Mitigating Risk: Smartphone Notifications, Adaptive Surveying, and Genetics

by

Mark Alan Fontana

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2015

Doctoral Committee:

Professor Miles Kimball, Chair
Associate Professor Daniel Benjamin, USC
Assistant Professor Joshua Hausman
Professor Edward Norton

Acknowledgments

I thank the University of Michigan School of Information's Socio-Technical Infrastructure for Electronic Transactions Multidisciplinary Doctoral Fellowship funded by NSF IGERT grant 0654014. For "Reconsidering Risk Aversion" (with Daniel Benjamin and Miles Kimball) we are grateful to NIH/NIA (R21-AG037741) for financial support, and thank Mike Gideon for helpful early conversations, Fudong Zhang for MATLAB guidance, and Bas Weerman for MMIC guidance. We also thank Yu (Leo) She, Xing Guo, Andrew Sung, and Jordan Kimball for excellent research assistance. For "The Genetics of Cigarette Excise Tax Responsiveness," the Health and Retirement Study genetic data is sponsored by the National Institute on Aging (grant numbers U01AG009740, RC2AG036495, and RC4AG039029) and was conducted by the University of Michigan. I also thank the Kilts-Nielsen Data Center at the University of Chicago Booth School of Business for their Retail Scanner data; I note that information on data availability and access is available at research.ChicagoBooth.edu/Nielsen .

TABLE OF CONTENTS

Acknowledgments . . . . . **ii**

List of Figures . . . . . **v**

List of Tables . . . . . **xvii**

Chapter

1 Smartphone Notifications and Smarter Living **1**

1.1 Abstract 1
1.2 Motivation 2
1.3 Sample 7
1.4 Four App-Based Smartphone Interventions 8
1.5 Methodological Issues 9
1.6 Average Treatment Effects 12
1.7 Additional Analyses 13
1.8 Results 15
1.8.1 Step Intervention #1 15
1.8.2 Step Intervention #2 18
1.8.3 Step Intervention #3 20
1.8.4 Sleep Intervention 22
1.9 Difference-in-Differences Robustness Checks 24
1.10 Conclusions 25
1.11 References and Works Consulted 26
1.12 Figures 34
1.13 Appendix: Demographic Mediators of Heterogeneity 61
1.13.1 Openness to Experience and Conscientiousness 61
1.13.2 Comfort with Adopting New Technology 63
1.13.3 Risk Aversion 64
1.13.4 Results 65

2 Reconsidering Risk Aversion * **66**

2.1 Abstract 66
2.2 Background and Motivation 67
2.3 Importance of Setting Portfolio Defaults and Savings Rates Appropriately 71
2.4 Frames and Axioms 72
2.5 Experiment Design and Methods 74

2.5.1 Design Considerations . . . 76
2.5.2 Subject Population . . . 82
2.5.3 Pre-Test . . . 83
2.5.4 Training Batteries . . . 83
2.5.5 Main Body Part 1: Elicitation of Untutored Preferences . . . 84
2.5.6 Psychological and Cognitive Batteries . . . 86
2.5.7 Main Body Part 2: Elicitation of Reasoned Preferences . . . 86
2.5.8 Inconsistency Checks . . . 87
2.5.9 Intransitivity Checks . . . 87
2.5.10 Follow-Up and Demographic Batteries . . . 88
2.6 Results . . . 89
2.6.1 Inconsistency Checks . . . 89
2.6.2 Intransitivity Checks . . . 94
2.7 Maximum Likelihood Estimation . . . 100
2.7.1 Nodewise Action Choice Frames . . . 101
2.7.2 Pairwise Complete Strategy Frames . . . 103
2.8 Conclusions and Plans for Further Research . . . 108
2.9 References and Works Consulted . . . 111
2.10 Appendix: Pilot Study . . . 115
2.11 Appendix: Flow of Inconsistency Checks . . . 116
2.12 Appendix: Pre-Test . . . 117
2.13 Appendix: Updating Conditional on Riskiness of Choices . . . 124
2.14 Appendix: MLE Diagnostic Tests . . . 124

3 The Genetics of Cigarette Excise Tax Responsiveness . . . 139

3.1 Abstract . . . 139
3.2 Motivation . . . 140
3.3 Related Literature . . . 144
3.4 Data . . . 145
3.4.1 Health and Retirement Study . . . 145
3.4.2 Genome-Wide Association Studies . . . 146
3.4.3 Cigarette Excise Taxes . . . 146
3.5 Summary Statistics . . . 146
3.6 Replication of Fletcher (2012) . . . 147
3.7 Population Stratification and rs2304297 . . . 148
3.8 Polygenic Scores and Cigarette Tax Response Heterogeneity . . . 149
3.9 Conclusions . . . 153
3.10 Figures . . . 155
3.11 References and Works Consulted . . . 170

LIST OF FIGURES

1.1 Step Intervention #1, Thanksgiving 2013. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. 35
1.2 Step Intervention #1, Thanksgiving 2013. Estimates from regression of daily steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 36
1.3 Step Intervention #1, Thanksgiving 2013. Estimates from regression of daily steps on the initial day of the intervention on treatment status, average steps on the preceding 6 days (standardized so it has mean 0 and variance 1), and an interaction term. 37
1.4 Step Intervention #1, Thanksgiving 2013. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles. 37
1.5 Step Intervention #1, Thanksgiving 2013. Comparison of the two treatments. Treatment12 is an indicator that takes on a value of 0 if the notification included the person's step average ("Stay at flock's front by meeting your X step average.") and a value of 1 if the message did not ("Stay at flock's front by maintaining your average today."). Column (1) shows mean difference in steps between the treatments. The dependent variable is total steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. 38
1.6 Step Intervention #1, Thanksgiving 2013. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the notifications (mutually exclusive categories). Note that the "ignore" category includes those who look at the app and ignore the notification and those who simply don't look at the app. Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects. 39
1.7 Step Intervention #1, Thanksgiving 2013. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally. 39

1.8 Step Intervention #1, Thanksgiving 2013. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. 40
1.9 Step Intervention #1, Thanksgiving 2013. Estimates from regression of an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 41
1.10 Step Intervention #1, Thanksgiving 2013. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the notification-based goal, only presented to the treatment group (average daily steps). Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. 42
1.11 Step Intervention #2, fourth day in the system. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects. 43
1.12 Step Intervention #2, fourth day in the system. Estimates from regression of daily steps on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 44
1.13 Step Intervention #2, fourth day in the system. Estimates from regression of daily steps on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday..., Column (7) is Saturday. 44
1.14 Step Intervention #2, fourth day in the system. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles. 45
1.15 Step Intervention #2, fourth day in the system. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects. 45
1.16 Step Intervention #2, fourth day in the system. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored initial notification. These estimates should not be interpreted causally. 46

1.17 Step Intervention #2, fourth day in the system. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects. 46
1.18 Step Intervention #2, fourth day in the system. Estimates from regression of an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 47
1.19 Step Intervention #2, fourth day in the system. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the notification-based goal, only presented to the treatment group (average daily steps over the preceding 2 days + 500 steps). Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects. 48
1.20 Step Intervention #2, fourth day in the system. Estimates from regression of an indicator for whether the person met the notification-based goal (average steps over the preceding 2 days + 500 steps) on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 49
1.21 Step Intervention #2, fourth day in the system. Estimates from regression of lagged indicator for whether the person met the notification-based goal (average steps over the preceding 2 days + 500 steps) on treatment status, for lags of 1 to 11 days. Regressions additionally controlling for state fixed effects, time fixed effects, or age, gender, BMI, and days since becoming a user (not shown) are essentially identical in terms of lagged treatment effect magnitudes and significances. 50
1.22 Step Intervention #3, after a particularly sedentary week. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects. 51
1.23 Step Intervention #3, after a particularly sedentary week. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects. 52

1.24 Step Intervention #3, after a particularly sedentary week. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally. 52
1.25 Sleep Intervention, after a particularly restless few days. Column (1) shows mean difference in minutes of sleep between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total minutes of sleep on the day of the initial notification. Columns (2) and (3) consider bed and wake times as the outcome variable (in minutes relative to 7pm the evening of the notification). Column (4) includes age, BMI, and an indicator for male gender as covariates. Column (5) includes state fixed effects. Column (6) includes date fixed effects. 53
1.26 Sleep Intervention, after a particularly restless few days. Estimates from regression of daily minutes of sleep on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 54
1.27 Sleep Intervention, after a particularly restless few days. Estimates from regression of minutes of sleep on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday... Column (5) is Thursday. Note this intervention was *not* fielded on Friday nor Saturday nights. 55
1.28 Sleep Intervention, after a particularly restless few days. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles. 55
1.29 Sleep Intervention, after a particularly restless few days. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects. 56
1.30 Sleep Intervention, after a particularly restless few days. Column (1) shows the mean difference in minutes of sleep between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally. 56
1.31 Sleep Intervention, after a particularly restless few days. Estimates from regression of lagged minutes of sleep on treatment status, for lags of 1, 2, 3, 4, 5, 6, and 14 days. Regressions additionally controlling for state fixed effects or age, gender, BMI, and days since becoming a user (not shown) are essentially identical in terms of lagged treatment effect magnitudes and significances. . . . 57

1.32 Sleep Intervention, after a particularly restless few days. Column (1) shows estimates from a regression of minutes of sleep per day averaged across all days the user remains active in the system (including the day of the initial notification and after) on first-day treatment status. Columns (2) and (3) additionally control for state fixed effects and gender, BMI, age, and days since becoming a user, respectively. 57
1.33 Sleep Intervention, after a particularly restless few days. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects. 58
1.34 Sleep Intervention, after a particularly restless few days. Estimates from regression of an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized. 59
1.35 Sleep Intervention, after a particularly restless few days. Estimates from regression of an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday... Column (5) is Thursday. Note this intervention was *not* fielded on Friday nor Saturday nights. 60

2.1 Complete Contingent Action Plan (1 question). Subjects are prompted with instructions: "Imagine you are currently 35 years old. You need to make three decisions. The decision between A and B takes place now. If you choose B, you also need to make two decisions that will lock in how you will invest at age 50. If you choose A, you do not need to make any more decisions. All decisions will affect how much money you will be able to spend each year during retirement (from age 65 on). In each decision, you will choose between two strategies: risky or conservative. Each conservative strategy will guarantee you a fixed amount to spend each year during retirement. Each risky strategy allows for possibly higher amounts. You do not need to choose the same kind of strategy in each decision." 75
2.2 Single Action in Isolation (2 questions). Subjects are prompted with instructions: "You will be asked to make decisions that will affect how much you will be able to spend each year during retirement (from age 65 on), imagining that you are currently 50 years old. You will choose between two strategies: risky or conservative. The conservative strategy will guarantee you a fixed amount to spend each year during retirement. Under the risky strategy, higher amounts are possible." 76

