this approach that individuals in behavioral models often do not act like expected-utility maximizers, and if it is ambiguous whether we should respect these nonstandard revealed preferences (as in Example 1.3), the impartial observer's preferences become ambiguous. Rather than probe this controversial question further, we next turn our focus to more applied questions.

Happiness-Metric Equivalent Variation? Assumptions 7 and 8 describes the key properties of the variable Z in this approach to welfare analysis. These assumptions do not require that the variable Z be money, and they are distinct from the structure usually used to derive equivalent variation in practice (budget constraints, expenditure functions etc). The variable Z could be identified with something in the real world other than money, such as a mental state.

4.4 Examples

We next illustrate via examples how prior work on behavioral welfare economics fits within our framework. In each example, we characterize what the "expected utility" formulation of the planner's preferences from Equation (4.2.5)/Proposition 4.2.2 looks like for a given set of beliefs ψ. This is done for expositional clarity to explore which kind of normative uncertainty a planner might face in some concrete examples. We turn to analyses of optimal policy under ambiguity and robustness in Section 4.5. We directly impose comparability across frames and return to the issues raised in Section 4.3 where they seem especially relevant.

4.4.1 Example 1: Biases Versus Strange Preferences

Let us introduce a general example in which the key intrapersonal question is whether some behavioral phenomenon arises due to a bias or a normative preference. Suppose the decision-making frame is some fixed frame θ^D and there is just one alternative frame denoted θ^A. Our representation of welfare from equation (4.2.5) becomes

$$w(x) = \psi(\theta^D)u(x, \theta^D) + [1 - \psi(\theta^D)]u(x, \theta^A). \tag{4.4.1}$$

With two frames, we denote disagreements using $V(x) \equiv u(x, \theta^D) - u(x, \theta^A)$. To relate our framework to prior work, let us re-write $w(x)$ using the definition of $V(x)$:

$$w(x) = u(x, \theta^D) - [1 - \psi(\theta^D)]V(x) \tag{4.4.2}$$

$$= u(x, \theta^A) + \psi(\theta^D)V(x). \tag{4.4.3}$$

$$u(x, \theta^D) = u(x, \theta^A) + V(x). \tag{4.4.4}$$

Prior work on behavioral frictions often uses a formulation like equation (4.4.3), where we think of $V(x)$ as the "behavioral" component of preferences, while "decision utility" takes a form like equation (4.4.4). The behavioral component $V(x)$ is typically a deviation from classical forms of preferences that may or may not be due to a bias. When $\psi(\theta^D) = 1$, for instance, the planner knows with certainty that V is a non-standard but normative preference rather than a bias. When $\psi(\theta^D) = 0$, the planner knows that V reflects a bias.

Next we refine this example by considering specific behavioral frictions from prior literature.

Example 1.1. Defaults. We now denote elements of $\mathcal{X}$ by (x, d), where the first element is a choice object and the second is the default. In a slight abuse of notation, we suppose both of these are drawn from a set of available options: $d, x \in X$. To nest the fixed cost model of default effects in Example 1 we specify

$$V(x, d) = -1\{x \neq d\}\gamma. \tag{4.4.5}$$

where γ is the fixed cost of choosing some option other than the default [see e.g. Carroll et al., 2009; Bernheim et al., 2015]. The fixed opt-out cost structure matches key empirical aspects of default effects, and this structure nests a variety of mechanisms by which default effects might influence behavior [Goldin and Reck, 2022b].[17] But depending on the mechanisms and our normative interpretation of them, the fixed cost may or may not be a normative cost.

[17] This empirical pattern is observed in widely varied contexts [Choi et al., 2006; Haggag and Paci, 2014; Altmann et al., 2013; Brown et al., 2013] but not everywhere [Brot-Goldberg et al., 2023]. The complexity and opacity of the Medicare Part D plans studied in Brot-Goldberg et al. [2023] suggests that another important factor might be individuals' understanding of the options they could get upon opting out. One could employ our overall normative approach to evaluate defaults while allowing for this possibility, but this is not nested in Example 1.1.

The alternative frame implied by (4.4.5) is the utility function individuals maximize when they opt out of the default and make an active choice. Drawing parallels between this example and the next, we label the utility function $u(x, \theta^A)$ "intrinsic utility." When $\psi(\theta^D) = 1$, we impose that γ reflects a welfare-relevant cost; when $\psi(\theta^D) = 0$, γ reflects a bias.[18]

In treating default adherence as a "biases versus strange preferences" question, we do not allow active choosers to make mistakes. This restriction is relaxed in the more general version of the model in Goldin and Reck [2022b], but doing so obviously requires introducing more frames than the two we posit here. An analogous limitation to Example 1 applies in general: in models of biases versus strange preferences, the modeller picks one behavioral factor to consider as a bias or a strange preference, and assumes away deviations from individual welfare maximization due to any other behavioral factor to obtain RP-coincidence (1.3).

Example 1.2. Reference Dependence. Reference dependence is the subject of a rich theoretical and empirical literature [Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Kőszegi and Rabin, 2006; Crawford and Meng, 2011; Thakral and Tô, 2021], including many policy-relevant applications [DellaVigna et al., 2017; Rees-Jones, 2018; Seibold, 2021]. A lack of consensus about whether to regard this phenomenon as a bias or a preference has hindered our ability to evaluate policy in these settings [O'Donoghue and Sprenger, 2018]. Reck and Seibold [2023] consider a model, nested by Example 1, in which the behavioral component of preferences $V(.)$ is a reference-dependent payoff featuring loss aversion.

We use a similar setup to the previous example, but replace the default d with a

[18] When a real cost is inflated above its true value, for instance due to present bias, we capture this possibility by $0 < \psi(\theta^D) < 1$. Convexifying the possible views of welfare in this way also effectively captures views of welfare according to which fixed costs are partially but not fully normative, e.g. models of present bias.

reference point $r \in X$. When researchers model reference-dependent choice, they introduce a utility function over x with classical properties labelled "intrinsic utility" or "consumption utility" [e.g. Kőszegi and Rabin, 2006], which is additively separable from a gain-loss payoff over (x, r). We can nest this in our biases versus strange preferences setup if we posit a naturally occuring frame in which the individual makes choices based on both intrinsic and gain-loss utility, and an alternative frame in which the individual makes choices based on intrinsic utility alone.[19] Following Kőszegi and Rabin [2006], we assume intrinsic utility is additively separable, so that we may write $u(x, r, \theta^A) = \sum_{i=1}^{N} u_i(x_i)$. For parameters $\Lambda_i > 0, \beta \in (0, 1]$, we specify a gain-loss payoff of the form

$$V(x, r) = -\sum_{i=1}^{N} 1\{x_i \leq r_i\} \Lambda_i \left[u_i(r_i) - u_i(x_i)\right]^{\beta}, \tag{4.4.6}$$

The individual only incurs a payoff along some dimension if they incur a loss, $x_i \leq r_i$. The parameter Λ_i governs the strength of loss aversion along dimension i, while the parameter β governs diminishing sensitivity. We separately consider the case without diminishing sensitivity ($\beta = 1$) and with it ($\beta < 1$) below.

This is a similar form to that proposed by Kőszegi and Rabin [2006] – gains and losses are evaluated over "utils" rather than units of each good – except that 1) we disregard gain domain payoffs where $x_i > r_i$ along some dimension, and 2) we allow the extent of loss aversion Λ_i to vary across dimensions i rather than being fixed. These choices are motivated by a more detailed analysis of forms of gain-loss utility in Reck and Seibold [2023].[20]

Examples 1.1 and 1.2 under $\beta = 1$ are both cases of the "Affine Categorical Thinking

[19] We borrow the term "naturally occuring frame" from Bernheim et al. [2015].

[20] Including a gain-domain payoff whose strength is governed an additional parameter, usually denoted by η [Tversky and Kahneman, 1991], would not change the results of interest to us provided that η_i is not too strong along any given dimension i, in a sense formalized in Reck and Seibold [2023], Appendix B6.

Model" from Ellis and Masatlioglu [2022]. The salience model of Bordalo et al. [2012] is also an Affine Categorical Thinking Model. One could adapt the approach we develop here to analyze welfare in this salience model, or another Affine Categorical Thinking Model. Not all such models can be nested within Example 1, but our overall approach is applicable to these models because they feature a family of intrapersonally comparable utility functions. Our next example does not fall within this class of models.

Example 1.3. Probability Re-Weighting. Starting with Kahneman and Tversky [1979], researchers have modelled deviations from expected utility theory due to the reweighting of objective probabilities [see also Prelec, 1998; Abdellaoui, 2000; Chateauneuf et al., 2007]. In a recent welfare analysis of state-run lotteries, Lockwood et al. [2023] present a model in which the main behavioral friction is probability re-weighting and it is ambiguous whether re-weighting reflects a bias or a normative preference. In particular, individuals' revealed preferences – identified empirically using demand responses to changes in lottery prizes – suggest their utility function puts excess weight on the jackpot payoff especially, i.e. more weight than expected utility requires given the extremely low probability of winning a jackpot. This finding suggests a particular form of probability re-weighting, and the main question they confront for welfare analysis is whether this jackpot payoff effect reflects a bias or a normative preference.

To nest their model in Example 1, we think of each component of $x = (x_1, ..., x_N)$ as the payoff that is realized for each realization of an uncertain state variable. Objective probability of each realized state is $\pi = (\pi_1, ..., \pi_N)$. Individuals re-weight each objective probability according to a function $f(\pi)$. Individuals are endowed with a Bernoulli utility function $\mu(x_n)$. Utility in the fixed decision-making frame is $u(x, \pi, \theta^D) = \sum_n f(\pi_n)\mu(x_n)$.

When $f(\pi) = \pi$ everywhere, we have classical expected utility maximization. If we view the vNM independence axiom as normative, then normative utility should coincide with expected utility. For an alternative frame θ^A in which the individual's choices respect the independence axiom, we have $u(x, \pi, \theta^A) = \sum_n \pi_n \mu(x_n)$. The disagreement between these two views of welfare is then, by our definition,

$$V(x, \pi) = \sum_n [\pi_n - f(\pi_n)]\mu(x_n). \tag{4.4.7}$$

Now with our framework, we can think of a planner who is uncertain about whether the excess weight on the jackpot payoff (and the resulting under-weighting of payoffs in other states) is normative. Lockwood et al. [2023] model the extent to which reweighting reflects a bias with a parameter that is isomorphic to $\psi(\theta^A)$ here.

4.4.2 Example 2: Present Focus

Our next example is motivated by prior work on present focus and the notion of intertemporal selves [e.g. Laibson et al., 1998; Caliendo and Findley, 2019]. The options are lifetime consumption plans: $\mathcal{X} = \mathbb{R}^T_+$, where T is the number of time periods. An option is now $x = (x_1, ..., x_T)$. The frame in this model is the vantage point from which individuals evaluate a consumption plan. We characterize individuals' preferences under commitment, i.e. we think of the individual selecting a full consumption plan in each period. We assume individuals are quasi-hyperbolic discounters as in Laibson [1997]. We also assume there is a period 0 in which the individual is entirely forward-looking, i.e. they do not consume or receive flow utility. For two parameters $\beta > 0$, $\delta \leq 1$, a flow utility function $\mu(x_t)$, and a vantage point $\tau = 0, ..., T$ we specify:

$$u(x, \tau) = \mathbb{1}\{\tau > 0\}\delta^\tau \mu(x_\tau) + \beta \sum_{t \neq \tau} \delta^t \mu(x_t). \qquad (4.4.8)$$

Note that with this formulation, we endow the period τ self with preferences over the prior selves' consumption; this is unconventional and we discuss the rationale for this modelling choice below. A commonly adopted approach to welfare analysis in models like this is to respect the revealed preferences of the period 0 self, sometimes called the "long-run view." The period 0 self is a classical exponential discounter, and in fact we find that a planner's welfarist objective based on formulation (4.4.8) has a representation along similar lines to the Biases versus Strange Preferences example. By construction, the planner's welfare function takes the following form:

$$\begin{aligned} w(x) &= \beta \sum_{t=1}^{T} \delta^t \mu(x_t) + \sum_{\tau=0}^{T} \psi(\tau)\mathbb{1}\{\tau > 0\}(1-\beta)\delta^\tau \mu(x_\tau) \\ &= u(x, 0) + (1 - \psi(0)) \sum_{\tau=1}^{T} \psi(\tau|\tau > 0)[u(x, \tau) - u(x, 0)] \qquad (4.4.9) \end{aligned}$$

This formulation for welfare resembles the formulation for welfare from Example 1, Equation (4.4.3): the first term is a utility function with classical features and the second is a deviation from classical preferences weighted by the planner's beliefs about the probability the individual has non-standard normative preferences $(1 - \psi(0))$. In this case, there is more than one alternative view because each of the period $\tau > 0$ selves could receive different normative weights (see also Example 3 below).

In the theory of social welfare functions, we frequently find an "anonymity" condition imposed on social welfare functions, which requires that no two individuals should have their utility differentially weighed [e.g. Maskin, 1978]. Anonymity does not seem useful for our model in general but it is intuitive to impose such a condition over the $\tau > 0$ selves. Doing so, we find a justification for the planner's adopting the long-run view of welfare, which *does not require assuming that present focus is a behavioral bias.*

Proposition 4.4.1. ***Intertemporal "Social" Welfare and the Long Run View.*** *In this model, if $\psi(\tau)$ is constant for $\tau > 0$, then for any $\psi(0)$, the planner's preferences coincide with the long-run view of welfare $u(x, 0)$.*

Proposition 4.4.1 presents a new justification for the long-run view of welfare, contributing to a debate in the literature about the normative justification (or lack thereof) for adopting the long-run view in welfare analysis [see discussion in Caliendo and Findley, 2019]. The intuition for the result is that even if present focus payoffs are normative, when they are aggregated across all the intertemporal selves that experience them, the intertemporal tradeoffs over present-focused payoffs are proportional to the intertemporal tradeoffs over long-run utility. However, the result relies on the assumption that the extent of present focus β is constant over time. This assumption rules out, for example, that individuals are present focused while they are young but not when they are old. In that case, the planner must make a material judgment about how to weigh the present focus payoff that is only present for the younger selves.

Remark on Intertemporal Selves Preferences Formulation. Holding x_s fixed for every $s < \tau$, the above generates the same choices as the conventional β-δ representation of preferences, i.e. $\tilde{u}(x_{t\geq\tau}, \tau) = \mathbb{1}\{\tau > 0\}\mu(x_\tau) + \beta \sum_{t=\tau+1}^{T} \delta^{t-\tau}\mu(x_t)$. The formulation differs from most prior work on intertemporal selves in that the period τ self is endowed with classically discounted preferences over consumption in periods $t < \tau$. This approach appears to fix multiple related issues with the intertemporal selves model identified by Bernheim and Rangel [2009], at the cost of being, admittedly, philosophically confusing.

What does it mean for the period τ self, who cannot go back in time to choose a different amount of consumption, to have preferences over past consumption? We assume the period-τ self agrees with most of their prior selves about intertemporal consumption tradeoffs, so that endowing this self with preferences over past consumption does not generate any new choice inconsistencies relative to those we find for observable, forward-looking choices. More formally, we assume that for any pair $\tau > 0$, $\tau' > 0$, if we consider two consumption plans x, x' such that $x_\tau = x'_\tau$ and $x_{\tau'} = x'_{\tau'}$ then we have $u(x, \tau) \geq u(x', \tau) \iff u(x, \tau') \geq u(x', \tau')$. Behaviorally, the period-τ and period-τ' selves make the same choices when we hold consumption in τ and τ' fixed in the menu. This approach works well for the β-δ model but appears to be less well-suited to more generic models of non-classical discounting.

Welfare with this formulation accords with BR-Dominance over committed choices. The setup is "rectangular" in that for any frame τ individuals have frame-dependent rational preferences over the entire option space; Bernheim and Rangel show that a lack of rectangularity in the naive application of "intertemporal-self Pareto optimality" leads to conceptual problems. If we naively write down a utility function in which the selves care only about current and future consumption, allocating all resources to the last-period self will always be an intertemporal-self Pareto optimum. But revealed preference does not suggest that the individual robustly prefers options that defer all consumption to the final period. We address this problem by adopting rectangular preferences using formulation (4.4.8) and considering preferences under commitment. Players in conventional games have deep structural preferences over outcomes influenced by the actions of prior movers even when they cannot choose to alter another player's actions; we make a similar assumption for the intrapersonal game here.

Why focus on revealed preferences under commitment to define welfare? Here, we are making an ex ante normative assumption that the consumption plan and not the process of choice (and the associated emotions, etc.) is sufficient for the evaluation of normative preferences. For example, we require that a naive agent who makes a plan and fails to adhere to it will not experience shame that alters their normative choices. This can also be viewed as an assumption about the set of potentially normative frames, which precludes revealed preferences in situations featuring non-commitment from being normative. This type of assumption, a version of which we make generally in defining normative preferences over options rather than menus, is criticized in Bernheim et al. [2024]. They also develop tools for relaxing it using additional information to identify the normative import of emotions. We do not know how our argument on the robustness of the long-run view would be altered by accounting for additional normative concerns due to choice processes and emotions. We defer a fuller treatment of this question to future work.

Finally, we remark on comparability of welfare across intertemporal selves. The units of utility with our formulation are determined by $\mu(x)$, which is cardinal (so that the individual can evaluate intertemporal tradeoffs) [Montiel Olea and Strzalecki, 2014]. We require an assumption that that the units of μ are the normative units for welfare analysis, but then comparisons of utility across τ with the formulation above are well-defined. For $\tau > 0$, level comparability also does not seem to be a problem: evaluating equation (4.4.8), we find that a constant consumption growth path generates the same level of utility for any $\tau > 0$. But when we compare $\tau = 0$ versus $\tau > 0$, examining equation (4.4.8), we find that the conventional level normalization, $\mu_\tau(0) = 0$ for every τ, implies that the present-focused self with $\beta < 1$ will always have more utility than the period 0 self due to the present-focused payoff that only the $\tau > 0$ self receives. Formally,

$$\forall \tau, \mu_\tau(0) = 0 \text{ and } \forall x \geq 0, \frac{d\mu_\tau(x)}{dx} \geq 0 \implies \forall x \geq 0, \min_{\psi \in \Delta(\Theta)} \psi(\tau) u(x,\tau) = u(x,0). \tag{4.4.10}$$

This logic implies that the globally ambiguity averse planner adopts the long run view of welfare. Whether this is another novel rationale for the long-run view or an artefact of a dubious level comparability assumption is debatable. This is a moot point when we introduce the anonymity condition from Proposition 4.4.1. If it is a problem, the solution requires specifying a consumption plan for which the level of utility is equal between $\tau = 0$ and $\tau > 0$ selves.

4.4.3 Example 3: Is a Feature of the Environment a Frame?

In Example 1.1, the default cannot be a frame by construction; the same is true of the reference point in Example 1.2. If we treat default adherence as a normative preference then by definition the default should not be a frame,[21] but if we do not, then it could be one and we might think of choices made given each default as coming from distinct frames rather than a unitary, naturally occurring frame. In the present example, we use a similar setup to Example 1.1, the frame has two components: $\theta = (\theta_1, \theta_2)$. The second component is a factor like the defualt ($\theta_2 \in X$),[22] and the first component, $\theta_1 \in \{0, 1\}$ indicates whether the second component can really be viewed as a frame ($\theta_1 = 1 \implies \theta_2/d$ is a frame), i.e. whether we obtain the frame exclusion condition from Observation 1.

[21] To see this, suppose two options and the individual chooses x_1 when x_1 is the default and x_2 when x_2 is the default (x_1 and x_2 are in the menu for both of these choices). When θ^D is the normative frame, this implies the individual's normative preference depends on which option is the default, which violates Frame Exclusion (Observation 1) if we regard the default itself as a frame.

[22] In this example, there are a continuum of frames, contrary to our initial setup with finite frames. We present an extension to continuous frames in Appendix D.5.

We express the utility function as $u(x, d, \theta_1, \theta_2)$ and we make two restrictions to capture our intuition. When $\theta_1 = 0$, saying feature d is not a frame requires that $u(x, d, 0, \theta_2)$ must be constant over θ_2, which we express with a utility function $u_0(x, d) \equiv u(x, d, 0, \theta_2)$ for any θ_2. If $\theta_1 = 1$, feature d is a frame so frame exclusion requires $u(x, d, 1, \theta_2)$ to be constant over d, which we express with a function $u_1(x, \theta_2) = u(x, d, 1, \theta_2)$. Denoting disagreements between the $\theta_1 = 0$ and $\theta_1 = 1$ cases by $V(x, d, \theta_2) = u_0(x, d) - u_1(x, \theta_2)$ and letting $\psi^0 = \sum_{\theta_2} \psi(0, \theta_2)$ be the total weight on $\theta_1 = 0$, we derive an identity similar to equation (4.4.2):

$$w(x) = u_0(x, d) - (1 - \psi^0) \sum_{\theta^2 \in D} \frac{\psi(1, \theta_2)}{1 - \psi^0} V(x, d, \theta_2). \qquad (4.4.11)$$

The weight ψ_0 is similar to $\psi(\theta^D)$ in Example 1. With this setup, we confront more ambiguity than in the biases versus strange preferences case from Example 1. For instance, if the planner knew with certainty that the effect of d on choices reflects a bias, then this resolves all ambiguity in Example 1 ($\psi(\theta^D) = 0$), but if analogously $\psi_0 = 0$ in (4.4.11), substantial ambiguity in welfare remains due to choice inconsistencies as d varies. We note that the most models of default effects considered in Bernheim et al. [2015] are nested in Example 1.1, but the anchoring model they consider resembles Example 3 under $\psi^0 = 0$. Understanding the difference between these examples clarifies why adopting the anchoring model generates more ambiguous welfare effects.

We do not engage deeply with models like Example 3 in the remainder of this paper, but this is done in the interest of providing simple illustrations of our robustness concept rather than our thinking that the approach applied by Example 1 is superior to the one implied by Example 3 for any particular behavioral phenomenon.

4.5 Robust Optimal Policy in Applications

In this section, we characterize the optimality of policies when there is normative uncertainty. For each case, we start by characterizing the optimal policy given a set of beliefs ψ about which frame is normative and a utilitarian objective like Equation (4.2.5). Then, we incorporate the case that the planner has some aversion to uncertainty, i.e. ambiguity aversion from Equation (4.2.6).

4.5.1 Setup

Before we turn to our examples, we translate the different formulations of the planner's preferences from Section 4.1 to notions of optimal policy:

Definition. A policy P^* is a ψ-*optimum* for $\psi \in \Delta(\Theta)$ if $P^* \in \arg\max_{P \in \mathcal{P}} E_\psi[u(x(P), \theta)]$.

For a given beliefs ψ, a ψ-optimum maximizes the planner's preferences in the case of risk-neutrality / no uncertainty-aversion.

Definition. For a given set of distributions $\Psi^* \subseteq \Delta(\Theta)$, a policy P^* is a *robust optimum* if

$$P^* \in \arg\max_{P \in \mathcal{P}} \min_{\psi' \in \Psi^*} E_{\psi'}[u(x(P), \theta)].$$

If Ψ^* is the closed and convex set of beliefs which define the planner's preferences in the ambiguity-averse case as in Section 4.2.2 then robustness corresponds to maximizing these preferences. We leave the particular value of ψ under risk or Ψ under ambiguity unspecified and illustrate how they matter for the optimum. In thinking about robust optima, we often adopt the intuition introduced by Hansen and Sargent [2001], thinking of an "evil agent" who, subject to the planner's chosen policy, picks a ψ to minimize welfare; to be a robust optimum, the policy is the best possible policy for welfare given the evil agent's reaction.

Definition. A policy P^* is a *globally robust optimum* if it is a ψ-optimum for all $\psi \in \Delta(\Theta)$.

Obviously, a globally robust optimum will also be a robust optimum for any $\Psi^* \subseteq \Delta(\Theta)$. Global robustness also has a straightforward relationship to BR-dominance:

Lemma 4.5.1. ***BR-Optimality and Global Robustness.*** *A policy $P^* \in \mathcal{P}$ is a globally robust optimum if and only if for every $P' \in \mathcal{P}$, for every $\theta \in \Theta$, $x(P^*) \succeq_\theta x(P')$.*

We generally focus on more global characterizations of optimal policies and defer perturbation-based characterizations to Appendix D.3. There, we show that a generalized version of the reduced-form welfare formula from Mullainathan et al. [2012] continues to hold under normative uncertainty: the welfare effect of a marginal reform equals its expected direct effect minus expected marginal internalities, both averaged across normative frames. We illustrate how this plays out in the examples in this section. A stylized corrective tax example (Section 4.5.2) illustrates the key intuition and how ambiguity aversion modifies optimality conditions.

4.5.2 Examples

Now we explore how the notions of optimality defined in Section 4.5.1 play out in some of our examples from Section 4.4.[23]

[23] Deriving characterizations of robust optimality under probabilistic uncertainty is straightforward but not very instructive beyond what we do here.

Optimal Defaults

We begin with the optimal defaults problem studied in Carroll et al. [2009]; Bernheim et al. [2015]; Chesterley [2017]; Goldin and Reck [2022b], and others. In the model we introduced in Example 1.1, the *intrinsic optimum* $x^* \equiv \arg\max_x u(x, \theta^A)$ is assumed to be known to the social planner.[24] We begin there, and then consider the case where the intrinsic optimum is not known, which which makes the planner's normative objective equivalent to aggregate welfare in Bernheim et al. [2015] and social welfare in Goldin and Reck [2022b]. Aggregation over potential intrinsic optima is interpreted in these papers as arising due to unobservable interpersonal heterogeneity rather than intrapersonal concerns.

Our illustrations of this model are simulations built on the assumption that the choice variable is one-dimensional, $x \in \mathbb{R}$. We suppose utility is approximately quadratic: $u(x, \theta^A) = -\frac{\alpha}{2}(x - x^*)^2$ for a known parameter $\alpha > 0$ and the intrinsic optimum x^*. For simplicity, we further assume the opt-out cost γ is known and x^* is either known to equal 0 or it follows a Gaussian distribution centered around 0. This pins down the shape of the welfare function, which we illustrate in Figure 4.1.

Figure 4.1a depicts welfare as a function of the default (under BIC), given a known intrinsic optimum. We plot welfare for varying weights on the frame θ^D, $\psi(\theta^D)$ from 0 to 1. Applying our definitions of optimality, we observe the following, which turn out to be true in full generality, i.e. without any of the restrictions introduced in the previous paragraph:

- The ψ-optimal defaults are the intrinsic optimum x^* and any default under

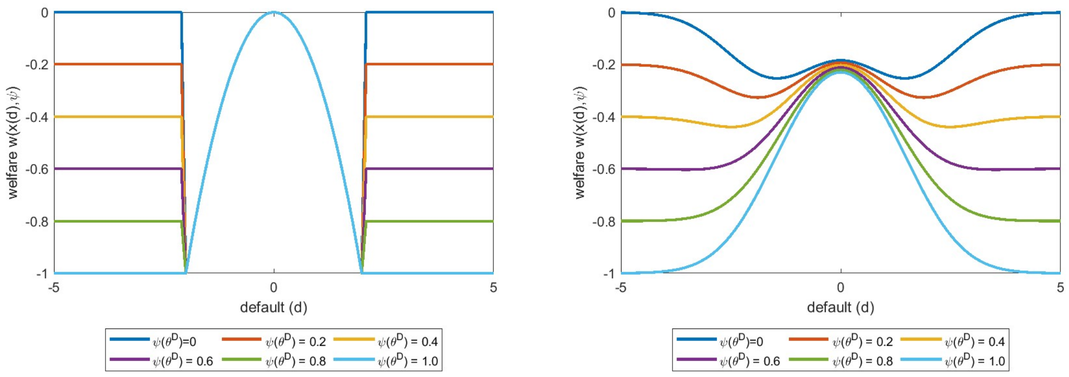
[24] The intrinsic optimum x^* is called the "ideal option" in earlier work on defaults [Bernheim et al., 2015; Goldin and Reck, 2022b] and the analogous option is called the "intrinsic optimum" in the reference dependence literature [Kőszegi and Rabin, 2006; Reck and Seibold, 2023]. In both cases, x^* does not depend on the default/reference point by construction; it obviously does depend on other aspects of the choice situation, e.g. prices. We suppose x^* is unique for simplicity.

which the individual chooses actively.[25]

- The intrinsic optimum x^* is the unique globally robust optimum.
- An active choice optimum is robust if and only if there is no ambiguity (Ψ^* is singleton) and $\psi(\theta^D) = 0$.

The fact that we find a globally robust optimum in this case clearly depends on the assumption that x^* is known, i.e. that the policymaker can set as the default the option that the individual would choose if they opt out. In this case, setting that option as the default is ensured to give the individual the best possible option and avoid any potentially normative opt-out cost. Relaxing the assumption that the planner knows x^*, we find the following characterization of robustness in the quadratic/Gaussian case:

Figure 4.1: Illustration of the Optimal Default



(a) Known Intrinsic Optimum

(b) Unknown Intrinsic Optimum

Proposition 4.5.1. ***Robust Optimal Defaults when the Intrinsic Optimum is Unknown***

[25] Formally, the set of ψ-optima is $\{d | d = x^* \text{ or } d < \underline{x} \text{ or } d > \overline{x}\}$, where $\underline{x}$ and $\overline{x}$ are the thresholds for active choice. These thresholds are equal to -2 and +2 in the illustration in Figure 4.1a.

- *The ψ-optima are the expected intrinsic optimum and the most extreme default possible in the positive or negative direction (henceforth the* extremum default*).*

- *None of the ψ-optima are globally robust.*

- *If the expected intrinsic optimum is ψ-optimal for some ψ in the interior of Ψ^*, the expected intrinsic optimum is the unique robust optimum.*

- *If the expected intrinsic optimum is not ψ-optimal for any $\psi \in \Psi^*$, the extremum default is the unique robust optimum.*

- *If the expected intrinsic optimum is ψ-optimal for some ψ on the boundary of Ψ^* but not the interior, both the expected intrinsic optimum and the extremum default are robust optima.*

Robust Control and the Optimal Default. Suppose $\Psi^* = B(\kappa, \psi)$ for some $\kappa > 0$ and some $\psi \in \Delta(\Theta)$.

If the extremum defualt is ψ-optimal, there is a threshold $\overline{\kappa}$ such that

- the extremum default is the unique robust optimum for $\kappa < \overline{\kappa}$, but

- the expected intrinsic optimum is the unique robust optimum for $\kappa > \overline{\kappa}$.[26]

If the expected intrinsic optimum is ψ-optimal, the expected intrinsic optimum is the robust optimum for any κ. The logic of the proof is illustrated by Figure 4.1b. When the intrinsic optimum is unknown, the default that maximizes welfare depends on the normative judgment about whether and to what extent the opt-out cost implied by revealed preferences, γ, is normative. The welfare-maximizing default is the expected intrinsic optimum when $\psi(\theta^D)$ is sufficiently large, while the extremum default maximizes welfare when $\psi(\theta^D)$ is sufficiently small. As such, a global robustness criterion like Bernheim and Rangel [2009] is inapplicable, provided that

[26] In the knife-edge case $\kappa = \overline{\kappa}$, both the extremum default and the expected intrinsic optimum are κ-ψ robust.

sufficiently extreme defaults (i.e. those where sufficiently many individuals make active choices) are feasible. Even so, there is still a sense in which the setting the expected intrinsic optimum as the default (i.e. minimizing opt-outs) is a more robust policy recommendation than an extremum default. As we can see in Figure 4.1b, the expected intrinsic optimum remains a local optimum as we vary normative weights, while the active choice policy becomes strictly worse when we put more normative weight on opt-out costs (because making an active choice requires incurring these costs). The intuition that this makes the the extremum default a less robust optimum appears in Goldin and Reck [2022b]; here we find that formalizing an approach to robustness allows us to capture that intuition.[27]

(Unconstrained) Optimal Reference Points

In this example, we employ a two-dimensional version of Example 1.2 and introduce some additional simplifying structure to derive a reduced-form representation of the planner's normative objective with some interesting commonalities to the previous example.

We suppose options are two-dimensional $x = (x_1, x_2)$, and that intrinsic utility is quasi-linear with x_2 the numeraire. The individual faces a budget constraint for given prices and income, with price of p_2 normalized to 1 and $p_1 = p$. The form of gain-loss utility follows from equation (4.4.6).

$$u(x_1, x_2, \theta^A) = \log(x_1) + x_2.$$

$$px_1 + x_2 = Z.$$

[27] Our proof of this result leverages the simplifying structure of our simulations, but the result generalizes. For less restrictive treatments of the optimal defaults problem, refer to Goldin and Reck [2022b]; Bernheim et al. [2015]; Bernheim and Gastell [2021].

$$V(x_1, x_2, r_1, r_2) = -1\{x_1 \leq r_1\}\Lambda_1[\log(r_1) - \log(x_1)]^\beta - 1\{x_2 \leq r_2\}\Lambda_2[r_2 - x_2]^\beta \tag{4.5.1}$$

We are interested in whether and when the planner might wish to induce the individual to use a different reference point. Evidence from the lab and the field suggests that policy reforms can indeed *shift reference points to some extent* [e.g. Homonoff, 2018; Rees-Jones, 2018; Seibold, 2021]. However, the full policy space $\mathcal{P}$ is difficult to characterize in applied settings where reference dependence appears to matter, because there is little consensus about how to model the formation of reference points. Here, we consider an environment with fixed prices and incomes, and we suppose that the planner can set the reference point at any point on the budget constraint: $\mathcal{P} = \{(r_x, r_y)|pr_x + r_y = Z\}$.[28] This confers a likely unrealistic amount of power on the planner to shape the individual's reference point, but with this approach we nevertheless find a thought-provoking characterization of robustness. To see why, first note that the model admits a reduced-form representation for welfare in terms of a single dimension of choice, x_1, that has some common features with the previous example. Assuming an interior solution, for fixed p and Z, we can re-write intrinsic utility as:

$$u(x_1, \theta^A) = \log(x_1) + Z - px_1.$$

$$V(x_1, r_1) = \begin{cases} -\Lambda_1[\log(r_1) - \log(x_1)]^\beta, & x_1 \leq r_1 \\ -\Lambda_2[px_1 - pr_1]^\beta, & x_1 > r_1. \end{cases} \tag{4.5.2}$$

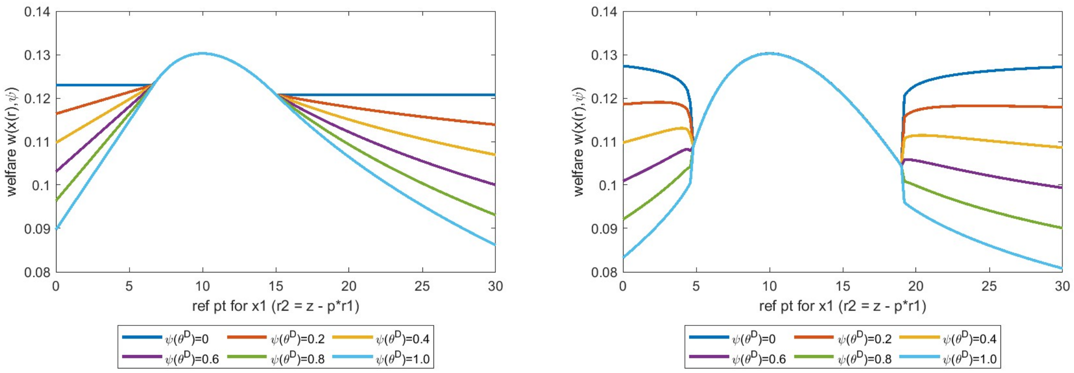
The *intrinsic optimum* is here characterized by $x_1^* = \frac{1}{p}$; for numeric illustration here we simply set $p = 0.1 \implies x_1^* = 10$.[29] To simulate welfare in the model, we

[28] If we relax the restriction that (r_1, r_2) must lie on the budget constraint, the globally robust optimum is for the planner to set the lowest reference point possible along each dimension; see Reck and Seibold [2023] Appendix B for further discussion.