2.3 Single Action with Backdrop (2 questions). Subjects are prompted with instructions: "You will be asked to make decisions that will affect how much you will be able to spend each year during retirement (from age 65 on), imagining that you are currently 50 years old. You will also be given information about how decisions you made when you were 35 turned out, that are beyond your control at this point. These grayed-out parts of the picture are things that could have happened, but you know for sure did not happen. You will choose between two strategies: risky or conservative. The conservative strategy will guarantee you a fixed amount to spend each year during retirement. Under the risky strategy, higher amounts are possible." 77
2.4 Two Contingent Actions with Backdrop (1 question). Subjects are prompted with instructions: "Imagine you are currently 35 years old, and have chosen risky decision B. You do not yet know how this decision has turned out. So, you need to make two decisions that will lock-in how you will invest at age 50. These decisions will affect how much money you will be able to spend each year during retirement (from age 65 on). In each decision, you will choose between two strategies: risky or conservative. Each conservative strategy will guarantee you a fixed amount to spend each year during retirement. Each risky strategy allows for possibly higher amounts. You do not need to choose the same kind of strategy in each decision." 78
2.5 Pairwise Choices Between Complete Strategies (10 questions). Subjects are prompted with instructions: "In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2. Each investment plan has a set of choices locked in along the way (at age 35 and age 50), shown by circled letters, that lead to possible levels of yearly spending during retirement (from age 65 on). Grayed-out parts are used to show things that can't happen if you choose that investment plan. Spinners show the chances of different outcomes. From a spinner, the chance of taking each fork is shown next to that fork. Each fork can lead either to a locked in choice, or directly to a level of yearly spending during retirement. Note that a path with a 50% chance at one fork, followed by another 50% chance at a later fork, means that there is only 25% chance of getting all the way to the end of that path. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page)." 79

2.6 Pairwise Choices Between Compound Lotteries (10 questions). Subjects are prompted with instructions: ”In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2, imagining that you are currently 35 years old. Each investment plan has different possible outcomes for how much you will be able to spend each year during retirement (from age 65 on). Note that some investment plans are named by one letter, ”A”; other investment plans are named by more than one letter, like ”BCE”. Spinners show the chances of different outcomes. From a spinner, the chance of taking each fork is shown next to that fork. Each fork can lead either to another spinner, or directly to a level of yearly spending during retirement. Note that a path with a 50% chance at one fork, followed by another 50% chance at a later fork, means that there is only a 25% chance of getting all the way to the end of that path. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page).” 80
2.7 Pairwise Choices Between Reduced Simple Lotteries (10 questions). Subjects are prompted with instructions: ”In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2, imagining that you are currently 35 years old. Each investment plan has different possible outcomes for how much you will be able to spend each year during retirement (from age 65 on). Note that some investment plans are named by one letter, like ”A” ; other investment plans are named by more than one letter, like ”BCE”. The picture shows the chance of each outcome happening next to the outcome (if the chance is less than 100%). Note that a 50% chance is twice as likely to happen as a 25% chance. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page). ” 81
2.8 Example of an intransitivity resolution. 88
2.9 MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. “Isolation” refers to Single Action in Isolation; “Backdrop” refers to Single Action with Backdrop; “Two Backdrop” refers to Two Contingent Actions with Backdrop; “Complete-1” refers to Complete Contingent Action Plan with all available data; and “Complete-4” refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame. 104
2.10 MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. “Isolation” refers to Single Action in Isolation; “Backdrop” refers to Single Action with Backdrop; “Complete-1” refers to Complete Contingent Action Plan with all available data; and “Complete-4” refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame. 105
2.11 MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. “Isolation” refers to Single Action in Isolation; and “Backdrop” refers to Single Action with Backdrop. All available data are used for each snapshot and frame. 106

2.12 MLE results for Nodewise Action Choices Frames, snapshots 1-5. Error bars denote standard errors. “Isolation” refers to Single Action in Isolation; “Backdrop” refers to Single Action with Backdrop; “Complete-1” refers to Complete Contingent Action Plan with all available data; and “Complete-4” refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame. 107
2.13 MLE results for Pairwise Strategy Choices Frames, snapshots 1-10. All available data are used for each snapshot and frame. Error bars denote standard errors. “Complete” refers to Pairwise Choices Between Complete Strategies; “Compound” refers to Pairwise Choices Between Compound Lotteries; and “Reduced” refers to Pairwise Choices Between Reduced Simple Lotteries. . . . 109
2.14 Example of an inconsistency check, initial question. 117
2.15 Example of an inconsistency check, conditional on answering ”It makes sense to have different choices.” 118
2.16 Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices.” 118
2.17 Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices” and now verifying updated preferences 119
2.18 Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices,” having verified updated preferences. 119
2.19 Example of an placebo inconsistency check, initial question. 119
2.20 Example of an placebo inconsistency check, conditional on answering ”It makes sense to have the same choice...” 120
2.21 Example of an placebo inconsistency check, initial question. 120
2.22 Example of an placebo inconsistency check, conditional on answering ”It makes sense to have different choices...” 121
2.23 Example of an placebo inconsistency check, conditional on answering ”It makes sense to have different choices...” and now verifying updated preferences. 121
2.24 Example of an placebo inconsistency check, conditional on answering ”It makes sense to have different choices...” and having verified updated preferences. 122
2.25 Diagnostic tests for μ, Single Action in Isolation, Snapshots 1-10. 126
2.26 Diagnostic tests for σ_x, Single Action in Isolation, Snapshots 1-10. 127
2.27 Diagnostic tests for σ_ϵ, Single Action in Isolation, Snapshots 1-10. 127
2.28 Diagnostic tests for μ, Single Action with Backdrop, Snapshots 1-10. 128
2.29 Diagnostic tests for σ_x, Single Action with Backdrop, Snapshots 1-10. 128
2.30 Diagnostic tests for σ_ϵ, Single Action with Backdrop, Snapshots 1-10. 129
2.31 Diagnostic tests for μ, Two Contingent Actions with Backdrop, Snapshots 1-10. 129
2.32 Diagnostic tests for σ_x, Two Contingent Actions with Backdrop, Snapshots 1-10. 130
2.33 Diagnostic tests for σ_ϵ, Two Contingent Actions with Backdrop, Snapshots 1-10. 130
2.34 Diagnostic tests for μ, Complete Contingent Action Plan-1, Snapshots 1-10. . . 131
2.35 Diagnostic tests for σ_x, Complete Contingent Action Plan-1, Snapshots 1-10. . 131
2.36 Diagnostic tests for σ_ϵ, Complete Contingent Action Plan-1, Snapshots 1-10. . 132
2.37 Diagnostic tests for μ, Complete Contingent Action Plan-4, Snapshots 1-10. . . 132

2.38 Diagnostic tests for σ_x, Complete Contingent Action Plan-4, Snapshots 1-10. . 133
2.39 Diagnostic tests for σ_ϵ, Complete Contingent Action Plan-4, Snapshots 1-10. . 133
2.40 Diagnostic tests for μ, Pairwise Choices between Complete Strategies, Snapshot 1. 134
2.41 Diagnostic tests for σ_x, Pairwise Choices between Complete Strategies, Snapshot 1. 134
2.42 Diagnostic tests for σ_ϵ, Pairwise Choices between Complete Strategies, Snapshot 1. 135
2.43 Diagnostic tests for μ, Pairwise Choices Between Compound Lotteries, Snapshot 1. 136
2.44 Diagnostic tests for σ_x, Pairwise Choices Between Compound Lotteries, Snapshot 1. 136
2.45 Diagnostic tests for σ_ϵ, Pairwise Choices Between Compound Lotteries, Snapshot 1. 137
2.46 Diagnostic tests for μ, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1. 137
2.47 Diagnostic tests for σ_x, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1. 138
2.48 Diagnostic tests for σ_ϵ, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1. 138

3.1 Histogram of polygenic scores based on GWA study estimates from Okbay et al. (2015), measuring genetic predisposition to high educational attainment (EA). 156
3.2 Histogram of polygenic scores based on GWA study estimates from the TAG Consortium (2010), measuring genetic predisposition to high daily cigarette consumption. 156
3.3 Federal, average state, and combined real (in 2000's dollars) cigarette excise taxes over time. The tax is levied per pack (20 cigarettes), expressed here in cents per pack. Note a dramatic increase in 2009 corresponding to provisions in the State Children's Health Insurance Program (SCHIP). 157
3.4 Fletcher (2012)'s specification along the extensive margin. OLS regression of smoking status on log cigarette excise tax rate, an indicator for having the GG genotype on rs2304297, an interaction term between log tax rate and the GG indicator, and covariates: birth year, birth year squared, an indicator for being female, indicators for self-reported race (with the omitted category being white), educational attainment (EA), being currently married, and income in thousands of dollars. Columns (1), (2), and (3) use the full HRS sample; Column (4) uses only self-reported whites. Standard errors are clustered by state and individual. 158

3.5 Fletcher (2012)'s specification along the intensive margin. OLS regression of cigarettes per day (CPD) on log cigarette excise tax rate, an indicator for having the GG genotype on rs2304297, an interaction term between log tax rate and the GG indicator, and covariates: birth year, birth year squared, an indicator for being female, indicators for self-reported race (with the omitted category being white), educational attainment (EA), being currently married, and income in thousands of dollars. Columns (1), (2), and (3) use the full HRS sample; Column (4) uses only self-reported whites. Standard errors are clustered by state and individual. 159
3.6 Fletcher (2012)'s specification. OLS regression of log cigarette excise tax rate on indicators for genotype on rs2304297 (CC as the omitted category), birth year, birth year squared, indicators for self-reported race (with the omitted category being white), educational attainment (EA), and income in thousands of dollars. Column (2) additionally includes wave fixed effects. Standard errors are clustered by state and individual. 160
3.7 OLS regressions testing for the extent to which population stratification drives variation in rs2304297. Column (1) regresses an indicator for the GG genotype on indicators for race (determined by the first two principal components (PCs) derived from the full HRS sample and all genotyped SNPs, according to HRS's quality control manual, with the omitted category being "Hispanic/Asian"). Column (2) regresses an indicator for the GG genotype on the first 10 principal components (PCs) of genotyped SNPs from the full HRS sample. Columns (3) and (4) repeat these analyses with the *number* of G alleles on rs2304297 as the dependent variable. 161
3.8 OLS regressions testing the predictive power of a polygenic score for smoking on smoking status. Column (1) regresses smoking status on the score. Column (2) regresses smoking status on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification. 162
3.9 OLS regressions testing the predictive power of a polygenic score for smoking on cigarettes per day (CPD). Column (1) regresses CPD on the score. Column (2) regresses CPD on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification. 163
3.10 OLS regressions testing the predictive power of a polygenic score for educational attainment (EA) on smoking status. Column (1) regresses smoking status on the score. Column (2) regresses smoking status on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification. 164

3.11 OLS regressions testing the predictive power of a polygenic score for educational attainment (EA) on cigarettes per day (CPD). Column (1) regresses CPD on the score. Column (2) regresses CPD on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification. 165
3.12 OLS regressions testing for cigarette excise tax response heterogeneity along the extensive margin with respect to genetic variation in a polygenic score for cigarette smoking. All analyses include only genetically European-decent HRS participants. Column (1) regresses smoking status on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual. 166
3.13 OLS regressions testing for cigarette excise tax response heterogeneity along the intensive margin with respect to genetic variation in a polygenic score for cigarette smoking. All analyses include only genetically European-decent HRS participants. Column (1) regresses cigarettes per day (CPD) on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual. 167

3.14 OLS regressions testing for cigarette excise tax response heterogeneity along the extensive margin with respect to genetic variation in a polygenic score for educational attainment (EA). All analyses include only genetically European-decent HRS participants. Column (1) regresses smoking status on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual. 168
3.15 OLS regressions testing for cigarette excise tax response heterogeneity along the intensive margin with respect to genetic variation in a polygenic score for educational attainment (EA). All analyses include only genetically European-decent HRS participants. Column (1) regresses cigarettes per day (CPD) on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual. 169

LIST OF TABLES

1.1 Summary of conditions that triggered each intervention, as well as the messages sent to users through the companion smartphone application, indicating exactly how users are nudged towards more steps or sleep. 9

1.2 Summary of number of days each intervention was implemented, and sample sizes for treatment and control groups. Note that a user only appears at most once in a given intervention. Sample sizes are based on data that has been trimmed only to include treatment and control group members who opened the app on the day of the intervention for "Step Intervention #2," "Step Intervention #3," and "Sleep Intervention." The data to do this trimming for "Step Intervention #1" was not available. Note also there are two slightly different "Step Intervention #1" treatment versions ("v1" and "v2") that differ in whether they present users with their exact historical daily step average. 10

1.3 Summary of initial-day treatment effects for the 4 interventions. Only "Step Intervention #3" had an insignificant treatment effect (all others were significant at the 0.0001 level, indicated by ***). Because of limited data availability on who opened the app, the "ignore" category for "Step Intervention #1" includes both those who looked at the notification and ignored it and those who never opened the app. In other interventions, the "ignore" category only includes the former, given availability of data on who opened the app. Therefore, the treatment effect for "Step Intervention #1" is an ITT and for all others are ATETs. 16

2.1 Welfare cost of investing with "wrong" preferences. Wealth loss (in percent) that is equivalent to mistaken behavior. True risk aversion, γ, is along the columns. Risk aversion used for portfolio choice, $\hat{\gamma}$, is along the rows. 72

2.2 Name of each axiomatic baby step (or simply "step") and the two frames associated with each. 74

2.3 Columns represent 6 potential sets of monetary amounts, associated with constant coefficients of relative risk aversion of 1.576, 2.958, 4.865, 7.184, 12.113, or 17.967, respectively. Descriptions of how these amounts are calculated can be found in the main text. 85

2.4 Percentage of respondents answering on a scale of 1-6, "To what extent do you agree with the following statement": "I enjoyed thinking through these choices," "Thinking through these choices was annoying," "Thinking through these choices made me feel stressed," "Thinking through these choices was frustrating." 89

2.5 Average inconsistency rates for untutored and reasoned preferences, and two-sided tests for differences in proportions (i.e. differences in inconsistency rates). "Total" denotes total potential inconsistencies for a given axiom. Inconsistency rates are calculated by averaging across subjects: total inconsistencies divided by total potential inconsistencies, by axiom. 90
2.6 Average inconsistency rates for untutored and reasoned preferences in waves 1 and 2, respectively, and two-sided tests for differences in proportions (i.e. differences in inconsistency rates) within wave 1 between untutored and reasoned preferences, between wave 1 reasoned preferences and wave 2 untutored preferences, and within wave 2 between untutored and reasoned preferences. Inconsistency rates are calculated by averaging across subjects: total inconsistencies divided by total potential inconsistencies, by axiom. 91
2.7 Percentage of respondents with any inconsistency. 93
2.8 Percentage of the time that subjects updated toward each frame, did not update, or swapped their choices for normal inconsistency checks. 94
2.9 Percentage of the time that subjects gave each of the following responses to "Why do you want to make different choices in these two situations?": (1) "The two situations are different enough that I want different choices", (2) "Some of the options are equally good to me, so it doesn't matter which one I choose", (3) "I chose how I thought the experimenters wanted me to chose", (4) "I don't know which options I prefer", (5) "I don't know or am confused", or (6) "Other." 95
2.10 Percentage of the time that subjects gave each of the following responses to "Why did you want to change your choices as you did?": (1) "I made a mistake when I first chose", (2) "Answering all of these questions made me change what I want", (3) "Some of the options are equally good to me, so it doesn't matter which one I choose", (4) "I chose how I thought the experimenters wanted me to chose", (5) "I don't know which options I prefer", (6) "I don't know or am confused", or (7) "Other." 96
2.11 Percentage of the time that subjects updated each frame, did not update, or updated both their choices for placebo inconsistency checks. 96
2.12 Percentage of the time that subjects gave each of the following responses to "Why do you want to make the same choices in these two situations?": (1) "The two situations are similar enough that I want to make the same choices", (2) "Some of the options are equally good to me, so it doesn't matter which one I choose", (3) "I chose how I thought the experimenters wanted me to chose", (4) "I don't know which options I prefer", (5) "I don't know or am confused", or (6) "Other." 97
2.13 Percentage of the time that subjects gave each of the following responses to "Why did you want to change your choices as you did?": (1) "I made a mistake when I first chose", (2) "Answering all of these questions made me change what I want", (3) "Some of the options are equally good to me, so it doesn't matter which one I choose", (4) "I chose how I thought the experimenters wanted me to chose", (5) "I don't know which options I prefer", (6) "I don't know or am confused", or (7) "Other." 98

2.14 For each axiom and inconsistency, the percentage of time each frame is associated with the riskier choice. For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2. 124
2.15 Conditional on Frame 1 being the riskier choice in an inconsistency, percentage of the time subjects' update toward each frame (or whether they choose not to update or to swap their choices). For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2. 125
2.16 Conditional on Frame 2 being the riskier choice in an inconsistency, percentage of the time subjects' update toward each frame (or whether they choose not to update or to swap their choices). For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2. 125

3.1 Summary Statistics, Static Variables 155
3.2 Summary Statistics, Time-Varying 155

CHAPTER 1

Smartphone Notifications and Smarter Living

1.1 Abstract

Obesity looms as one of America's top public health crises. Recent advances in mobile computing have facilitated an interest in mobile health (or simply "mHealth") as a means of improving the timeliness of healthcare information and interventions (Klasnja and Pratt, 2012). In the realm of behavioral economics, loss aversion and reference-dependent utility highlight the importance of reference points in influencing individual decision-making (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991, 1992; Kahneman et al., 1991). A host of related work indicates that expectations and goals can set these reference points (e.g. Koszegi and Rabin 2006, 2007, 2009). However, much extant behavioral research on goals is either theoretical or experimental, and little attention has been given to whether temporary goals suggested through mobile health platforms outside of a laboratory might influence behavior. Existing mHealth studies in the domain of physical activity and sleep are multifaceted in terms of strategies, environments, and samples (Klasnja and Pratt, 2012), rending cross-study comparison difficult, further stymied by small sample sizes and therefore questionable robustness. It is unclear how best to set these goals in practice, and for how long their impact might persist. Toward these ends, using data on a subsample of hundreds of thousands of users of one particular commercially available wearable activity tracker, I apply tools from behavioral economics to interpret and analyze 4 randomized smartphone-based goal-setting interventions meant to spur either increased daily steps or earlier bedtimes. These notifications, and the temporary goals they proffer (and hence reference points they attempt to set), were customized to either the specific day of the year on which they were sent or a user's recent activity. They were purposely chosen to constitute modest improvements over expectations and recent performance, and hence were not overly ambitious (i.e. typically much lower than the default goal and only slightly higher than activity in preceding days). Their power is in framing *not* meeting the temporary,

modest goal as a loss, beneficially exploiting loss aversion to propel improvement. "Step Intervention #1" targeted lack of physical activity on Thanksgiving by suggesting a user try to meet her daily step average. "Step Intervention #2" targeted a user's fourth day in the system, suggesting she try to surpass her 2-day average by 500 steps ($\sim$0.3 miles). "Step Intervention #3" targeted lack of activity after a particularly sedentary week by encouraging a user to surpass her recent (lower) average by 500 steps. Finally, "Sleep Intervention" targeted lack of sleep after a particularly restless week and encouraged a user to get to bed earlier. Initial day treatment effects were significant for "Step Intervention #1," "Step Intervention #2," and "Sleep Intervention," engendering $\sim$170 and $\sim$160 additional steps per successful step intervention ($\sim$0.1 miles), respectively, and $\sim$5 minutes additional rest for the sleep intervention. Only the sleep intervention continued to have significant treatment effects beyond the initial day, engendering about a minute more reported sleep per night for as long as a user remained in the system. These results were robust to using an indicator for whether a user met her *default* goal of 10,000 steps or 8 hours of sleep that day as the outcome variable (instead of steps or minutes of sleep). They were also robust instead to using an indicator variable for whether a user met her *temporary, notification-based* goal (although this specification showed additional evidence of responsiveness to both "Step Intervention #2" and "Step Intervention #3"). There was substantive treatment effect heterogeneity with respect to day of the week: Thursdays and Saturdays drove the effectiveness of "Step Intervention #2"; Tuesdays, Sundays, and Wednesdays drove the effectiveness of "Sleep Intervention" (note this intervention was not fielded on Friday nor Saturday nights). I find limited treatment effect heterogeneity with respect to gender, age, and BMI. There was, however, substantial heterogeneity in challenge acceptance with respect to demographics, as women, younger users, and higher BMI users were more likely to opt into the challenges. These acceptance effects were stable across days of the week. Overall, my results are consistent with users' reference points being temporarily updated by modestly ambitious notification-based goals. These goals positively exploit loss aversion by framing *not* meeting the temporary, achievable improvement as a loss, promoting physical activity in the short-term and increased reported sleep in the medium-term.

1.2 Motivation

Obesity looms as one of America's top public health crises. Using data from the 2011-2012 National Health and Nutrition Examination Survey, Ogden et al. (2014) report that more than one-third of American adults were obese[1]. According to the Centers for Disease

[1] Obese is defined as having a BMI greater than or equal to 30.

Control and Prevention (CDC), citing estimates from Finkelstein et al. (2009), total annual medical costs associated with obesity in the U.S. were $147 billion in 2008, amounting to approximately $1,429 more in medical expenditures per obese person[2]. Using an instrumental variables approach that addresses the endogeneity of weight, Cawley and Meyerhoefer (2011) estimate that the per-person cost of obesity is almost twice as high.

The rise of the obesity epidemic has seen the concurrent rise of mobile computing as a ubiquitous part of daily existence. Nearly two-thirds of Americans now own a smartphone (Smith, 2015). The last few years has also seen a spectacular rise in the use of wearable activity trackers (or simply "wearables"). These devices, typically worn on the wrist, enable their users to track physical activity (e.g. steps and sleep) and observe the extent to which they meet goals related to these outcomes through interactive smartphone applications (or simply "apps"). These advancing in mobile computing have facilitated an interest in mobile health (or simply "mHealth") as a means of improving the availability and timeliness of healthcare information and interventions. Klasnja and Pratt (2012) provide an excellent overview of mHealth work up until that point, nothing mHealth platforms are useful given the widespread adoption of computationally powerful mobile devices, people's tendency to carry these devices everywhere, and the ease at which mobile devices can capture personal and contextual information. Quoting briefly from Klasnja and Pratt (2012), these technologies have sought to encourage physical activity; promote healthier diets; monitor symptoms for diseases and disorders as broad-reaching as heart disease, diabetes, asthma, HIV, and cancer; remind patients about taking their medication and attending appointments; help with smoking cessation; and encourage sunscreen application.

Setting aside metabolic syndromes and physical disabilities, the formula for maintaining a healthy weight is theoretically straightforward: healthier diet, more physical activity, and a more consistent sleep schedule (Andersen, 1999; Patel and Hu, 2008; Patel et al., 2006). Actual commitment to forming and fulfilling goals related to these outcomes, of course, is more difficult, given difficulties with self-control and a human tendency toward present-bias[3]. Bryan et al. (2010) describe two varieties of commitment devices meant to help overcome these difficulties and help fulfill desired future behavior: "hard" devices[4]

[2]See also: http://www.cdc.gov/obesity/data/adult.html

[3]A related issue is over optimism: for example, DellaVigna and Malmendier (2006) use data from U.S. health clubs to show that people are optimistic about their future gym usage and end up overpaying with a one-time yearly fee relative to a per visit fee.