[29] Varying prices is a straightforward extension building on our work in the next subsection. Introducing price variation requires specifying how such variation affects the reference point.

suppose $\Lambda_1 = \Lambda_2 = 0.5$, we set $Z = 10$. We express welfare in equivalent variation units relative to a baseline where $x_1 = r_1$,[30] and further normalize this as a share of income for interpretability.[31]

Figure 4.2: Illustration of Optimal Reference Points



(a) Without Diminishing Sensitivity ($\beta = 1$)

(b) With Diminishing Sensitivty ($\beta = 0.5$)

Figure 4.2 plots welfare as a function of the reference point for x_1, where the reference point for good 2 is then pinned down by the budget constraint. In the first panel, we rule out diminishing sensitivity ($\beta = 1$), and in the second we include it, supposing $\beta = 0.5$. Without diminishing sensitivity, we find that the intrinsic optimum is the globally robust optimum, as in the defaults model under known x^*, and in this case it is also the unique ψ-candidate optimum for any ψ. That the optimal reference point is the intrinsic optimum when $\psi(\theta^D) = 1$ is key to the Preferred Personal Equilibrium concept of Kőszegi and Rabin [2007]. In a non-stochastic environment, selecting a Preferred Personal Equilibrium from the set of Personal Equilibria is

[30] Under our quasilinariy assumption using any baseline where $x_2 \geq r_2$ ($\iff$ $x_1 \leq r_1$) would give the same quantitative expressions for welfare, but outside the quasilinear case, using $x = r$ as the baseline ensures we obtain Assumption 8. Intuitively, because loss aversion modifies marginal value of a dollar according to revealed preferences under θ^D but not θ^A, it seems most sensible to compare money-metric welfare under these two frames from a baseline where the individual incurs no losses.

[31] In other words, we plot $\tilde{w}(x, \psi) = \frac{E_\psi[u(x,\theta)]-Z}{Z}$.

equivalent to solving our planner's problem under $\psi(\theta^D) = 1$. In fact, we observe here that the planner – or an individual setting a reference point to maximize the welfare of their future, reference-dependent self as in Fudenberg and Levine [2006] – would also want to set the intrinsic optimum as the reference point for any $\psi(\theta^D)$ [see Reck and Seibold, 2023, for further discussion].

We observe that welfare behaves differently in three domains in both panels of Figure 4.2a. To understand why, first observe that when the reference point is on the budget constraint, the individual can either consume the reference point itself, a bundle with $x_1 < r_1$ and $x_2 > r_2$ (a loss over good 1) or a budle with $x_1 > r_1$ and $x_2 < r_2$ (a loss over good 2). When the reference point falls around the intrinsic optimum of $x_1^* = 10$, the individual chooses the reference point, so $V(x, r) = 0$ because there are no gains or losses, and their welfare is peaked around 10 because this is the intrinsic optimum. At a very high reference point for good 1, the individual chooses to consume some $x_1 > x_1^*$ to reduce their losses over good 1 due to Loss Aversion. Similarly at a very low reference point for good 1, the individual consumes more x_2 to reduce losses in x_2 and therefore consumes less of good 1 than x_1^*. Without diminishing sensitivity x_1 is constant over r in the latter two cases, so under $\psi(\theta^D) = 0$, welfare is flat. When $\psi(\theta^D) > 0$, changing the reference point has direct welfare effects, by increasing the magnitude of losses, and consequently, welfare falls further as r moves to further extremes and the losses grow.

We introduce diminishing sensitivity in Figure 4.2b. Here, welfare unsurprisingly behaves similarly in the domain where $x = r$ but we find non-monotonicity in welfare outside this domain: more extreme reference points appear to be more desirable for small $\psi(\theta^D)$. The intuition here is similar to penalty defaults above: under diminishing sensitivity, as losses grow to an extreme, the individual stops trying to avoid losses. When the losses themselves receive little welfare weight (($\psi(\theta^D)$ is near zero), this is desirable, because the planner believes that the individual *should* stop trying to avoid losses. However, when loss aversion is viewed more as a normative preference ($\psi(\theta^D)$ is near 1) the direct negative welfare effect of imposing extreme losses on the individual makes extreme reference points highly undesirable. Like extremum defaults, reference points that generate extreme losses are desirable under $\psi(\theta^D) = 0$ but this desirability is *not robust*. Based on what we found in Figure 4.1, it is obvious that if we introduced some uncertainty about the intrinsic optimum, we could even get an extreme reference point to be ψ-optimal for some sufficiently small $\psi(\theta^D)$, but this will tend not to be robust just as in Proposition 4.5.1.

Our notion of robustness plays out very similarly in the defaults and reference points models (compare Figure 4.1b and Figure 4.2b). We generalize the common features of this example in Proposition D.1.3, in the Appendix. Intuitively, in both examples, setting the default or reference point at the (expected) intrinsic optimum minimizes disagreements about welfare across frames. The general proposition presents a sufficient condition for a ψ-optimum to be a robust optimum, which also involves minimizing such disagreements.

Corrective Taxation

Let us consider optimal corrective taxation in the biases versus strange preferences example. Suppose for simplicity that we are in the quasi-linear environment from the previous example, and introduce a nonlinear tax on good 1 according to a tax schedule $T(x_1)$, which is fully incident on consumers. Utility under the alternative/classical preferences frame θ^A is

$$u(x_1, \theta^A) = \mu(x_1) + Z - px_1 - T(x_1) + R$$

where the sub-utility function $\mu(x_1)$ is twice differentiable, increasing, and concave. The variable R is rebated revenue from the corrective tax, which is determined by the simple government budget constraint $R = T(x)$. Suppose further that the tax is unrelated to the behavioral friction, so $V(x)$ is invariant to T; this rules out misperception of tax incentives.[32]

Expressing the disagreement $V(x) = u(x, \theta^D) - u(x, \theta^A)$ as a function of x_1 (leveraging the budget constraint as in the previous example), and assuming paternalistic risk neutrality, we write

$$w(x) = u(x_1, \theta^A) + \psi(\theta^D)V(x_1)$$

Following the same logic as Mullainathan et al. [2012], the ψ-optimal corrective tax is

$$T^*(x; \psi) = [1 - \psi(\theta^D)]V(x) + C, \tag{4.5.3}$$

where C is a constant pinned down by the government budget constraint. This can be understood by taking a derivative with respect to x_1. Where T^* is differentiable with respect to x_1, we find

$$\frac{\partial T(x_1; \psi)}{\partial x_1} = [1 - \psi(\theta^D)]\frac{dV(x_1)}{dx_1},$$

equating the marginal tax rate with the expected marginal internality (see equation D.3.1).[33] What about robust optima? For a given amount of x_1 chosen by the

[32] This is a common assumption in prior work on corrective taxation, but there are many proposed theories of tax misperception in which the assumption is obviously violated. Integrating the theory of corrective taxes for internalities with the theory of tax misperceptions is beyond the scope of this paper.

[33] To prove that this describes the optimal tax, observe that with this schedule, the individual's choice of x_1 optimizes the planner's expectation for their welfare over x_1:

$$u(x_1, \theta^D) = \mu(x_1) + Z - px_1 - T(x_1) + R + V(x_1) = u(x_1, \theta^A) + \psi(\theta^D)V(x_1) = w(x_1).$$

individual (under BIC in the frame θ^D), we note that welfare is increasing in $\psi(\theta^D)$ if $V(x_1) > 0$ and decreasing if $V(x_1) < 0$. From this observation we can prove the following:

Proposition 4.5.2. *Let* $\underline{\psi} \equiv \min_{\psi \in \Psi^*} \psi(\theta^D)$, *and let* $\overline{\psi} \equiv \max_{\psi \in \Psi^*} \psi(\theta^D)$. *The robust optimal marginal tax rate given* Ψ^* *is*

$$\frac{dT^*(x_1)}{dx_1} = \begin{cases} [1-\underline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) > 0 \\ [1-\overline{\psi}]\frac{dV(x_1)}{dx_1} & V(x_1) < 0 \\ 0 & V(x_1) = 0. \end{cases} \tag{4.5.4}$$

The intuition is as follows: the ambiguity averse planner wishes to set a tax rate that is optimal in the worst case scenario for normative preferences. When $V(x_1) > 0$ at some chosen x_1, by construction $u(x_1, \theta^D) > u(x_1, \theta^A)$, so the worst-case scenario places maximal weight on the "biases" frame θ^A and minimal weight on the "strange preferences" frame θ^D. When $V(x_1) < 0$, we find the opposite.

How corrective taxation plays out therefore depends on whether the level of utility is higher in the biases or strange preferences case. For example, if we consider the corrective taxation of addictive goods Gruber and Köszegi [2001]; O'Donoghue and Rabin [2006]; Allcott et al. [2019], ambiguity arises over the question of whether the individual is rationally or harmfully addicted. In the rational case, the optimal corrective tax is zero because addiction is not a bias, while in the harmful case, the optimal correcive tax seeks to mitigate over-consumption of the addictive good. If we assume that the level of utility is lower when the individual is harmfully addicted, which seems highly intuitive, then 4.5.2 suggests the optimal corrective tax will be shaded toward the harmfully addicted case. In this setting, valuing robustness leads the planner to select a *more* paternalistic policy, in contrast to the defaults and reference points examples.

In the next section, we briefly examine the joint optimality of a two-dimensional policy involving a nudge and a corrective tax.

Nudges

Perhaps the most well-known example of behavioral public policy is the use of nudges [Thaler and Sunstein, 2009]—interventions designed to influence behavior without restricting choice or imposing significant costs.

In this section, we examine the welfare effects of nudges under conditions of normative ambiguity, as a way to formalize longstanding concerns surrounding these policies. We build on the framework of Allcott et al. [2022] to evaluate a nudge intended to "debias" behavior. As a motivating example, we consider the case of a potentially controversial nudge; graphic warning labels on cigarette packets. We also introduce some interpersonal heterogeneity in order to illustrate how our framework interacts with the interpersonal heterogeneity that is central in studies like Allcott et al. [2022].

Setup. A unit mass of individuals are deciding whether to buy a product (cigarette packet). Their value is $v \sim F$, the price is p and utility is quasi-linear. All individuals have heterogeneous bias $\gamma \in \mathbb{R}^+$ which shifts demand but not welfare. The government has access to a nudge (intensity $\sigma \in \mathbb{R}^+$) which works to de-bias γ - each individual's demand *reduces* with sensitivity $\tau \in \mathbb{R}^+$ to the nudge, but their welfare is not affected apart from through the direct per-σ psychic cost of the nudge,

$I \in \mathbb{R}^+$.[34] We simplify the setup by assuming a perfectly-competitive supply-side. The planner can levy a tax t which increases the price paid to $p + t$. Allcott et al. [2022] adopt the following formulations of preferences:

$$\text{Decision Utility: } u(\text{buy}, \theta^D) = v + \underbrace{\gamma}_{\text{“Bias”}} - \underbrace{\sigma\tau}_{\text{Nudge}} - p$$

$$\text{Experienced Utility: } u(\text{buy}, \theta^E) = v - p - \underbrace{\sigma I}_{\text{Psychic Cost}}$$

Normative Ambiguity: θ^E might not be the normative frame, as Allcott et al. [2022] assume. We consider two sources of ambiguity: the consumer might not have biased consumption initially (γ is welfare-relevant), also the planner cannot know whether there exists a psychic cost I (I does not affect behavior). This intuition suggests five potentially normative frames:

$$\begin{aligned}
\text{Decision Utility: } u(\text{buy}, \theta^D) &= v + \gamma - \sigma\tau - p \\
\text{Classical Utility: } u(\text{buy}, \theta^C) &= v - p \\
\text{Psychic-Cost Utility: } u(\text{buy}, \theta^P) &= v - p - \sigma I \\
\text{No Initial Bias Utility: } u(\text{buy}, \theta^U) &= v + \gamma - p \\
\text{Psychic-cost + No Initial Bias Utility: } u(\text{buy}, \theta^{UP}) &= v + \gamma - \sigma I - p
\end{aligned}$$

Seeing as the direct psychic effect of the nudge is captured by I, we rule out θ^D as the normative frame ($\psi(\theta^D) = 0$): $\sigma\tau$ cannot be welfare relevant.

We denote the two sources of planner uncertainty (and according beliefs) as follows:

[34] We restrict to the case of $I > 0$, $\gamma > 0$ and $\tau > 0$ for simplicity although in general I could also be a negative cost, some people might under-consume cigarettes and the graphic label could cause some consumers to be more likely to buy. Note that in the process of imposing that the nudge reduces consumption, our notation now differs from Allcott et al. [2022] in that τ is the refers to how much demand *reduces* with a $\sigma = 1$ nudge.

let $\psi_1 = \mathbb{P}[\gamma \text{ is normative}]$ and $\psi_2 = \mathbb{P}[I \text{ is normative}]$. Then a expected-utilitarian planner has preferences:

$$w(\text{buy}) = v + \psi_1\gamma - \psi_2\sigma I - p \text{ for each individual}$$
$$W = \int_{v \geq p+t+\sigma\tau-\gamma} v + \psi_1\gamma dF - [p + \psi_2\sigma I]\, Q^*$$

where $Q^* = \mathbb{P}[v \geq p + t + \sigma\tau - \gamma]$.

$\psi-$**optima.** One of the key results of Allcott et al. [2022] is that with heterogeneous bias γ, taxes should correct average bias and nudges should only be used to reduce variance. How do these conclusions change in the case of normative uncertainty / ambiguity? We start with the expected-utility formulation of a risk-neutral planner.

Following an analogous logic to Section 4.5.2, the optimal tax with heterogeneous bias γ and heterogeneous τ indeed corrects average bias (net of nudge effects and psychic-costs):

$$t^* = \mathbb{E}_m[(1 - \psi_1)\gamma - \sigma\tau + \sigma\psi_2 I] \tag{4.5.5}$$

where $\mathbb{E}_m$ integrates across the marginal consumers: $\{(v, \gamma, \tau)|v = p + \sigma\tau - \gamma\}$. Demand is defined as $D(p) = \mathbb{P}[v > p + \sigma\tau - \gamma]$, let $D'_p < 0$ be its derivative. When the tax is set optimally, Allcott et al. [2022] show:[35]

$$\begin{aligned}\frac{dW}{d\sigma} &= \frac{1}{2}\frac{d}{d\sigma}Var_m[(1-\psi_1)\gamma - \sigma\tau]D'_p - \psi_2 I\\ &= \{\sigma Var_m[\tau] - (1-\psi_1)Cov_m[\gamma,\tau]\}\underbrace{D'_p}_{<0} - \psi_2 I\end{aligned}$$

Let us consider the case of no psychic cost $\psi_2 = 0$. This yields the intuition that nudges can only increase welfare under normative ambiguity if $(1-\psi_1)Cov_m[\gamma,\tau] \geq \sigma Var_m[\tau]$ i.e. only if how sensitive demand decreases in response to the nudge (τ) is positively correlated with initial over-consumption (γ), mediated by $\mathbb{P}$[no initial over-consumption in the first place]. Unsurprisingly, if $\psi_1 = 1$ (no initial bias), $\frac{dw}{d\sigma}\big|_{\sigma=0} < 0$ which means the optimal tax and nudge are both 0.

In general, the $\psi-$optimal nudge σ^* satisfies $\frac{dw}{d\sigma} = 0$ at an interior optimum. This yields:

$$\sigma^* = \frac{(1-\psi_1)Cov_m[\gamma,\tau] + \frac{\psi_2 I}{D'_p}}{Var_m[\tau]}$$

$\sigma^* \geq 0$ only if $Cov_m[\gamma,\tau] \geq 0$ which means that σ^* is weakly decreasing in ψ_2 and ψ_1: the planner prefers a lower nudge if they were less sure people were initially over-consuming cigarettes and if they are more sure there is a psychic cost of a graphic label.

Ambiguity. For simplicity, we consider the case of global robustness where the planner wishes to solve $\max_{t,\sigma}\min_{\psi_1,\psi_2} W(\psi_1,\psi_2) = \int_{v\geq p+t+\sigma\tau-\gamma} v + \psi_1\gamma dF - [p + \psi_2\sigma I]\, Q^*$.

[35] Their result is without normative uncertainty, but the proof is almost identical.

We focus first on the inner minimization. Since we assumed $\gamma \geq 0$, the evil agent chooses to minimize $\psi_1 = 0$ to minimize welfare. Furthermore, the evil agent wants to maximize $\psi_2 = 1$ to maximize the psychic cost and hence minimize welfare.

What does this imply about robust policy? Plugging in $\psi_1 = 0, \psi_2 = 1$ to the formulae:

$$\begin{aligned} t^*_{\text{Robust}} &= \mathbb{E}_m[\gamma - \sigma^* \tau + \sigma^* I] \\ \sigma^*_{\text{Robust}} &= \frac{Cov_m[\gamma, \tau] + \frac{I}{D'_p}}{Var_m[\tau]} \end{aligned}$$

at an interior solution, and $\psi^* = 0, t^* = \mathbb{E}_m[\gamma]$ otherwise.

Ambiguity reduces the desirability of nudging compared to the case where the planner knows $\psi_2 < 1$. The planner faces a tradeoff in selecting the nudge between reducing the variance of the bias toward over-consumption γ and imposing psychic costs. In the worst-case scenario, the latter receives maximal weight. However, the overall effect of ambiguity on the optimal nudge is unclear, because of ambiguity over the extent of bias toward over-consumption $(1-\psi_1)$. In the worst-case scenario, the extent of such bias is maximized, so the value of reducing the variance of biases becomes larger than in the case of known $\psi_1 > 0$, which increases the optimal nudge intensity. The robustly optimal tax t^* is decreasing in ψ_1 and increasing in ψ_2 which means that robustness pushes the planner to be paternalistic and set the highest tax possible whilst subsuming the average effect of the nudge.

4.6 Discussion: Identifying Normative Weights.

Much of the recent literature on behavioral welfare economics, including work we drew on in our examples in Section 4.5, contains analysis that attempts to resolve the uncertainty about normative preferences that is primitive in our model. Using the model in Example 1.3, for instance, Lockwood et al. [2023] seek to identify the normative weight on the probability-weighting term ($\psi(\theta^D)$) with additional data analysis using a Counterfactual Normative Consumer identification strategy. A formal account of all of the different ways one might justify an identification strategy for normative weights is beyond the scope of this paper. Instead we discuss some intriguing similarities between these prior approaches to intrapersonal welfare analysis and strategies to identify interpersonal welfare weights.

The simplest approach to pinning down ψ is to assume the answer from some normative principle, e.g. by aggregating welfare using inverse-consumption weights in interpersonal problems. Such a treatment of intrapersonal welfare is found, for example, in the assumption by O'Donoghue and Rabin [2006] that the "long-run view" of welfare is normative, which is derived from the normative principle that inter-temporal preferences should be time-consistent – Bernheim [2009] provides a contrary perspective. A more sophisticated version of this approach is the "Counterfactual Normative Consumer" approach [Allcott and Taubinsky, 2015; Goldin and Reck, 2020; Allcott et al., 2019; Lockwood et al., 2023]. This approach leverages information on the revealed preferences of "debiased" individuals or experts, assuming 1) experts are not subject to framing effects, 2) we can observe the experts' revealed preferences, and 3) expertise is independent of preferences. We discuss this approach and how entertaining failures of these assumptions might lead the planner to employ our proposed approaches to robustness in Appendix D.6. The counterfactual normative consumer models typically conceive of normative uncertainty continuously; we develop an extension of our theory to a continuum of frames in Appendix D.5. One insight that emerges from this analysis is that while the $\psi(\theta)$ term in our expressions for welfare is a probability, under a linearity assumption, the $\psi(\theta)$ term in our expressions is isomorphic to the frame-dependent weights used in e.g. Bernheim et al. [2015] and Lockwood et al. [2023].

Second, researchers use revealed preference methods that approximate the "veil of ignorance" or "impartial observer" thought experiment. In the ideal version of this experiment, an individual, being aware of framing effects but not subject to them (being aware of interpersonal inequality but having not been assigned a type), makes choices that require them to trade off welfare under various frames (types). Under some assumptions, choices in such an ideal experiment are implied by choices in feasible experiments. For example, Capozza and Srinivasan [2023] experimentally estimate interpersonal welfare weights by having participants make choices to reveal their willingness-to-pay to transfer income from person A to person B, varying the incomes of A and B. In Allcott and Kessler [2019], the authors implement a meta-choice approach by eliciting the willingness to pay to be nudged and interpreting this elicitation as if individuals know how to evaluate the potentially biased choices they will make after getting the nudge or not. Allcott et al. [2022] use a willingness to pay experiment to identify the psychic cost of a graphic warning label in a similar way.

Lastly, many recent studies use implemented policies such as tax schedules [Hendren, 2020; Lockwood and Weinzierl, 2016] and transfer policies [Hendren and Sprung-Keyser, 2020] to reveal social welfare weights. The idea is to use the chosen policy to reverse-engineer the weights which would have meant that policy was optimal. This is an interesting approach which could be applied to behavioral problems. For example, if a social planner sets a "penalty-default" in our defaults example, they reveal that their intrapersonal welfare weight sets $\psi(\theta^D) \approx 0$. More ambitiously, we can imagine inferring policymakers' normative judgments about present focus from the design of illiquid/mandated savings policies, which is similar to the central exercise in Beshears et al. [2020].

The important point for us is that these methods all require untestable normative judgments. As such, we view welfare analysis in our framework – analyzing how uncertainty about normative judgments matters for optimal policy using our notions of robustness – as a useful complement to tools that help us identify the appropriate normative judgment. If there is even a little room for doubt about the validity of the approaches summarized here, our framework provides a way to assess the importance of such doubts for optimal policy.

4.7 Conclusion

The core argument of our paper is that a primary obstacle to the development of behavioral welfare economics – the question of how to do welfare analysis when we get conflicting information from revealed preferences – is the intrapersonal analogue of an older and more familiar problem: interpersonal comparisons of utility. We exploit the parallel between interpersonal and intrapersonal problems to develop criteria for the welfarist evaluation of policy in the presence of uncertainty about an individuals' normative preferences, and we explored what insights the resulting criteria might provide, in general and in the context of specific examples.

Showing that welfare in intrapersonal and intrapersonal problems can be modelled very similarly could be interpreted optimistically or pessimistically, depending on one's views about how economists typically approach interpersonal comparisons. From a pragmatic perspective, our results give applied researchers the tools to conduct welfare analysis when they wish to respect some revealed preferences but are unsure how to resolve ambiguity in revealed preferences. Specifically, we provide applied researchers the conceptual tools to separate empirical quantities that are informative for policy (e.g. what is the magnitude of potential internalities, how does behavior respond to a policy reform, etc) from normative judgments about how exactly these quantities map to an optimal policy. Within the framework, we can explore how disagreements about normative judgments and ambiguity map to disagreements or ambiguity in optimal policies. The approach has parallel limitations to interpersonal welfare, requiring restrictions on the richness of the set of frames/types and comparability of welfare across frames/types. We do not expect to spark universal consensus about how to overcome these limitations, but we adapt tools from interpersonal problems in order to analyze intrapersonal welfare cautiously and to clarify the necessary normative assumptions.

We identify some opportunities for future work. From a theoretical perspective, a few generalizations of our results are available, upon overcoming some technical challenges. One could formalize the derivation of a normative criteria without endowing the planner with primitive beliefs, e.g. using the approach like that of Savage [1954] or Maskin [1978]. More ambitiously, it might be possible to extend our framework to accommodate some models of limited attention, as discussed in Section 4.1.2, or models in which the process of choosing introduces potentially normative concerns [Bernheim et al., 2024], as discussed in our present focus example. One could also derive better ways to think about comparability and/or to discipline the set of frames using behavioral decision theory, like Ellis and Masatlioglu [2022]. From a more applied perspective, our work on how our notions of robustness play out in specific models barely scratches the surface of what is possible.

Bibliography

Abdellaoui, M. (2000): "Parameter-Free Elicitation of Utility and Probability Weighting Functions," *Management science*, 46, 1497–1512. [*Cited on page 102.*]

Abramson, B., J. Boerma, and A. Tsyvinski (2024): "Macroeconomics of Mental Health," Tech. rep., National Bureau of Economic Research. [*Cited on page 8.*]

Alatas, V., R. Purnamasari, M. Wai-Poi, A. Banerjee, B. A. Olken, and R. Hanna (2016): "Self-targeting: Evidence from a field experiment in Indonesia," *Journal of Political Economy*, 124, 371–427. [*Cited on page 6.*]

Allcott, H., D. Cohen, W. Morrison, and D. Taubinsky (2022): "When do" Nudges" Increase Welfare?" Tech. rep., National Bureau of Economic Research. [*Cited on pages 122, 123, 124, 125, and 128.*]

Allcott, H. and J. B. Kessler (2019): "The welfare effects of nudges: A case study of energy use social comparisons," *American Economic Journal: Applied Economics*, 11, 236–276. [*Cited on page 128.*]

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019): "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *Quarterly Journal of Economics*, 23, 1557–1626. [*Cited on pages 70, 121, 127, 231, and 232.*]

Allcott, H. and D. Taubinsky (2015): "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market," *American Economic*

Review, 105, 2501–38. [*Cited on pages 70, 73, 83, 127, and 223.*]

Alloy, L. B. and A. H. Ahrens (1987): "Depression and pessimism for the future: biased use of statistically relevant information in predictions for self versus others." *Journal of personality and social psychology*, 52, 366. [*Cited on pages 57, 63, and 204.*]

Alon, S. and G. Gayer (2016): "Utilitarian preferences with multiple priors," *Econometrica*, 84, 1181–1201. [*Cited on page 91.*]

Altmann, S., A. Falk, and A. Grunewald (2013): "Incentives and Information as Driving Forces of Default Effects," Working paper. [*Cited on page 99.*]

Anders, J. and C. Rafkin (2022): "The Welfare Effects of Eligibility Expansions: Theory and Evidence from SNAP." *SSRN.* [*Cited on page 205.*]

Angelucci, M. and D. Bennett (2024a): "Depression, Pharmacotherapy, and the Demand for a Preventive Health Product," *Available at SSRN 4808853.* [*Cited on page 7.*]

——— (2024b): "The economic impact of depression treatment in india: Evidence from community-based provision of pharmacotherapy," *American economic review*, 114, 169–198. [*Cited on page 7.*]

APA (2013): *Diagnostic and statistical manual of mental disorders: DSM-5*, vol. 5, American psychiatric association Washington, DC. [*Cited on page 1.*]

Arulsamy, K. and L. Delaney (2022): "The impact of automatic enrolment on the mental health gap in pension participation: Evidence from the UK," *Journal of Health Economics*, 102673. [*Cited on page 7.*]

Baily, M. N. (1978): "Some aspects of optimal unemployment insurance," *Journal of public Economics*, 10, 379–402. [*Cited on page 16.*]

Baranov, V., S. Bhalotra, P. Biroli, and J. Maselko (2020): "Maternal depression, women's empowerment, and parental investment: evidence from a randomized controlled trial," *American economic review*, 110, 824–59. [*Cited on page 8.*]

Barker, N., G. T. Bryan, D. Karlan, A. Ofori-Atta, and C. R. Udry (2021): "Mental Health Therapy as a Core Strategy for Increasing Human Capital: Evidence from Ghana," *American Economic Review: Insights, Forthcoming.* [*Cited on page 8.*]

Bell, E., J. Christensen, P. Herd, and D. Moynihan (2022): "Health in Citizen-State Interactions: How Physical and Mental Health Problems Shape Experiences of Administrative Burden and Reduce Take-Up," *Public Administration Review.* [*Cited on pages 1 and 7.*]

Benkert, J.-M. and N. Netzer (2016): "Informational Requirements of Nudging," Working paper. [*Cited on pages 73 and 81.*]

Berkhout, E., P. Koot, and N. Bosch (2019): "Gebruik (en niet-gebruik) van toeslagen in Nederland [Take-up (and non-take-up) of benefits in the Netherlands]," . [*Cited on page 23.*]

Bernheim, B. D. (2009): "Behavioral Welfare Economics," *Journal of the European Economic Association*, 7, 267–319. [*Cited on page 127.*]

——— (2016): "The good, the bad, and the ugly: A unified approach to behavioral welfare economics1," *Journal of Benefit-Cost Analysis*, 7, 12–68. [*Cited on page 80.*]

Bernheim, B. D., A. Fradkin, and I. Popov (2015): "The Welfare Economics of Default Options in 401(k) Plans," *American Economic Review*, 105, 2798–2837. [*Cited on pages 77, 99, 101, 109, 112, 115, and 128.*]

Bernheim, B. D. and J. M. Gastell (2021): "Optimal Default Options: The Case for Opt-Out Minimization," Nber working paper 28254. [*Cited on page 115.*]

Bernheim, B. D., K. Kim, and D. Taubinsky (2024): "Welfare and the Act of Choosing," Working paper, National Bureau of Economic Research. [*Cited on pages 107 and 130.*]

Bernheim, B. D. and A. Rangel (2009): "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *Quarterly Journal of Economics*, 124, 51–104. [*Cited on pages 70, 71, 73, 75, 76, 80, 81, 106, 114, and 218.*]

Bernheim, B. D. and D. Taubinsky (2018): "Behavioral Public Economics," in *Handbook of Behavioral Economics: Applications and Foundations*, Elsevier, vol. 1, 381–516. [*Cited on pages 73 and 80.*]

Beshears, J., J. J. Choi, C. Clayton, C. Harris, D. Laibson, and B. C. Madrian (2020): "Optimal illiquidity," Nber working paper no. 27459. [*Cited on page 129.*]

Bharadwaj, P., M. M. Pai, and A. Suziedelyte (2017): "Mental health stigma," *Economics Letters*, 159, 57–60. [*Cited on page 2.*]

Bhat, B., J. de Quidt, J. Haushofer, V. H. Patel, G. Rao, F. Schilbach, and P.-L. P. Vautrey (2022): "The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes," Tech. rep., National Bureau of Economic Research. [*Cited on page 7.*]

Bierman, E. J., H. C. Comijs, F. Rijmen, C. Jonker, and A. T. Beekman (2008): "Anxiety symptoms and cognitive performance in later life: results from the longitudinal aging study Amsterdam," *Aging and Mental Health*, 12, 517–523. [*Cited on pages 7 and 56.*]

Bloom, D. E., E. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, A. Hamandi, M. Mowafi, et al. (2012): "The global economic burden of noncommunicable diseases," Tech. rep., Program on the Global Demography of Aging. [*Cited on page 1.*]

Bordalo, P., N. Gennaioli, and A. Shleifer (2012): "Salience Theory of Choice Under Risk," *The Quarterly journal of economics*, 127, 1243–1285. [*Cited on page 102.*]

Brewer, M., T. Dang, and E. Tominey (2022): "Universal Credit: Welfare Reform and Mental Health," . [*Cited on page 66.*]

Bronchetti, E., J. Kessler, E. Magenheim, D. Taubinsky, and E. Zwick (2023): "Is attention produced optimally?" *Econometrica*, 91, 669–707. [*Cited on page 81.*]

Brot-Goldberg, Z., T. Layton, B. Vabson, and A. Y. Wang (2023): "The Behavioral Foundations of Default Effects: Theory and Evidence from Medicare Part D," *American Economic Review*, 113, 2718–2758. [*Cited on pages 99 and 226.*]

Brown, Z., N. Johnstone, I. Haščič, L. Vong, and F. Barascud (2013): "Testing the Effect of Defaults on the Thermostat Settings of OECD Employees," *Energy Economics*, 39, 128–134. [*Cited on page 99.*]

Caliendo, F. N. and T. S. Findley (2019): "Commitment and Welfare," *Journal of Economic Behavior & Organization*, 159, 210–234. [*Cited on pages 103 and 105.*]

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326. [*Cited on pages 46, 48, 55, 180, and 203.*]

Capozza, F. and K. Srinivasan (2023): "Who Should Get Money? Estimating Welfare Weights in the US," . [*Cited on page 128.*]

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on causal effects in a generalized regression kink design," *Econometrica*, 83, 2453–2483. [*Cited on pages 43, 46, 178, 179, 183, and 186.*]

Carroll, G. D., J. J. Choi, D. Laibson, B. C. Madrian, and A. Metrick (2009): "Optimal Defaults and Active Decisions," *Quarterly Journal of Economics*, 124, 1639–1674. [*Cited on pages 99 and 112.*]

Chateauneuf, A., J. Eichberger, and S. Grant (2007): "Choice under Uncertainty with the Best and Worst in Mind: Neo-Additive Capacities," *Journal of Economic Theory*, 137, 538–567. [*Cited on page 102.*]

Chesterley, N. (2017): "Defaults, Decision Costs and Welfare in Behavioural Policy Design," *Economica*, 84, 16–33. [*Cited on page 112.*]

Chetty, R. (2008): "Moral hazard versus liquidity and optimal unemployment insurance," *Journal of political Economy*, 116, 173–234. [*Cited on page 16.*]

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and Taxation: Theory and Evidence," *American Economic Review*, 99, 1145–1177. [*Cited on page 81.*]

Choi, J. J., D. Laibson, B. C. Madrian, and A. Metrick (2006): "Saving for Retirement on the Path of Least Resistance," in *Behavioral Public Finance: Toward a New Agenda*, ed. by E. J. McCaffery and J. Slemrod, Russell Sage Foundation. [*Cited on page 99.*]

Christian, C., L. Hensel, and C. Roth (2019): "Income shocks and suicides: Causal evidence from Indonesia," *Review of Economics and Statistics*, 101, 905–920. [*Cited on page 8.*]

Cook, J. B. and C. N. East (2024): "Work Requirements with No Teeth Still Bite: Disenrollment and Labor Supply Effects of SNAP General Work Requirements," Tech. rep., National Bureau of Economic Research. [*Cited on page 38.*]

Crawford, V. P. and J. Meng (2011): "New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 101, 1912–32. [*Cited* *on page 100.*]

Cronin, C. J., M. P. Forsstrom, and N. W. Papageorge (2024): "What good are treatment effects without treatment? mental health and the reluctance to use talk therapy," *Review of Economic Studies*, rdae061. [*Cited on pages 2, 8, and 66.*]

Danesh, K., J. T. Kolstad, J. Spinnewijn, and W. D. Parker (2024): "The Chronic Disease Index: Analyzing Health Inequalities Over the Lifecycle," Tech. rep., National Bureau of Economic Research. [*Cited on page 29.*]

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 112, 2851–2883. [*Cited on pages 70 and 83.*]

d'Aspremont, C. and L. Gevers (1977): "Equity and the informational basis of collective choice," *Review of Economic Studies*, 44, 199–209. [*Cited on page 91.*]

de Chaisemartin, C. and X. D'Haultfœuille (2023): "Two-Way Fixed Effects and Difference-in-Differences Estimators with Heterogeneous Treatment Effects and Imperfect Parallel Trends," *Available at SSRN.* [*Cited on page 35.*]

De Finetti, B. (1937): "La prévision: ses lois logiques, ses sources subjectives," in *Annales de l'institut Henri Poincaré*, vol. 7, 1–68. [*Cited on page 88.*]

Debreu, G. (1959): "Topological methods in cardinal utility theory," in *Mathematical Methods in Social Sciences*, ed. by K. J. Arrow, S. Karlin, and P. Suppes, Stanford University Press, 16–26. [*Cited on page 71.*]

DellaVigna, S., A. Lindner, B. Reizer, and J. F. Schmieder (2017): "Reference-Dependent Job Search: Evidence from Hungary," *Quarterly Journal of Economics*, 132, 1969–2018. [*Cited on page 100.*]

Deshpande, M. and Y. Li (2019): "Who is screened out? Application costs and the targeting of disability programs," *American Economic Journal: Economic*

Policy, 11, 213–48. [*Cited on page 6.*]

Deshpande, M. and L. M. Lockwood (2022): "Beyond health: Nonhealth risk and the value of disability insurance," *Econometrica*, 90, 1781–1810. [*Cited on pages 8 and 51.*]

Einav, L. and A. Finkelstein (2011): "Selection in insurance markets: Theory and empirics in pictures," *Journal of Economic perspectives*, 25, 115–38. [*Cited on page 16.*]

Ellis, A. and Y. Masatlioglu (2022): "Choice with Endogenous Categorization," *The Review of Economic Studies*, 89, 240–278. [*Cited on pages 102 and 130.*]

Ellsberg, D. (1961): "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, 75, 643–669. [*Cited on pages 72 and 89.*]

Evans, V. C., G. L. Iverson, L. N. Yatham, and R. W. Lam (2014): "The relationship between neurocognitive and psychosocial functioning in major depressive disorder: a systematic review," *The Journal of clinical psychiatry*, 75, 17306. [*Cited on pages 7, 51, 57, 63, and 204.*]

Fehr, D., G. Fink, and B. Jack (2022): "Poor and Rational: Decision-Making under Scarcity," *Journal of Political Economy*, forthcoming. [*Cited on page 7.*]

Finkelstein, A. and M. J. Notowidigdo (2019): "Take-up and targeting: Experimental evidence from SNAP," *The Quarterly Journal of Economics*, 134, 1505–1556. [*Cited on pages 3, 6, 10, 19, 36, 63, 151, 204, and 205.*]

Fleurbaey, M. and F. Maniquet (2011): *A Theory of Fairness and Social welfare*, vol. 48, Cambridge University Press. [*Cited on page 72.*]

Fudenberg, D. and D. K. Levine (2006): "A Dual-Self Model of Impulse Control," *American Economic Review*, 1449–76. [*Cited on page 118.*]

Ganong, P. and S. Jäger (2018): "A permutation test for the regression kink design," *Journal of the American Statistical Association*, 113, 494–504. [*Cited on* *pages 49 and 199.*]

Gelber, A., D. Jones, D. W. Sacks, and J. Song (2020): "Using non-linear budget sets to estimate extensive margin responses: Evidence and method from the earnings test," *American Economic Journal: Applied Economics.* [*Cited on page 178.*]

Giannella, E., T. Homonoff, G. Rino, and J. Somerville (2023): "Administrative burden and procedural denials: Experimental evidence from SNAP," Tech. rep., National Bureau of Economic Research Cambridge, MA. [*Cited on page 6.*]

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of mathematical economics*, 18, 141–153. [*Cited on pages 72, 89, 216, 217, and 218.*]

Godard, M., P. Koning, and M. Lindeboom (2022): "Application and award responses to stricter screening in disability insurance," *Journal of Human Resources.* [*Cited on page 2.*]

Goldin, J. and D. Reck (2020): "Revealed Preference Analysis with Framing Effects," *Journal of Political Economy*, 126, 2759–95. [*Cited on pages 127, 231, and 232.*]

——— (2022a): "Optimal defaults with normative ambiguity," *Review of Economics and Statistics*, 104, 17–33. [*Cited on page 57.*]

——— (2022b): "Optimal Defaults with Normative Ambiguity," *Review of Economics and Statistics*, 104, 17–33. [*Cited on pages 73, 81, 99, 100, 112, 115, 226, and 229.*]

Gross, J. J. and R. F. Muñoz (1995): "Emotion regulation and mental health." *Clinical psychology: Science and practice*, 2, 151. [*Cited on pages 7, 51, and 56.*]

Gruber, J. and B. Köszegi (2001): "Is addiction "rational"? Theory and evidence," *The Quarterly Journal of Economics*, 116, 1261–1303. [*Cited on page 121.*]

Haggag, K. and G. Paci (2014): "Default Tips," *American Economic Journal: Applied Economics*, 6, 1–19. [*Cited on page 99.*]

Haller, A. and S. Staubli (2024): "Measuring the Value of Disability Insurance from Take-Up Decisions," . [*Cited on pages 2, 8, 18, and 204.*]

Hammar, Å. and G. Årdal (2009): "Cognitive functioning in major depression-a summary," *Frontiers in human neuroscience*, 3, 26. [*Cited on pages iii, 7, and 56.*]

Hammond, P. J. (1976): "Equity, Arrow's Conditions, and Rawls' Difference Principle," *Econometrica*, 793–804. [*Cited on pages 91 and 218.*]

Hansen, L. P. and T. J. Sargent (2001): "Robust control and model uncertainty," *American Economic Review*, 91, 60–66. [*Cited on pages 90, 110, and 218.*]

——— (2008): *Robustness*, Princeton university press. [*Cited on pages 218 and 224.*]

Harsanyi, J. C. (1955): "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, 63, 309–321. [*Cited on pages 72, 91, and 97.*]

Haushofer, J. (2019): "Is there a Psychological Poverty Trap?" Tech. rep., Working Paper. [*Cited on page 66.*]

Haushofer, J., R. Mudida, and J. Shapiro (2020): "The comparative impact of cash transfers and psychotherapy on psychological and economic well-being," *NBER working paper*, 7. [*Cited on page 50.*]

Heekelaar, M. (2021): "Beschikbare en benodigde financiële middelen voor de Participatiewet: Analyse," . [*Cited on page 35.*]

Hendren, N. (2016): "The policy elasticity," *Tax Policy and the Economy*, 30, 51–89. [*Cited on page 61.*]

——— (2020): "Measuring economic efficiency using inverse-optimum weights," *Journal of public Economics*, 187, 104198. [*Cited on page 128.*]

Hendren, N. and B. Sprung-Keyser (2020): "A unified welfare analysis of government policies," *The Quarterly Journal of Economics*, 135, 1209–1318. [*Cited on pages 6, 15, 59, 61, 91, 94, 128, and 153.*]

Homonoff, T. and J. Somerville (2021): "Program Recertification Costs: Evidence from SNAP," *American Economic Journal: Economic Policy*, 13, 271–98. [*Cited on page 6.*]

Homonoff, T. A. (2018): "Can Small Incentives Have Large Effects? The Impact of Taxes Versus Bonuses on Disposable Bag Use," *American Economic Journal: Economic Policy*, 10, 177–210. [*Cited on page 116.*]

Hyman, S., D. Chisholm, R. Kessler, V. Patel, and H. Whiteford (2006): "Mental disorders," *Disease control priorities related to mental, neurological, developmental and substance abuse disorders*, 1–20. [*Cited on page 56.*]

Inspectie SZW (2021): "Niet-gebruik van de algemene bijstand: Een onderzoek naar de omvang, kenmerken, langdurigheid en aanpak," *Den Haag: Inspectie SZW*. [*Cited on page 25.*]

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47, 263–92. [*Cited on pages 100 and 102.*]

Kaplow, L. and S. Shavell (2001): "Any non-welfarist method of policy assessment violates the Pareto principle," *Journal of Political Economy*, 109, 281–286. [*Cited on pages 71, 77, 85, and 211.*]

Kaur, S., S. Mullainathan, S. Oh, and F. Schilbach (2021): "Do Financial Concerns Make Workers Less Productive?" Tech. rep., National Bureau of Economic Research. [*Cited on page 7.*]

Kessler, R. C., P. Berglund, O. Demler, R. Jin, D. Koretz, K. R. Merikangas, A. J. Rush, E. E. Walters, and P. S. Wang (2003): "The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R)," *jama*, 289, 3095–3105. [*Cited on pages 7, 51, and 57.*]

Kleven, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464. [*Cited on pages 47 and 156.*]

Kline, P. and C. R. Walters (2019): "On Heckits, LATE, and numerical equivalence," *Econometrica*, 87, 677–696. [*Cited on page 19.*]

Knight, F. H. (1921): *Risk, uncertainty and profit*, Houghton Mifflin. [*Cited on pages 72 and 88.*]

Ko, W. and R. A. Moffitt (2024): "Take-up of social benefits," *Handbook of Labor, Human Resources and Population Economics*, 1–42. [*Cited on pages 1, 3, 4, and 13.*]

Köbberling, V. and P. P. Wakker (2003): "Preference foundations for nonexpected utility: A generalized and simplified technique," *Mathematics of Operations Research*, 28, 395–423. [*Cited on pages 86 and 212.*]

Kőszegi, B. and M. Rabin (2006): "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics*, 121, 1133–65. [*Cited on pages 100, 101, and 112.*]

——— (2007): "Reference-Dependent Risk Attitudes," *American Economic Review*, 97, 1047–73. [*Cited on page 117.*]

Kromhout, M., N. Kornalijnslijper, and M. de Klerk (2018): "Summary changing care and support for people with disabilities," . [*Cited on page 40.*]

Krueger, A. B. and B. D. Meyer (2002): "Labor supply effects of social insurance," *Handbook of public economics*, 4, 2327–2392. [*Cited on page 48.*]

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 443–477. [*Cited on pages 103 and 229.*]

Laibson, D. I., A. Repetto, J. Tobacman, R. E. Hall, W. G. Gale, and G. A. Akerlof (1998): "Self-Control and Saving for Retirement," *Brookings Papers on Economic Activity*, 1998, 91–196. [*Cited on page 103.*]

Landais, C., A. Nekoei, P. Nilsson, D. Seim, and J. Spinnewijn (2021): "Risk-based selection in unemployment insurance: Evidence and implications," *American Economic Review*, 111, 1315–1355. [*Cited on pages 204 and 205.*]

Lichtenstein, S. and P. Slovic (2006): *The Construction of Preference*, Cambridge University Press. [*Cited on page 80.*]

Lockwood, B. B., H. Allcott, D. Taubinsky, and A. Sial (2023): "What Drives Demand for State-Run Lotteries? Evidence and Welfare Implications," Nber working paper no. 28975. [*Cited on pages 73, 102, 103, 127, and 128.*]

Lockwood, B. B., A. Sial, and M. Weinzierl (2021): "Designing, Not Checking, for Policy Robustness: An Example with Optimal Taxation," *Tax Policy and the Economy*, 35, 1–54. [*Cited on page 219.*]

Lockwood, B. B. and M. Weinzierl (2016): "Positive and normative judgments implicit in US tax policy, and the costs of unequal growth and recessions," *Journal of Monetary Economics*, 77, 30–47. [*Cited on page 128.*]

Lopes, F. V., B. Ravesteijn, T. Van Ourti, and C. Riumallo-Herl (2023): "Income inequalities beyond access to mental health care: a Dutch nationwide record-linkage cohort study of baseline disease severity, treatment intensity, and mental health outcomes," *The Lancet Psychiatry*, 10, 588–597. [*Cited on page 24.*]

Lund, C., A. Breen, A. J. Flisher, R. Kakuma, J. Corrigall, J. A. Joska, L. Swartz, and V. Patel (2010): "Poverty and common mental disorders in low and middle income countries: A systematic review," *Social science & medicine*, 71, 517–528. [*Cited on page iii.*]

Mani, A., S. Mullainathan, E. Shafir, and J. Zhao (2013): "Poverty impedes cognitive function," *Science*, 341, 976–980. [*Cited on page 7.*]

Martin, L., L. Delaney, and O. Doyle (2023a): "Everyday administrative burdens and inequality," *Public Administration Review.* [*Cited on page 7.*]

Martin, L., L. Delaney, O. Doyle, et al. (2023b): "The Distributive Effects of Administrative Burdens on Decision-Making," *Journal of Behavioral Public Administration*, 6. [*Cited on pages 7, 63, and 204.*]

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012): "Revealed Attention," *American Economic Review*, 102, 2183–2205. [*Cited on page 80.*]

Maskin, E. (1978): "A Theorem on Utilitarianism," *The Review of Economic Studies*, 45, 93–96. [*Cited on pages 91, 104, and 130.*]

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of econometrics*, 142, 698–714. [*Cited on pages 47 and 196.*]

McFadden, D. (1981): "Econometric Models of Probabilistic Choice"," . [*Cited on page 19.*]

McGarry, K. (1996): "Factors determining participation of the elderly in supplemental security income," *Journal of Human Resources*, 331–358. [*Cited on page 48.*]

Milgrom, P. and I. Segal (2002): "Envelope theorems for arbitrary choice sets," *Econometrica*, 70, 583–601. [*Cited on page 223.*]

Miller, R. and N. Bairoliya (2023): "Health, longevity, and welfare inequality of older americans," *Review of Economics and Statistics*, 105, 1145–1160. [*Cited on page 51.*]

Miller, S., E. Rhodes, A. W. Bartik, D. E. Broockman, P. K. Krause, and E. Vivalt (2024): "Does Income Affect Health? Evidence from a Randomized Controlled Trial of a Guaranteed Income," Tech. rep., National Bureau of

Economic Research. [*Cited on page 8.*]

Ministerie van SZW (2015): "Participatiewet," https://wetten.overheid.nl/BWBR0015703/2015-01-01/. [*Cited on pages 5, 23, 25, and 177.*]

——— (2022): "Participatiewet in Balans: uitkomsten beleidsanalyse," https://www.rijksoverheid.nl/documenten/kamerstukken/2022/06/21/bijlage-rapport-participatiewet-in-balans. [*Cited on pages 5, 38, 39, 57, 166, and 167.*]

Mogstad, M., A. Torgovitsky, and C. R. Walters (2024): "Policy evaluation with multiple instrumental variables," *Journal of Econometrics*, 105718. [*Cited on page 19.*]

Mongin, P. and M. Pivato (2021): "Rawls's difference principle and maximin rule of allocation: a new analysis," *Economic Theory*, 71, 1499–1525. [*Cited on pages 91 and 218.*]

Montiel Olea, J. L. and T. Strzalecki (2014): "Axiomatization and Measurement of Quasi-Hyperbolic Discounting," *Quarterly Journal of Economics*, 129, 1449–1499. [*Cited on page 107.*]

Muilwijk-Vriend, S., C. Tempelman, L. Kroon, M. Lammers, R. Ponds, C. van Woerkens, and P. Koning (2019): *Gezondheidsproblemen in WW en bijstand*, SEO Economisch Onderzoek. [*Cited on page 31.*]

Mullainathan, S., J. Schwartzstein, and W. J. Congdon (2012): "A reduced-form approach to behavioral public finance," *Annu. Rev. Econ.*, 4, 511–540. [*Cited on pages 7, 70, 73, 83, 111, 120, 208, and 223.*]

Mullainathan, S. and E. Shafir (2013): *Scarcity: Why having too little means so much*, Macmillan. [*Cited on page 51.*]

Nichols, A. L. and R. J. Zeckhauser (1982): "Targeting transfers through restrictions on recipients," *The American Economic Review*, 72, 372–377. [*Cited* *on pages 1 and 6.*]

O'Donoghue, T. and M. Rabin (2006): "Optimal Sin Taxes," *Journal of Public Economics*, 90, 1825–1849. [*Cited on pages 70, 83, 121, and 127.*]

O'Donoghue, T. and C. Sprenger (2018): "Reference-Dependent Preferences," in *Handbook of Behavioral Economics: Applications and Foundations*, Elsevier, vol. 1, 1–77. [*Cited on page 100.*]

Pei, Z., D. S. Lee, D. Card, and A. Weber (2022): "Local polynomial order in regression discontinuity designs," *Journal of Business & Economic Statistics*, 40, 1259–1267. [*Cited on page 186.*]

Prelec, D. (1998): "The Probability Weighting Function," *Econometrica*, 66, 497–527. [*Cited on page 102.*]

Rafkin, C., A. Solomon, and E. Soltas (2023): "Self-Targeting in US Transfer Programs," *Available at SSRN 4495537.* [*Cited on pages 3, 6, 15, 19, 63, 204, and 205.*]

Rambachan, A. and J. Roth (2023): "A more credible approach to parallel trends," *Review of Economic Studies*, 90, 2555–2591. [*Cited on page 36.*]

Rawls, J. (1971): *A Theory of Justice*, Belknap Press. [*Cited on page 218.*]

Reck, D. and A. Seibold (2023): "The Welfare Economics of Reference Dependence," Nber working paper no. 31381. [*Cited on pages 73, 100, 101, 112, 116, 118, and 229.*]

Rees-Jones, A. (2018): "Quantifying Loss-Averse Tax Manipulation," *Review of Economic Studies*, 85, 1251–78. [*Cited on pages 100 and 116.*]

Rees-Jones, A. and D. Taubinsky (2018): "Taxing Humans: Pitfalls of the Mechanism Design Approach and Potential Resolutions," *Tax Policy and the Economy*, 32, 107–133. [*Cited on pages 70 and 83.*]

Ridley, M., G. Rao, F. Schilbach, and V. Patel (2020): "Poverty, depression, and anxiety: Causal evidence and mechanisms," *Science*, 370, eaay0214. [*Cited on pages 1 and 66.*]

Roos, A.-F., M. Diepstraten, R. Douven, et al. (2021): "When Financials Get Tough, Life Gets Rough?: Problematic Debts and Ill Health," Tech. rep., CPB Netherlands Bureau for Economic Policy Analysis Hague, the Netherlands. [*Cited on page 29.*]

Roth, C., P. Schwardmann, and E. Tripodi (2024): "Misperceived effectiveness and the demand for psychotherapy," . [*Cited on page 8.*]

Russo, A. (2023): "Waiting or Paying for Healthcare: Evidence from the Veterans Health Administration," Tech. rep., Working Paper. [*Cited on page 18.*]

Saez, E. and S. Stantcheva (2016): "Generalized Social Marginal Welfare Weights for Optimal Tax Theory," *American Economic Review*, 106, 24–45. [*Cited on page 91.*]

Savage, L. J. (1954): *The Foundations of Statistics*, Wiley. [*Cited on page 130.*]

Schmidt, L., L. Shore-Sheppard, and T. Watson (2021): "The Effect of Safety Net Generosity on Maternal Mental Health and Risky Health Behaviors," Tech. rep., National Bureau of Economic Research. [*Cited on page 8.*]

SCP (2019): "Eindevaluatie van de Participatiewet," . [*Cited on pages 5, 35, 39, and 165.*]

Seibold, A. (2021): "Reference Points for Retirement Behavior: Evidence from German Pension Discontinuities," *American Economic Review*, 111, 1126–65. [*Cited on pages 100 and 116.*]

Sen, A. (1976): "Welfare Inequalities and Rawlsian Axiomatics," *Theory and Decision*, 7, 243–262. [*Cited on pages 72 and 91.*]

——— (1986): "Social choice theory," *Handbook of Mathematical Economics*, 3, 1073–1181. [*Cited on pages 71 and 91.*]

——— (1997): *On economic inequality*, Oxford university press. [*Cited on page 91.*]

——— (1999): *Development as Freedom*, New York: Oxford University Press. [*Cited on pages 8 and 51.*]

——— (2008): "The idea of justice," *Journal of human development*, 9, 331–342. [*Cited on pages 8 and 51.*]

Sen, A. K. (1970): *Collective Choice and Social Welfare*, Holden-Day. [*Cited on pages 91 and 218.*]

Serena, B. L. (2024): "The Causal Effect of Scaling up Access to Psychotherapy," . [*Cited on page 8.*]

Shah, A. K., S. Mullainathan, and E. Shafir (2012): "Some consequences of having too little," *Science*, 338, 682–685. [*Cited on page 7.*]

Shahn, Z. (2023): "Subgroup Difference in Differences to Identify Effect Modification Without a Control Group," *arXiv preprint arXiv:2306.11030.* [*Cited on page 35.*]

Shepard, M. and M. Wagner (2022): "Do ordeals work for selection markets? Evidence from health insurance auto-enrollment," Tech. rep., National Bureau of Economic Research. [*Cited on pages 6 and 10.*]

Sher, I. (2023): "Generalized Social Marginal Welfare Weights Imply Inconsistent Comparisons of Tax Policies," Working paper, arxiv:2102.07702. [*Cited on page 91.*]

Shreekumar, A. and P.-L. Vautrey (2021): "Managing Emotions: The Effects of Online Mindfulness Meditation on Mental Health and Economic Behavior," . [*Cited on page 7.*]

Silver, D. and J. Zhang (2022): "Impacts of Basic Income on Health and Economic Well-Being: Evidence from the VA's Disability Compensation Program," Tech. rep., National Bureau of Economic Research. [*Cited on pages 8 and 51.*]

Solmi, M., J. Radua, M. Olivola, E. Croce, L. Soardo, G. Salazar de Pablo, J. Il Shin, J. B. Kirkbride, P. Jones, J. H. Kim, et al. (2022): "Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies," *Molecular psychiatry*, 27, 281–295. [*Cited on page 51.*]

Sugden, R. (2004): "The opportunity criterion: consumer sovereignty without the assumption of coherent preferences," *American economic review*, 94, 1014–1033. [*Cited on page 82.*]

Thakral, N. and L. T. Tô (2021): "Daily Labor Supply and Adaptive Reference Points," *American Economic Review*, 111, 2417–43. [*Cited on page 100.*]

Thaler, R. H. and C. R. Sunstein (2009): *Nudge: Improving decisions about health, wealth, and happiness*, Penguin. [*Cited on page 122.*]

Tversky, A. and D. Kahneman (1991): "Loss Aversion in Riskless Choice: A Reference-Dependent Model," *Quarterly Journal of Economics*, 106, 1039–61. [*Cited on pages 100 and 101.*]

van der Veen, R. (2019): "Basic income experiments in the Netherlands?" *Basic Income Studies*, 14. [*Cited on pages 35 and 166.*]

Verlaat, T. and A. Zulkarnain (2022): "Evaluatie experimenten Participatiewet: effecten op brede baten," . [*Cited on page 23.*]

Von Neumann, J. and O. Morgenstern (1953): "Theory of Games and Economic Behavior," in *Theory of Bames and Economic Behavior*, Princeton University Press. [*Cited on pages 87 and 214.*]

Wakker, P. (1984): "Cardinal Coordinate Independence for Expected Utility," *Journal of Mathematical Psychology*, 28, 110–117. [*Cited on pages 71 and 72.*]

Wakker, P. P. and H. Zank (1999): "A Unified Derivation of Classical Subjective Expected Utility Models through Cardinal Utility," *Journal of Mathematical Economics*, 32, 1–19. [*Cited on pages 72 and 88.*]

Weymark, J. A. (1991): "A reconsideration of the Harsany-Sen debate on Utilitarianism," Cambridge University Press, 255–320. [*Cited on page 72.*]

WHO (2022): "World mental health report: transforming mental health for all," . [*Cited on pages iii and 1.*]

Woodford, M. (2020): "Modeling imprecision in perception, valuation, and choice," *Annual Review of Economics*, 12, 579–601. [*Cited on page 19.*]

Wooldridge, J. M. (2019): "Correlated random effects models with unbalanced panels," *Journal of Econometrics*, 211, 137–150. [*Cited on page 167.*]

Wu, D. and B. D. Meyer (2023): "Certification and Recertification in Welfare Programs: What Happens When Automation Goes Wrong?" Tech. rep., National Bureau of Economic Research. [*Cited on page 6.*]

Zeckhauser, R. (2021): "Strategic sorting: the role of ordeals in health care," *Economics & Philosophy*, 37, 64–81. [*Cited on page 3.*]

Appendix A

Appendix to Chapter 1

Let θ have a type-specific indirect utility functions: $u_\theta(c, y)$ is increasing in consumption c and decreasing in earned income y. Income depends on take-up but is fixed otherwise:[1] let $y_\theta^{SA=1}$ refer to income earned if on social assistance and $y_\theta^{SA=0}$ if not. All income (including benefits) is taxed at marginal tax rate τ. Thus, $v_\theta(B)$ is given by:

$$v_\theta(B) \triangleq u_\theta\left((1-\tau)\cdot\left[y_\theta^{SA=1}+B\right], y_\theta^{SA=1}\right) - u_\theta\left((1-\tau)\cdot y_\theta^{SA=0}, y_\theta^{SA=0}\right) \quad \text{(A.0.1)}$$

Thus, value is the net-utility gain from social assistance and comes from two main sources. First, if $y_\theta^{SA=0} \leq y_\theta^{SA=1}+B$, θ derives utility from the top-up in consumption $(1-\tau)y^{SA=0} \rightarrow (1-\tau)\cdot\left[y_\theta^{SA=1}+B\right]$. Second, if $y^{SA=1} < y^{SA=0}$, θ also derives value from a lowered cost of working when supported by social assistance. Importantly,

[1] The assumption of no labour supply responses follows Finkelstein and Notowidigdo [2019] and simplifies the theoretical analysis. In the Netherlands, social assistance tops income up to a social minimum. Therefore, conditional on receipt, income ≈ 0 for many people. This means that the decision in practice can be reasonably approximated to take-up SA (and earn low/no income) vs do not take-up SA (and earn income).

heterogeneous value across types does not only come from different y_θ, the utility functions u_θ also differ.

Note that eligibility then is defined as $y \leq \bar{y}$ where $y = SA \cdot y^{SA=1} + (1 - SA) \cdot y^{SA=0}$.

Proof of Proposition 1.2.1. Social welfare is defined as follows.

$$W = \int \lambda_\theta \mathcal{U}_\theta d\mu$$

Using the chain rule: $\frac{dW}{d\Lambda} = \frac{\partial W}{\partial \Lambda} + \frac{\partial W}{\partial B} \cdot \frac{\partial B}{\partial \Lambda}$, and using the Leibniz rule to differentiate under the integral gives Equation (1.2.2). Here, the Envelope Theorem implies the behavioural welfare effect is 0. For example,

$$\begin{aligned}\frac{d\,\mathcal{U}_\theta}{d\Lambda} &= \frac{d}{d\Lambda} \int_{-\infty}^{\varepsilon_\theta^*} [v_\theta(B) - \kappa_\theta(\Lambda) - \varepsilon]\ \ dF(\varepsilon) \\ &= \frac{d\varepsilon_\theta^*}{d\Lambda} \cdot \underbrace{[v_\theta(B) - \kappa_\theta(\Lambda) - \varepsilon_\theta^*]}_{=0 \text{ by defn of } \varepsilon_\theta^*} + \int_{-\infty}^{\varepsilon_\theta^*} [-\kappa_\theta'(\Lambda)]\, dF(\varepsilon)\end{aligned}$$

The above step is the Envelope Theorem at work.

$$= -\kappa_\theta'(\Lambda) \cdot F(\varepsilon_\theta^*)$$

Similarly, $\frac{d\,\mathcal{U}_\theta}{dB} = v_\theta'(B) \cdot F(\varepsilon_\theta^*)$. Therefore:

$$\frac{dW}{d\Lambda} = \int \lambda_\theta \mathbb{P}[SA]_\theta \left[v_\theta'(B) \cdot \frac{dB}{d\Lambda} - \kappa_\theta'(\Lambda) \right] d\mu$$

Let G be the government's budget. Budget neutrality implies $\frac{dG}{d\Lambda} = 0$. Using the chain and Leibniz rule again, and dropping θ subscripts:

$$\begin{aligned}\frac{dG}{d\Lambda} = \int \big[&\tau(y^{SA=0} - y^{SA=1}) \\ &+ (1-\tau)\cdot B\big]\cdot\frac{\partial\mathbb{P}[SA]}{\partial\Lambda} + \big[\tau(y^{SA=0} - y^{SA=1}) + (1-\tau)\cdot B\big]\cdot\frac{\partial\mathbb{P}[SA]}{\partial B}\cdot\frac{dB}{d\Lambda} \\ &+ (1-\tau)\cdot\mathbb{P}[SA]\cdot\frac{dB}{d\Lambda}d\mu = 0\end{aligned}$$

Rearranging gives Equation (1.2.3). □

A.1 MVPF Formulae

The MVPF measures the ratio of the direct welfare effect to beneficiaries of a policy, divided by the cost to the government. Direct welfare effects are written in the units of each types' willingness-to-pay. Hendren and Sprung-Keyser [2020] show that the composite policy increasing Λ (B adjusts) is social-welfare improving, if the gains from increasing spending on dB exceed the losses from reducing spending through an increase $d\Lambda$.

Let η_θ denote each individual's social marginal utility of income. Therefore, $\eta_\theta = \lambda_\theta \cdot v'_\theta$: social marginal utility is equal to social marginal welfare weight × individual marginal utility of income. Let $WTP^P_\theta = \frac{d\mathcal{U}_\theta}{dP}\cdot\frac{1}{v'_\theta}$ be θ's willingness-to-pay for a policy P: the direct welfare effect divided by the marginal utility of income.

Proposition A.1.1. *[Hendren and Sprung-Keyser, 2020] Let $\bar{\eta}_P$ be the average social marginal utility of the beneficiaries a policy P:*

$$\bar{\eta}_P = \int \eta_\theta \frac{WTP^P_\theta}{\int WTP^P_\theta d\mu} d\mu \tag{A.1.1}$$

The composite policy experiment of a budget-neutral increase in Λ *financing an increase in* B *is good for welfare* W *iff:*

$$\bar{\eta}_{dB} \cdot MVPF_{dB} > \bar{\eta}_{d\Lambda} \cdot MVPF_{d\Lambda} \tag{A.1.2}$$

where:

$$\bar{\eta}_{dB} = \int \eta_\theta d\mu \tag{A.1.3}$$

$$\bar{\eta}_{d\Lambda} = \int \eta_\theta \frac{\kappa'_\theta / v'_\theta}{\int \kappa'_\theta / v'_\theta d\mu} d\mu \tag{A.1.4}$$

and the MVPF of an increase in ordeals is given by Equation (A.1.5).

$$MVPF_{d\Lambda} = \frac{\overbrace{-\int \lambda_\theta \cdot \mathbb{P}[SA]_\theta \cdot \frac{\kappa'_\theta(\Lambda)}{v'_\theta(B)} \ d\mu}^{\textit{Direct Effect} <0}}{\underbrace{\int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} d\mu}_{\textit{Behavioral Revenue Effect} <0}} \tag{A.1.5}$$

and the MVPF of an increase in benefit level is given by Equation (A.1.6).

$$MVPF_{dB} = \frac{\overbrace{\int \lambda_\theta \cdot \mathbb{P}[SA]_\theta \ d\mu}^{\textit{Direct Effect} >0}}{\underbrace{(1-\tau) \cdot \int \mathbb{P}[SA]_\theta \ d\mu}_{\textit{Mechanical Revenue Effect} >0} + \underbrace{\int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial B} d\mu}_{\textit{Behavioral Revenue Effect} >0}} \tag{A.1.6}$$

The direct effect of an increase in ordeals $d\Lambda$ is that it imposes dis-utility on infra-marginal individuals κ'_θ. Written in terms of € cost, this is $\frac{\kappa'_\theta}{v'_\theta}$. Increasing barriers saves the government money through lower take-up, corresponding to the denominator. The direct effect of an increase in benefit level dB is that it transfers €1 of benefits to all infra-marginal individuals. The government has to pay for the mechanical extra program cost, as well as the new-entrants. See Appendix C.1 for how to calculate these formulas when sufficient statistics are estimated on the eligible population.

Proof of Proposition A.1.1. From the proof of Proposition 1.2.1,

$$\frac{\partial W}{\partial \Lambda} = -\int \lambda_\theta \mathbb{P}[SA]_\theta \kappa'_\theta d\mu \tag{A.1.7}$$

$$\frac{\partial W}{\partial B} = \int \lambda_\theta \mathbb{P}[SA]_\theta v'_\theta d\mu \tag{A.1.8}$$

$$\frac{\partial G}{\partial \Lambda} = \int FE_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} d\mu \tag{A.1.9}$$

$$\frac{\partial G}{\partial B} = (1-\tau)\int \mathbb{P}[SA]_\theta \, d\mu + \int FE \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial B} d\mu \tag{A.1.10}$$

The first two equations follow by the Envelope theorem, as in the proof of Proposition 1.2.1. Dividing yields the MVPF formulas.