[4]For example: Charness and Gneezy (2009) investigate the impact of an intervention aimed at promoting gym visits per month. Not only did their scheme encourage attendance while monetary incentives were in place, but the effects on attendance persisted for many weeks even after the financial incentives disappeared. They noted substantial improvement in health outcomes such as weight and waist size. Using a large-scale workplace field experiment, Royer et al. (2012) show that workers respond to financial incentives, but long-term effects were modest unless the treatment had been combined with some commitment contract.

harness financial incentives to encourage goal commitment, while "soft" devices harness psychological disappointment. My focus here is on "soft" devices. Makers of wearable and mHealth technologies are in a unique position to conduct experiments that attempt to nudge their users toward healthier lives using such soft devices. These firms hold troves of personal health data and can easily customize these nudges according to users' recent activity (or lack thereof). Their experimental platforms are advantageous given their pre-existing user bases (which are typically large), as well as their customizability and ability to be quickly modified to changing conditions.

In the realm of behavioral economics, loss aversion (i.e. concavity over gains and convexity over losses), reference-dependent utility, and diminishing sensitivity (i.e. increased sensitivity to changes closer to the reference point), highlight the importance of reference points in influencing individual decision-making (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991, 1992; Kahneman et al., 1991). These concepts formalize the psychological tendency of humans to detect and respond to relative changes. Indeed, loss aversion has been observed in a broad range of domains (e.g. golf in Pope and Schweitzer, 2011; football and domenstic violence in Card and Dahl 2011). However, the process through which reference points are formed and updated is an area of active research. The status quo is one potential source (e.g. Kahneman et al., 1990). On the other hand, Koszegi and Rabin (2006, 2007, 2009) argue recent expectations and beliefs endogenously drive reference point formation, although updating is often slow, even in the face of quickly changing beliefs. Experimental evidence for expectations and beliefs driving reference point formation and hence behavior abounds (Abeler, 2011; Ericson et al., 2011). Goals can also explicitly act as reference points via loss aversion[5] to help overcome present-bias and self-control problems (Heath et al., 1999; Suvorov and van de Ven, 2008; Koch and Nafziger, 2011), yet excessively high goals can be counterproductive (Matthey et al, 2007). However, much relevant research on goals is either theoretical or experimental, and little is known about whether temporary goals suggested through mobile health platforms outside of a laboratory might influence behavior via loss aversion and reference-dependent utility. Moreover, it is unclear how best to set these goals, and for how long their impact might persist.

In terms of direct experimental evidence of the effectiveness of mHealth platforms, Klasnja and Pratt (2012) highlight that broad evaluation is difficult because of the diverse, multifaceted nature of relevant strategies, outcomes, and environments. With respect to interventions that aim to increase the accessibility of health information about physical ac-

[5]Hsiaw (2013) uses an optimal stopping problem to show that loss aversion is not a necessary condition for non-binding goals to act as reference points and help overcome present-bias (when commitment to goals is sufficiently high). In Hsiaw's framework, time-consistent agents can be harmed by goal commitment.

tivity (via health messages, reminders, and glanceable displays), there exist a multitude of small studies with sample sizes only in the hundreds (or even fewer), often focused only on one particular subpopulation, many in the form of feasibility and pilot studies[6]. Many focus on the effect of having an mHealth platform or app rather than designing particular interventions to target goal formation conditional on having that technology. A meta-analysis by Lewis et al. (2015) reviews 11 articles that investigate physical activity interventions implemented through wearable devices and their smartphone companion apps. These interventions varied substantially from each other in terms of both subject pools and designs, ranging from the introduction of a wearable device, emails, and companion apps; to specific text messages that were meant to target behavior; to counseling over the telephone and in-person. They conclude that more high-quality randomized experiments are necessary to evaluate how best to design such interventions and which demographic groups might be best served by them.

The mHealth literature on sleep is even sparser. Rather than evaluating interventions, most studies focus on the difficulties of measurement itself or specific sleep disorders. Jalali and Bigelow (2015) provide an overview of current technologies with a focus on insomnia; Kishimoto et al. (2006) focus on the difficulty of measuring sleep posture; Behar et al. (2015) focus on deriving an algorithm for automatically detecting sleep apnea. The most relevant study to my endeavor evaluates a randomized intervention aimed at improving the sleep behavior of airline pilots (Van Drongelen et al., 2014). However, their intervention only included the introduction of an app with tailored advice, compared to a control group that were directed to a website with standard advice about fatigue. They find the treatment group improved along self-reported dimensions of fatigue, sleep quality, strenuous physical activity, and snacking behavior. These conclusions, however, cannot be extended far beyond the domain of airline pilots.

In the hopes of unlocking an mHealth-based solution to obesity, I apply tools from behavioral economics to proprietary data from hundreds of thousands of users of one particular commercially available wearable activity tracker to interpret and analyze whether 4 interventions delivered through a smartphone were able to promote physical activity and healthier sleep schedules by temporarily altering users' goals and hence their reference points. To my knowledge, these interventions outpace prior studies' sample sizes by many

[6]For example, Cadmus-Bertram et al. (2015) use a sample of 31 postmenopausal, overweight women to evaluate the effect of a randomized 16-week web-based self-monitoring intervention that included a wearable device, instructional session, and follow-up, comparing their treatment group to a group using a standard pedometer. Thorndike et al. (2014) use a sample of 104 medical residents to evaluate the effect of a randomized 6-week intervention that gave feedback about steps and energy consumed, comparing their results to the impact of using the same wearable without any interactive feedback.

orders of magnitude. The evaluation of a sleep intervention aimed at promoting early bedtimes is almost entirely novel. These notifications, and the temporary goals they proffer (and hence reference points they attempt to set), were customized to either the specific day of the year on which they were sent or a user's recent patterns of physical activity. They represent an innovation relative to prior studies that lack: (1) the ability to target a specific day of the year with sufficient statistical power and (2) comprehensive baseline and recent measures of activity on which to base suggestions. These interventions are fairly straightforward to implement and constitute a low-cost means of potentially mitigating the obesity epidemic in America. Although there are substantive fixed costs in setting up a system to implement such interventions, these costs pale in comparison to the costs of obesity cited above. Moreover, once such a system is in place, the marginal cost of an additional notification is essentially zero.

Interventions were simple: treated users were greeted on the home page of the wearable's smartphone app with a short message challenging them to a new one-day step goal or bedtime goal. Users who accepted the challenge were pushed a notification to the home screen of their smartphones later in the day, reminding them of their commitment. Users who accepted the notification were also able to continue accepting the challenge on subsequent days. Again, temporary goals were specifically chosen to target either the specific day of the year (e.g. Thanksgiving) or how a user had been recently performing. They were purposely chosen to constitute fairly modest improvements over expectations and recent performance, and hence were not overly ambitious (i.e. typically much lower than the default goal and only slightly higher than activity in preceding days). Their power (and novelty) is in framing *not* meeting the temporary goal as a loss, beneficially exploiting loss aversion. If the new reference point were set too high (relative to recent activity), diminishing sensitivity implies that people would not be particularly responsive. However, by setting the reference point only slightly higher than recent activity, diminishing sensitivity implies that people should be responsive.

First, "Step Intervention #1" targeted lack of physical activity on a typically gluttonous holiday: Thanksgiving. The notification suggested maintaining one's typical step average, further nothing the average user walked 1,873 fewer steps ($\sim$1.1 miles) during the prior Thanksgiving. Second, "Step Intervention #2" targeted a user's fourth day in the system by suggesting a temporary goal that was 500 steps ($\sim$0.3 miles) higher than her average steps in the two prior days. Third, "Step Intervention #3" targeted activity during a particularly sedentary week. The notification told a user what her step average had been over the preceding week, noting it was below average, and suggesting a goal 500 steps ($\sim$0.3 miles) higher than this recent, relatively sedentary average. Finally, "Sleep Intervention" targeted

a user's bedtime during a particularly restless week, suggesting a bedtime that, combined with a user's recent wake-up times, would get them within 30 minutes of 8 hours of sleep. Initial day treatment effects were significant for "Step Intervention #1," "Step Intervention #2," and "Sleep Intervention," engendering $\sim$0.1 miles more walking per successful step intervention and 5 minutes additional rest for the sleep intervention. "Step Intervention #3" showed no significant treatment effect. Only the sleep intervention continued to have significant treatment effects beyond the initial day, engendering about a minute more reported sleep per night for as long as a user remained in the system. This semi-permanent increase appears to have been driven by the initial notification itself, rather than subsequent follow-up notifications.

These results were robust to using as the outcome variable (instead of steps or minutes of sleep) an indicator for whether a user met her *default* goal of 10,000 steps or 8 hours of sleep that day. Results were also robust to using instead an indicator variable for whether a user met her *temporary, notification-based* goal. In addition to replicating the results of the main and default specifications, the notification-based goal specification shows additional evidence of responsiveness to both "Step Intervention #2" (lasting responsiveness to the temporary goal for 6 days, including the initial day) and "Step Intervention #3" ($\sim$0.3% increased probability of meeting the temporary goal for 2 days, including the initial day).

Overall, my results are consistent with users' reference points being temporarily updated by modest, notification-based goals. These goals positively exploit loss aversion by framing *not* meeting the temporary improvement as a loss, promoting physical activity in the short-term and increased reported sleep in the medium-term.

1.3 Sample

My sample consists of a random subset of iPhone or Android smartphone users who purchased (or were gifted) a particular brand's consumer wearable activity tracker during 2013 and 2014. The sample was overrepresented[7] by people who were (1) healthier in terms of BMI, (2) city-dwelling, (3) in their mid-30's, and (4) presumably wealthier (given they purchased or were gifted a wearable activity tracker). The sample is also presumably more motivated in terms of living healthy lives. All members of the samples I analyze use identical technologies and apps. This enables me to avoid any confounding that might arise

[7]Compared to the U.S. Census Bureau's Current Population Survey (Annual Social and Economic Supplement, 2012). The Census indicates that 1.8% of the population is underweight (BMI less than 18.5), 31.2% of the population is healthy weight (BMI between 18.5 and 25.0), 34.0% overweight (BMI between 25 and 30.0), and 33.0% obese (BMI greater than or equal to 30). My wearable sample was overrepresented by underweight and healthy individuals but underrepresented by obese individuals.

from differences in software, hardware, or experimental platform. These differences cannot be ruled out as driving differences in treatment responses among different groups in comparing prior, smaller studies aimed at specific-subpopulations.

1.4 Four App-Based Smartphone Interventions

Interventions were triggered by either the day of the year or a certain pattern of recent step or sleep activity. A random subset of users who would otherwise qualify for the treatment were *not* sent an intervention (i.e. the control group, or the untreated group, or the group of withheld users)[8]. Intervention notifications were sent to the home screen of the wearable's companion app in the wee hours of the morning on the day of the intervention, challenging the user to a new one-day goal. Treated users were able to accept, decline, or ignore the challenge (and could easily navigate other features of the app while ignoring the notification/challenge, which remained on the home screen of the app). See Table 1.1 for an overview of triggering conditions for each intervention and the content of the notifications. See Table 1.2 for the sizes of treatment and control groups, as well as the number of days each intervention was fielded. Note "Step Intervention #1" had two different treatment groups that only differed slightly in the wording of the notification.