□

A.2 Identification

In this section, I set out how to identify the relationship between $f_\varepsilon(v_\theta - \kappa_\theta)$ across types using take-up levels and a first-order Taylor approximation. The key case is when $\mathbb{P}[SA]_\theta \neq \mathbb{P}[SA]_{\tilde{\theta}}$. For argument's sake - suppose that we are considering two types $\theta = L, H$.

This proposition requires some additional structure:

Let indirect utility $u_\theta(c,y) = v_\theta \cdot c - \frac{n_\theta}{1+1/e} \cdot \left(\frac{y}{n_\theta}\right)^{1+1/e}$: quasi-linear utility with scaling factor v–denoting the marginal value of income–and isoelastic disutility of labour, as in e.g. Kleven [2016]. Individuals then differ based on their value of money, and their ability n_θ. For simplicity, Frisch elasticities are the same across types. In this case, $y^{SA=0} = \arg\max u\left((1-\tau)y, y\right) = n \cdot v \cdot (1-\tau)^e$. Suppose also that $\kappa(\Lambda) = \kappa_1 \cdot \Lambda + \kappa_0$. Therefore,

$$SA = 1 \iff u\left((1-\tau) \cdot \left(B + y^{SA=1}\right), y^{SA=1}\right) - \kappa_1 \cdot \Lambda + \kappa_0 - \varepsilon \geq u\left((1-\tau)y^{SA=0}, y^{SA=0}\right) \tag{A.2.1}$$

Then:

Proposition A.2.1. *Identification of $f_L \triangleq f_\varepsilon(v_L - \kappa_L)$ in terms of $f_H \triangleq f_\varepsilon(v_H - \kappa_H)$ is given by:*

$$\mathbb{P}[SA]_L - \mathbb{P}[SA]_H \approx \left(\Psi \frac{\partial \mathbb{P}[SA]_L}{\partial B} + \Lambda \frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda}\right) \cdot \left(\frac{f_L}{f_H} - 1\right) \tag{A.2.2}$$

where $\Psi = B + y^{SA=1} - \frac{y^{SA=0}}{1+e} - \frac{\left(y^{SA=1}\right)^{1+1/e}}{\left(y^{SA=0}\right)^{1/e}(1+e)}$.

Note that if the LHS = 0, the RHS will imply that $f_L = f_H$ as long as $\Psi \frac{\partial \mathbb{P}[SA]_L}{\partial B} \neq \Lambda \frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda}$.

Proof.

$$v(B) = u\left((1-\tau) \cdot (B + y^{SA=1}, y^{SA=1}\right) - u\left((1-\tau)y^{SA=0}, y^{SA=0}\right)$$

First, by Taylor's theorem:

$$\mathbb{P}[SA]_L - \mathbb{P}[SA]_H = F(v_L - \kappa_L) - F(v_H - \kappa_H) \approx [v_L - v_H - (\kappa_L - \kappa_H)] \cdot \underbrace{f(v_H - \kappa_H)}_{f_H}$$

Goal: approximate $v_L - v_H$ and $\kappa_L - \kappa_H$ using take-up responses to changes in B and Λ.

Given the structural assumptions, $v(B) = v \cdot (1-\tau)\{B + y^{SA=1} - y^{SA=0}\} - \frac{n}{1+1/e} \cdot \left(\frac{y^{SA=1}}{n}\right)^{1+1/e} + \frac{n}{1+1/e} \cdot \left(\frac{y^{SA=0}}{n}\right)^{1+1/e}$. But since $y^{SA=0} = n \cdot v \cdot (1-\tau)^e$, this means:

$$v(B) = v \cdot (1-\tau) \cdot \underbrace{\left\{ B + y^{SA=1} - \frac{y^{SA=0}}{1+e} - \frac{\left(y^{SA=1}\right)^{1+1/e}}{\left(y^{SA=0}\right)^{1/e} 1+e} \right\}}_{\triangleq \Psi} \tag{A.2.3}$$

Note that: $v'(B) = v \cdot (1-\tau)$ in this setting. Finally, I assume $\kappa(\Lambda) = \kappa_1 \cdot \Lambda + \kappa_0$ where $\kappa_1 = \kappa'(\Lambda)$. To match the empirical application, assume income is fixed across types.

$$\begin{aligned} F(v_L - \kappa_L) - F(v_H - \kappa_H) &\approx [(v'_L(B) - v'_H(B)) \cdot \Psi - (\kappa'_L(\Lambda) - \kappa'_H(\Lambda)) \cdot \Lambda - \Delta\kappa_0] \cdot f_H \\ &= \left(\frac{\partial \mathbb{P}[SA]_L}{\partial B} \cdot \frac{f_H}{f_L} - \frac{\partial \mathbb{P}[SA]_H}{\partial B} \right) \cdot \Psi \\ &\quad + \left(\frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda} \cdot \frac{f_H}{f_L} - \frac{\partial \mathbb{P}[SA]_H}{\partial \Lambda} \right) \cdot \Lambda - \alpha \end{aligned}$$

by Equations (1.3.1) and (1.3.2) and where $\alpha = f_H \cdot \Delta\kappa_0$. Note that when the LHS $= 0$, we know that $f_L = f_H$. Therefore, $\alpha = \left(\frac{\partial \mathbb{P}[SA]_L}{\partial B} - \frac{\partial \mathbb{P}[SA]_H}{\partial B}\right) \cdot \Psi + \left(\frac{\partial \mathbb{P}[SA]_L}{\partial \Lambda} - \frac{\partial \mathbb{P}[SA]_H}{\partial \Lambda}\right) \cdot \Lambda$. Rearranging gives Equation (A.2.2).

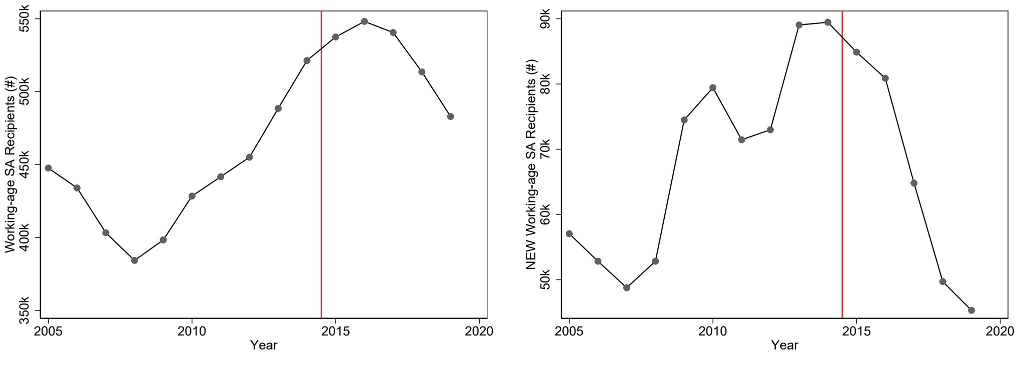
□

Appendix B

Appendix to Chapter 2

B.1 Context and Data

This section contains summary statistics about the data - comparing the general population to those eligible for social assistance. Pseudocode for my calculation of eligibility is presented in Appendix B.1.1



(a) Caseload

(b) Inflow

Figure B.1: Take-up of SA over time

Notes: Take-up of SA (%) is plotted over time for 2005 - 2021. Both absolute caseload and inflow are shown. Two time periods are defined by an important policies: before/after the Participation Act of 2015 as discussed above.

Socio-economic Demographics	General Population	Eligible
Gender (%)		
Woman	49.9	53.8
Man	50.1	46.2
Education (%)		
Primary School	5.4	26.7
High School	31.8	46.8
Bachelor's	14.3	6.0
Masters-PhD	8.5	2.6
Unknown	40.1	17.9
Main source of Income (%)		
Employment or Civil Service Job	63.2	8.9
Director-shareholder	2.2	0.1
Self-employment	9.9	4.6
Other Job	0.2	0.0
Unemployment Insurance	2.0	2.5
Disability Insurance	5.5	6.5
Social Assistance	4.3	55.3
Other Benefits	1.9	12.9
Pension	3.8	1.3
Student Aid	0.6	3.3
Other (not active or without income)	6.1	4.7
Household Composition (%)		
Single person household	17.8	45.6
Couple without children	26.8	11.1
Couple with children	45.1	20.1
Single parent	6.4	19.6
Couples and parents with flatmates	2.1	1.9
Other shared households	1.0	1.6
Other Information		
Age	46.4 (11.0)	45.0 (11.3)
Foreign-born (%)	16.4 (37.0)	42.5 (49.4)
Household Std. Disposable Income (€)	66,949.4 (73,978.0)	13,125.2 (2,795.6)
Household Net Worth (€)	169,760.0 (4,227,453.1)	-5,497.5 (85,933.0)
Contracted Hours (per year)	1,509.7 (602.6)	471.1 (451.0)
Eligible (%)	6.6 (24.8)	100.0 (0.0)
Receipt of Social Assistance (%)	5.1 (21.9)	60.0 (49.0)

Table B.1: Summary Statistics for General and Eligible Populations

(Mental) Health Information	**General Population** Mean (SD)	**Eligible** Mean (SD)
General Health		
All Care Spending (€)	2,037.4 (7,181.0)	3,711.6 (11,015.0)
Physical Chronic Conditions (count)	0.67 (1.13)	1.03 (1.44)
Mental Health (admin)		
Mental Healthcare Spending (€)	274.3 (3,237.2)	1,055.9 (6,892.6)
Psychotropic Medication (%)	10.3 (30.3)	24.7 (43.1)
Anti-psychotics (%)	2.1 (14.4)	8.4 (27.7)
Anxiolytics (%)	2.2 (14.7)	8.0 (27.1)
Anti-depressants (%)	7.6 (26.6)	16.1 (36.7)
Hypnotics and Sedatives (%)	1.2 (11.1)	4.5 (20.7)
ADHD Medication (%)	0.7 (8.5)	1.7 (12.8)
Mental Health Hospitalizations (%)	0.05 (2.1)	0.12 (3.5)
Deaths by Suicide (%)	0.01 (1.2)	0.05 (2.3)
Mental Health (survey)		
Loneliness (0-11)	2.64 (3.14)	5.51 (3.82)
Life Control (7-35)	27.13 (5.09)	22.36 (5.72)
Kessler-10 Psychological Distress (10-50)	15.69 (6.43)	22.24 (9.82)

Table B.2: Summary Statistics for General and Eligible Populations

B.1.1 Eligibility Pseudocode

Algorithm 1 Eligibility Calculation

```
Procedure CalculateIncome(calculation_type)
  if (calculation_type == "Yearly")
    Income = income from work, assets & benefits.
    Deduct taxes & national insurance contributions
  else if (calculation_type == "Monthly")
    Gross Income = monthly employment income (spolis).
      Gross Income $\mapsto$ Add yearly income from business, assets, sickness/disability benefits /12
      Gross Income $\mapsto$Add unemployment benefits over periods with no employment income
    Deductions = payroll taxes + national insurance contributions + employee insurance contributions
      Deductions $\mapsto$ Add other taxes (not on bijstand income)

Procedure DefineFamilies()
  Households = as in household income data (rinpersoonkern).
  Co-residents = people living at same address
  Families = $\leq 2$ adult Co-Residents in same Household, plus children.

Procedure CostSharing()
  Cost-sharers = adults
  Remove students (age 21-30) not receiving student grants
  Threshold = threshold ( # Cost-sharers in Family)
  Add norm-adjustment for all singles pre-2015.

Procedure CheckEligibility()
  Set Eligible = "Yes" if Income $\leq$ Threshold, wealth $\leq$ wealth limit, and house value $\leq$ house limit.
  Set Eligible = "No" if age $< 21$ or striking or living outside NL or in institutional hh or {age 21-27 student not receiving student grants}
```

B.2 Average Take-up Levels: Additional Material

$\hat{\beta}$: SA receipt regressed on $\mathbb{1}\{\text{Dispensed psychotropic drug}\}$, coefficients relative to good mental health (no dispensation) (p.p.)	(1)
ADHD	0.0459 (0.170)
Anti-Depressant	1.412*** (0.0506)
Hypnotic/Sedative	0.0719 (0.0845)
Anxiolytic	-0.0859 (0.066)
Anti-Psychotic	-1.399*** (0.0701)
Year, age and gender FEs	✓
Lagged income controls	✓
Lagged work-status FEs	✓
Individual FEs	
All other controls	✓
Observations (people-years)	5,187,572
R^2	0.650
Baseline mean	62.45

Standard errors in parentheses
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table B.3: SA Receipt vs Mental Health (different conditions)

Notes: Coefficients of social assistance take-up regressed on psychopharmacology dispensation fixed effects (by type: ADHD medications, anti-depressants, hypnotics/sedatives, anti-anxiety medications and anti-psychotics). Point estimates added to the control mean, with 95% confidence intervals. Lagged controls include income, wealth, education, work status, household composition, municipality, year, age, sector fixed effects, physical health, and benefits schedule. Eligible population from 2011 to 2020. Standard errors clustered at the municipality level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\hat{\beta}$: Receipt of SA poor vs good MH (p.p.)	3.072***	0.491	-0.819*	1.429***	0.540***	1.984***	0.911***
	(0.810)	(0.699)	(0.362)	(0.095)	(0.071)	(0.065)	(0.0498)
Year, age and gender FEs		✓	✓	✓	✓		✓
Lagged income controls			✓	✓	✓		✓
Lagged work-status FEs				✓	✓		✓
Individual FEs						✓	✓
All other controls					✓		✓
Observations (people-years)	5,671,855	5,671,855	5,187,572	5,187,572	5,187,572	5,361,899	5,014,850
R^2	0.001	0.045	0.161	0.640	0.650	0.001	0.059
Baseline mean	59.97	59.97	62.45	62.45	62.45	60.07	62.00

Standard errors in parentheses
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table B.4: SA Receipt vs Mental Health: progressively adding controls

Notes: Results of a regression of receipt of social assistance on mental health status (measured by dispensation of psychotropic meds). First column shows the results with no controls. Second column shows results adding year, age and gender fixed effects. Third column shows results adding lagged income controls. Fourth column shows results adding lagged hh composition, education, municipality, wealth and work-status controls. Fifth column shows results adding sector, physical health and benefits schedule controls. Sixth column shows results with individual fixed effects only (no controls). Seventh column shows results with individual fixed effects and all controls. The sample contains the calculated eligible for SA in 2011-2020. Standard-errors are clustered at the municipality-level.

	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{\beta}$: Receipt of SA poor vs good MH (p.p.)	0.540***	1.695***	1.767***	0.533	2.513***	0.191
	(0.071)	(0.086)	(0.527)	(0.575)	(0.673)	(0.450)
Year, age and gender FEs	✓	✓	✓	✓	✓	✓
Lagged income controls	✓	✓	✓	✓	✓	✓
Lagged work-status FEs	✓	✓	✓	✓	✓	✓
Individual FEs						
All other controls	✓	✓	✓	✓	✓	✓
Observations (people-years)	5,187,572	5,162,351	14,402	12,718	6,514	3,690,830
R^2	0.650	0.650	0.690	0.695	0.746	0.639
Baseline mean	62.45	62.66	64.34	63.89	64.71	62.78

Standard errors in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table B.5: SA Receipt vs Mental Health (different measures)

Notes: Coefficients of social assistance take-up regressed on mental health status–indicators of: psychotropic drugs (1), mental healthcare (2), severe surveyed psychological distress (3)/loneliness (4)/lack of control over own life (5), or mental health hospitalisation (6). Point estimates and standard errors shown. Lagged controls include income, wealth, education, work status, household composition, municipality, year, age, sector fixed effects, physical health, and benefits schedule. Eligible population from 2011 to 2020 (2011-2017 for hospitalisations). Around 2% of the general population are surveyed. Standard errors clustered at the municipality level.

B.3 Barrier Screening Effects: Additional Material

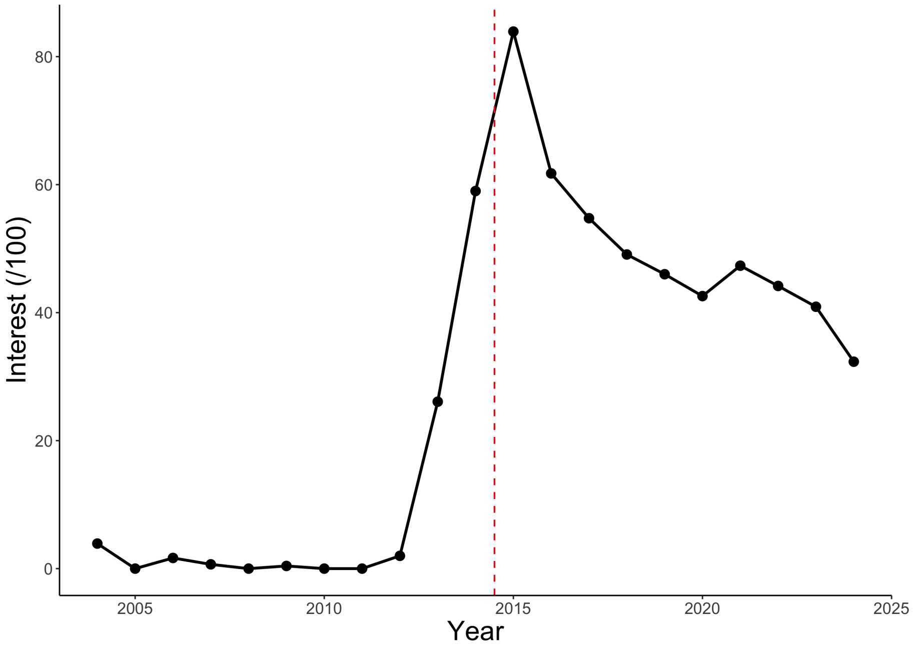


Figure B.2: Coverage of the "Participation Act" over time

Notes: Google Trends for *Participatiewet*, the Dutch translation of the "Participation Act" over time in the Netherlands.

B.3.1 Detail on the Participation Act

I argue that the Participation Act increased barriers to accessing social assistance. The policy intensified the obligations associated with receiving social assistance and incentivised municipalities to reduce caseload which they did via (threat of) sanctions.

SCP [2019], a report evaluating the Participation act, contains the results from a survey of 80 municipalities which asked representatives how often they impose obligations, and for each type of obligation how many impose these more after the introduction of the Participation Act. An overview of the results are shown in Table B.6.

Obligation	**Percent of Impose**	**Percent More Since PA15**
Language	76.5	69.4
Work	93.8	26.3
Accept Jobs	95.1	19.5
Register	48.1	20.5
Move	13.6	54.5
Commute 3 hours	29.6	50.0
Acquire skills	75.3	24.6
Clothes	63.0	49.3
Quid-pro-quo	87.7	56.8

Table B.6: Obligations

Notes: Percentage of municipalities surveyed who impose the full list of obligations (Column 2) and who impose obligations more often since the Participation Act (Column 3). The different obligations in-full are: achieving a good command of the Dutch language, work re-integration, register as a job seeker, being willing to relocate municipality, being prepared to travel a distance with a total travel time of 3 hours per day to find work, acquiring and retaining knowledge and skills necessary for acquiring wealth, wearing the correct clothing in work/volunteering, and quid-pro-quo unpaid voluntary work.

Table B.6 shows that many municipalities say they intensified the various obligations. No surveyed municipality said that they imposed obligations less often. Indeed, van der Veen [2019] state that "the PA introduced a much stricter regime of entitlement conditions, involving mandatory participation in 're-integration' activities [... and] introduced an important element of workfare, the so-called 'quid-pro-quo'".

There are plans in the Netherlands to repeal the Participation Act. Ministerie van SZW [2022] makes the case for a "Participation Act in Balance". The authors work with (former) social assistance recipients, municipalities and other experts to suggest that the obligations associated with the 2015 are too strict. They state:

> "Applying for social assistance is experienced by various experts as complex, tedious and too long. A negative tone [by the municipality] is also

> mentioned, threatening action from the outset and a creating a sense of mutual distrust. At the same time, citizens experience a high degree of dependence on the government. A feeling of shame prevails that they have to make use of social assistance, even though in situations they simply cannot (temporarily) do otherwise. People definitely understand the need for monitoring and enforcement, but the way in which this is done now is drastic. A small event can have major consequences. People do not always feel heard or treated as an equal person. Fear also arises. This can create a barrier to applying for assistance, even when the need is great."[1]

B.3.2 Results

Formally, the sample-selection issue can be framed as follows. Let $\mathbf{e}_i = (e_{i1}, ...e_{iT})$ where $e_{it} \in \{0, 1\}$ denotes eligibility. Let $\mathbb{X}_{it}$ be all explanatory variables (and $\mathbf{X}_i$ similarly). Essentially, we only "observe" $(\mathbf{X}_{it},\ SA_{it})$ for i, t such that $e_{it} = 1$ - i.e. only these observations are included in the regression. Wooldridge [2019] shows that the necessary identification assumption in this setting is given by Equation (B.3.1).

$$\mathbb{E}[\varepsilon_{it} \big| \mathbf{X}_i, \eta_i, \mathbf{e}_i] = 0 \tag{B.3.1}$$

However, note that eligibility is a (non-linear) function of observables: $e_{it} \triangleq \phi(y_{it}, \bar{y}_i, ...)$. Therefore, controlling for $y_{it}, \bar{y}_i$ etc implies that selection is fully determined by observables. I.e. the standard assumption $\mathbb{E}[\varepsilon_{it} \big| \mathbf{X}_i, \eta_i] = 0$ is sufficient. In this case, it is particularly important to check that the time-varying controls are not driving the results.

[1] Translated from page 8 of Ministerie van SZW [2022]

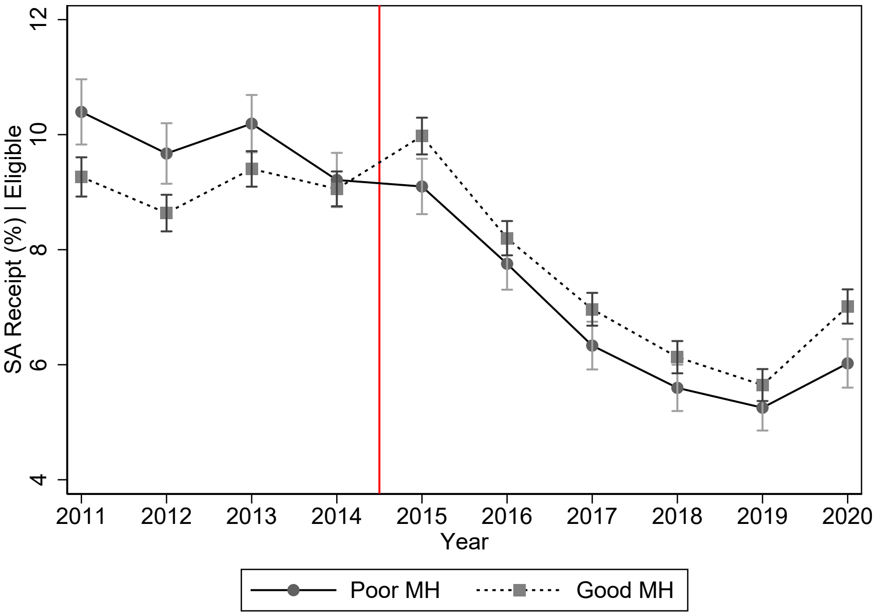


Figure B.3: SA Receipt and Inflow over time

Notes: Evolution of inflow of social assistance over time, split by people with poor mental health in the pre-period vs those with good mental health in this period. Raw means and respective 95% confidence intervals are shown. The introduction of the Participation Act in 2015 is shown by the red vertical line. Standard-errors are clustered at the level of municipality of residence in 2013.

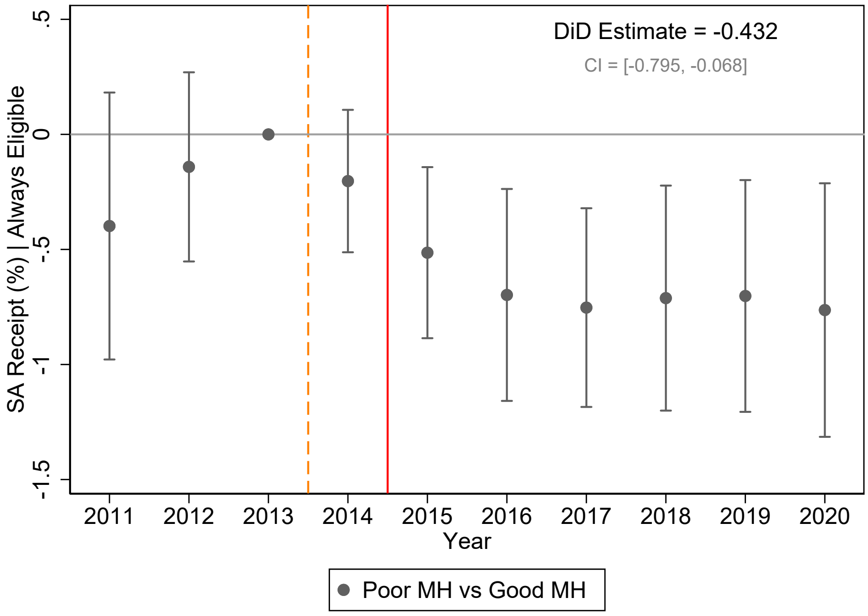


Figure B.4: Effects of Participation Act for the Always-Eligible

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is always-eligible middle-age couples and poor mental health is defined by prescription of psychopharma in the pre-period. Controls include individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. The TWFE estimate $\hat{\delta}$ in the regression $SA_{it} = \alpha + \eta_i + \gamma_t + \delta \cdot \mathbb{1}\{t \geq 2013\} \times \text{Poor MH}_i + X'_{it}\theta + \varepsilon_{it}$ is also shown. Standard-errors are clustered at the level of municipality of residence in 2013.

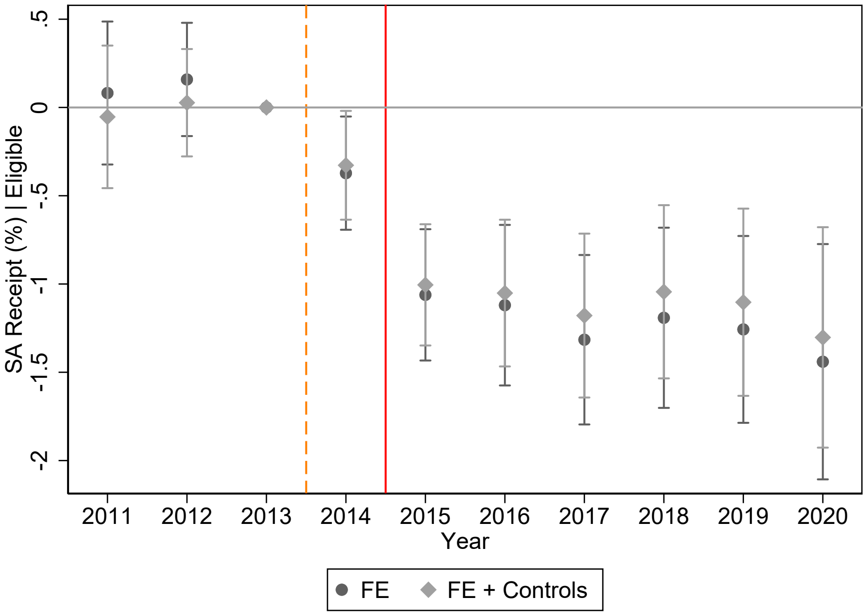


Figure B.5: Effects of Participation Act with/without controls

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in the pre-period. $\hat{\delta}_t$ are shown for two specifications - one with no time-varying controls (only individual FEs), and one with all time-varying controls - individual fixed effects, income, education and muncipality, hh composition and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

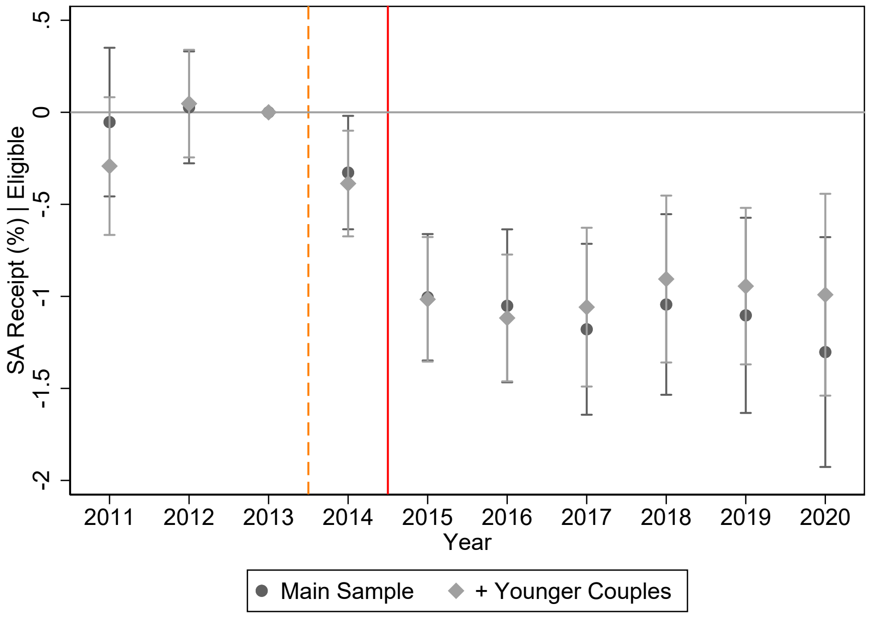


Figure B.6: Expanded Sample: Including Younger Couples

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in the pre-period. $\hat{\delta}_t$ are shown for two specifications – one with the standard analysis population, and the other with additionally including younger couples. Controls include individual fixed effects, income, education and municipality, household composition, and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

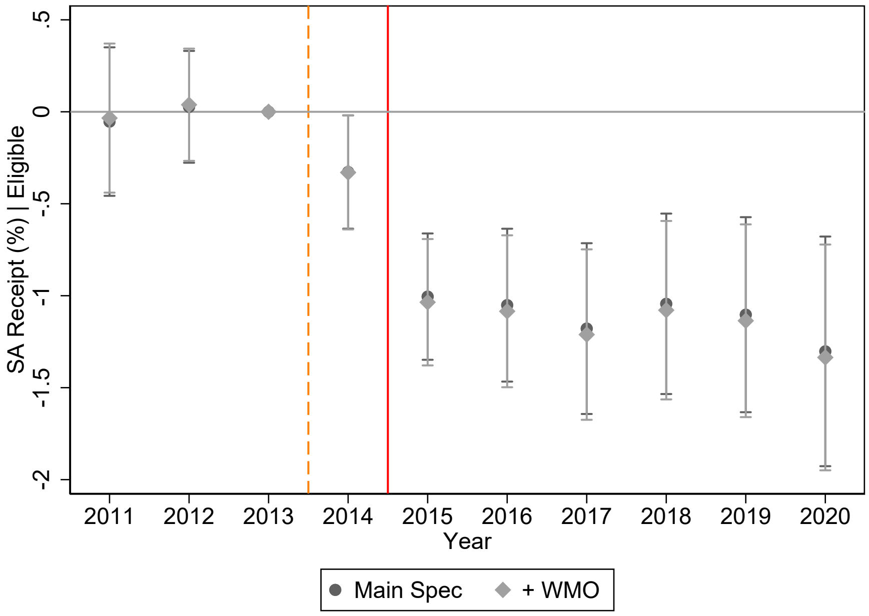


Figure B.7: Additional Control: Use of WMO Home Support

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in the pre-period. $\hat{\delta}_t$ are shown for two specifications – one with the main time-varying controls (individual fixed effects, income, education and municipality, household composition, and sector fixed effects), and one adding controls for use of home support via the WMO. Standard-errors are clustered at the level of municipality of residence in 2013.

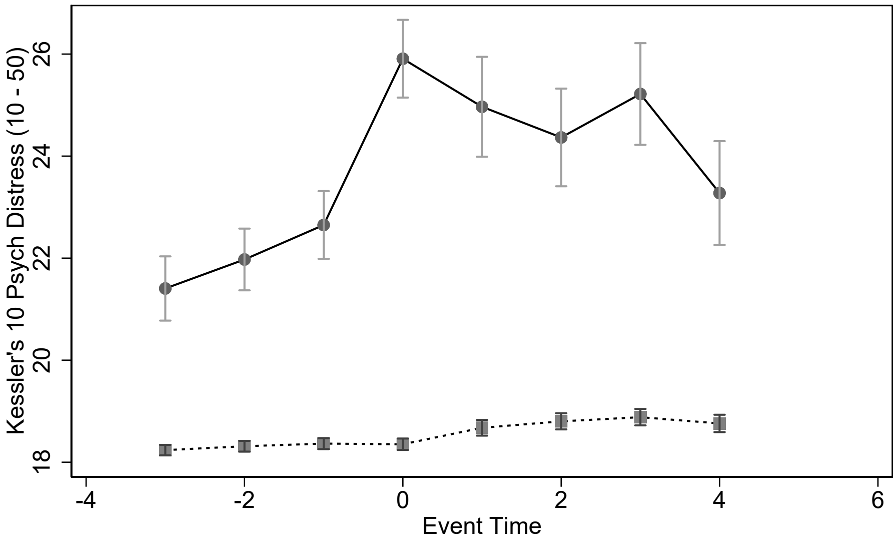


Figure B.8: Event-Time Dynamics of Subjective Mental Health (K10)

Notes: This plot shows the mean subjective mental health (measured by Kessler's 10 Psychological Distress) for two groups: one group is prescribed psychopharma for the first time in Event Time 0, the other group has no prescriptions for all event times $t \leq 0$. Standard-errors are clustered at the level of municipality of residence in 2013.

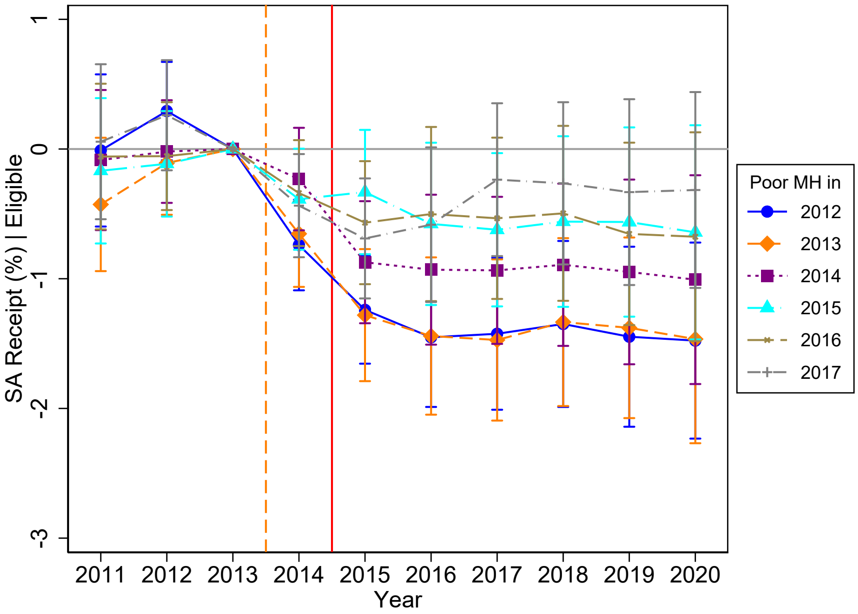


Figure B.9: Robustness to Timing of Poor Mental Health Definition

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples and poor mental health is defined by prescription of psychopharma in the pre-period. $\hat{\delta}_t$ are shown using alternative definitions of Poor $\text{MH}_i = \mathbb{1}\{\text{Prescribed Psychopharma in year } y\}$, for $y \in \{2012, ..., 2017\}$. Controls include individual fixed effects, income, education and municipality, household composition, and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

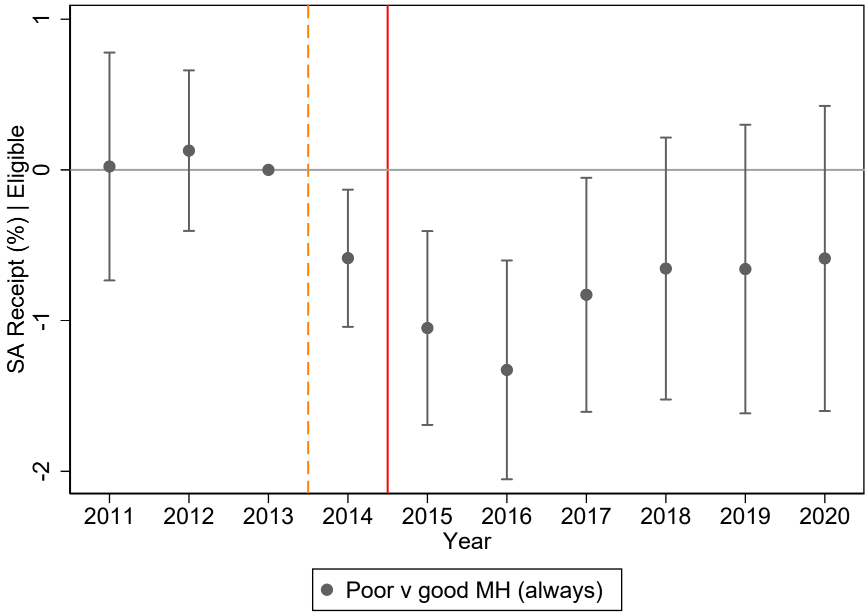


Figure B.10: Always Poor Mental Health (2011–2020)

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is eligible middle-age couples. Poor MH_i is defined as individuals prescribed psychopharma in every year from 2011 to 2020, compared to those with good mental health throughout. Controls include individual fixed effects, income, education and municipality, household composition, and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

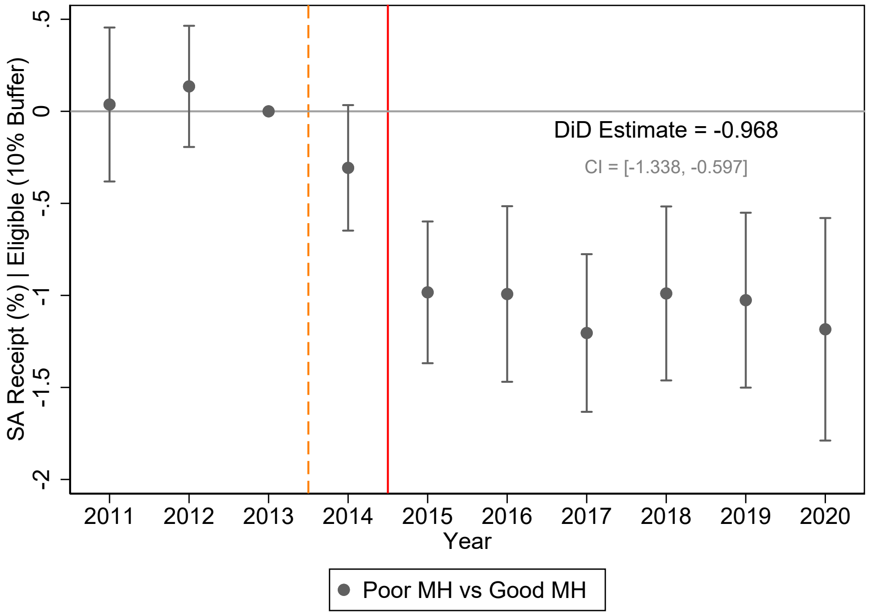


Figure B.11: Buffer Sample: Income Below 90% of Threshold

Notes: Estimates $\hat{\delta}_t$ from Equation (2.3.1) showing the heterogeneous treatment effects of an increase in ordeals on rate-of-receipt by baseline mental health. The analysis population is limited to individuals with income below 90% of the eligibility threshold. Poor $\text{MH}_i = \mathbb{1}\{\text{Prescribed Psychopharma in pre-period}\}$. Controls include individual fixed effects, income, education and municipality, household composition, and sector fixed effects. Standard-errors are clustered at the level of municipality of residence in 2013.

B.4 Benefit Take-up Effects: Additional Material

This section contains additional material relating to the RKD estimation of the effect of changes in benefit level on SA receipt (heterogeneously by mental health).

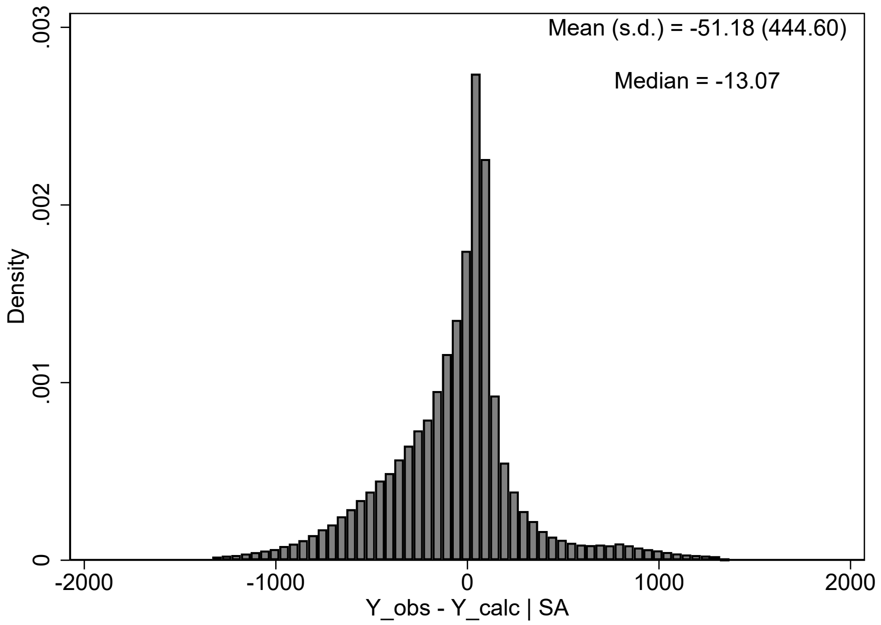


Figure B.12: Measurement Error of Income Concept

Notes: Histogram of $Y_{\text{true}} - Y_{\text{calc}}$ for the analysis population of the RKD.

B.4.1 Theory

Income is exempted "insofar as, in the judgment of the [municipality], it contributes to [their] employment opportunities" [Ministerie van SZW, 2015]. Appendix B.4.3 contains some descriptive facts about income-exemptions. This complicates matters because now, B is no longer deterministic (it depends on case-worker leniency) and $\frac{dB}{dy} \neq 1$ necessarily. Let the true benefits schedule be denoted $B = b(Y, \nu)$ where ν captures noncompliance with policy formula due to exemptions.

To properly re-scale the reduced-form estimates, we need to know how B depends on Y ex-ante. However, there is selection into social assistance with respect to exemptions. This makes sense because applicants receive more money with an exemption vs without, holding income fixed. Figure B.13 shows that observed benefits conditional on receipt departs from the benefits schedule, particularly at and above the threshold. In this region, applicants really only take-up social assistance if they receive an exemption. Selection on exemptions implies ex-post benefits received $\mathbb{E}[B|SA, Y = y]$ is not a good proxy for the ex-ante schedule $\mathbb{E}[B|Y = y]$.

I impute the benefits schedule using a theoretical approach.[2] I recover the ex-ante schedule from the ex-post schedule using Bayes-rule and average receipt. This rescaling exercise explained in-depth in Appendix B.4.3. While we may be worried about the endogeneity of using receipt in this calculation, I obtain similar results when I assume a less-flexible form for the probability of exemption - i.e. that it is constant w.r.t. y. In this case, the imputation does not depend on the full take-up function by income.

Figure B.14 shows the results of the process to impute the ex-ante benefits schedule, heterogeneously by baseline mental health (measured by lagged psychopharma dispensations). People with poor mental health receive more exemptions that those without - presumably because they have larger costs of working and this incentivises the municipality promote re-integration more.

I use the generalized non-separable model of Card et al. [2015]: receipt of SA is a function of benefit level B, income Y and error term ε: $\mathbb{P}[SA] = p(B, Y, \varepsilon)$. Let I_X be the support of random variable X which is potentially multi-dimensional, in which case represents a product space.

Proposition B.4.1. *(Card et al. [2015]) Under regularity, smooth effect of income,*

[2] Gelber et al. [2020] also use imputation for the first-stage of their RKD. The key idea, as in their paper, is that this imputation generates measurement error in the first-stage as well. The Card et al. [2015] framework can account for this measurement error.

y, first stage and non-negligible population at the kink, smooth density, smooth probability of no measurement error and monotonicity:

(a) $\mathbb{P}[\varepsilon \leq e, \nu \leq v | Y = y]$ *continuously differentiable in* y^* *at* $y^* = \bar{y}$ $\forall (e, v) \in I_{\varepsilon,\nu}$.

(b)

$$\frac{\lim_{\xi \to \bar{y}^+} \frac{d\mathbb{P}[SA|Y^*=y^*]}{dy^*}\Big|_{y^*=\xi} - \lim_{\xi \to \bar{y}^-} \frac{d\mathbb{P}[SA|Y^*=y^*]}{dy^*}\Big|_{y^*=\xi}}{\lim_{\xi \to \bar{y}^+} \frac{d\mathbb{E}[B^*|Y^*=y^*]}{dy^*}\Big|_{y^*=\xi} - \lim_{\xi \to \bar{y}^-} \frac{d\mathbb{E}[B^*|Y^*=y^*]}{dy^*}\Big|_{y^*=\xi}} = \int \frac{\partial \mathbb{P}[SA \mid B = b(\bar{y}, v), Y = \bar{y}, \varepsilon = e]}{\partial B} \cdot \varphi(e, v)\ dF_{\varepsilon,\nu}(e, v) \tag{B.4.1}$$

where weighting function

$$\varphi(e, v) = \frac{\mathbb{P}[U_Y = 0 | Y = \bar{y}, \varepsilon = e, \nu = v]\big(b_1^+(v) - b_1^-(v)\big)\frac{f_{Y|\varepsilon=e,\nu=v}(\bar{y})}{f_Y(\bar{y})}}{\int \mathbb{P}[U_Y = 0 | Y = \bar{y}, \varepsilon = e, \nu = \omega]\big(b_1^+(v) - b_1^-(v)\big)\frac{f_{Y|\varepsilon=e,\nu=\omega}(\bar{y})}{f_Y(\bar{y})} dF_\nu(\omega)} \tag{B.4.2}$$

The fuzzy RKD estimates a weighted average of marginal effects of B on $\mathbb{P}[SA]$ with weights $\varphi(e, v)$. The intuition is as follows. $\varphi(e, v)$ has three main components. $\frac{f_{Y|\varepsilon=e,\nu=v}(\bar{y})}{f_Y(\bar{y})}$ is the weight in a sharp RKD and reflects the relative likelihood an individual is located at the kink. $b_1^+(v) - b_1^-(v)$ reflects size of the kink: the fuzzy RKD upweights people with larger kinks. $\mathbb{P}[U_Y = 0 | Y = \bar{y}, \varepsilon = e, \nu = v]$ reflects the probability that the assignment variable is correctly measured at threshold.

The Card et al. [2015] identification assumptions are stated in full in Appendix B.4.4. Two are key to my setting. (a) the density of Y^* is continuously differentiable at the threshold $\bar{y}$, (b) the benefits-schedule is continuous $\implies$ $\mathbb{P}[\text{Exemption}|Y = y]$ continuous at $\bar{y}$.

Figure B.24 and Figure 2.12 show no evidence for non-smoothness of the distribution of income. Discontinuous $\mathbb{P}[\text{Exemption}|Y = y]$ would imply discontinuous

$\mathbb{E}[B|SA, Y = y]$ at the threshold. However, Figure B.13 exhibits no such discontinuity. Moreover, there are no conditions in the law which state income below/above the threshold should be exempted differently.

B.4.2 Estimation

I use monthly data for the regression kink design because eligibility is based on the previous month's income, making granular analysis crucial. While the data provide detailed monthly information on labor income and contracted hours, income from other benefits is only available yearly, which motivates my sample restrictions:

Sample Restrictions: I restrict the sample to individuals working more than zero hours and whose primary income is from work, to avoid notches in the benefit schedule (e.g., disability benefits) tied to the social assistance (SA) eligibility threshold. This threshold corresponds to the social minimum, which links to other government programs. Therefore, individuals who derive all their income from other benefits are ineligible for SA and are excluded. The typical person at the threshold earns most of their income from work/self-employment, with potential supplementary benefits, making them likely to move above or below the threshold at any point.

I further restrict the sample to singles before 2015, as misclassification near the threshold is more common for couples, and limit the period after the Participation Act to ensure the analysis is unaffected by changes in ordeal requirements.

Specification: I estimate a standard fuzzy RKD specification, using local linear regression. I use a Calonico et al. [2014] (hereafter, CCT) robust bandwidth of approximately €60. For the CCT bandwidth selection algorithm, I do not use regularization. This is because the CCT framework is not designed to efficiently identify heterogeneous RKDs nor account for measurement error. Both would suggest the use of a larger bandwidth.[3] The non-regularized CCT bandwidth delivers a larger bandwidth and has the same asymptotic properties as with regularization. The specification is as follows, where the IV estimate $\frac{\hat{\beta}_1}{\hat{\delta}_1}$ measures $\frac{\partial \mathbb{P}[SA|Y=\bar{y}]}{\partial B}$. I cluster standard-errors at the municipality level.

B.4.3 Income Exemptions

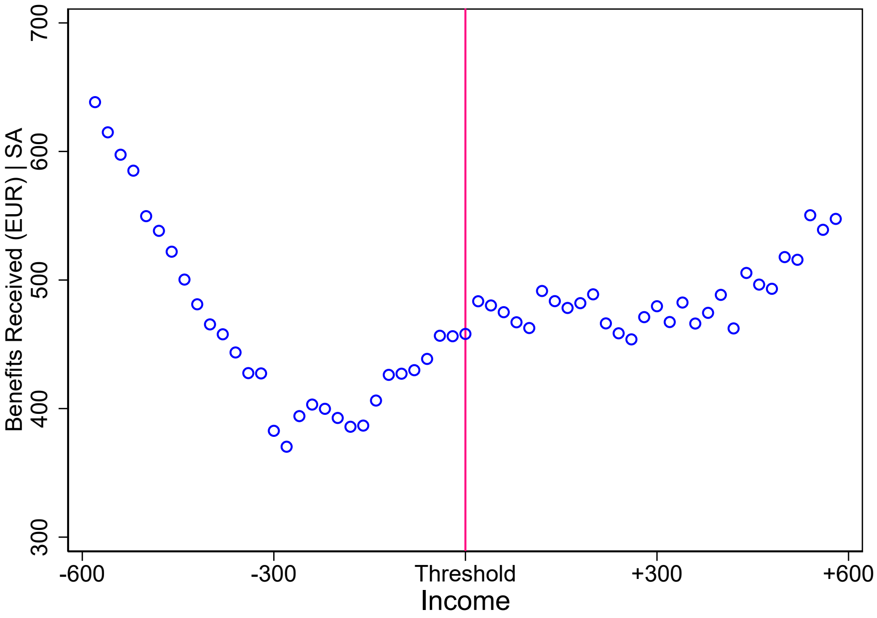


Figure B.13: $\mathbb{E}[B|SA, Y = y]$

Notes: Plot of benefits received conditional on receipt, averaged within income slice (€10 bins). A window of €1000 either side of the threshold is shown.

I model the unobserved benefits schedule as Equation (B.4.3).

[3] Indeed, the CCT robust bandwidth without regularization performs poorly in simulations (see Appendix B.4.5).

$$B = b(y,\nu) = \begin{cases} \bar{y} & \text{if exemption, } \nu = 1 \\ \max\{\bar{y} - y, 0\} & \text{if exemption, } \nu = 0 \end{cases} \tag{B.4.3}$$

where $\nu = 1$ with probability $p(y)$. This approach is motivated by the fact that $\mathbb{E}[B|SA, Y = y] \approx \bar{y}$ for $y \geq \bar{y}$. People with income above the threshold are not eligible for any benefits unless they receive an exemption, therefore $\mathbb{E}[B|SA, Y = y]$ is a good measure of benefits received conditional on exemption when $y \geq \bar{y}$. I allow for the possibility that exemptions can vary in reduced-form likelihood throughout the income distribution.

Proposition B.4.2 (Benefits-Schedule Imputation). *Suppose that the benefits-formula is given by Equation (B.4.3). Then,* $\mathbb{E}[B|Y = y] = p(y) \cdot \bar{y} + \big(1 - p(y)\big) \cdot \max\{\bar{y} - y, 0\}$ *where:*

$$p(y) = \begin{cases} \frac{\big(\mathbb{E}[B|SA,Y=y] - (\bar{y}-y)\big) \cdot \mathbb{P}[SA|Y=y]}{y \cdot \mathbb{P}[SA|Y=y,\nu=1]} & \textit{if } y \leq \bar{y} \\ \frac{\mathbb{E}[B|SA,Y=y] \cdot \mathbb{P}[SA|Y=y]}{\bar{y} \cdot \mathbb{P}[SA|Y=y,\nu=1]} & \textit{if } y \geq \bar{y} \end{cases} \tag{B.4.4}$$

The proof is a simple application of Bayes-rule. I proxy $\mathbb{P}[SA|Y = y, \nu = 1] \approx \mathbb{P}[SA|Y = 0]$: the take-up rate conditional on exemption is equal to the take-up rate for people who have no income ($\approx 100\%$).

Proof of Proposition B.4.2. Let $\mathbb{E}_y \triangleq \mathbb{E}[\cdot|Y = y]$ and $\mathbb{P}_y \triangleq \mathbb{P}(\cdot|Y = y)$

$$\begin{aligned}
\mathbb{E}[B|SA, Y = y] &= \mathbb{E}_y[B|SA] \\
&= \frac{\mathbb{E}_y[B \cdot \mathbb{1}\{SA\}]}{\mathbb{P}_y[SA]} \\
\mathbb{E}_y[B \cdot \mathbb{1}\{SA\}] &= \mathbb{E}_y[B \cdot \mathbb{1}\{SA\} \cdot \mathbb{1}\{\nu = 1\}] + \mathbb{E}_y[B \cdot \mathbb{1}\{SA\} \cdot \mathbb{1}\{\nu = 0\}] \\
&= \bar{y} \cdot \mathbb{P}_y[SA \cap \nu = 1] + \max\{\bar{y} - y, 0\} \cdot \mathbb{P}_y[SA \cap \nu = 0] \\
&= \bar{y} \cdot \mathbb{P}_y[SA \cap \nu = 1] + \max\{\bar{y} - y, 0\} \cdot \big[\mathbb{P}_y[SA] - \mathbb{P}_y[SA \cap \nu = 1]\big]
\end{aligned}$$

Note that $\mathbb{P}_y[\nu = 1] = p(y)$.

$$\begin{aligned}
&= \bar{y} \cdot p(y) \cdot \mathbb{P}_y[SA|\nu = 1] + \max\{\bar{y} - y, 0\} \cdot \big[\mathbb{P}_y[SA] - p(y) \cdot \mathbb{P}_y[SA|\nu = 1]\big] \\
&= \begin{cases} [\bar{y} - y] \cdot \mathbb{P}_y[SA] + y \cdot p(y) \cdot \mathbb{P}_y[SA|\nu = 1] & \text{if } y \leq \bar{y} \\ \bar{y} \cdot p(y) \cdot \mathbb{P}_y[SA|\nu = 1] & \text{if } y \geq \bar{y} \end{cases}
\end{aligned}$$

$$\text{Therefore, } \mathbb{E}_y[B|SA] = \begin{cases} \frac{[\bar{y}-y] \cdot \mathbb{P}_y[SA] + y \cdot p(y) \cdot \mathbb{P}_y[SA|\nu=1]}{\mathbb{P}_y[SA]} & \text{if } y \leq \bar{y} \\ \frac{\bar{y} \cdot p(y) \cdot \mathbb{P}_y[SA|\nu=1]}{\mathbb{P}_y[SA]} & \text{if } y \geq \bar{y} \end{cases}$$

Rearranging for $p(y)$ gives the expression in Equation (B.4.4).

□

B.4.4 Card et al. [2015] assumptions for validity of fuzzy RKD

1. **Regularity:** (ε, ν) has bounded support. $p(\cdot, \cdot, \cdot)$ is **continuous** and partially differentiable w.r.t. first and second arguments. $p_1(b, y, e)$ continuous.

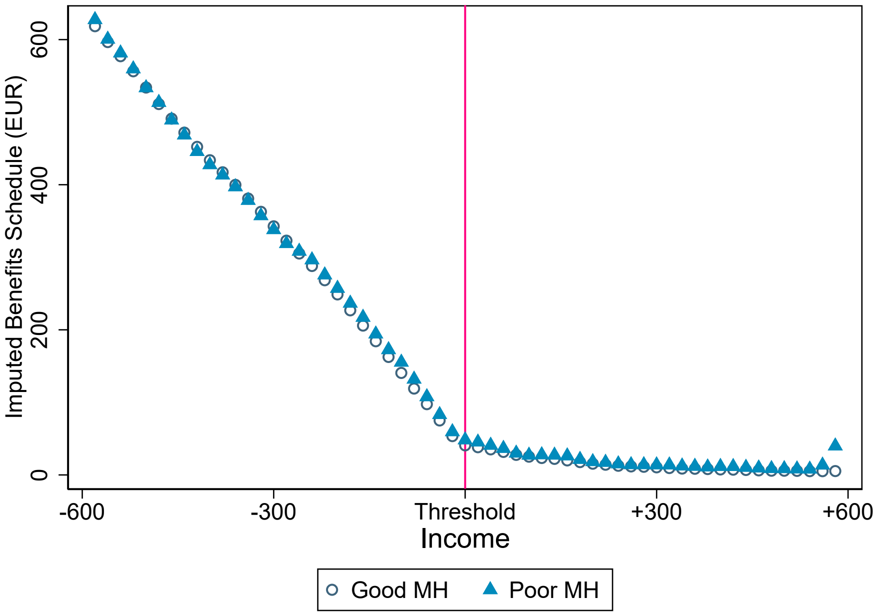


Figure B.14: Benefits Schedule Imputation by Mental Health

Notes: Results of Imputation from Proposition B.4.2, calculated separately for people dispensed anti-depressants in the year previously (poor mental health) versus those who were not (good mental health).

2. **Smooth effect of Y:** $p_2(b, y, e)$ is continuous.

3. **First Stage and Nonnegligible Population at Kink:** $b(y, v)$ continuous and $b_1(y, v)$ continuous apart from at $y = \bar{y}$. Positive mass at kink.

4. **Smooth Density:** Density of Y is continuously differentiable

5. **Smooth Probability of No Measurement Error:** $\mathbb{P}[U_Y = 0, U_B = 0|Y = y, \varepsilon, \nu]$ and partial derivative w.r.t. y is continuous.

6. **Monotonicity:** Either $b_1^+(v) \geq b_1^-(v)$ for all v or $b_1^+(v) \leq b_1^-(v)$ for all v.

There are two conditions for identification specific to my context worth highlighting: Assumption 9 and Assumption 10

Assumption 9 (No 0-censoring)**.**

(a) Take-up is not censored to 0 below threshold:

$$\forall \mathbb{P}[SA|B = b, Y \leq \bar{y}] > 0 \tag{B.4.5}$$

(b) Take-up is not censored to 0 above threshold:

$$\exists \Delta > 0 \ s.t. \ \mathbb{P}[SA|Y = y] > 0 \ \forall y \in [\bar{y}, \bar{y} + \Delta] \tag{B.4.6}$$

Assumption 10 (Continuous probability of exemption)**.**

$$\mathbb{P}[Exemption|Y = y] \ continuous \ at \ \bar{y} \tag{B.4.7}$$

Without both parts of Assumption 9, the numerator of the estimand in Equation (B.4.1) will be 0, while without one part only, regularity is violated. In my sample, around 8% of people receiving social assistance have $Y_{\text{true}} > \bar{y}$. Assumption 10 is a corollary of $b(y, v)$ being continuous.

B.4.5 Estimation Choices

To assess the performance of the CCT robust bandwidth in my context, I perform simulation analyses on a simplified version of the model set out in Chapter 1. The motivation for these analyses is that the frameworks are not designed for (i) measurement error and (ii) efficiently detecting heterogeneous RKD effects.

Setup

I simulate a million individuals which are characterised by ability $Y \sim U[500, 1500]$. This corresponds to their income. I set a fixed cost to be $\kappa = 150$ for everyone.

Choice error $\varepsilon = \frac{U_1+U_2}{2}$ where $U_j \overset{\text{i.i.d.}}{\sim} U[-200, 200]$. I.e. ε follows a symmetric triangular distribution centered around 0. The threshold $\bar{y} = 1000$ for everyone. Benefits schedule $B(y)$ is programmed as $B(y) = \max\{\bar{y} - \nu \cdot y, 0\}$ where exemption $\nu \in \{0, 1\}$ and $\mathbb{P}[\nu = 1|Y = y] \equiv 0.1$. Individual y takes up iff:

$$B(y) \geq \kappa \tag{B.4.8}$$

In the case of measurement error, I let $Y^* = Y + U_Y$ where $U_Y \sim N(0, 100)$. I then run the CCT robust bandwidth and RKD analyses exactly as in the main analysis. Specifically, I impute the benefits schedule as in Proposition B.4.2.

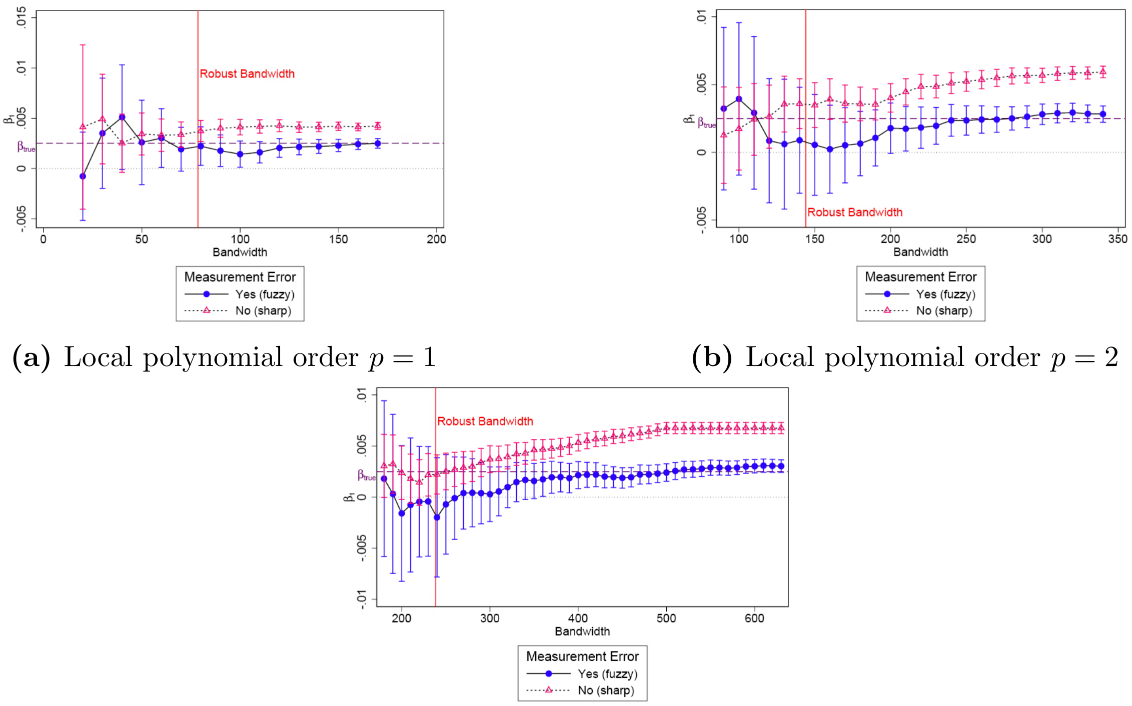
Results

Polynomial order: applying rules-of-thumb from Pei et al. [2022] suggests a linear estimator. Furthermore, simulations show that with measurement-error - linear estimators out-perform higher order polynomials at the CCT robust optimal bandwidth. This result echoes Card et al. [2015] who suggests that the CCT bandwidths can be too small for RKDs.

Bandwidth: for linear estimation, CCT bandwidths seem to perform well, but estimates become noisy for lower values with measurement error. For the identification of heterogeneous effects under measurement error, CCT performs poorly: I now assume that half of my simulated individuals have value $\alpha = 1$, and half $\alpha = 2$. Individuals take-up iff:

$$\alpha \cdot B(y) \geq \kappa \tag{B.4.9}$$

and rate of receipt $\mathbb{P}[SA|Y = y] = F_\varepsilon\big(\alpha \cdot B(y) - \kappa\big)$. I estimate the RKDs separately



(a) Local polynomial order $p = 1$

(b) Local polynomial order $p = 2$

(c) Local polynomial order $p = 3$

Figure B.15: RKD Polynomial Simulations

Notes: Results of simulations showing estimates from RKDs using different bandwidths and different local polynomial orders. In each, the CCT robust bandwidth is shown.

for $\alpha = 1, 2$ and test for a difference in the RKD estimates at different bandwidths. The estimates are shown in Figure B.16. The plot shows that the CCT bandwidth performs poorly (noisy and biased estimate of the heterogeneous RKD), whereas the estimators converge to the true effect for larger bandwidths.

Other: use standard triangular kernel.

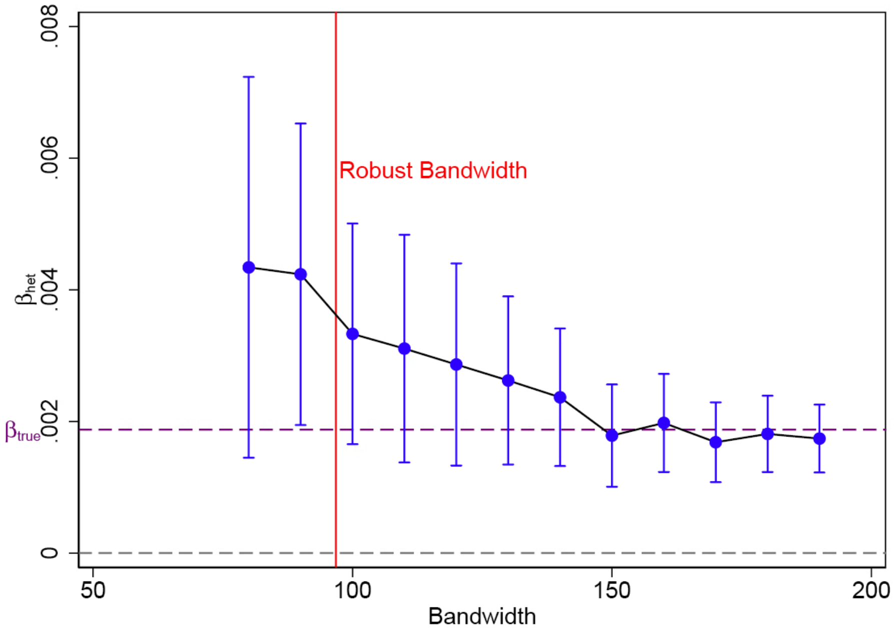


Figure B.16: Heterogeneous RKD Simulation

Notes: Results of simulations showing estimates from heterogeneous RKD ($\alpha = 1$ vs 2) using different bandwidths. CCT robust bandwidth is shown.

B.4.6 Results

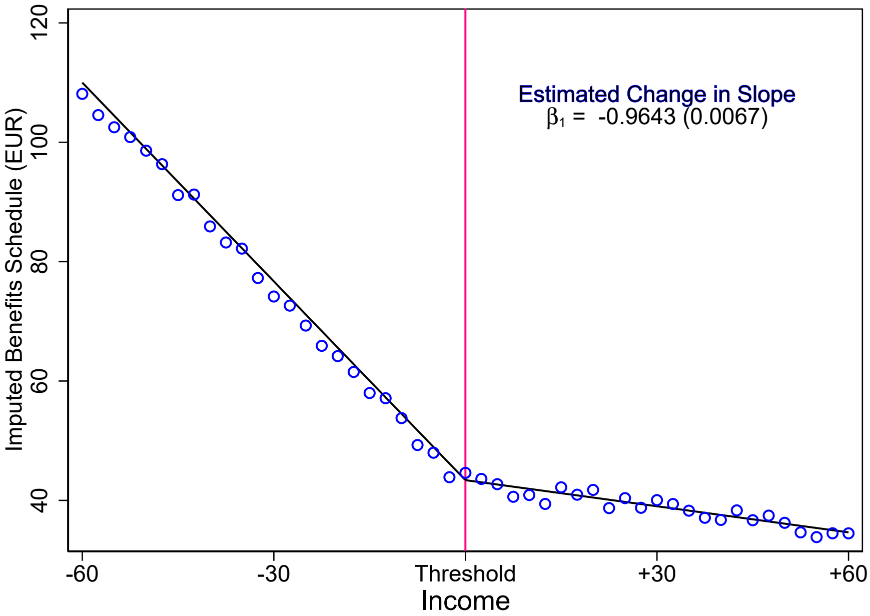


Figure B.17: RKD First Stage

Notes: Average imputed benefits schedule within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as the estimated change in slopes following the regression kink design. Standard-errors are clustered at the municipality level.