If a treated user accepted the challenge on the home screen of the smartphone app, the user was sent a push notification reminder of their commitment to the home screen of their smartphone (not to the home screen of the app) at either 4pm for each of the step interventions or an hour before promised bedtime for "Sleep Intervention." If and when the intervention was accepted and completed, the user received another push notification congratulating them. Therefore, a user received a total of 3 notifications if they accepted and achieved the goal, 2 notifications if they accepted and did not achieve the goal, and 1 notification just for being in the treatment group.

A user could be part of at most one treatment or one control group once per week. After accepting an intervention, and either achieving or failing the challenge, a user could also have agreed to "chain," and thus take on the challenge again the next day. This could continue indefinitely, where a "chain" (if it lasted more than a week) overrode subsequent potential treatment or control group assignment with respect to other interventions. The treatment was therefore only multi-day to the extent that the user continued to accept app-based challenges. It is important to highlight that these subsequent notifications do *not* have another randomized control group associated with them. All subsequent-day analyses

[8]Balancing tests confirm no significant differences between control and treatment groups with respect to step and sleep behavior leading up to the intervention, age, BMI, gender, and days as a user.

Intervention	Conditions	Message
Step Intervention #1	Thanksgiving 2013.	We slow down to give thanks. Users averaged 1,873 fewer steps last Thanksgiving. [v1: Stay at flock's front by meeting your X step average.] [v2: Stay at flock's front by maintaining your average today.]
Step Intervention #2	4th day as a user and 2-day average steps is more than 500.	Aim for X steps today? That's 500 more than your 2 day average. We'll keep track of your journey.
Step Intervention #3	Activity in at least 4/7 past days, user's 7-day non-zero daily step average is 80% or less than average.	You haven't been your active self lately. Your 7 day step average of average of X is less than usual. Start fresh with 500 extra steps today?
Sleep Intervention	User's 3-day average total sleep is 20% less than their sleep goal (with sleep logged all 3 nights). It is not a Friday nor Saturday.	You've been turning in late recently. Remember, your brain needs plenty of pillow-time to sort new information. Get in bed by [(average wakeup time) - (sleep goal in hours) - (30 minutes)] tonight?

Table 1.1: Summary of conditions that triggered each intervention, as well as the messages sent to users through the companion smartphone application, indicating exactly how users are nudged towards more steps or sleep.

are therefore based on the original control group randomization from the initial day of the intervention.

1.5 Methodological Issues

There are a number of methodological issues worth keeping in mind before presenting my results. First, with respect to the sleep intervention, it is important to understand exactly how sleep is measured. The process was not completely automatic. Users were foremost encouraged to track their sleep by putting the device into sleep mode right before going to sleep and putting it back into awake mode after waking up. However, the wearable was sophisticated enough to guess when a person was asleep even if sleep mode was never activated. In this case, the user was sent a push notification to the home screen of their smart-

Intervention	Days Implemented	Treatment	Control
Step Intervention #1	1	17,509+17,131	17,988
Step Intervention #2	22	23,626	24,230
Step Intervention #3	194	556,127	138,597
Sleep Intervention	19	63,361	15,472

Table 1.2: Summary of number of days each intervention was implemented, and sample sizes for treatment and control groups. Note that a user only appears at most once in a given intervention. Sample sizes are based on data that has been trimmed only to include treatment and control group members who opened the app on the day of the intervention for "Step Intervention #2," "Step Intervention #3," and "Sleep Intervention." The data to do this trimming for "Step Intervention #1" was not available. Note also there are two slightly different "Step Intervention #1" treatment versions ("v1" and "v2") that differ in whether they present users with their exact historical daily step average.

phone in the morning (after the device detected the person was awake) asking whether they were indeed asleep during the proposed time. Users were able to verify the proposed sleep duration or edit it to fix any mistakes. For any particular observation, I cannot tell which method was used to generate that data point. Related, observed treatment effects from the sleep intervention could be driven by any combination of at least three mechanisms: (1) increasing actual sleep; (2) better reminding users to put the device into sleep mode, thereby only better capturing existing sleep; and (3) encouraging users to over report, i.e. putting the device into sleep mode for longer than they actually slept. I am unable to delve further into which of these three forces drive observed treatment effects. I therefore simply discuss the effect of the interventions on *reported* sleep.

Second, it is important to highlight that causal estimates derived from these interventions are not directly comparable across different interventions given different selection criteria. Once a user's recent activity matches the pattern for a given intervention, she is either randomly given the intervention or not. However, there *is* selection into the intervention itself (i.e. the triggering conditions), thereby limiting external validity.

Third, although I am eager to interpret positive treatment effects for step interventions as evidence of increased physical activity, it is possible that there are substitution effects between steps and other unmeasured physical activities. For example, after receiving one of our step interventions, it is possible a person decided *not* to go bike riding (as she might usually) and instead only walked a bit more. To the extent that these substitution effects exist, our causal estimates overestimate how much the interventions promote overall physical activity. Because we cannot measure all physical activity or exercise a user might engage in, I can do little to directly test how much of an issue these substitution effects might have been in practice. In contrast, substitution across time within the categories

sleeping or walking, e.g. I walk more today so I might slack off with respect to walking tomorrow, can indeed be investigated. I find no evidence of such temporal substitution.

Fourth, and related to the previous point, it is possible that interventions have effects across time on *other* activities not directly targeted by the interventions. For example, sleeping a little more one night might positively or negatively impact steps the next day (a well-rested person might be more active, but sleeping more literally takes time away from potentially walking). For another example, walking more today might make someone more tired, leading them to sleep more that evening (the effect could also work in the other direction). Toward these ends, we could use step treatment status today as an instrument for measuring the impact of steps today on sleep tonight. Similarly, we could use sleep treatment status tonight as an instrument for measuring the impact of sleep tonight on steps tomorrow. While I believe these are exciting ideas, these analyses are better suited for follow-up studies. For the remainder of this paper, I therefore ignore cross-activity substitution effects.

Finally, it is important to acknowledge that the interventions investigated herein were not necessarily ideal from a survey design perspective. For example, Liao et al. (2015) develop a just-in-time micro-randomized paradigm for mobile interventions in which each participant might be part of thousands of randomizations located sequentially in time, enabling the detection of proximal treatment effects. The interventions I describe must be taken as given and do not employ such sophisticated randomization strategies. While simplicity facilitates communication and understanding of my findings, no doubt one could have tested a more aggressive or complex set of interventions that varied, for example, the exact wording of the notifications, the difficulty of the associated challenge, or the number and frequency of notifications. One could have also better designed the interventions to directly inquire about prior reference points, facilitating even better personal customization. Randomizing the modesty of the notifications' suggestions (i.e. the extent of the push suggested by the challenges) would have also been advantageous. It is likely that some designs could have yielded significantly larger[9] average treatment effects. It is therefore remarkable that despite the unobtrusive, light-handed nature of the interventions, we are able to find significant effects on reference points and behavior.

[9]This highlights the tradeoff between researcher-designed experiments, which requires timely and costly data collection but allow for greater control over experimental design and sample representativeness, and corporate-designed experiments, which offer much greater statistical power and timeliness at the cost of control over experiment design and recruitment.

1.6 Average Treatment Effects

Randomization allows for straightforward evaluation of the causal impact of the interventions. Following standard notation for the analysis of treatment effects, I let $y_i(1)$ and $y_i(0)$ be the behavior of individual i under treatment and control, respectively. Let T_i be an indicator for treatment status. My baseline analysis simply uses OLS to regress y_i on T_i:

$$y_i = \beta_0 + \beta_1 T_i + \epsilon_i \tag{1.1}$$

More specifically, i indexes all users who met conditions in Table 1.1 from both treatment and control groups; y_i denotes the outcome variable targeted by the treatments, either current daily steps or minutes of sleep. $\hat{\beta}_1$ therefore measures the mean difference in outcomes between treatment and control, i.e. $E(y_i(1)|T_i = 1)$–$E(y_i(0)|T_i = 0)$. Because T_i is randomized, we can further assume $E(y_i(0)|T_i = 1) = E(y_i(0)|T_i = 0)$ and $E(y_i(1)|T_i = 0) = E(y_i(1)|T_i = 1)$. We can therefore also assume that average treatment effects (ATE) equal average treatment effects on the treated (ATET), which also equal average treatment effects on the control (ATEC). For all interventions except "Step Intervention #1," I trim both treatment and control groups to users who actually opened the app on the day the notification was first sent. Note that users cannot view the notification without opening the app. Data on whether a user opened the app or not was *not* available in 2013[10]. The difference amounts to measuring an intent-to-treat (ITT) effect in "Step Intervention #1" (from Thanksgiving 2013) but an average treatment effect on the treated (ATET) in the other interventions (from 2014). In fact, for interventions where both sorts of treatment effects can be measured, ITTs are all only slightly lower than ATETs. I have omitted these results for brevity; I only report ITTs when the meta-data is unavailable, i.e. for "Step Intervention #1."

I test the robustness of these estimates by adding covariates including age, gender, BMI, as well as state and date fixed effects. In follow-up analyses measuring any lingering impact of the initial notification days later, I use y_i taken from the desired number of days following the initial intervention. It is again important to highlight that any subsequent notifications after the initial day do *not* have another randomized control group associated with them. All subsequent-day analyses are therefore based on the original control group randomization from the initial day of the intervention.

Prior literature on mHealth interventions for physical activity and sleep cannot strongly inform our prior on the size of $\hat{\beta}_1$. As argued above, this owes to the multifaceted differ-

[10] On a small number of random days in 2014 (sometimes only impacting Android devices and not iPhones) this data is also not available. I've simply thrown away all observations from these days from both platforms.

ences between past experimental designs and the one presented here, as well as the general lack of research on sleep interventions. The following example serves to highlight the difficulty of comparing mHealth interventions on different platforms and with different samples to predict the magnitude of treatment effects. Wang et al. (2015) reports the effect of smartphone notifications on steps, finding that 3 text messages per day for 6 weeks engendered 1,266 extra steps per day only during the first week. Their setup differs from my interventions in several ways, all of which push our expectations of $\hat{\beta}_1$ below 1,266 extra steps per day for a week. My interventions consisted of only a single in-app notification (which was not pushed to the home screen of the smartphone and hence required opening the app to view it) and a single push notification to the home screen of the smartphone (conditional on accepting); notifications on subsequent days were only possible if a user accepted the initial in-app notification and elected to continue "chaining" the challenges. This is in contrast with Wang et al. (2015)'s guaranteed 3 text messages per day. Their notifications were also more salient given all 3 were pushed to the home screen of the smartphone. Moreover, their 3 text messages were guaranteed to be repeated everyday for 6 weeks. Therefore, I expected significant, albeit smaller effects than Wang et al. (2015) on the day of the notification itself, and small, if any, effects on the days following the initial notification. Prior research unfortunately provides little in the way of expectations for the performance of my sleep intervention.

Initial day treatment effects were significant for "Step Intervention #1," "Step Intervention #2," and "Sleep Intervention," engendering $\sim$170 and $\sim$160 additional steps per successful step intervention ($\sim$0.1 miles), respectively, and $\sim$5 minutes additional rest for the sleep intervention. Only the sleep intervention continued to have significant treatment effects beyond the initial day, engendering about a minute more reported sleep per night for as long as a user remained in the system[11].

1.7 Additional Analyses

Our large sample enabled several additional follow-up analyses that were not possible in previous, smaller studies. First, I explore heterogeneous treatment effects with respect to gender, age, BMI, days as a user, day of week, and recent average activity. Note that information on age, gender, and BMI is based on self-reported data entered by users through the app[12]. There was substantive treatment effect heterogeneity with respect to day of the

[11]The interventions do not have any impact along the extensive margin, i.e. treated users do not end up logging activity for more days than control users. In other words, there is no differential attrition from the interventions.