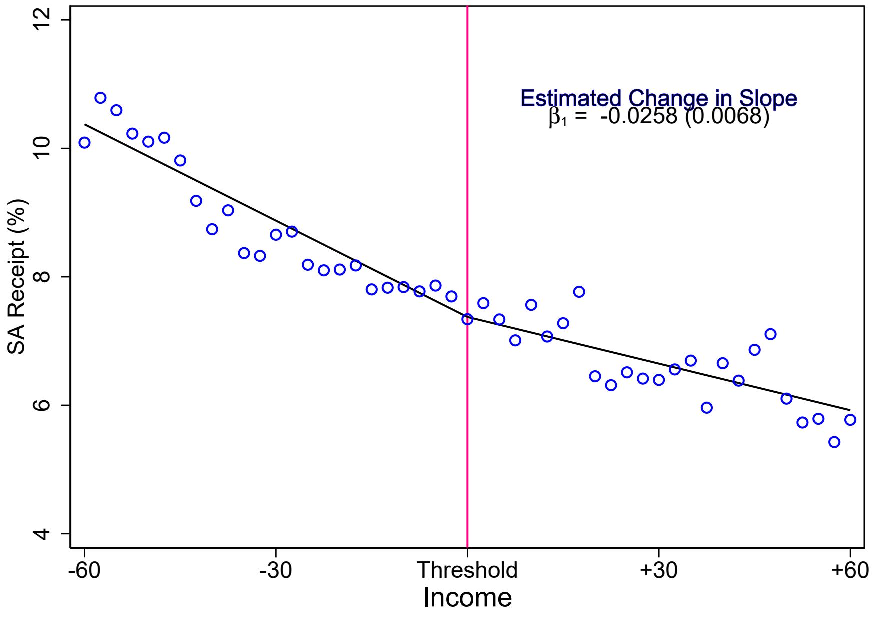


Figure B.18: Benefit Take-up Effects

Notes: Average rate of receipt within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as the estimated change in slopes following the regression kink design. Standard-errors are clustered at the municipality level.

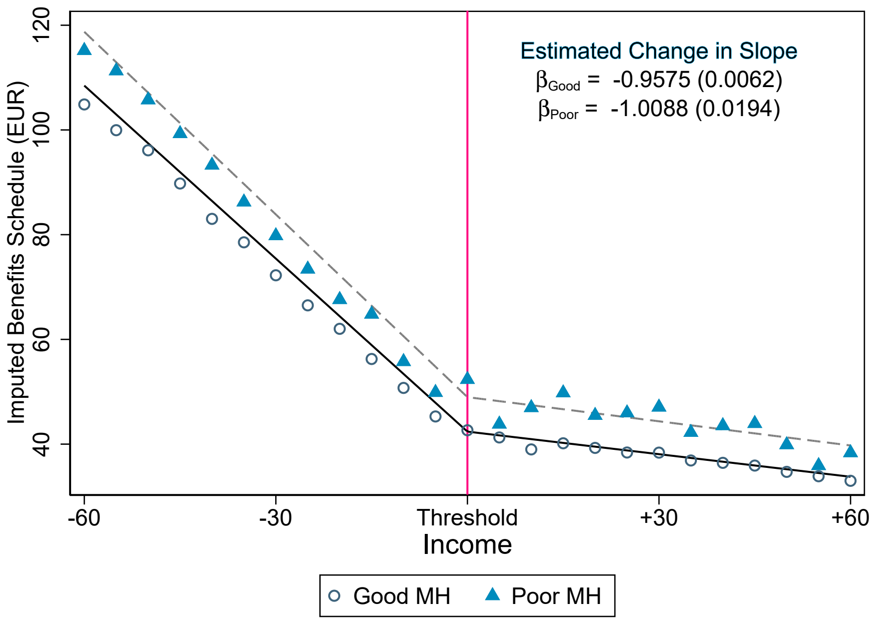


Figure B.19: RKD First Stage by Mental Health

Notes: Average imputed benefits schedule within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as receiving anti-depressants in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as the estimated change in slopes following the regression kink design. Standard-errors are clustered at the municipality level.

Effects on $\mathbb{P}[SA]$	**Reduced Form**				**IV**			
	Overall		Heterogeneous by MH		Overall		Heterogeneous by MH	
	Raw	+ Controls	Raw	+ Controls	Raw	+ Controls	Raw	+ Controls
Income - Threshold *Overall: everyone, Het: good MH*	-0.0242***	-0.0186***	-0.0241***	-0.0174***	-0.0203***	-0.0153***	-0.0209***	-0.0144**
	(0.00386)	(0.00304)	(0.00394)	(0.00329)	(0.00481)	(0.00385)	(0.00494)	(0.00421)
Income - Threshold (het) *Het: poor vs good MH*			0.00058	-0.00703			0.00506	-0.00513
			(0.00938)	(0.00826)			(0.0119)	(0.0104)
$\min\{\text{Income - Threshold}, 0\}$ *Overall: everyone, Het: good MH*	-0.0258***	-0.0207***	-0.0218**	-0.0187**				
	(0.00684)	(0.00585)	(0.00720)	(0.00653)				
$\min\{\text{Income - Threshold}, 0\}$ (het) *Het: poor vs good MH*			-0.0290	-0.0136				
			(0.0182)	(0.0164)				
Benefits *Overall: everyone, Het: good MH*					0.0213***	0.0267***	0.0227**	0.0194**
					(0.00600)	(0.00707)	(0.00751)	(0.00677)
Benefits (het) *Het: poor vs good MH*							0.0503**	0.0317*
							(0.0173)	(0.0145)
Observations (people-months)	487,475	448,307	487,475	448,307	487,475	448,307	487,475	448,307
R^2	0.002	0.225	0.003	0.226	0.001	0.226	0.001	0.226
Regressors	2	354	5	339	2	548	5	474

Standard errors in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table B.7: Full RKD Benefit Take-up Effect Results

Notes: Estimates from the regression kink design using a €60 bandwidth. Columns 2–5 show reduced-form estimates; Columns 6–9 show IV estimates using the imputed benefits schedule as a first stage. Columns 2 and 3 estimate the slope of social assistance receipt with respect to income, without and with controls (month, year, age, gender, wealth, education, municipality, household composition, sector fixed effects). Columns 4 and 5 show heterogeneous effects by mental health. "Income – Threshold" refers to individuals with good mental health; "Income – Threshold (het)" captures the difference for poor mental health. Standard errors clustered at the municipality level. Sample: singles, 2011–2014, primarily earning from work. See text for details.

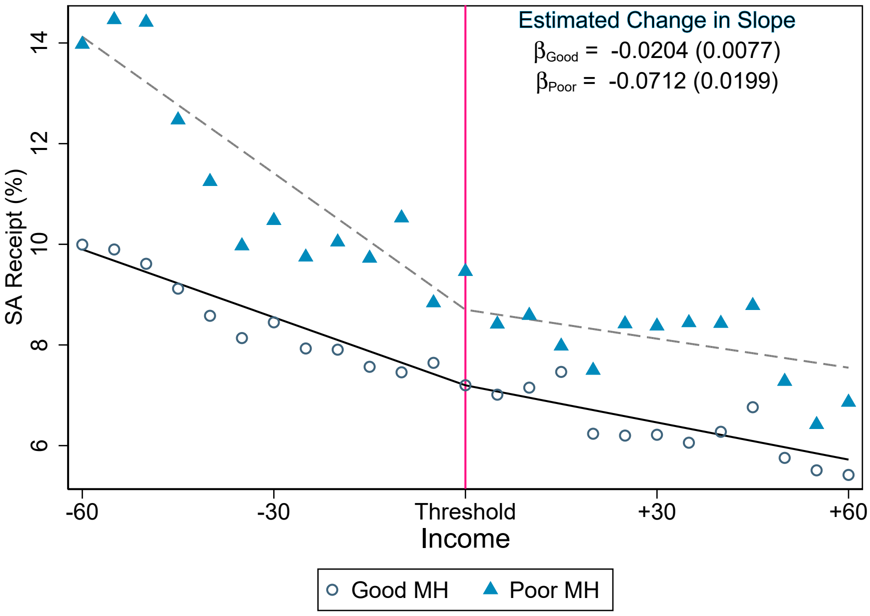


Figure B.20: Benefit Take-up Effects by 𝟙{Anti-Depressants}

Notes: Average rate of receipt within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as receiving anti-depressants in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as estimated change in slopes from the regression kink design. Standard-errors are clustered at the municipality level.

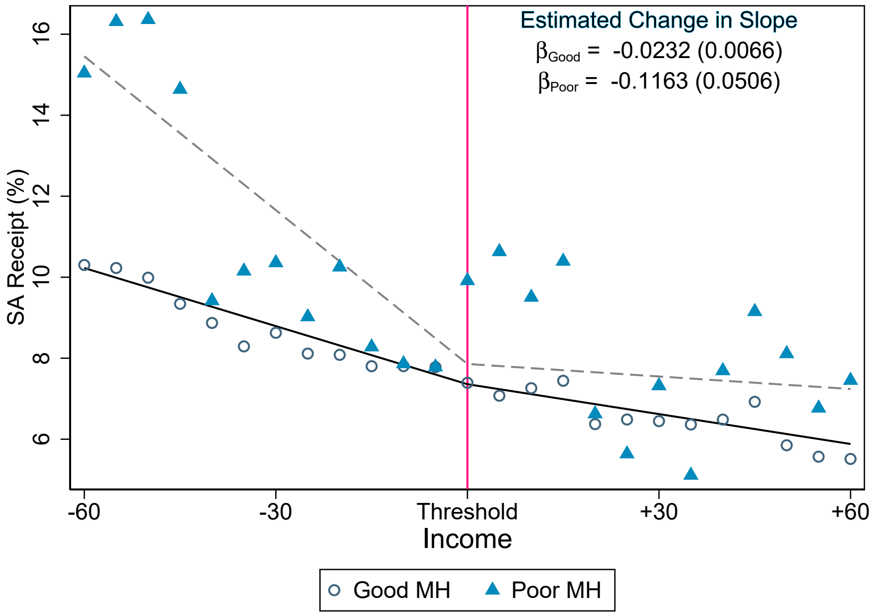


Figure B.21: Benefit Take-up Effects by 𝟙{Anti-Psychotics}

Notes: Average rate of receipt within income slice in a small window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as receiving anti-psychotics in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as estimated change in slopes from the regression kink design. Standard-errors are clustered at the municipality level.

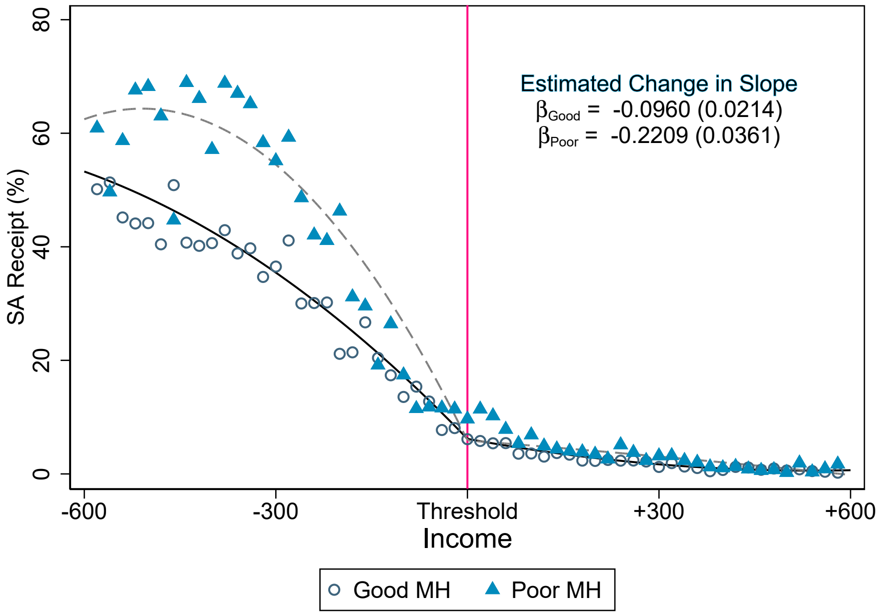
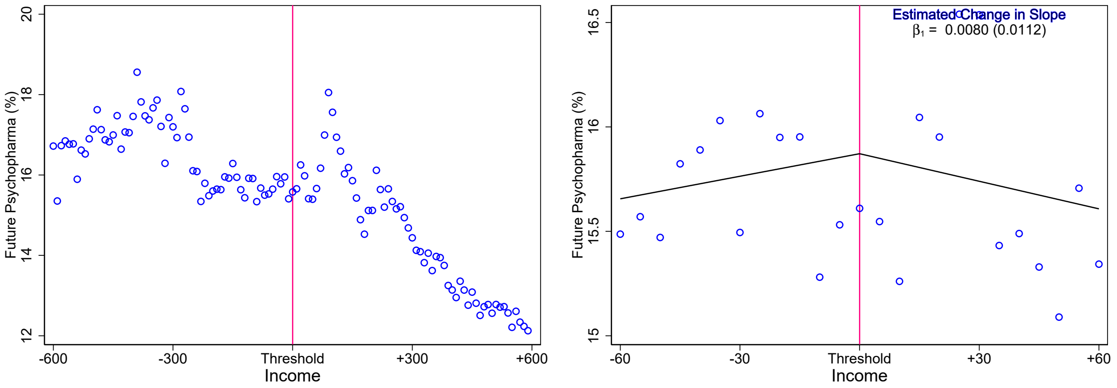


Figure B.22: Benefit Take-up Effects by Surveyed Mental Health

Notes: Average rate of receipt within income slice in a large window of income either side of the eligibility threshold. Income in this plot is monthly. Poor mental health is defined as reporting severe psychological distress in the survey in 2012. The sample contains single employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions.



(a) Overview

(b) Zoomed

Figure B.23: Use of Psychotropic Drugs around Eligibility Threshold

Notes: Average dispensations of psychotropic drugs in the future year within current income slice in a window around the eligibility threshold. Income in this plot is monthly. The sameple contains single employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions.

B.4.7 Validity of RKD

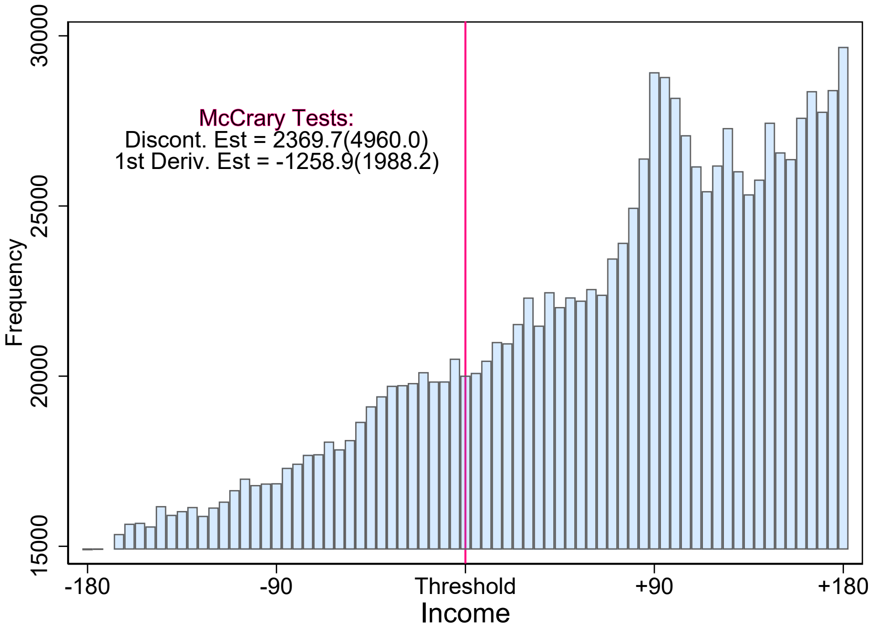


Figure B.24: Density around Eligibility Threshold

Notes: Density of income around the eligibility threshold. McCrary [2008] tests for discontinuity in levels and slopes around the threshold are shown. Income in this plot is monthly. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions.

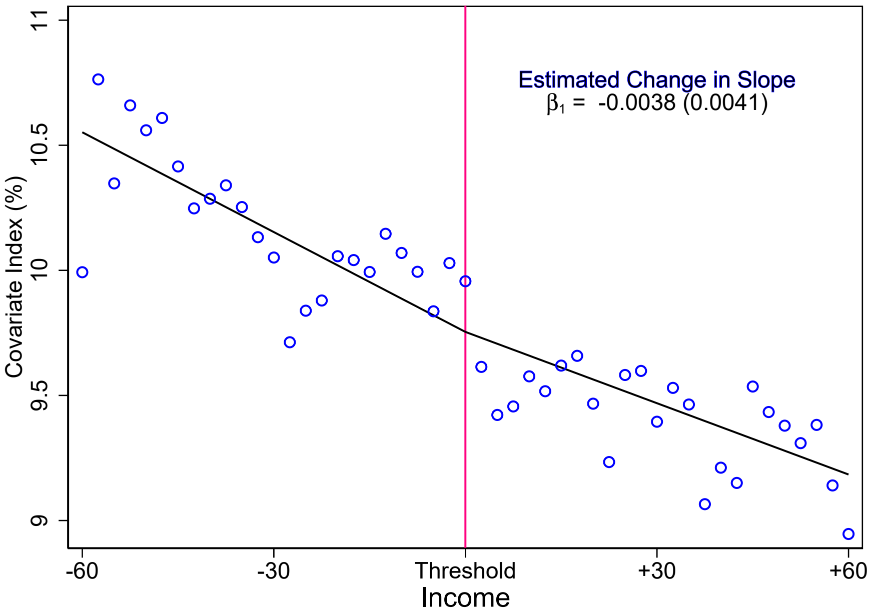


Figure B.25: RKD Covariate Test

Notes: Covariate Test: plot shows fitted values of a regression of social assistance take-up on all pre-determined controls used throughout this paper including income, education, hh composition, municipality FEs. These fitted values form a "Covariate Index" which is binned. An RKD estimate with income as the running variable is also shown. Income in this plot is monthly. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Standard-errors are clustered at the municipality level.

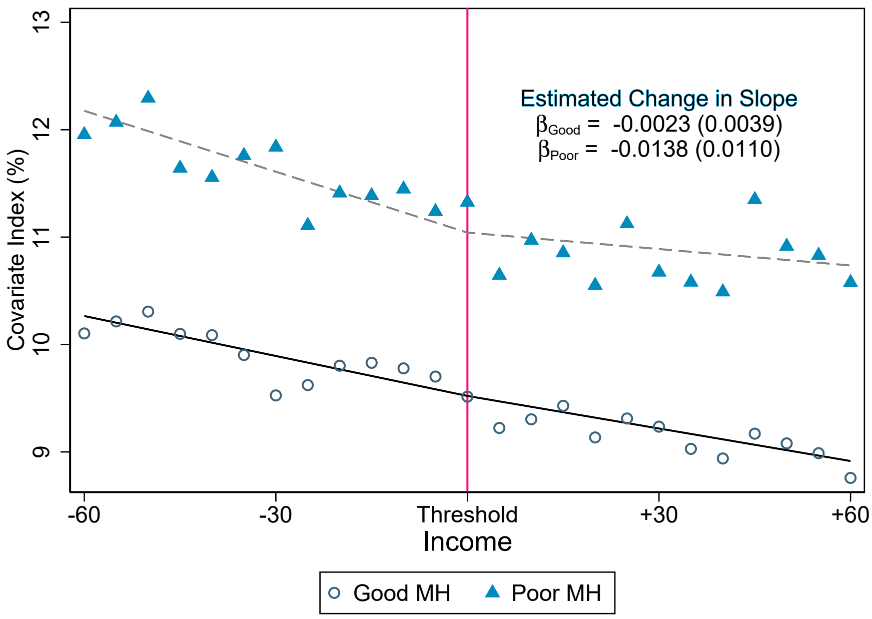


Figure B.26: RKD Covariate Test by Mental Health

Notes: Covariate Test: plot shows fitted values of a regression of social assistance take-up on all pre-determined controls used throughout this paper including income, education, hh composition, municipality FEs. These fitted values form a "Covariate Index" which is binned. An RKD estimate with income as the running variable is also shown. Separated by mental health. Income in this plot is monthly. Poor mental health is defined as receiving psychopharma in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Regression lines are shown following Section 2.4.1, as well as the estimated change in slopes following the regression kink design. Standard-errors are clustered at the municipality level.

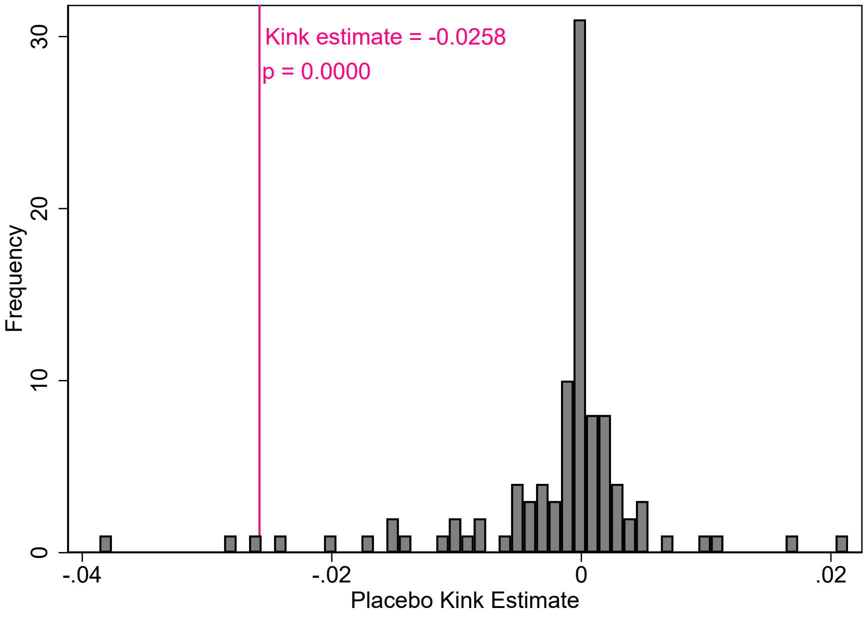


Figure B.27: RKD Permutation Test

Notes: Results of permutation test à la Ganong and Jäger [2018]. I estimate RKDs on 100 placebo kinks in the range $[\bar{y} - 600, \bar{y} + 600]$ and plot a histogram of the estimates. A binomial test is used to check whether the true estimate is an outlier. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Standard-errors are clustered at the municipality level.

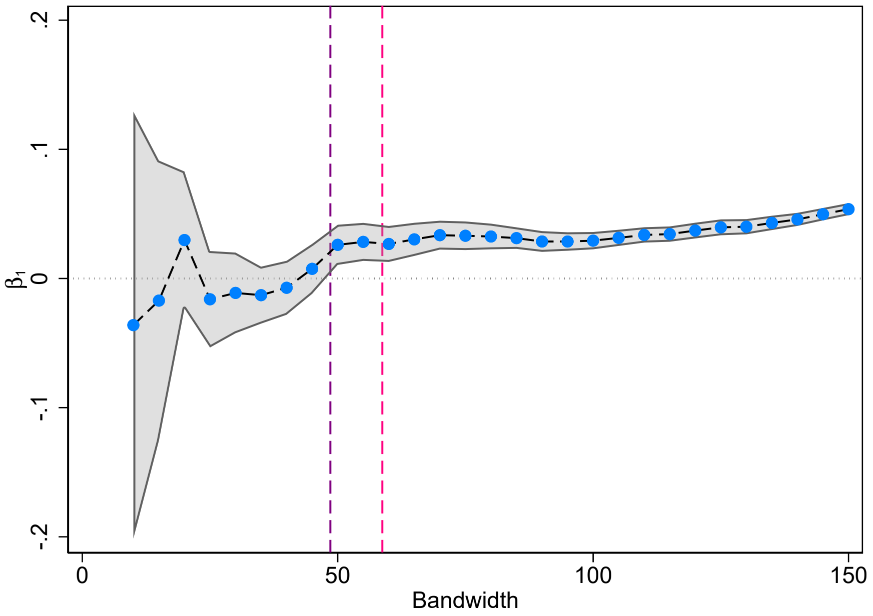


Figure B.28: RKD Bandwidth Test

Notes: Results of test of sensitivity to changes in bandwidth. I estimate RKDs changing the bandwidth, with the CCT robust bandwidth displayed. The lower purple dashed line indicates the CCT robust bandwidth with regularization, and the upper pink dashed line indicates the CCT robust bandwidth without regularization. This plot shows the estimates and confidence intervals. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Standard-errors are clustered at the municipality level.

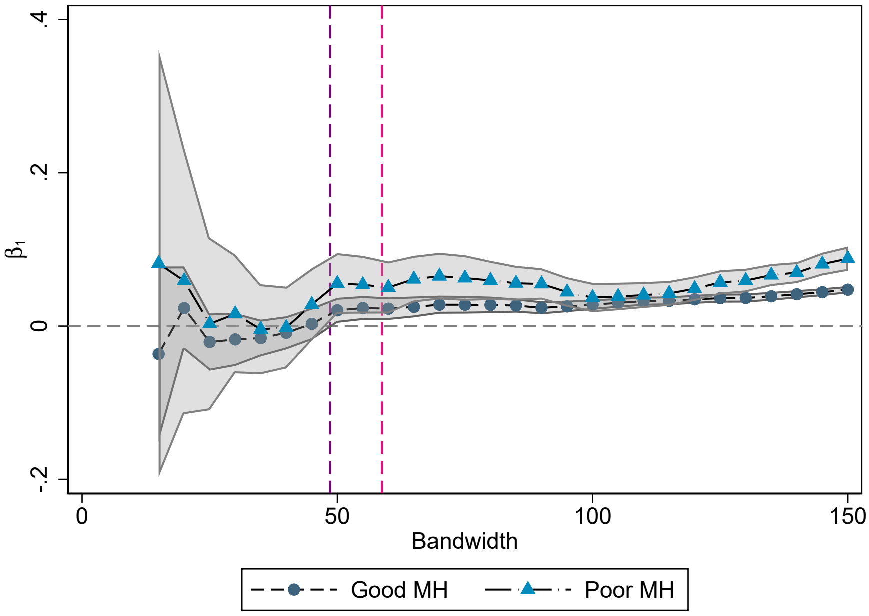


Figure B.29: RKD Bandwidth Test by Mental Health

Notes: Results of test of sensitivity to changes in bandwidth. I estimate heterogeneous RKDs changing the bandwidth, with the CCT robust bandwidth displayed. The lower purple dashed line indicates the CCT robust bandwidth with regularization, and the upper pink dashed line indicates the CCT robust bandwidth without regularization. This plot shows the estimates and confidence intervals. Poor mental health is defined as receiving psychopharma in the year previously. The sample contains singles employees, years 2011-2014. See Section 2.4.1 for details on sample restrictions. Standard-errors are clustered at the municipality level.

Appendix C

Appendix to Chapter 3

C.1 Eligibility

Throughout the empirical analysis, I focus on take-up levels and responses among the *eligible* population. This is because I am interested in take-up *behaviour* across types, and not underlying eligibility. However, the theoretical framework above does not model eligibility directly. Indeed, the government budget constraint as defined in Equation (1.2.1) makes much more sense if it holds for θ in the general population, and not the eligible. In reality, the ineligible fund benefits for the recipients, and not the eligible non-takers.

Proposition C.1.1 shows that identifying take-up levels and responses for the eligible population is sufficient for the general population as long as $\mathbb{P}[SA|\text{Ineligible}] = 0$.

Proposition C.1.1. *Assume* $\mathbb{P}[SA|\mathit{Ineligible}] = 0$*. Then:*

$$\mathbb{P}[SA]_\theta = \mathbb{P}[SA \mid \mathit{Eligible}]_\theta \cdot \mathbb{P}[\mathit{Eligible}]_\theta \tag{C.1.1}$$

and take-up responses to policy X *are given by:*

$$\frac{\partial \mathbb{P}[SA]_\theta}{\partial X} = \frac{\partial \mathbb{P}[SA \mid \mathit{Eligible}]_\theta}{\partial X} \cdot \underbrace{\left(\frac{\mathbb{P}[\mathit{Eligible}]_\theta}{1 - \mathbb{P}[SA \mid \mathit{Eligible}]_\theta \cdot \mathbb{P}[\mathit{Ineligible} \mid \mathit{No}\ SA]_\theta}\right)}_{EE_\theta\text{: } \mathit{Effective\ Eligibility}_\theta} \tag{C.1.2}$$

Proposition C.1.1 follows from Bayes Rule, the fact that eligibility is determined by $y \leq \bar{y}$ where $y = SA \cdot y^{SA=1} + (1 - SA) \cdot y^{SA=0}$ and from the fact we have assumed no labor supply responses to dB or $d\Lambda$. The intuition is as follows: we need to adjust for baseline incomplete take-up and the fact that ineligible people can still be on the margin of take-up (if they were just indifferent between earning income above the threshold and switching to earning income below the threshold and receiving social assistance) when mapping conditional take-up responses to the general population.

How should we implement Proposition C.1.1 when calculating welfare effects? When integrating against average take-up levels, Bayes Rule $\rightarrow \int \mathbb{P}[SA]_\theta \cdot H_\theta d\mu = \mathbb{P}[\text{Eligible}] \cdot \int \mathbb{P}[SA \mid \text{Eligible}]_\theta \cdot H_\theta d\mu_{\text{Eligible}}$. Where μ_{Eligible} is the conditional density of types θ. Similarly, Bayes Rule $\rightarrow \int \frac{\partial \mathbb{P}[SA]_\theta}{\partial X} \cdot H_\theta d\mu =$

$\mathbb{P}[\text{Eligible}] \cdot \int \frac{\partial \mathbb{P}[SA|\text{Eligible}]_\theta}{\partial X} \cdot \frac{1}{1 - \mathbb{P}[SA \mid \text{Eligible}]_\theta \cdot \mathbb{P}[\text{Ineligible} \mid \text{No } SA]_\theta} \cdot H_\theta d\mu_{\text{Eligible}}$.

C.2 Identification

Sufficient Statistics	Method	Estimated Value
$\left(\frac{\partial \hat{\mathbb{P}}[SA]_H}{\partial B}, \frac{\partial \hat{\mathbb{P}}[SA]_L}{\partial B}\right)$	RKD	$(0.000227, 0.000503)$
v'_H $f(v_H - \kappa_H)$	Normalization RKD$_H$ + v'_H	1 0.000227
$f(v_L - \kappa_L)$	$\hat{\mathbb{P}}[SA]_H = \hat{\mathbb{P}}[SA]_L$ + shortcut: $f(\cdot)_L = f(\cdot)_H$	0.000227
v'_L	RKD$_L$ + $f(v_L - \kappa_L)$	2.2
$\left(\frac{\partial \hat{\mathbb{P}}[SA]_H}{\partial \Lambda}, \frac{\partial \hat{\mathbb{P}}[SA]_L}{\partial \Lambda}\right)$	(Diff, Diff-in-Diff)	$(-0.014, -0.023)$
(κ'_H, κ'_L)	(Diff-in-)Diff + $f(\cdot)_L = f(\cdot)_H$	$(98, 161)$

Table C.1: Table summarising the calibration of the key sufficient statistics

For the identification, I use the point-estimates of the take-up response to a change in benefit level using the regression kink design with Calonico et al. [2014] robust bandwidth. These point estimates, $(0.000227, 0.000503)$ for good and poor mental health, respectively are smaller than those estimated on the larger bandwidth of €600 either side of the threshold as in Figure 2.11, $(0.000778, 0.00145)$ for good and poor mental health, respectively. If I alternatively use these larger estimates for the calibration, I find that $v'_L = 1.86$, $\kappa'_H = 28.6$ and $\kappa'_L = 46.9$. These estimates imply:

$$MVPF_{dB} = 0.47$$

$$MVPF_{d\Lambda} = 0.71$$

and increasing barriers is concluded as 48% more effective than increasing benefits, although both $MVPF$s are below 1. Since regression kink design estimates are intended to be local to the kink, the preferred estimate is the one using the smaller, robust bandwidth because the shape of the take-up function away from the threshold is affected by unobservables.

C.3 Relaxing Modelling Assumptions

Relaxing additivity involves assuming $SA = 1 \iff v_\theta(B, \varepsilon) > \kappa_\theta(\Lambda, \varepsilon)$. Given monotonicity in ε, behaviour will still follow a threshold-rule. Marginal entrants have ε^*_θ which satisfies the implicit equation $v_\theta(B, \varepsilon^*_\theta) = \kappa_\theta(\Lambda, \varepsilon^*_\theta)$. Without additivity, a bounding argument can be made about $\frac{dW}{d\Lambda}$ following Haller and Staubli [2024].

Relaxing independence of ε leads to the following. Suppose we used the model of Rafkin et al. [2023] where $v'_\theta(B)$ is independent of θ conditional on income, but $\varepsilon_\theta \sim F_\theta$. Then, same average take-up levels combined with the difference-in-differences results would suggest $f_{\varepsilon_L}(v_L - \kappa_L) = 1.65 \times f_{\varepsilon_H}(v_H - \kappa_H)$, inconsistent with the regression kink design results. Using Finkelstein and Notowidigdo [2019]'s model, κ'_θ are assumed to be opportunity costs of time, the only reason why need would vary across individuals with the same income is due to misperceptions of the benefit level and $\varepsilon_\theta \sim F_\theta$. My results would then suggest $\kappa'_L = w_L =$ €11.7 and $\kappa'_H = w_H =$ €13.7, where w_θ is θ's wage. Then, the regression kink design estimates would suggest $v'_L = 1.8 \times v'_H$. This would imply that people with poor mental health have an easier time overcoming barriers, and are substantially *less* pessimistic about the benefit level. Both of these results contradict psychological evidence [Martin et al., 2023b; Evans et al., 2014; Alloy and Ahrens, 1987].