[12]Users entered their weight and height from which BMI was calculated.

week: Thursdays and Saturdays drove the effectiveness of "Step Intervention #2"; Tuesdays, Sundays, and Wednesdays drove the effectiveness of "Sleep Intervention" (note this intervention was not fielded on Friday nor Saturday nights by design). On the other hand, I find limited treatment effect heterogeneity with respect to gender, age, BMI, days as a user, and recent average activity.

I also evaluate who was most likely to accept the interventions, with respect to gender, age, BMI, days as a user, and evaluate the extent to which these effects were consistent across days of the week. There was substantive heterogeneity in challenge acceptance: women, younger users, and higher BMI users were more likely to opt into the challenges. Interestingly, unlike main treatment effects, acceptance effects were consistent across days of the week. Number of days as a user was also robustly associated with propensity to accept the notifications; I must be careful to control for selective attrition (as a proxy for motivation) in these analyses. Without further controls, those who had been in the system the longest would tend to be the most motivated (by virtue of the fact that relatively less motivated users would have already discontinued use), and hence were most likely to accept a notification's challenge. However, to correct for this selection issue, we can control for total days the user ends up being a user for (only known ex post). After this control, more experienced users were slightly less likely to accept a notification's challenge.

See "Appendix: Demographic Mediators of Heterogeneity" for additional analyses of how these findings mesh with predictions inferred from relationships between broader concepts plausibly related to responding to smartphone notifications (openness to experience, comfort with technology, conscientiousness, and risk aversion) and demographics.

I also delve into quantile treatment effects with respect to the distribution of daily steps and minutes of sleep. This enables us to see how different ends of the activity and sleep distributions are impacted by the interventions. For example, it is conceivable that people who sleep very little might have responded to being reminded to get to bed early much more than those who sleep closer to 8 hours.

Moreover, instead of investigating the impact of the interventions directly on a continuous measure of steps or minutes of sleep, I also analyze two threshold specifications meant to judge the extent to which two potential reference points were salient. These act as robustness checks on my main specification.

The first threshold specification uses an indicator variable as the outcome variable for whether a person reached the *default* goal (i.e. not set by the intervention) of 10,000 steps or 8 hours of sleep. This tests for the saliency of the default goal in the treatment group versus the control group. These goals are especially ambitious given the fact that the default goals are typically much higher than recent activity. Results are entirely consistent with the main

specification. "Step Intervention #1," "Step Intervention #2," and "Sleep Intervention" engendered $\sim$1.0%, $\sim$1.1%, and 1.5% increased probabilities, respectively, of meeting the default goal. "Step Intervention #3" continued to have no impact. The effect of "Sleep Intervention" again lingered for as long as the user remained in the system.

The second threshold specification uses an indicator as the outcome variable for whether a person reached the *temporary, notification-based* goal. This tests for the saliency of the goal suggested by the intervention, which, by definition, was only revealed to treated users, and was designed not to be overly ambitious relative to recent activity. Results are again consistent with those found in the main specification, although this specification showed additional evidence of responsiveness to both "Step Intervention #2" (lasting responsiveness to the temporary goal for 6 days, including the initial day) and "Step Intervention #3" ($\sim$0.3% increased probability of meeting the temporary goal for 2 days, including the initial day). Unfortunately, missing data disallowed replication of this goal-based threshold analysis for "Bedtime Steps." However, because this intervention had a lasting, direct impact on sleep, both with respect to the continuous outcome variable and the threshold relative to the default goal, this follow-up analysis is less important.

1.8 Results

Most step notifications were effective on the initial day of the intervention, engendering $\sim$0.1 miles worth of walking on that day (the exception was "Step Intervention #3"). The sleep intervention, "Sleep Intervention," engendered about 5 minutes of sleep on the initial day, as well as a permanent additional minute of reported sleep per night as long as the person remained a user. See 1.3 for these initial day treatment effect estimates, as well as the percentage of users who accepted, declined, and ignored[13] the notifications.

1.8.1 Step Intervention #1

"Step Intervention #1," fielded on Thanksgiving 2013, caused those who received the notification to walk on average 173.1 steps more on the day of the initial notification (Figure 1.1). This point estimate is stable to the inclusion of demographic covariates and state fixed effects.

[13]It is important to remember that for "Step Intervention #1," the "ignore" group includes both people who did not open the app as well as people who opened the app but did not explicitly accept nor decline (because data was not available from 2013 denoting who opened the app on a particular day). For all other interventions, the "ignore" group only included people who opened the app but did not explicitly accept nor decline. However, results for these 3 interventions do not differ substantially when we also include people who did not open the app

Intervention	Treatment Effect	Accept	Decline	Ignore
Step Intervention #1	173.1*** steps	13.8%	2.3%	83.9%
Step Intervention #2	161.7*** steps	26.2%	6.6%	67.2%
Step Intervention #3	12.4 steps	11.8%	3.6%	84.6%
Sleep Intervention	4.94*** minutes	22.5%	4.6%	72.9%

Table 1.3: Summary of initial-day treatment effects for the 4 interventions. Only "Step Intervention #3" had an insignificant treatment effect (all others were significant at the 0.0001 level, indicated by ***). Because of limited data availability on who opened the app, the "ignore" category for "Step Intervention #1" includes both those who looked at the notification and ignored it and those who never opened the app. In other interventions, the "ignore" category only includes the former, given availability of data on who opened the app. Therefore, the treatment effect for "Step Intervention #1" is an ITT and for all others are ATETs.

Quantile regressions at the 5th, 25th, 50th, 75th, and 95th percentiles of the step distribution indicate treatment effects were insignificant at low ends of the distribution, most strongly significant at the median, still significant but lower in magnitude at the 75th percentile, and marginally significant (but larger in magnitude) at the 95th percentile (Figure 1.4).

I augment my baseline specification to include gender, BMI, age, and total days as a user (the latter three standardized) as well as an interaction term for each with treatment status (Figure 1.2). The only covariate that showed evidence of mediating treatment effects was BMI: the treatment was most effective for those with lower BMI, on the order of 20 steps per standard deviation of BMI. I additionally test for heterogeneity with respect to average steps over the six days leading up to the notification, see Figure 1.3. Those with higher pre-treatment step averages were more impacted by the intervention: a one-standard deviation increase in preceding average steps (around 3,790 steps) was associated with a 92-step larger treatment effect.

Recall that "Step Intervention #1" actually consisted of two different treatment groups (which are simply lumped together in the above analyses). The first encouraged a user to maintain his or her daily average and included what that average actually had been ("Stay at flock's front by meeting your X step average."), while the second omitted the latter bit of information ("Stay at flock's front by maintaining your average today."). However, the effect on steps between these treatments was statistically indistinguishable even after controlling for age, BMI, and gender, as well as state fixed effects (Figure 1.5.

13.8% of users explicitly accepted the initial notification. 2.3% of users declined the initial notification. 83.9% of users ignored the initial notification (or did not open the app). Users who accepted the notification tended to be female, younger, and higher BMI, and

had been users for longer at the time of the intervention (Figure 1.6). Given that users who remained in the system were likely to be particularly motivated (by virtue of the fact that less motivated users would naturally drop from the sample), I additionally control for the total eventual number of days for which each user ended up being in the system (which is only known ex post given we must wait for all users to attrite to calculate this number). Point estimates with respect to age, gender, and BMI remain stable, but the sign of the effect of total days in the system at the time of the intervention switches: more experienced users were less likely to accept the notification. Interestingly, I find these results involving propensity to accept the notification also hold for the other three interventions.

I also split the treatment group into those who chose to accept, decline, or ignore the notification on the day of the initiation notification and calculate differences in mean steps between each of these groups and the control group (Figure 1.7). These results cannot be interpreted causally. Indeed, people sort themselves into these categories based on demographics, the kind of day they expect to have ("There is no way I'm going to make this goal, so I am not going to try" sort of logic), their motivation on that day, the weather, etc. Acceptance is significantly associated with over 2,235 steps more per person relative to the control group; declining with 811 fewer steps; ignoring with 139 fewer steps.

Recall that contingent upon accepting the initial notification, a user was given the same notification the next day (i.e. the day after Thanksgiving), which they may or may not again accept. This "chain" continued as long as a user continued to accept challenges. Conditional on accepting the first notification, 45.2% of users accepted the second. Conditional on accepting the second, 32.5% accepted the third. The remaining percentages (up to the sixth) were 43.4%, 51.5%, 65.6%, and 70.9%, respectively. However, on these subsequent days, there were no new randomizations (i.e. no new control groups from which we randomly withheld notifications). I can therefore only compare the activity of the treatment group X days after the initial notification to the control group X days after the initial notification. At lags of 1, 2, 3, 4, 5, 6, and 14 days, there are no statistically significant effects. This remains the case if I include demographics or state fixed effects as covariates (as well as if I include relevant treatment effect interactions). I also consider lagged treatment effect heterogeneity with respect to average steps over the 6 days preceding the initial intervention, yielding no significant results. Consistent with these findings, there is no significant difference in total number of steps (across all future days) or total steps per day (again across all future days) between treatment and control groups (with or without the full gamut of controls).

The default goal threshold specifications, which use an indicator variable for whether a user reaches 10,000 steps as the dependent variable, are broadly consistent with my main

specifications. Being in the treatment caused a 1.01% increased probability of reaching 10,000 steps on the day of the notification (Figure 1.8, robust to the inclusion of demographic covariates and state fixed effects). However, heterogeneity in treatment effects with respect to BMI disappears (Figure 1.9). This is intuitive: higher BMI people walk less on average and therefore struggle more to meet 10,000 steps, even with the nudge of the intervention. There is no lasting impact in terms of likelihood of exceeding the default goal on subsequent days.

The intervention-based goal threshold specifications, which use an indicator variable for whether a user reaches their daily average (i.e. the goal suggested by "Step Intervention #1"), are also consistent with the main specification. Being in the treatment causes a 1.76% increased probability of reaching one's daily average (Figure 1.10, also robust to inclusion of demographics and state fixed effects). There is only marginally suggestive evidence of heterogeneity with respect to BMI. There is also no evidence of any lasting impact of the intervention in terms of likelihood of exceeding the notification-based goal.

1.8.2 Step Intervention #2

"Step Intervention #2," which was fielded on a user's fourth day in the system, caused those who received the notification and opened the app to walk on average 161.7 steps more on the day of the initial notification than their control group counterparts who also opened the app (Figure 1.11). This result is robust to including demographic covariates and state fixed effects. Including date fixed effects decreases the magnitude of the treatment effect somewhat, indicating day-specific effects were important.

Quantile regressions at the 5th, 25th, 50th, 75th, and 95th percentiles of the step distribution indicate that treatment effects were fairly stable and significant throughout the distribution (slightly stronger at the median), except at the very upper end of the distribution, where they were insignificant (Figure 1.14). This finding is intuitive: users who are most active are less likely to be influenced by a small nudge.

I augment my baseline specification to include gender, age, and BMI (the latter two standardized) as well as an interaction term for each with treatment status (Figure 1.12). I cannot consider days as a user for this intervention because it was always pushed on a user's fourth day in the system. The treatment was slightly more effective for older users.