Assumption (ii) follows in the case that types are one-dimensional [Landais et al., 2021]. However, note that to maximise internal validity of the quasi-experimental design, sample restrictions are made both in Section 2.3 and Section 2.4. In Section 2.3, I focus on couples, as for them, the Participation Act was a change in ordeals only, and not also a change in benefit level. Note that the majority of individuals in this sample have income much below the threshold. In Section 2.4, I focus on singles, as I mis-classify couples more than singles, and in the RKD analysis, measurement error is much more consequential, because I zoom into a small window around the threshold. Moreover, I restrict to employees as I observe monthly income for this group. The samples for the two instruments are quite different, as confirmed by Tables C.2 and C.3, and the within-sample compliers may be even more different across instruments (as in Landais et al. [2021]). This is an important caveat.

However, my framework is flexible enough to relax this assumption under structural assumptions. Note that **Step 2** of the identification method can be applied separately. Therefore, the result that people with poor mental health have a 2× higher need and 64% higher cost than those with good mental health (relatively speaking) still holds.

Below, I employ some additional structure in order to use the correlation test to identify net value − cost, which then allows for the quantification of all sufficient statistics without maintaining Assumption (ii). I find that in the structural model, the probability of being marginal to a barrier instrument less than $1/10^{\text{th}}$ to that of a benefits instrument - this only pushes the welfare comparison that I explore in the next section *more* to the side of reducing barriers.

Details: Assume linearity: $v_\theta(B) = v_\theta \cdot B$ and $\kappa_\theta(\Lambda) = \kappa_\theta \cdot \Lambda$. Note that one or other of these assumptions is assumed throughout in Anders and Rafkin [2022], Finkelstein and Notowidigdo [2019] and Rafkin et al. [2023]. In this case: $\hat{\mathbb{P}}[SA]_L \approx \hat{\mathbb{P}}[SA]_H \implies v_L \cdot B - \kappa_L \cdot \Lambda = v_H \cdot B - \kappa_H \cdot \Lambda$. This means that $\kappa_L - \kappa_H = \frac{v_L \cdot B - v_H \cdot B}{\Lambda}$. Recall that $\frac{\partial \hat{\mathbb{P}}[SA]_H}{\partial B} = 0.000227$ and $\frac{\partial \hat{\mathbb{P}}[SA]_L}{\partial B} = 2.2 \times 0.000227$. These estimates imply $\kappa_L - \kappa_H = \frac{(2.2-1) \times 874.54}{\Lambda}$.

Let $f_\varepsilon^{d\Lambda} = \alpha \cdot f_\varepsilon^{dB}$ - i.e. assume that the ratio of the probability of being marginal to a benefits-level instrument over probability of being marginal to an ordeal change is constant across mental health types. In this case, $\frac{\partial \hat{\mathbb{P}}[SA]_L - \hat{\mathbb{P}}[SA]_H}{\partial \Lambda} = -(\kappa_L - \kappa_H) \cdot \alpha \cdot 0.000503$, as $\frac{\partial \hat{\mathbb{P}}[SA]_H}{\partial B} = f_\varepsilon^{dB}$. Rearranging for α,

$$\hat{\alpha} = \frac{-\Lambda \frac{\partial \hat{\mathbb{P}}[SA]_L - \hat{\mathbb{P}}[SA]_H}{\partial \Lambda}}{0.000503 \times 1058.7} \tag{C.3.1}$$

Therefore as long as we can estimate the heterogeneous semi-elasticity $-\Lambda \frac{\partial \hat{\mathbb{P}}[SA]_L - \hat{\mathbb{P}}[SA]_H}{\partial \Lambda}$, we are done.

I use Table B.6 to calibrate the percent change in ordeals $\frac{\partial \Lambda}{\Lambda} = 22.1\%$ - which comes from treating the final column as percent changes in each of the scores (second column) where the score cannot exceed 100%. Therefore,

Socio-economic Demographics	Middle-Age Couples	Single Employees
Gender (%)		
Woman	51.9	57.4
Man	48.1	42.6
Education (%)		
Primary School	3.5	12.3
High School	31.6	46.9
Bachelor’s	3.5	5.0
Masters-PhD	1.9	1.9
Unknown	23.7	33.8
Main Source of Income (%)		
Employment or Civil Service Job	6.5	88.6
Director-shareholder	0.1	0.2
Self-employment	5.2	0.8
Other Job	0.1	0.0
Unemployment Insurance	2.4	2.5
Disability Insurance	8.9	1.7
Social Assistance	55.3	4.6
Other Benefits	9.9	0.4
Pension	2.2	1.0
Student Aid	0.2	0.2
Other (not active or without income)	9.2	0.0
Household Composition (%)		
Single Person Household	13.6	65.5
Couple Without Children	40.8	1.7
Couple With Children	42.1	3.8
Single Parent	13.6	26.6
Couples and Parents with Flatmates	2.1	1.3
Other Shared Households	1.4	1.1
Other Information		
Age	54.5 (5.9)	46.5 (7.8)
Foreign-born (%)	56.5 (49.6)	30.2 (45.9)
Household Std. Disposable Income (€)	5,731 (12,767)	21,739 (11,350)
Household Net Worth (€)	-5,309 (103,775)	-4,627 (51,309)
Contracted Hours (per year)	553 (498)	1,416 (552)
Eligible (%)	100.0 (0.0)	27.2 (44.5)
Receipt of Social Assistance (%)	60.4 (48.9)	9.4 (29.2)

Table C.2: Summary Statistics for Middle-Age Couples and Single Employees

$$\begin{aligned} -\Lambda \frac{\partial \hat{\mathbb{P}}[SA]_L - \hat{\mathbb{P}}[SA]_H}{\partial \Lambda \qquad 206} &= \frac{0.09}{0.221} = 0.040 \\ \implies \hat{\alpha} &= \frac{0.040}{0.53} = 0.075 \end{aligned}$$

(Mental) Health Information	Middle-Age Couples Mean (SD)	Single Employees Mean (SD)
General		
All Care Spending (€)	3,502 (9,850)	1,997 (6,818)
Physical Chronic Conditions (count)	1.61 (1.66)	0.73 (1.14)
Mental Health (admin)		
Mental Healthcare Spending (€)	381 (3,675)	14 (35)
Psychotropic Medication (%)	22.4 (41.7)	13.9 (34.6)
Anti-psychotics (%)	5.7 (23.1)	2.37 (15.2)
Anxiolytics (%)	7.1 (25.6)	2.57 (15.8)
Anti-depressants (%)	16.2 (36.8)	9.42 (29.2)
Hypnotics and Sedatives (%)	4.0 (19.5)	1.23 (11.0)
ADHD Medication (%)	0.5 (6.8)	0.64 (8.0)
Mental Health Hospitalizations (%)	0.08 (2.8)	0.06 (2.4)
Deaths by Suicide (%)	0.02 (1.4)	0.01 (1.1)
Mental Health (survey)		
Loneliness (0-11)	5.01 (3.77)	4.44 (3.75)
Life Control (7-35)	22.06 (5.73)	–
Kessler-10 Psychological Distress (10-50)	22.02 (9.81)	18.90 (8.11)

Table C.3: Summary Statistics for Middle-Age Couples and Single Employees (Mental Health)

In particular, $f^{d\Lambda} < f^{dB}$ - which only pushes in the direction of $MVPF_{d\Lambda} > MVPF_{dB}$.

C.4 Welfare Effects

Calculating social marginal utilities of beneficiaries of benefit and barrier change instruments:

$$\begin{aligned}
\bar{\eta}_{dB} &= 0.25 \times 2.07 + 0.73 \times 1 \\
&\approx 1.29 \\
\bar{\eta}_{d\Lambda} &= 0.25 \times 2.07 \times \frac{36.2/2.07}{0.25 \times 36.2/2.07 + 0.73 \times 20.2} + 0.73 \times 1 \times \frac{20.2}{0.25 \times 36.2/2.07 + 0.73 \times 20.2} \\
&\approx 1.26
\end{aligned}$$

C.5 Robustness to Bias

Suppose a share ψ of $\kappa_\theta(\Lambda)$ is a true cost, and $(1-\psi)$ is a hassle cost, which affects behaviour but not welfare. Then: $\mathbb{P}[SA]_\theta = F_\varepsilon\left[v_\theta(B) - \kappa_\theta(\Lambda)\right]$ still, but:

$$\mathcal{U}_\theta = \int_{-\infty}^{\varepsilon_\theta^*} \left[v_\theta(B) - \kappa_\theta(\Lambda) + MI_\theta - \varepsilon\right] dF(\varepsilon) \tag{C.5.1}$$

where $\varepsilon_\theta^* = v_\theta(B) - \kappa_\theta(\Lambda)$ and $MI_\theta = (1-\psi) \cdot \kappa_\theta(\Lambda)$ is the marginal internality [Mullainathan et al., 2012]. Note that since the true cost $\psi \cdot \kappa \leq \kappa$, behaviour over-states the ordeal-cost, so take-up is too low relative to the private optimum. This means that a marginal increase in Λ has an extra negative behavioural welfare cost coming from people moving further away from the private optimum. A marginal increase in B has an extra positive behavioural welfare gain coming from the internality correction. This is shown in Proposition C.5.1.

Proposition C.5.1. *First order welfare effects when perceived cost differs from true cost.*

$$\frac{d\mathcal{U}_\theta}{d\Lambda} = -\psi \cdot \kappa_\theta'(\Lambda) \cdot \mathbb{P}[SA]_\theta + MI_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial \Lambda} \tag{C.5.2}$$

$$\frac{d\mathcal{U}_\theta}{dB} = v_\theta'(B) \cdot \mathbb{P}[SA]_\theta + MI_\theta \cdot \frac{\partial \mathbb{P}[SA]_\theta}{\partial B} \tag{C.5.3}$$

Proof.

$$\mathcal{U}_\theta = \int_{-\infty}^{\varepsilon_\theta^*} \left[v_\theta(B) - \kappa_\theta(\Lambda) - \varepsilon\right] dF(\varepsilon) + \int_{-\infty}^{\varepsilon_\theta^*} MI_\theta \; dF(\varepsilon)$$

which means that, by the Leibniz integral rule:

$$\frac{d\mathcal{U}_\theta}{d\Lambda} = -\kappa_\theta'(\Lambda) \cdot F(\varepsilon_\theta^*) + 0 + (1-\psi)\kappa_\theta(\Lambda)\frac{\partial F(\varepsilon_\theta^*)}{\partial \Lambda} + (1-\psi)\kappa_\theta'(\Lambda) \cdot F(\varepsilon_\theta^*)$$

where the 0 comes from $\varepsilon_\theta^* = v_\theta(B) - \kappa_\theta(\Lambda)$ - this is the Envelope Theorem at play.

Rearranging gives Equation (C.5.2). Similarly,

$$\frac{d\mathcal{U}_\theta}{dB} = v_\theta'(B) \cdot F(\varepsilon_\theta^*) + 0 + (1-\psi)\kappa_\theta(\Lambda)\frac{\partial F(\varepsilon_\theta^*)}{\partial B}$$

and there is no final term because MI_θ is independent of B. □

These first order effects imply new MVPF formulas for the welfare effect of changing benefits and barriers. The fiscal externalities are unchanged - since they depend on behaviour only. However, the direct welfare effects reflect Equations (C.5.2) and (C.5.3).

Corollary C.5.1.1. *With bias:*

$$MVPF_{d\Lambda} = \frac{-\psi \cdot \int \lambda \cdot \frac{\kappa'(\Lambda)}{v'(B)} \mathbb{P}[SA] d\mu - (1-\psi) \cdot \int \lambda \frac{\kappa(\Lambda)}{v'(B)} \frac{\partial \mathbb{P}[SA]}{\partial \Lambda} d\mu}{\int FE \cdot \frac{\partial \mathbb{P}[SA]}{\partial \Lambda} d\mu} \tag{C.5.4}$$

$$MVPF_{dB} = \frac{\int \lambda \mathbb{P}[SA] d\mu + (1-\psi) \cdot \int \lambda \frac{\kappa(\Lambda)}{v'(B)} \frac{\partial \mathbb{P}[SA]}{\partial B} d\mu}{\int FE \cdot \frac{\partial \mathbb{P}[SA]}{\partial B} d\mu} \tag{C.5.5}$$

C.5.1 Calibration

How does bias affect the quantification of welfare effects? This requires us to evaluate the size of MI_θ, the marginal internality for each type. According to the theory,

$$MI_\theta = (1-\psi) \cdot \kappa_\theta(\Lambda) \tag{C.5.6}$$

Note that the marginal internality depends on *average* ordeal-costs, rather than marginal ordeal-costs. In order to evaluate this term, I make the linearization $\kappa_\theta(\Lambda) = \kappa_\theta \cdot \Lambda$. Therefore, evaluating the new $MVPF$ formulas requires taking a stance on what Λ is. As discussed in Appendix C.3, qualitative evidence from municipalities suggests the percent change in Λ due to to the Participation Act is an increase of 22.1%. Further, I assume that the Participation Act represented an absolute change in Λ of 1 unit. Therefore, $\Lambda = 1/0.221 = 4.52$. For example, Λ could represent number of hours spent on obligations, and κ_θ is the welfare cost per hour spent. When $\kappa_\theta(\Lambda) = \kappa_\theta \cdot \Lambda$, $\kappa_\theta = \kappa'_\theta(\Lambda)$.

Therefore, given the estimates from Chapter 3:

$$MI_L = (1 - \psi) \cdot 4.52 \cdot 161 = (1 - \psi) \cdot 728$$

$$MI_H = (1 - \psi) \cdot 4.52 \cdot 98 = (1 - \psi) \cdot 442$$

These estimates mean that we can quantify how large the $MVPF$ formulas are for different values of ψ. For $\psi = 1$ - the $MVPF$ are as Chapter 3. What if ordeal-costs were a pure bias which affects behaviour only but not welfare? Then:

$$MVPF_{d\Lambda}^{\psi=0} = 0.30$$

$$MVPF_{dB}^{\psi=0} = 0.96$$

$MVPF_{d\Lambda}^{\psi=0} < MVPF_{d\Lambda}^{\psi=1}$ as there is no direct welfare effect of the increase in barriers. $MVPF_{d\Lambda}^{\psi=0} \neq 0$, however, because of the negative behavioural welfare effect. $MVPF_{dB}^{\psi=0} > MVPF_{dB}^{\psi=1}$ because of the internality correction that an increase in benefits provides. Finally, we can quantify the level of bias ψ^* required to reverse the welfare ordering $MVPF_{d\Lambda} > MVPF_{dB}$. This turns out to be $\psi^* \approx 35\%$. That is to say, the government needs to be confident that at least 56% of the as-if ordeal-costs are purely a bias in order to reverse the welfare conclusions. Alternatively, as long as people don't over-estimate the size of the cost by a factor of 3, then the welfare conclusions remain robust. Finally, note that $d\Lambda$ is unsurprisingly more sensitive to bias than dB.

Appendix D

Appendix to Chapter 4

D.1 Theory Appendix

In this section we present some assumptions, results and proofs from our general model.

D.1.1 Setup

Proof of Observation 1. If $(x, \theta) \succeq_* (x', \theta)$, then by Assumption 1.3, we have $x \succeq_{\theta^*} x'$. Applying Assumption 1.3 again, this implies $(x, \theta') \succeq_* (x', \theta')$. □

Proof of Observation 2. By Assumption 1.3, a normative frame $\theta^* \in \Theta$ exists such that $x \succeq_{\theta^*} x'$. Since $x \succeq_\theta x'$ for all $\theta \in \Theta$, it follows that $x \succeq_{\theta^*} x'$, and thus Assumption 1.3 implies $x \succeq_* x'$. □

D.1.2 Unknown Normative Frame

Proof of Proposition 4.2.1. This argument is due to Kaplow and Shavell [2001]. We observe that $\succ_w$ does not have the proposed representation if and only if there are two options x, x' such that for every θ, $x \sim_\theta x'$ but $w(x) \neq w(x')$. Toward a contradiction, suppose we find two such x, x'; without loss of generality $x \succ_w x'$.

Starting from x', construct x'' by increasing the good x_n from Assumption 3 by a small amount $\delta > 0$. By continuity (5.2), if δ is sufficiently small we must have $x \succ_w x''$. But for every θ, $x'' \succ_\theta x' \sim_\theta x$, so BR-dominance require $x'' \succsim_w x$. This establishes sufficiency of our assumptions for representation (4.2.4); necessity is easily verified. □

Proof of Proposition 4.2.2. We use the utility function whose existence is assumed in Assumption 6 to construct an outcome space comprising the output of the utility function. The fact that we obtain a subjective expected utility representation of the planner's preferences then follows from Theorem 5 of Köbberling and Wakker [2003]: their solveability pre-condition is implied by continuity (Assumption 2 and Assumption 5.2) and the richness of our option set (Assumption 4). For the set of axioms that is equivalent to the existence of a subjective expected utility representation with uniqueness up to affine transformation with at least two non-null frames/states in their theorem, Assumption 5.1 and Assumption 5.2 ensure the *weak ordering* condition, *monotonicity* is ensured by weak BR-dominance (Assumption 5.3), *tradeoff consistency* is ensured by Assumption 6. Their Archimedean axiom is satisfied when there is more than one frame by Assumption 4; with more technical work we conjecture that one could weaken 4 to something approaching their Archimedean axiom. Proposition 4.2.1 requires that the utility function in our representation of the planner's preferences is a representation of $\succeq_\theta$. □

D.1.3 Expected Utility Representation of Planner's Preferences Using Von Neumann & Morgenstern's Theory

Re-defining the Planner's Objective. Adapting classical Expected Utility Theory to this setting requires us to conceive of counterfactuals that describe situations in which the planner attaches different weights to each frame. We introduce the notion of an intrapersonal lottery to capture this. The primitive components of such a lottery are an option $x \in \mathcal{X}$, the state space Θ, and a distribution $\psi \in \Delta(\Theta)$. The outcomes of a lottery entail consuming a particular x in a particular state θ. We conceive of a lottery $L(x, \psi)$ in terms of a vector of weights/probabilities $\big(\psi(\theta_1), ..., \psi(\theta_{|\Theta|})\big)$ and a vector of outcomes $\big((x, \theta_1), ..., (x, \theta_{|\Theta|})\big)$. Compound lotteries entail mixtures of weights: for $p \in [0, 1]$ and two distributions ψ_1, ψ_2 we describe these using the notation

$$pL_1(x) + (1 - p)L_2(x) = L(x, p\psi_1 + (1 - p)\psi_2),$$

where $L_n(x) = L(x, \psi_n)$.

We abuse notation slightly by denoting the planner's preferences over lotteries by $\succeq_w$. Now, the planner's preferences are defined not only over what option the individual consumes but also over the planner's normative beliefs: $\succsim_w$ is a binary relation on the set of lotteries $\mathcal{L}$. We strengthen Assumption 5 as follows:

Assumption 11. ***Expected Utility Assumptions Over Intrapersonal Lotteries.***

Assumption 11.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 11.2. ***Continuity.*** *For any $L \in \mathcal{L}$, the sets $\{L' \in \mathcal{L} : L' \succsim_w L\}$ and $\{L' \in \mathcal{L} : L' \precsim_w L\}$ are closed.*

Assumption 11.3. ***Strong BR-Dominance.*** *For any ψ, x, if $x \succsim_\theta x'$ for every θ, then $L(x, \psi) \succsim_w L(x', \psi)$. If, additionally, there exists θ such that $x \succ_\theta x'$ and $\psi(\theta) > 0$, then $L(x, \psi) \succ_w L(x', \psi)$.*

Assumption 11.4. ***Independence.*** *For any x, any $L_1(x), L_2(x), L_3(x) \in \mathcal{L}$, and any $p \in [0, 1]$*

$$L_1(x) \succsim_w L_2(x) \implies pL_1(x) + (1-p)L_3(x) \succsim_w pL_2(x) + (1-p)L_3(x). \quad \text{(D.1.1)}$$

Proposition D.1.1. *Maintain Assumptions 1, 2 and 3. Then Assumption 11 holds if and only if there is a function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that $u(x, \theta)$ represents individual preferences $\succeq_\theta$ for every θ, and the planner's preferences $\succeq_w$ are represented by*

$$w(x; \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) u(x, \theta^*). \quad \text{(D.1.2)}$$

Moreover, u is continuous and unique up to positive affine transformation.

Proof. Assumptions 11.1, 11.2, and 11.4 are the axioms of classical expected utility over the outcomes (x, θ). We therefore obtain a payoff function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that the planner's preferences take the expected utility form $w(x, \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) u(x, \theta)$. That $u(x, \theta)$ must be a representation of $\succeq_\theta$ follows from 11.3 by the same logic as Proposition 4.2.1; the strong form of BR-dominance is required to rule out the degenerate case where u is constant over x. This establishes sufficiency of Assumption 11 for the desired representation of $\succeq_w$; necessity is easily verified. □

Assumptions 11.1, 11.2, and 11.4 are the axioms of classical expected utility over the outcomes (x, θ). We therefore obtain a payoff function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ such that the planner's preferences take the expected utility form $w(x, \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) u(x, \theta)$.

That $u(x, \theta)$ must be a representation of $\succeq_\theta$ follows from 11.3 by the same logic as Proposition 4.2.1; the strong form of BR-dominance is required to rule out the degenerate case where u is constant over x. This establishes sufficiency of Assumption 11 for the desired representation of $\succeq_w$; necessity is easily verified.

Discussion of Proposition D.1.1 The proof of this proposition is a straightforward adaptation of the Expected Utility Theorem of Von Neumann and Morgenstern [1953]. The way that the planner trades off risk according to the independence assumption 11.4 implies a cardinalization of utility, which is the function $u(x, \theta)$ in (D.1.2).

D.1.4 Paternalistic Risk Aversion

Although we find that some fully comparable utility function that is suitable for evaluation of intrapersonal tradeoffs must exist under the assumptions of Proposition 4.2.2 or Proposition D.1.1, these results do not not shed light on which particular representation of frame-dependent preferences we ought to suppose is fully comparable across frames in any given setting or model. The following Corollary to these propositions, in which we consider a specific welfare metric $v(x, \theta)$ that represents ordinal preferences, proves useful for thinking about comparability and the planner's risk preferences. In the main text, the only candidate for the function v we considered was money-metric equivalent variation and we discussed paternalistic risk aversion in terms of risk preferences over monetary payoffs. Here, we take a slightly more general approach.

Definition. We say that two utility functions $u(x, \theta)$ and $v(x, \theta)$ exhibit *ordinal level comparability* if for any (x, θ) and (x', θ'),

$$u(x, \theta) \geq u(x', \theta') \iff v(x, \theta) \geq v(x', \theta').$$

Corollary D.1.1.1. *Consider a utility function $u(x, \theta)$ that gives the representation in Proposition 4.2.2 or Proposition D.1.1. Under the assumptions of either proposition, for any function $v(x, \theta)$ that exhibits ordinal level comparability with $u(x, \theta)$, there is a transformation $\omega : \mathbb{R} \to \mathbb{R}$ such that*

$$w(x; \psi) = \sum_{\theta^* \in \Theta} \psi(\theta^*) \omega(v(x, \theta)). \tag{D.1.3}$$

Moreover, ω is strictly increasing, continuous, and unique up to positive affine transformation.

Proof. This result obviously follows from Proposition D.1.1.1 and the definition of ordinal level comparability. □

The function ω converts the welfare metric $v(x, \theta)$ into the cardinal units the planner uses to conduct welfare comparisons across frames. In main text Section 4.3.2, we introduce conditions under which money-metric equivalent variation provides a representation of individual preferences that exhibits ordinal level comparability with cardinal utility, in which case ω should account for any non-linearity of the individual's preference for money. In the main text, we denoted the transformation ω for monetary payoffs by u_ζ.

The Value of Robustness with Probabilistic Uncertainty To understand the value of the robustness of welfare across frames in this case, it is instructive to impose some smoothness on the transformation ω from Corollary D.1.1.1.

Corollary D.1.1.2. ***Variance Representation*** *Assume the function ω from representation (D.1.3) is twice differentiable. Then up to second-order Taylor approximation of ω, the planner's objective is*

$$w(x,\psi) \approx \omega\Big(E_\psi[v(x,\theta)]\Big) + \frac{\omega''\Big(E_\psi[v(x,\theta)]\Big)}{2} \cdot Var_\psi\Big[v(x,\theta)\Big] \tag{D.1.4}$$

where $E_\psi[v(x,\theta)] = \sum_{\theta\in\Theta}\psi(\theta)v(x,\theta)$ and $Var_\psi\Big[v(x,\theta)\Big] = \sum_{\theta\in\Theta}\psi(\theta)\Big[v(x,\theta) - E_\psi[v(x,\theta)]\Big]^2$.

Proof. For ease of notation shorten $v = v(x, \theta)$, a random variable with respect to θ for given x, and $\overline{v} = E_\psi[v(x, \theta)]$, a deterministic number for given x, ψ. Using a Taylor Expansion of ω around $\overline{v}$ we find

$$\omega(v) \approx \omega(\overline{v}) + \omega'(\overline{v})\cdot(v-\overline{v}) + \omega''(\overline{v})\cdot(v-\overline{v})^2$$
$$\implies \mathbb{E}_\psi\big[\omega(v)\big] \approx \underbrace{\omega(\overline{v})}_{\text{Fixed Number}} + \omega'(\overline{v})\cdot\underbrace{\mathbb{E}_\psi[v-\overline{v}]}_{=0} + \frac{\omega''(\overline{v})}{2}\cdot\mathbb{E}_\psi[v-\overline{v})]^2$$

The result follows, as $w(x,\psi) = \mathbb{E}_\psi\big[\omega(v(x,\theta))\big]$. □

We say that the planner's preferences exhibit *paternalistic risk aversion over v* if $\omega'' < 0$ and *paternalistic risk neutrality over v* if $\omega'' = 0$ (this can be converted to statements about primitive preferences over normative lotteries using standard expected utility theory). Corollary D.1.1.2 implies that under paternalistic risk aversion over a welfare metric v with probabilistic uncertainty, the planner values robustenss. Under paternalistic risk neutrality over v, the planner's objective coincides with expected welfare according to v, i.e. $E_\psi[v(x,\theta)]$. But under paternalistic risk aversion over v, the variance of the welfare metric v across normative frames begins to matter (up to second-order approximation), and in particular welfare is decreasing in this variance. When, according to the welfare metric v, there is more disagreement in revealed preferences across frames about welfare under some policy P_0 compared to an alternative P_1, and mean welfare is similar between the two, paternalistic risk aversion over v suggests that P_0 is less desirable than P_1. Unlike the notion of robustness we present in the main text for ambiguity aversion, whether this notion of robustness is relevant is specific to the welfare metric we have in mind. Proposition D.1.1 tells us that under our assumptions, there will always be some measure of welfare $u(x,\theta)$ over which the planner's preferences exhibit paternalistic risk neutrality.

If ω is a homogeneous transformation (i.e. the planner's preferences exhibit *scale invariance* over the welfare metric v), we find a familiar functional form for ω. For a parameter $\eta \in \mathbb{R}$, we have

$$\omega(v) = \begin{cases} \frac{v^{1-\eta}}{1-\eta}, & \eta \neq 1 \\ \log(v) & \eta = 1. \end{cases} \tag{D.1.5}$$

Paternalistic risk aversion over v further implies $\eta > 0$, and η is of course the Arrow-Pratt coefficient of relative (paternalistic) risk aversion.

D.1.5 Ambiguity

We say that a lottery $L(x,\psi)$ is *constant over u* if for the given x, $u(x,\theta) = u(x,\theta')$ for any θ,θ', i.e. if it generates a constant payoff for every normative frame. Abandoning Assumption 11.4, we introduce conditions on the planner's preferences drawn from Gilboa and Schmeidler [1989].

Assumption 12. ***Ambiguity Aversion Assumptions.***

Assumption 12.1. ***Rationality.*** *$\succsim_w$ is complete and transitive on $\mathcal{L}$.*

Assumption 12.2. ***Continuity.*** *For any $L \in \mathcal{L}$, the sets $\{L' \in \mathcal{L} : L' \succsim_w L\}$ and $\{L' \in \mathcal{L} : L' \precsim_w L\}$ are closed.*

Assumption 12.3. ***Certainty Independence.*** *There is a representation $u(x,\theta)$ such that for any x, any pair $L_1(x), L_2(x) \in \mathcal{L}$, any lottery $L_3^c(x)$ that is constant over u, and any $p \in (0,1)$,*

$$L_1(x) \succsim_w L_2(x) \implies pL_1(x) + (1-p)L_3^c(x) \succsim_w pL_2(x) + (1-p)L_3^c(x).$$

Assumption 12.4. ***Weak BR-dominance.*** *For any $x, x' \in \mathcal{X}$ and any $\psi \in \Delta(\Theta)$, if $x \succsim_\theta x'$ for every θ, then $L(x,\psi) \succsim_w L(x',\psi)$.*

Assumption 12.5. ***Uncertainty Aversion.*** *For any x, any pair $L_1(x), L_2(x)$, and $p \in (0,1)$,*

$$L_1(x) \sim_w L_2(x) \implies pL_1(x) + (1-p)L_2(x) \succsim_w L_1(x).$$

Assumption 12.6. ***Non-degeneracy.*** *There exists $L, L' \in \mathcal{L}$ such that $L \succ_w L'$.*

Comments: Relative to Assumption 11, Assumption 12.3 weakens Assumption 11.4 so that it only holds when we mix a given pair of lotteries with a constant lottery. Assumption 12.6 rules out the degenerate case where the planner is indifferent across all policies/options.

Proposition D.1.2. ***MaxMin Welfare Under Ambiguity Aversion.*** *Maintain Assumptions 1, 2 and 3. Assumption 12 holds if and only if there exist a function $u : \mathcal{X} \times \Theta \to \mathbb{R}$ and a set $\Psi^* \subseteq \Delta(\Theta)$ such that $u(x,\theta)$ represents $\succsim_\theta$ for every θ, Ψ^* is closed and convex, and the planner's preferences $\succsim_w$ are represented by*

$$w(x) = \min_{\psi \in \Psi^*} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x,\theta^*)) \right\}. \tag{D.1.6}$$

Proof. Take the representation of individual preferences whose existence is implied by 12.3 and denote this $\tilde{u}$. Observe that BR-dominance implies Gilboa and Schmiedler's weak monotonicity condition over realizations of $\tilde{u}(x,\theta)$ for this representation. Theorem 1 of Gilboa and Schmeidler [1989] then implies there is a strictly increasing transformation $\omega(\tilde{u})$ such that the planner's preferences are represented by $w(x) = \min_{\psi \in \Psi^*} \{\sum_{\theta^*} \psi(\theta^*)\omega(\tilde{u}(x,\theta^*))\}$. The result follows, as $u \equiv \omega(\tilde{u})$ is also a representation of individual preferences by construction. □

Forms of Ambiguity and Related Alternative Axiomatizations

Gilboa and Schmeidler [1989] defer the structure of the set Ψ^* to applications, apart from the requirement that Ψ^* is closed and convex [for a useful discussion, see

Hansen and Sargent, 2001]. Following our discussion above about interpretations of the set of frames and the weights themselves, we envision two potential approaches. The first is more global: we could define a subset of the set of frames $\Theta^* \subseteq \Theta$, and let $\Psi^* = \Delta(\Theta^*)$. This approach is very similar in spirit to the concept of a "welfare-relevant domain" in Bernheim and Rangel [2009], and seems suitable when the planner has no philosophically acceptable way of specifying a unique set of welfare weights. The second approach is more local and drawn from the literature on robust control [e.g. Hansen and Sargent, 2008]: the planner begins with a specific distribution ψ that represents their best guess about the correct normative weights, and accounts for ambiguity in a neighborhood of this distribution. In this case, Ψ^* could be a ball of distributions around the best guess ψ whose radius is determined by a tolerance parameter $\kappa \geq 0$: $\Psi^* = B(\psi, \kappa) \equiv \{\psi' \in \Delta(\Theta) \text{ s.t. } ||\psi' - \psi|| \leq \kappa\}$. This approach seems more applicable in the case where the planner uses a statistical model like the "counterfactual normative consumer" approach discussed below to identify welfare weights, but nevertheless confronts ambiguity because the underlying model may be misspecified. We return to this last idea in Section 4.6 and Appendices D.5 and D.6.

Global MaxMin Criteria. In the case where normative ambiguity is more globally conceived of over the set $\Psi^* = \Delta(\Theta^*)$ for a subset of "welfare-relevant" frames Θ^*, the max-min expected welfare criterion from (4.2.6) becomes a more global max-min criterion:

$$\min_{\psi \in \Delta(\Theta^*)} \left\{ \sum_{\theta^*} \psi(\theta^*) u(x, \theta^*)) \right\} = \min_{\theta \in \Theta^*} u(x, \theta) \tag{D.1.7}$$

Building on social welfare theory, rather than using Assumption 12, one could derive this criterion this using an analogue of Rawls' [1971] Difference Principle: assume there is some representation of $\succeq_\theta$, u, such that $x \succeq_w x'$ if and only if the individual prefers x to x' in the frame $\theta \in \Theta^*$ in which they are the least well-off according to u. This obviously implies a criterion like equation (D.1.7).[1] Intuitively, with this more global type of ambiguity, endowing the planner with a direct preference for equity across frames (without any notion of intrapersonal lotteries) leads to the same place as giving the planner a preference to hedge in Assumption 12.5 together with a global notion of ambiguity.[2]

All of the objectives discussed above intersect at a global robustness criterion, which obtains under extreme ambiguity, or extreme paternalistic risk aversion with prob-

[1] Respecting strict BR-dominance requires a slight modification of the Difference Principle – the "Equity" axiom from Sen [1970] and Hammond [1976] – to break ties under indifference in the least well-off frame. Then we would obtain a lexicographic max-min criterion.

[2] For a deeper discussion of the theory of ambiguity aversion due to Gilboa and Schmeidler [1989] and Rawlsian social welfare, see Mongin and Pivato [2021].

abilistic uncertainty. Formally, we define the *global max-min* criterion as the one implied by equation (4.2.6) for $\Psi^* = \Delta(\Theta)$. The global max-min criterion is the closest analogue to Rawlsian social welfare in our framework.