I also split the analysis by day of the week (Figure 1.13). Results indicate effects on Thursdays and Saturdays were responsible for the average treatment effects (spurring 274.8 and 402.1 steps on the initial day of the notification, respectively). I am unable to think of any particularly sound intuition for the Thursday finding. Saturday's finding is

more intuitive: with more free time, already-motivated people are particularly responsive. This effect likely does not exist on Sunday because users sense the immanency of Monday and the waning of the weekend's freedom, and are therefore less responsive to suggestions of increased physical activity.

I next test for heterogeneity with respect to average steps over the two days leading up to the notification (i.e. their first full two days in the system). The relevant interaction term is insignificant (results omitted).

26.2% of users explicitly accepted the initial notification. These users again tended to be female, younger, and higher BMI (Figure 1.15). 6.6% of users declined the notification. 67.2% of users ignored the notification (but indeed looked at the app). These results are completely consistent with those for "Step Intervention #1." Although I cannot analyze heterogeneity with respect to days as a user (given this notification is always pushed on a user's fourth day), I include additional specifications controlling total eventual number of days for which each user ends up being in the system for consistency with other interventions' analyses. Interestingly, despite heterogeneity in treatment effects with respect to day of the week, results involving propensity to accept the notification are consistent across the week.

I again split the treatment group into those who chose to accept, decline, or ignore the initial notification, showing differences in mean steps between each of these groups and the control group (Figure 1.16). These estimates cannot be interpreted causally for the same reasons as in "Step Intervention #1." Acceptance was significantly associated with 859 steps more per person relative to the control group (point estimates for declining and ignoring are negative, but insignificant).

As before, contingent upon accepting the challenge on the initial day, a user was given the opportunity to accept the same challenge the next day. The "chain" continued as long as a user continued to accept challenges. Conditional on accepting the first notification, 31.6% of users accepted the second. Conditional on accepting second, 32.7% accept the third. Conditional on accepting the third challenge, 43.8% accept fourth. However, on day 8 as a user, i.e. potentially the fifth day of the intervention, users who were initially in the treatment group were *all* sent the notification as though they had continued to chain to that point. 18.9% accepted this notification. Conditional on accepting this notification, 36.1% accepted the next day; 37.5% accepted on the subsequent day. On day 11 as a user, i.e. potentially the eighth day of the intervention, all users were again pushed the intervention as though they had been chaining. 17.1% accepted this notification. Users continue to chain in similar proportions. However, because there were no new randomly withheld notifications at any point during this process (except on initial day of the notification), I

can only compare the activity of the treatment group X days after the initial notification to the control group X days after the initial notification. At lags of 1, 2..., 13 days, there are no statistically distinguishable effects. In contrast to the successful initial notification sent initially on day 4, the follow-up notifications sent to all active users on days 8 and 11 were ineffective. These conclusions continue to hold if I include state fixed effects, or gender, age, and BMI as covariates (with interaction terms or not). They also hold if I control for average steps over 2 days prior to the initial intervention (again including an interaction term or not). Consistent with these findings, there is no significant difference in total number of steps (across all future days) or total steps per day (again across all future days) between treatment and control groups (with or without the full gamut of controls). I omit these null findings for brevity.

The default goal threshold specifications are broadly consistent with the main specification. Being in the treatment caused a 1.19% increased probability of reaching 10,000 steps (Figure 1.17). Results are robust to the inclusion of demographic covariates, state fixed effects, and date fixed effects. However, heterogeneous treatment effects with respect to age disappear at a 5% significance level (Figure 1.18).

The notification-based threshold specifications are also broadly consistent with the main specification, albeit with additional evidence of responsiveness. Being in the treatment caused a 2.91% increased probability of reaching the temporary goal (Figure 1.19). This is robust to the inclusion of demographic covariates, state fixed effects, and date fixed effects. Heterogeneity with respect to age remains, indicating a one standard deviation increase in age is associated with a 0.12% increased likelihood of exceeding the goal. (Figure 1.20). These specifications, unlike the main or default specifications, show lasting responsiveness for up to *five* days after the initial notification, though with declining effect sizes relative to the initial day (Figure 1.21. This is consistent with the temporary reference point being salient among treatment group members for a several days even sans an improvement in average steps on those days.

1.8.3 Step Intervention #3

"Step Intervention #3," fielded after a particularly sedentary week, did not spur activity among those who received the notification and opened the app on the initial day of the notification as compared to their control group counterparts who also opened the app (Figure 1.22). Treatment effects remain insignificant if I include demographic covariates, state fixed effects, or date fixed effects. Including main and interaction terms for gender, age, BMI, and days as a user also yielded insignificant estimates. Splitting the analysis by day

of the week did not indicate any particular day had significant treatment effects. There were also no significant treatment effects when I split the sample into seasons of the year. Moreover, testing for heterogeneity with respect to average steps over the seven days leading up to the notification did not yield anything significant. Given that the exact pattern of activity leading up to the intervention might matter, I also tried stratifying the sample by how many of the seven days leading up to the intervention had exactly zero recorded activity; each stratum had insignificant treatment effects. This remained true if I instead stratified by how many of the seven days leading up to the intervention had fewer than 500 steps. I omit these null results for brevity.

11.8% of users explicitly accepted the intervention, while 3.6% of users declined the notification and 84.6% ignored the notification (but indeed looked at the app). Users who accepted the notification again tended to be female, younger, higher BMI, and were in the system longer at the time of the intervention (Figure 1.23). Point estimates with respect to age, gender, and BMI remain stable when I additionally control for total eventual number of days for which each user ended up being in the system (i.e. controlling for selection), but the sign of the effect of total days in the system at the time of the intervention switches: more experienced users were less likely to accept the notification after controlling for tenure. Results with respect to propensity to accept are consistent across seasons of the year and again across days of the week.

I next split the treatment group into those who chose to accept, decline, or ignore the notification, showing differences in mean steps between each of these groups and the control group (Figure 1.24). Again, these results cannot be interpreted causally. Acceptance is significantly associated with 1106 more steps per person relative to the control group; declining 478 more steps; and ignoring 156 fewer steps.

Chaining for this intervention worked just like "Step Intervention #2" except there were no days on which the notifications were re-sent to everyone regardless if they had been chaining to that point. Unsurprisingly, there are no significant lagged treatment effects.

The default goal threshold specifications also yielded nothing regardless of the inclusion of covariates and interactions (both for the day of the initial notification and subsequent days). Stratifying the sample by the number of days with zero recorded activity over the preceding week also yielded nothing. However, the notification-based goal threshold specifications show marginally significant effects on the initial day (0.30% increased probability of meeting the goal) and one day later (0.29% increased probability of meeting the goal). These effects were driven entirely by people who had *zero* days of missing data in the week preceding the intervention (with effect sizes of 0.80% and 0.59%, respectively, at $p < 0.0001$). Taken at face value, these results are consistent with the temporary reference

point being salient among particularly active treatment group members for a couple days even without engendering a direct improvement in average steps in that group.

1.8.4 Sleep Intervention

"Sleep Intervention," which targeted an earlier bedtime after a particularly restless few days, spurred an additional 4.94 minutes of reported sleep on the evening of the initial notification among those who received the notification and opened the app as compared to their control group counterparts who also opened the app (Figure 1.25). This effect was almost entirely driven by earlier bedtimes (not waking up later). Treatment effects remain stable with the inclusion of demographic covariates, state fixed effects, and date fixed effects.

Quantile regression at the 5th, 25th, 50th, 75th, and 95th percentiles of steps indicate that treatment effects are insignificant at the extremes of the distribution and are strongest at the 75th percentile (see Figure 1.28). In other words, people who already tended to get a bit more sleep (but not too much) were most responsive.

I detect no treatment effect heterogeneity with respect to gender, age, nor BMI (Figure 1.26). I also test for heterogeneity with respect to average minutes of sleep over the seven evenings leading up to the notification; the interaction term is insignificant. I next split the analysis by day of the week, excluding Fridays and Saturdays, since notifications were never sent on those days (Figure 1.27). Mondays and Thursdays had insignificant treatment effects. Mondays are usually particularly stressful as people recover from being back to work, so it is unsurprising that the intervention was unsuccessful. The lack of responsiveness is particularly surprising on Thursday given we observed particularly strong responsiveness on Thursdays for "Step Intervention #2." Tuesdays had the strongest treatment effects (7.10 minutes), followed by Sundays (6.93 minutes) and Wednesdays (4.25 minutes). Although strong effects on Sunday are fairly intuitive given people generally catch up on sleep before the start of the new workweek, strong effects on Tuesdays and Wednesdays are a bit more surprising.

22.5% of users accepted the notification, while 4.6% of users declined the notification and 72.9% of users ignored the notification (but indeed looked at the app). Users who accepted were again likely to be female, younger, higher BMI, and in the system longer at the time of the intervention (see Figure 1.29). Point estimates with respect to age, gender, and BMI remain stable when I additionally control for total eventual number of days each user ends up being in the system, but the sign of the effect of total days in the system at the time of the intervention again switches. These results are stable across days of the week

and consistent with all prior discussed interventions.

After splitting the treatment group into accepters, decliners, and ignorers, and again noting the following cannot be interpreted causally, I find acceptance is associated with 24.4 more minutes of sleep and declination with 11.6 fewer minutes of sleep (Figure 1.30, with an insignificant coefficient for ignorers).

Chaining for this intervention works just like "Step Intervention #3." Among those who accepted the initial notification, 6.7% accepted the notification on the second day. Conditional on accepting the notification on the second day, 12.8% accepted on the third day. For subsequent days the percentages were 39.7%, 57.3%, and 46.0%, respectively. However, unlike the other interventions, there were significant longitudinal effects (Figure 1.31). One day lagged from the treatment still saw the treatment group reporting on average 3.32 minutes more sleep than the control group; 2 days lagged saw the treatment group reporting 2.12 minutes more sleep; 3 days lagged, 3.55 minutes more; 4 days lagged, 4.03 minutes more; 5 days lagged, 1.62 minutes more (marginally insignificant); 6 days lagged, 2.35 minutes more (back to highly significant); and 14 days lagged, 1.88 minutes more (only significant at the 5% level). Results are robust to the inclusion of state fixed effects, demographic covariates (gender, age, and BMI), and total days as a user at the time of the intervention (not reported for brevity). Including these covariates and their interactions with initial treatment status yield no heterogeneity in lagged treatment effects (also not reported). There was also no heterogeneity with respect to lagged treatment effects and average minutes of sleep over the seven evenings preceding the initial notification. Amazingly, the treatment appears to have a semi-permanent impact on reported sleep behavior, engendering 1.19 minutes of additional reported sleep per evening, averaged across all evenings the user remains active in the system starting with the evening of the initial notification (Figure 1.32. This semi-permanent effect is robust to the inclusion of state fixed effects and demographic covariates.

Two forces could drive these semi-permanent effects: (1) the "chaining" notifications might continually spur earlier bedtimes, or (2) the initial notification alone might spur a persistently earlier bedtime. Given low acceptance rates associated with the chaining notifications, the latter is probably the relevant mechanism. Indeed, only 13 users chain to a seventh notification, while persistent treatment effects are observed well after a week. To directly test this hypothesis, I re-ran the longitudinal analysis excluding treatment group members who ever chain after the initial notification (i.e. anyone who accepts and repeats the challenge). Point estimates and significance levels are almost identical to those in the main specifications, indicating the chaining itself was not the relevant mechanism, but rather the initial notification had a lasting impact on behavior. Of course, as caveated above,

we cannot rule out reporting effects.