Corollary D.1.2.1. ***Intersection of Various Objectives at Global Max-Min***

- *If $\Psi^* = B(\kappa, \psi)$, for any ψ, the planner's objective in (4.2.6) coincides with the global max-min criterion for $\kappa > 1$.*
- *If $\Psi^* = \Delta(\Theta^*)$, the planner's objective in (4.2.6) and/or (D.1.7) coincides with the global max-min criterion for $\Theta^* = \Theta$.*
- *Given a welfare metric v under scale invariance over v for the parameter η and probabilistic uncertainty with $\psi(\theta) > 0$ for very $\theta \in \Theta$, the planner's objective – Equation (D.1.3) with the functional form in equation (D.1.5) – approaches the global max-min criterion as $\eta \to \infty$.*[3]

Proof. The first two claims are obvious from equation (4.2.6) and (D.1.7). The last has a well-known analogue in the nesting of Rawlsian welfare functions in the family of generalized utilitarian welfare functions taking the form in equation (D.1.5). □

Partial Characterization of Robust Optimality. Our next result provides a sufficient condition for a policy that is a ψ-optimum for $\psi \in \Psi^*$ to also be a robust optimum. Note that this is not a full characterization as we do not obtain necessity; the condition nevertheless builds intuition and proves useful in applications below. To state the condition we use the cardinal *disagreement* in welfare between some frame θ and the decision-making frame θ^D:

$$V(x, \theta, \theta^D) = u(x, \theta^D) - u(x, \theta).$$

Note that in the main text, when there is only one frame besides θ^D, we suppress the second and third inputs as these are always equal to θ^A and θ^D respectively; below, only θ^D is fixed and suppressed.

Proposition D.1.3. ***Sufficient Condition for a ψ-Optimum to be a Robust Optimum.*** *Let $P^* \in \mathcal{P}$ be a $\psi-$optimum for some $\psi \in \Delta(\theta)$. Then, for any $\Psi^* \subseteq \Delta(\Theta)$ such that $\psi \in \Psi^*$, P^* is a robust optimum if*

$$P^* \in \arg\min_{P \in \mathcal{P}} \max_{\psi' \in \Psi^*} \sum_{\theta \in \Theta} \big(\psi'(\theta) - \psi(\theta)\big) \cdot V\big(x(P, Z, \theta^D), \theta, \theta^D\big). \tag{D.1.8}$$

[3] The interpersonal analogue of this is a well-known result about Rawlsian social welfare functions; see also Lockwood et al. [2021].

Proof. By supposition,

$$\begin{aligned}
P^* &\in \arg\min_{P\in\mathcal{P}} \max_{\psi'\in\Psi^*} \sum_{\theta\in\Theta} \big(\psi'(\theta)-\psi(\theta)\big)\cdot V\big(x(P,Z,\theta^D),\theta\big) \\
&= \arg\min_{P\in\mathcal{P}} \left\{ \sum_{\theta\in\Theta} \psi(\theta)\cdot V\big(x(\theta^D,P),\theta,P\big) - \min_{\psi'\in\Psi^*} \sum_{\theta\in\Theta} \psi'(\theta)\cdot V\big(x(\theta^D,P),\theta,P\big) \right\} \\
&= \arg\min_{P\in\mathcal{P}} \Bigg\{ u(x(P,Z,\theta^D)),\theta^D) - \sum_{\theta\in\Theta} \psi(\theta)\cdot V\big(x(P,Z,\theta^D),\theta\big) \\
&\qquad - \min_{\psi'\in\Psi^*} \Big[u(x(P,Z,\theta^D),\theta^D) - \sum_{\theta\in\Theta} \psi'(\theta)\cdot V\big(x(P,Z,\theta^D),\theta\big) \Big] \Bigg\} \\
&= \arg\min_{P\in\mathcal{P}} \left\{ W(P,Z,\theta^D;\psi) - \min_{\psi'\in\Psi^*} W(P,Z,\theta^D;\psi') \right\} \\
&\iff \forall P\in\mathcal{P}, W(P^*,Z,\theta^D;\psi) - \min_{\psi'\in\Psi^*} W(P^*,Z,\theta^D;\psi') \leq W(P,Z,\theta^D;\psi) - \min_{\psi'\in\Psi^*} W(P,Z,\theta^D;\psi')
\end{aligned}$$

However, $W(P^*,Z,\theta^D;\psi) \geq W(P,Z,\theta^D;\psi)$ as P^* is a $\psi-$optimum. We therefore obtain

$$\min_{\psi'\in\Psi^*} W(P^*,Z,\theta^D;\psi') - \min_{\psi'\in\Psi^*} W(P,Z,\theta^D;\psi') \geq W(P^*,Z,\theta^D;\psi) - W(P,Z,\theta^D;\psi) \geq 0 \tag{D.1.9}$$

So P^* is a robust optimum.

Note that above we suppressed the dependence between θ^D and P in writing out the steps of the proof above. But on inspection, we can see that each step of the proof obtains when θ^D depends non-trivially on P. □

The condition from Proposition D.1.3 for a given ψ-optimum to be robust is more likely to be met when disagreements about welfare evaluated at that policy are not too large and the set Ψ^* over which the planner evaluates robustness is a relatively close neighborhood around the relevant distribution ψ. We note that θ^D can depend arbitrarily on P in Equation (D.1.8).

D.2 Comparability

Proof of Lemma 4.3.1. Suppose first that $x \succ_{\theta^*} x(P_0, Z_0, \theta_0^D)$, i.e. $u(x, \theta^*) > u(x(P_0, Z_0, \theta_0^D), \theta^*)$. Assumption 7.1 ensures that $u(x(P_0, Z_0 + \zeta, \theta_0^D), \theta^*)$ is a strictly increasing function in ζ; Assumption 7.2 ensures this function is continuous. Assumption 7.3 implies that there is some $\hat{\zeta}$ such that $u(x(P_0, Z_0 + \hat{\zeta}, \theta_0^D), \theta^*) > u(x, \theta^*)$. The result follows from the Intermediate Value Theorem – note that $u(x, \theta^*)$ is continuous by Assumption 2. The same logic applies where $x \prec_{\theta^*} x(P_0, Z_0, \theta_0^D)$, and in the case of indifference, $\zeta = 0$.

Having established existence and uniqueness, that $\zeta(x, \theta^*)$ represents $u(x, \theta^*)$ follows from

$$\begin{aligned} x \succeq_{\theta^*} x' &\iff x(P_0, Z_0 + \zeta(x;\theta^*), \theta_0^D) \succeq_{\theta^*} x(P_0, Z_0 + \zeta(x';\theta^*), \theta_0^D) \text{ by definition \& transitivity} \\ &\iff \zeta(x;\theta^*) \geq \zeta(x';\theta^*) \text{ by Assumption \ Assumption 7.1.} \end{aligned}$$

□

Proof of Proposition 4.3.1. The result follows the exact same logic as the proof of Proposition 4.2.1, but we use small amounts of Z to construct BR-dominant options rather than small amounts of the good described by Assumption 3 (which is no longer required). □

Proof of Lemma 4.3.2. First we prove that level comparability implies Assumption 8. Assuming ordinal level comparability, Assumption 8.1 follows from the observation that $\zeta(x_0, \theta) = \zeta(x_0, \theta') = 0$ by construction, so then ordinal level comparability implies $u(x_0, \theta) = u(x_0, \theta')$. The second condition then follows from Assumption 7.1 (strict monotonicity over money).

Second we prove that Assumption 8 implies ordinal level comparability. Take any x, x', θ, θ'. Using Assumption 8.1 we have

$$u(x, \theta) \geq u(x', \theta') \iff u(x, \theta) - u(x_0, \theta) \geq u(x', \theta') - u(x_0, \theta').$$

Using the definition of equivalent variation, suppressing the baseline input, we have

$$\begin{aligned} &u(x, \theta) - u(x_0, \theta) \geq u(x', \theta') - u(x_0, \theta') \\ \iff &u(x(P_0, Z_0 + \zeta(x, \theta)), \theta) - u(x_0, \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x_0, \theta'). \end{aligned}$$

Now using Assumption 8.2, we find

$$u(x(P_0, Z_0 + \zeta(x,\theta)), \theta) - u(x_0, \theta) \geq u(x(P_0, Z_0 + \zeta(x', \theta')), \theta') - u(x_0, \theta')$$

$$\iff \zeta(x,\theta) \geq \zeta(x,\theta').$$

□

Proof of Proposition 4.3.2. Lemma 4.3.1 implies the equivalent variation exists, is unique and represents $\succeq_\theta$ for some baseline. Lemma 4.3.2 implies that this representation satisfies ordinal level comparability with the planner's cardinal utility function. Then applying Proposition 4.3.2 gives the result. □

D.3 Perturbations

In this section, we explore a perturbation approach to evaluating policy reforms in our framework. We consider one-dimensional policy variation, supposing $\mathcal{P} = \mathbb{R}$ for simplicity. We assume all the derivatives necessary to apply the perturbation approach exist. Expressing the planner's welfare under BIC as $W(P, Z, \theta^D) = w(x(P, Z, \theta^D))$ we are interested in $\frac{\partial W}{\partial P}$, or equivalently the change in welfare dW that results from a marginal policy perturbation dP.

In order to illustrate how the envelope theorem plays out in many applications, we modify our setup slightly. We suppose that in reduced-form we can conceive of every option $\tilde{x} \in \mathcal{X}$ as a component the individual can choose and a fixed feature like a default: $\tilde{x} = (x, P)$. Here, we assume there is a bijection such that every value of the fixed feature corresponds to a value of the policy P, so we might as well denote the fixed feature by P.

As the set of frames is finite, we suppose the frame θ^D is unaffected by policy; this can be relaxed with the continuous-frames extension developed in Appendix D.5. As θ^D is fixed throughout this section, we express disagreements as $V(x, \theta) = u(x, P, \theta^D) - u(x, P, \theta)$.

We present three characterizations, one leveraging Proposition 4.2.2, where we think of u as a utility function that is fully comparable (Paternalistic Risk Neutrality), one based on Proposition 4.3.2/Equation (4.3.4), where we think of u as a money-metric utility function that is ordinally level comparable to cardinal utility with diminishing utility over money $u''_\zeta < 0$ (Paternalistic Risk Aversion), and one rooted in Proposition D.1.2/Equation (4.2.6) (Ambiguity Aversion) given a fully comparable utility function.

D.3.1 Under Risk Neutrality

We begin with the planner's objective under probabilistic uncertainty about the normative frame and risk-neutrality. Applying the envelope theorem of Milgrom and Segal [2002] under θ^D, we find[4]

$$\frac{\partial W(P,Z,\theta^D)}{\partial P} = \underbrace{E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right]}_{\text{Direct Effect}} - \underbrace{\frac{\partial x(P,Z,\theta^D)}{\partial P}}_{\text{Beh. Resp.}} \cdot \underbrace{(1-\psi(\theta^D))E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial x}\middle| \theta \neq \theta^D\right]}_{\text{Marginal Internality}} \tag{D.3.1}$$

The term $\frac{\partial u(x,P,\theta)}{\partial P}$ in equation (D.3.1) is the partial derivative of $u(x,P,\theta)$ with respect to its second argument – the *direct effect* of varying P. All terms are evaluated at the status quo (P,Z,θ^D), where $x = x(P,Z,\theta^D)$.

In a model in which normative preferences are known – i.e. a model with a singular alternative frame θ^A, as in Example 1, and weight $\psi(\theta^A) = 1$ – this derivation matches the reduced-form characterization of welfare in Mullainathan et al. [2012].[5] Here, we extend this characterization to accommodate an unkown normative frame. Under risk-neutrality, we find similar set of terms as Mullainathan et al. [2012], but we replace the direct effect and marginal internality under a known normative frame with their expected values under an uncertain normative frame. This focuses applied analysis of normative ambiguity on specific questions: how do frames differ in their implied direct effects and marginal internalities?

How do disagreements about welfare shape the effects in (D.3.1)? We can see the answer for the internality term in equation (D.3.1). For the direct effect term, we observe that

$$E_\psi\left[\frac{\partial u(x,P,\theta)}{\partial P}\right] = \frac{\partial u(x,P,\theta^D)}{\partial P} - [1-\psi(\theta^D)]E_\psi\left[\frac{\partial V(x,P,\theta)}{\partial P}\middle| \theta \neq \theta^D\right]. \tag{D.3.2}$$

D.3.2 Under Paternalistic Risk Aversion

How do disagreements in money-metric welfare matter for policy evaluation? Assuming ordinal level comparability of equivalent variation $\zeta(x,P,\theta)$ (suppressing the baseline parameters) and diminishing marginal utility of money $u''_\zeta < 0$, we find

[4] Note that P is one-dimensional by assumption here. When x is multidimensional, the second term of this expression should be regarded as a dot product of the vectors $\frac{\partial x}{\partial P}$ and $\frac{\partial V}{\partial x}$.

[5] With a known normative frame it is also not necessary to account for potential differences in the value of a dollar across frames in characterizing when a local perturbation improves welfare, so we could freely use any equivalent variation representation of preferences in the application of the perturbation approach [see also Allcott and Taubinsky, 2015].

an intuitive representation that leverages the mean-variance characterization from Corollary D.1.1.2. The variance of welfare equals the variance of the disagreement with θ^D. We express disagreements here as V_ζ to highlight these are in the units of ζ (dollars):

$$Var_\psi\Big[\zeta(x,P,\theta)\Big] = Var_\psi\Big[V_\zeta(x,P,\theta)\Big].$$

Denoting mean indirect utility at (P,Z,θ^D) by $\overline{W}_\zeta(P,Z,\theta^D;\psi) = E_\psi[\zeta(x(P,Z,\theta^D),P)]$, we find that up to second-order approximation of u_ζ,

$$\frac{\partial W(P,Z(P),\theta^D)}{\partial P} \approx u'(\overline{W}_\zeta)\frac{\partial \overline{W}_\zeta}{\partial P} + \frac{u''(\overline{W}_\zeta)}{2}\cdot\frac{dVar_\psi\Big[V_\zeta(x(P,Z,\theta^D),P,\theta)\Big]}{dP}. \quad \text{(D.3.3)}$$

By construction, equations (D.3.1) and (D.3.2) above characterize the effect of dP on mean welfare, $\frac{\partial \overline{W}_\zeta}{\partial P}$ in the first term. The second term then captures how disagreements matter when the planner is risk averse. The characterization is intuitive (and arguably obvious from Corrollary D.1.1.2): given $u''_\zeta < 0$ a policy reform that increases the variance of disagreements about money-metric welfare is less desirable, holding the effect on expected welfare $\overline{W}_\zeta$ fixed.

Remark: Setting Aside Money Metrics. The above characterization holds for any welfare metric under ordinal level comparability and paternalistic risk aversion, but we stated it in terms of money-metric utility to emphasize the relationship with prior work (and diminishing marginal utility over money is intuitive). We do not engage with the money-metric welfare concept going forward in this appendix. We simply assume a utility function that is comparable across frames. We do consider the variance of a given utility function over frames and typically interpret this under paternalistic risk aversion. This can be interpreted in terms of money-metric utility as a welfare metric, or more broadly as any utility function rooted in Proposition 4.3.2.

D.3.3 Under Paternalistic Ambiguity Aversion

Now we turn to the ambiguity averse objective from Proposition D.1.2. We let

$$\psi^*(\theta,P) \equiv \arg\min_{\psi\in\Psi^*} E_\psi[W(P,Z,\theta^D;\theta)].$$

Following Hansen and Sargent [2008], on develop intuition by thinking of ψ^* as being chosen by an "evil agent" who minimizes welfare given the planner's choice of policy.

When ψ^* is differentiable in P, we find

$$\frac{\partial W}{\partial P} = \frac{\partial \overline{W}(P, Z, \theta^D; \psi^*)}{\partial P}. \tag{D.3.4}$$

This welfare effect is the same as $\frac{\partial \overline{W}}{\partial P}$ above (direct effects and behavoiral effects multiplied by marginal internalities) but mean welfare is evaluated over the welfare-minimizing distribution ψ^*. Re-optimization by the evil agent $(\partial\psi^*/\partial P)$ does not have a first-order welfare effect as a consequence of the envelope theorem where it applies.[6]

D.3.4 Examples

Under probabilistic uncertainty in Example 1, we find that the variance of utility over frames is quadratic in the disagreement bewen the two frames, V:

$$Var_\psi(V) = \psi(\theta^D)(1 - \psi(\theta^D))V(x, P)^2. \tag{D.3.5}$$

Evaluating the change in welfare from a policy reform due to the change in variance – the second term from equation (D.3.3) – we find:

$$\frac{\omega''(\overline{W})}{2} \cdot \frac{dVar_\psi\Big[V(x(P, Z, \theta^D), P)\Big]}{dP} = u''_\zeta(\overline{W})Var_\psi(V) * \frac{1}{V}\frac{dV}{dP} \tag{D.3.6}$$

where V and dV/dP are evaluated at $(x(P, Z, \theta^D), P)$. This is a reduced-form expression that carries some intuition. Note that the last term resembles a semi-elasticity; this term is positive when V moves away from zero following a marginal change in P. The importance of disagreements for policy evaluation depends on 1) the degree of paternalistic risk aversion over our measure of welfare (u''_ζ), 2) the extent of disagreement in the status quo $(Var_\psi(V))$, and 3) the change in the magnitude of disagreement generated by the reform.

The character of the $\frac{1}{V}\frac{dV}{dP}$ term depends on more specific features of the model. Let us illustrate this in Example 1.1. To obtain differentiability we introduce some unobserved heterogeneity (conventional uncertainty about the individual's type) so that instead of $V = -1\{x \neq d\}\gamma$, we have $V = -Pr[x \neq d]\gamma$.[7] The right-hand side

[6] Where ψ^* is discontinuous or non-differentiable in P, the envelope theorem does not apply and we require a more global approach to fully characterize optimal policy.

[7] We acknowledge this is informally construed in the interest of avoiding extra notation. We continue to assume that γ is uniform for simplicity, so the unobserved heterogeneity should involve

of (D.3.6) becomes

$$u''_\zeta(\overline{W})\underbrace{\psi(\theta^D)(1-\psi(\theta^D))Pr[x \neq d]^2\gamma^2}_{Var_\psi(V)}\left\{\frac{1}{Pr[x \neq d]}\frac{\partial Pr[x \neq d]}{\partial d}\right\} \tag{D.3.7}$$

The last term in this expression is the semi-elasticity of opt-outs with respect to a change in the default [see also Brot-Goldberg et al., 2023]. A reform of the default rule that increases opt-outs will be less desirable when the planner values robustness, to an extent governed by the other terms in the expression. In Example 1.2, the analogous semi-elasticity term is a weighted semi-elasticity of losses across various dimensions, where the weights depend on the strength of loss aversion in each dimension.

Under ambiguity aversion, the evil agent selects the $\psi \in \Psi^*$ that puts maximal weight on the frame in which welfare is lowest: when $V < 0$, ψ^* places maximal weight on θ^D and where $V > 0$, the evil agent places maximal weight on θ^A. As $V \leq 0$ everywhere in Examples 1.1 and 1.2 – note that this is an implication of the assumption that the level of utility is the same across frames where $x = d$ or $x = r$ – the evil agent always places maximal weight on θ^D in these models. By similar reasoning to the probabilistic uncertainty case, this will make policies where opt-outs are frequent less desirable in Example 1.1, and it will make policies where losses relative to the reference point are larger less desirable in Example 1.2.

D.4 Examples

This section contains proofs from Section 4.4 and Section 4.5.

Proof of Proposition 4.4.1. With constant weights for $\tau > 0$, $\psi(\tau|\tau > 0) = \frac{1}{T}$, and equation (4.4.9) simplifies as follows:

$$\begin{aligned} w(x) &= \beta\sum_{t=1}^{T}\delta^t\mu(x_t) + \frac{1-\psi(0)}{T}(1-\beta)\sum_{\tau=1}^{T}\delta^\tau\mu(x_\tau), && \text{(D.4.1)} \\ &= \left[\beta + (1-\beta)\frac{1-\psi(0)}{T}\right]\sum_{t=1}^{T}\delta^t\mu(x_t), && \text{(D.4.2)} \end{aligned}$$

which is a constant multiple of $u(x, 0)$. □

other preference parameters. See Goldin and Reck [2022b] for a more thorough treatment of the question of interpersonal heterogeneity in this setting.

Proof of Lemma 4.5.1. Suppose $x(P^*)$ BR-dominates any other $x(P')$. Then global optimality of P^* follows from the monotonicity of expected welfare. For the other direction, suppose P^* does not BR-dominate some P', i.e. there is some θ' strictly better off under P' than P^*. Let $\psi(\theta) = 1\{\theta = \theta'\}$. As P^* is not a ψ-optimum for this ψ, it cannot be globally optimal. □

Proof of Proposition 4.5.1. In the defaults case, the policy parameter is 1-d: $P = d$, the default option. As discussed previously, this is usually thought of as an example of Bias vs Strange Preferences - where under θ^D, the as-if cost implied by behaviour is normative, and under θ^A it is a pure bias. $\psi(\theta^D)$, which we abbreviate to just $\psi = \mathbb{P}[\theta = \theta^D]$.

$$u(x, \theta^D, d) = u(x) - \gamma \cdot \mathbb{1}\{x \neq d\} \tag{D.4.3}$$

$$u(x, \theta^A, d) = u(x) \tag{D.4.4}$$

Therefore, $V(x, d) = -\gamma \cdot \mathbb{1}\{x \neq d\}$ and welfare $W(d) = u\big(x(d)\big) - \psi^D \cdot \gamma \cdot \mathbb{1}\{x(d) \neq d\}$.

As a simple example, let $u(x) = -\frac{\alpha}{2}(x - x^*)^2$ where x^* is unknown. $x, x^* \in X$, the choice set which is $X \subset \mathbb{R}$ and defaults at the max and min of X force the consumer to choose actively. $\psi \in [0, 1]$.

1. First, show that the expected intrinsic optimum d_{min} default is a $\kappa - \psi$ robust optimum for any ψ making d_{min} a candidate optimum. This $\psi \approx 1$. So, $B(\kappa, \psi) = [\psi - \kappa, min(\psi + \kappa, 1)]$.

   Recall $W(d, \psi') = -\frac{\alpha}{2}(x(d) - x^*)^2 - \psi' \cdot \gamma \cdot \mathbb{1}\{x(d) \neq d\}$. Since $\gamma > 0$...

   $$\arg \min_{\psi' \in B(\kappa, \psi)} W(d_{min}, \psi') = min(\psi + \kappa, 1) \tag{D.4.5}$$

   The evil agent wants to make the opt-out cost as large as possible so chooses ψ' as large as possible. Therefore, the $\kappa - \psi$ robust optimum is defined by...

   $$d^* = \arg\max_d -\frac{\alpha}{2}(x(d) - x^*)^2 - min(\psi + \kappa, 1) \cdot \gamma \cdot \mathbb{1}\{x(d) \neq d\} \tag{D.4.6}$$

   Since d_{min} is a candidate optimum for ψ, it is also a candidate optimum for $\psi' > \psi$ since under those judgements the opt-out cost is strictly more likely to be normative - suggesting that minimizing opt-outs will be better. Therefore, d_{min} is a $\kappa - \psi$ robust optimum for any κ.

2. Now show that the penalty default is only a $\kappa - \psi$ robust optimum for small κ. Let ψ be the judgement which makes the penalty default a candidate optimum ($\psi \approx 0$). Then, $B(\kappa, \psi) = [max(0, \psi - \kappa), \psi + \kappa]$. Similarly to the minimizing opt-outs example, the evil agent wants to maximise ψ' and so sets...

$$\arg \min_{\psi' \in B(\kappa,\psi)} W(d_{pen}, \psi') = \psi + \kappa \tag{D.4.7}$$

By definition, $x(d_{pen}) = x^*$ and the individual opts-out for sure, therefore...

$$\min_{\psi' \in B(\kappa,\psi)} W(d_{pen}, \psi') = 0 - \gamma(\psi + \kappa) \tag{D.4.8}$$

Consider an alternative policy $\bar{d} = \mathbb{E}[x^*]$, i.e. the minimizing opt-out default. From before, we know that...

$$\min_{\psi' \in B(\kappa,\psi)} W(\bar{d}, \psi') = \underbrace{-\frac{\alpha}{2}(\mathbb{E}[x^*] - x^*)^2}_{=-\Lambda \text{ fixed w.r.t. } \kappa} - (\psi + \kappa) \cdot \gamma \cdot \underbrace{\mathbb{P}\{x(\bar{d}) \neq \bar{d}\}}_{=p \text{ small}} \tag{D.4.9}$$

Therefore, $\bar{d} \succ d_{pen}$ if

$$-\Lambda - (\psi + \kappa) \cdot \gamma \cdot p > -\gamma(\psi + \kappa)$$
$$\iff \gamma \cdot (\psi + \kappa) \cdot (1 - p) > \Lambda$$
$$\iff \kappa > \frac{\Lambda}{\gamma \cdot (1 - p)} - \psi \triangleq \bar{\kappa}$$

where $\bar{\kappa}$ is most likely > 0 given $\psi \approx 0$. I.e. d_{pen} is only a $\kappa - \psi$ robust optimum for at most $\kappa < \bar{\kappa}$. Importantly, note that $\bar{\kappa}$ is **decreasing** in $\gamma = V(x(d_{pen}), d_{pen})$.

□

Proof of Proposition 4.5.2. This result is derived in the discussion in the main text preceding the statement of the proposition. □

D.5 Convex Sets of Frames and Interpretation of Normative Weights

D.5.1 Continuous frames and convex hulls

Here we discuss an extension of our model in which the frame space is conceived of as the convex hull of a finite set of frames. The extension serves to generalize our results and clarify the relationship of our work to the counterfactual normative consumer approach to behavioral welfare analysis, taken up in the next section of this Appendix.

Basics

In the main body we have a set of N frames $\Theta = \{\theta_1, ..., \theta_N\}$. We can think of these as parameters of a (cardinal) utility function $u(x, \theta)$. Suppose each θ_n implies a set of N real-valued preference parameters $\theta_n = (\theta_{n1}, ..., \theta_N)$. Let $\tilde{\Theta}$ be the convex hull of Θ – the set of all convex combinations of elements of Θ. We note that if each of the elements of Θ is non-trivial – no component θ_n can be expressed as the convex combination of other components – then the dimensionality of $\tilde{\Theta}$ must be equal to the number of elements of Θ. Here we assume non-triviality and note that trivial frames can be thought of as elements of $\tilde{\Theta}$ rather than Θ by the logic below.

Linearity and Equivalence to Previous Objectives

By construction, for each $\tilde{\theta} \in \tilde{\Theta}$, there exists a unique weighting function $\psi : \Theta \to \mathbb{R}$ such that $\tilde{\theta} = \sum_{\theta \in \Theta} \psi(\theta)\theta$ and $\sum_{\theta \in \Theta} \psi(\theta) = 1$.

Definition. A utility function is *frame-wise linear* if for any weighting function ψ,

$$u(x, \psi_1\theta_1 + ...\psi_n\theta_N) = \sum_{\theta \in \Theta} \psi(\theta)u(x, \theta).$$

Framewise linearity is arguably a strong restriction but we find it in applied work. Utility is linear in the χ parameter in Lockwood et al's lottery paper, and the π parameter in Goldin and Reck [2022b] and Reck and Seibold [2023]. It is also met in the quasi-hyperbolic discounting model of Laibson [1997] – utility is linear in the present focus parameter β from Example 2). In all of these models, we obtain this equivalence between normative weights on discrete frames and the convex hull of frames under linearity.

If the utility function is frame-wise linear, then for any $\tilde{\theta} \in \tilde{\Theta}$ we have a weighting function ψ such that

$$u(x, \tilde{\theta}) = \sum_{\theta \in \Theta} \psi(\theta) u(x, \theta).$$

This has an obvious equivalence to the utilitarian representation from Propositions 4.2.2 and D.1.1. We discuss this in the main text in reviewing the relationship of our work to the counterfactual normative consumer model.

- If the set of potential utility functions is captured by the discrete set Θ, ψ's are Bayesian beliefs about the probability $\theta \in \Theta$ is normative
- If the set of potential utility functions is captured by the continuous set $\tilde{\Theta}$, under framewise linearity, ψ's are weights derived from the planner's Bayesian beliefs over a convex and continuous set of potential preference parameters.

We could combine the two concepts: if we maintain frame-wise linearity, further introducing a pdf over $\tilde{\theta}$ to capture subjective/Bayesian beliefs about which frame is normative leads to a criterion of the same form, but with a more nuanced interpretation of the weighting function. Under the independence assumption, for any pdf over $\tilde{\theta}$, we can find weights on the discrete set Θ such that the planner's objective maximizes expected welfare over Θ given these weights.

D.6 Counterfactual Normative Consumers and Identification of Normative Weights

Continuing with the setup introduced in the previous appendix section, we now discuss the relationship of our work with the counterfactual normative consumer (CNC) approach. The basic idea here is to suppose that the normative weights are implied by the planner's information, $I \in \mathcal{I}$, so let us express the weights as $\psi(\theta, I)$ such that for fixed I, $\psi \in \Delta(\Theta)$. To be clear, we are using the second interpretation of the weights in ψ above here, assuming framewise linearity.

We consider a stylized version of the counterfactual normative consumer to abstract from interpersonal heterogeneity. Suppose the planner knows about the choices of one other individual. Let us call this other individual the expert and let us call the decision-maker for whom the planner is trying to set an optimal policy the main decision-maker or main DM. The CNC research design is justified by the following assumptions:

- **CNC0 (Knowledge of Expert)**: the expert's revealed preferences are constant over frames, and RP-coincidence holds for the expert.
- **CNC1 (Observation of Expert)**: the preferences of the expert are known/observed in at least one frame.
- **CNC2 (Similarity of Expert and Main DM)**: the main DM and the expert have the same preferences in the normative frame.

Discussion. We observe that CNC0 and CNC1 imply that the planner knows the expert's normative preferences, or their choices in the normative frame. We might assume this directly, but stating the assumptions this way emphasizes that we do not need to observe the expert's choices in the normative frame specifically. For instance, the planner might observe choices in the same decision-making frame θ^D as the main decision-maker and *assume* the expert's choices are constant over frames on the basis of a survey bias proxy. So adopting CNC0 and CNC1 matches how the CNC approach is implemented in practice. In Goldin and Reck [2020], experts are assumed to be those who choose consistently across frames, while in Allcott et al. [2019], experts are identified via a survey bias proxy; both of these require RP-coincidence for the expert – Goldin and Reck label this the consistency principle, Allcott et al. et al impose it when the specify normative benchmarks for their bias proxies. Both CNC0 and CNC2 have untestable normative aspects. Generally, CNC2 can be thought of as the assumption that enables extrapolation from information on the preferences of experts to the preferences of others. Implementations of the CNC approach in practice tend to impose CNC2 via a statistical independence assumption, modelling some interpersonal heterogeneity we do not include here for simplicity, and typically assuming CNC2 conditional on a set of observables.

In any case, the implication of these assumptions is that the planner can infer the normative frame $\tilde{\theta} \in \tilde{\Theta}$ for the main DM by observing the expert's choices, and $\tilde{\theta}$ identifies a weighting function $\psi(\theta)$ on Θ such that the planner's (utilitarian) objective represents normative preferences $\succ^*$ with certainty. That the objective given known preferences takes the same form as the ones we have studied can be viewed as a consequence of the linearity assumption.

This idealized version of CNC is therefore a model in which true/normative preferneces are known given the planner's information, which includes CNC0, CNC1, and CNC2. Moreover, because the weighting function implied by the expert's preferences is unique, we can say that knowledge of the expert's preferences *point-identifies* the appropriate normative weights.

The robustness concepts we develop in our paper help us think through policy problems in which the planner believes that CNC0, CNC1, or CNC2 might fail. For the sake of illustration, let us consider how these assumptions might fail in the context of sugary drink consumption in Allcott et al. [2019]. In this model, the expert is an individual with similar characteristics to the main decision-maker; the expert has no self-reported problems with self-control, no present bias according to a survey bias proxy, and good knowledge about the health risks of consuming sugary drinks.

1. **Frame misspecification.** CNC0 could fail if the set of frames is mis-specified, so that normative choices are not constant across what the researchers consider as frames (this can be formalized along the lines of Example 3). For instance, the payoff due to impulsiveness or lack of self-control could be normatively relevant for the main decision-maker even though the "expert" does not experience these payoffs. (In other words, the possibility here is that the "survey bias proxy" is not capturing a bias but a normative preference).

2. **Expert misspecification.** Alternatively, CNC0 might fail because the "expert" is not completely debiased, for instance because they do struggle a little bit with self-control even though they report having no difficulties with it.

3. **Noisy choices.** CNC1 might fail if the experts revealed preferences are not perfectly observed. For example, we might only have a noisy estimate of how much soda the expert consumes.

4. **Selection Bias.** CNC2 might fail if being an expert is correlated with preferences. This could obviously happen because of frame misspecification above, but setting this aside, the possibility is that experts may be a statistically selected group of individuals, so that they have different preferences from non-expert decision-makers. For instance, those with especially high knowledge about the health consequences of consuming sugary drinks could have gained this knowledge because they especially value good health, and this could lead them to consume less sugary drinks regardless of how well-informed they are. This type of possibility suggests Roy-type selection into expertise that would threaten the (conditional) independence required by CNC2 [Goldin and Reck, 2020].

We do not claim all of these failures are material in the context of corrective taxes on sugary drinks. Allcott et al. [2019] work to defend against many of these concerns, even though the importance of robustness for optimal policy is not formalized in their model.

How we evaluate robustness when employing the CNC identification strategy seems to depend on which of the failures listed above we have in mind. The case of noisy expert choices (Failure 3) seems to suggest thinking about robustness in terms of probabilistic uncertainty; the relevant variance could then be derived from $u(x, \theta)$ for given x and the standard error for our estimate weights in the normative frame. The possibility of (unstructured) expert misspecification or selection bias suggests a local max-min criterion, using a neighborhood $\Psi^* = B(\theta^*, \kappa)$ around our estimate of θ^*, as in the robust control literature. The possibility of frame mis-specification, however, suggests a more global robustness concept, as this involves more philosophical questions about whether the influence of some factor like self-control is due to a framing effect (see also Example 3). We emphasize that these are only suggestions for the appropriate robustness concept to apply to entertain potential failures of these assumptions. For instance, one could also think of expert misspecification in probabilistic terms. Ultimately, the question how we think about uncertainty about the validity of these assumptions, and preferences for how to accommodate it.