The default goal threshold specifications, which use an indicator variable for whether a user reaches 8 hours of sleep as the dependent variable, are broadly consistent with the main specification. Being in the treatment caused a 1.54% increased probability of reaching 8 hours of sleep (Figure 1.33). Results are robust to the inclusion of additional covariates, state fixed effects, and date fixed effects. There continue to be no heterogeneous treatment effects with respect to gender, age, BMI, or total days as a user (Figure 1.34. However, except for Wednesday, heterogeneities with respect to day of the week persist (Figure 1.35). Unfortunately, as mentioned above, missing data disallowed undertaking the goal-based threshold analysis for this intervention. However, the fact that this intervention had a lasting, direct impact on sleep, both with respect to the continuous outcome variable and the threshold relative to the default goal, makes the robustness check in this instance less interesting.

1.9 Difference-in-Differences Robustness Checks

I also employ a difference-in-differences approach to evaluate the robustness of my initial-day average treatment effects. I first evaluate the within-person difference between steps or minutes of sleep on the initial day of the notification and steps or sleep on the day prior. I do this separately for treatment and control groups and take averages, finally computing the difference-in-differences. Results for all interventions are consistent with my main specification's findings.

For "Step Intervention #1," we expect people to walk less on the first day of the notification compared to the day prior. In fact, people in the treatment walked 360 steps on average less on the day of the notification compared to the day before. People in the control walked 434 steps less on average. Both differences are highly significant. The difference-in-differences, 74 steps, has a p-value of 0.06. This is a bit smaller than the main specification's ITT. Users in "Step Intervention #2" walked 54 steps more on average on the day of the notification compared to the day before (p-value of approximately 0.05). People in the control walked 91 fewer steps (p-value less than 0.01). The difference-in-differences, 145, is highly significant and only a little smaller than the main specification ATET. Users in "Step Intervention #3" walked 4,348 steps more on the day of the notification compared to the day before; people in the control 4,346 steps more. This speaks to people naturally "bouncing back" after a few low-step days, i.e. mean reversion (regardless of receiving a notification or not). The difference-in-differences is insignificant. This null finding is consistent with my main specification, which found an insignificant ATET even

on the initial day of the notification. Finally, for "Sleep Intervention," recall we targeted users who had recently not been sleeping well. People in the treatment slept 52 minutes more on the day of the notification as compared to the day prior; people in the control 46 minutes more. The difference-in-differences (6 minutes) is highly significant. This again speaks to people naturally "bouncing back," i.e. mean reversion (this time with respect to minutes of sleep). However, unlike "Step Intervention #3," the "Sleep Intervention" notification furthered that recovery. Notice that the 6-minute effect size is also consistent with the main specification's ATET.

1.10 Conclusions

Smartphone notifications fielded outside of a laboratory setting can indeed harness loss aversion and reference-dependent utility to promote pro-health behavior. Temporary goals were specifically chosen to target either the specific day of the year (e.g. Thanksgiving) or how a user had been recently performing. Moreover, these goals were chosen to constitute fairly modest improvements over expectations and recent performance, and hence were not excessively ambitious (i.e. typically much lower than the default goal and only slightly higher than activity in preceding days). Their power is in framing *not* meeting the temporary goal as a loss. If the new reference point were set too high (relative to recent activity), diminishing sensitivity implies that people would not be particularly responsive. However, setting the reference point only slightly higher than recent activity facilitated responsiveness to the temporary goal. Indeed, treatment effects for "Step Intervention #1" and "Step Intervention #2" were on the order of 0.10 miles on the initial day of the notification, but were not persistent, while "Step Intervention #3" showed no significant effect on physical activity on either initial or subsequent days. Initial treatment effects for "Sleep Intervention" were on the order of 5 minutes of additional rest that night. These sleep-based effects persisted for as long as a user remained in the system, engendering approximately an additional minute reported sleep per night (driven by the initial notification itself rather than follow-up notifications). Analyses investigating the interventions' impact on the probability of exceeding either the default goal or the notification-based goal were broadly consistent with my main findings. The latter showed additional evidence of responsiveness for "Step Intervention #2" (for six days, including the initial notification) and "Step Intervention #3" (for two days, including the initial notification), indicating salience of the reference point even sans an improvement in behavior. Moreover, there was limited treatment effect heterogeneity with respect to gender, age, and BMI, but substantial treatment effect heterogeneity with respect to day of the week. However, users who were female,

younger, and higher BMI were robustly more likely to accept the challenges proffered by the interventions across all days of the week.

To my knowledge, these results constitute the largest mHealth study of physical activity and sleep, and the first to channel tenants of reference-dependent utility and loss aversion toward improving behavior. Results are encouraging, indicating that a small number of smartphone notifications which propose modestly ambitious goals relative to a particular day or pattern of recent activity can temporarily alter reference points. These goals positively exploit loss aversion by framing *not* meeting the temporary improvement as a loss, promoting physical activity in the short-term and increased reported sleep in the medium-term. More broadly, my results indicate that applying tools from behavioral economics to other sorts of mHealth interventions is a promising avenue for future research.

1.11 References and Works Consulted

Abeler, Johannes, et al. "Reference points and effort provision." The American Economic Review (2011): 470-492.

Allen, Eric J., et al. Reference-dependent preferences: Evidence from marathon runners. No. w20343. National Bureau of Economic Research, 2014.

Andersen, R. E. (1999). Exercise, an active lifestyle, and obesity. The Physician and Sportsmedicine, 27(10), 41-50.

Anderson, L. R., & Mellor, J. M. (2008). Predicting health behaviors with an experimental measure of risk preference. Journal of health economics, 27(5), 1260-1274.

Barsky, R. B., Thomas F. Juster, Miles S. Kimball and Matthew D. Shapiro (1997)," Preference Parameters and Behavioral Heterogeneity: an Experimental Approach in the health and Retirement Study. Quarterly Journal of Economics,112(2), 537-580.

Behar, J., Roebuck, A., Shahid, M., Daly, J., Hallack, A., Palmius, N., ... & Clifford, G. D. (2015). SleepAp: An automated obstructive sleep apnoea screening application for smartphones. Biomedical and Health Informatics, IEEE Journal of, 19(1), 325-331.

Broos, A. (2005). Gender and information and communication technologies (ICT) anxiety: Male self-assurance and female hesitation. CyberPsychology & Behavior, 8(1), 21-31.

Bryan, Gharad, Dean Karlan, and Scott Nelson. ”Commitment devices.” Annu. Rev. Econ. 2.1 (2010): 671-698.

BoothKewley, S., & Vickers, R. R. (1994). Associations between major domains of personality and health behavior. Journal of personality, 62(3), 281-298.

Borghans, L., & Golsteyn, B. H. (2006). Time discounting and the body mass index: Evidence from the Netherlands. Economics & Human Biology, 4(1), 39-61.

Borghans, L., Heckman, J. J., Golsteyn, B. H., & Meijers, H. (2009). Gender differences in risk aversion and ambiguity aversion. Journal of the European Economic Association, 7(23), 649-658.

Brummett, B. H., Babyak, M. A., Williams, R. B., Barefoot, J. C., Costa, P. T., & Siegler, I. C. (2006). NEO personality domains and gender predict levels and trends in body mass index over 14 years during midlife. Journal of research in personality, 40(3), 222-236.

Burke, L. E., Styn, M. A., Sereika, S. M., Conroy, M. B., Ye, L., Glanz, K., ... & Ewing, L. J. (2012). Using mHealth technology to enhance self-monitoring for weight loss: a randomized trial. American journal of preventive medicine, 43(1), 20-26.

Cadmus-Bertram, L. A., Marcus, B. H., Patterson, R. E., Parker, B. A., & Morey, B. L. (2015). Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. American journal of preventive medicine.

Card, David, and Gordon B. Dahl. ”Family violence and football: The effect of unexpected emotional cues on violent behavior.” The Quarterly Journal of Economics 126.1 (2011): 103.

Cawley, J., & Meyerhoefer, C. (2012). The medical care costs of obesity: an instrumental variables approach. Journal of health economics, 31(1), 219-230.

Chapman, B. P. (2009). Can the influence of childhood SES on men and women's adult body mass be explained by adult SES or personality? Findings from a national sample. Health psychology: official journal of the Division of Health Psychology, American Psy-

chological Association, 28(4), 419.

Chapman, B. P., Duberstein, P. R., Sörensen, S., & Lyness, J. M. (2007). Gender differences in Five Factor Model personality traits in an elderly cohort.Personality and Individual Differences, 43(6), 1594-1603.

Charness, G., & Gneezy, U. (2009). Incentives to exercise. Econometrica, 77(3), 909-931.

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. Psychological assessment, 4(1), 5.

Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (CREATE). Psychology and aging, 21(2), 333.

Davis, C., Shapiro, C. M., Elliott, S., & Dionne, M. (1993). Personality and other correlates of dietary restraint: An age by sex comparison. Personality and Individual Differences, 14(2), 297-305.

Del Giudice, M., Booth, T., & Irwing, P. (2012). The distance between Mars and Venus: Measuring global sex differences in personality. PloS one, 7(1), e29265.

DellaVigna, S., & Malmendier, U. (2006). Paying not to go to the gym. The American Economic Review, 694-719.

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. Handbook of experimental economics results, 1, 1061-1073.

Ericson, Keith M. Marzilli, and Andreas Fuster. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." The Quarterly Journal of Economics 126.4 (2011): 1879-1907..

Finkelstein, Eric A., et al. "Annual medical spending attributable to obesity: payer-and service-specific estimates." Health affairs 28.5 (2009): w822-w831.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. Journal of economic literature, 351-401.

Handbook of Causal Analysis for Social Research. New York, NY: Springer, 2013.

He, Jun, and Lee Freeman. "Are men more technology-oriented than women? The role of gender on the development of general computer self-efficacy of college students." AMCIS 2009 Proceedings (2009): 672.

Heath, Chip, Richard P. Larrick, and George Wu. "Goals as reference points."Cognitive psychology 38.1 (1999): 79-109.

Hsiaw, Alice. "Goal-setting and self-control." Journal of Economic Theory 148.2 (2013): 601-626.

Huffman, A. H., Whetten, J., & Huffman, W. H. (2013). Using technology in higher education: The influence of gender roles on technology self-efficacy.Computers in Human Behavior, 29(4), 1779-1786.

Ikeda, S., Kang, M. I., & Ohtake, F. (2010). Hyperbolic discounting, the sign effect, and the body mass index. Journal of health economics, 29(2), 268-284.

Jalali, Leila, and Philip Bigelow. "Current Status and Future Trends of Wireless and Mobile Health Technologies in Sleep Medicine: Insomnia Case Study."Mobile Health. Springer International Publishing, 2015. 129-144.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. The journal of economic perspectives, 193-206.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. Journal of political Economy, 1325-1348.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263-291.

Klasnja, P., & Pratt, W. (2012). Healthcare in the pocket: mapping the space of mobile-phone health interventions. Journal of biomedical informatics, 45(1), 184-198.

Kimball, M. S., Sahm, C. R., & Shapiro, M. D. (2008). Imputing risk tolerance from survey responses. Journal of the American statistical Association,103(483), 1028-1038.

Khwaja, A., Silverman, D., & Sloan, F. (2007). Time preference, time discounting, and smoking decisions. Journal of health economics, 26(5), 927-949.

Kishimoto, Y., Akahori, A., & Oguri, K. (2006, September). Estimation of sleeping posture for M-Health by a wearable tri-axis accelerometer. In Medical Devices and Biosensors, 2006. 3rd IEEE/EMBS International Summer School on (pp. 45-48). IEEE.

Koch, Alexander K., and Julia Nafziger. "Selfregulation through Goal Setting*."The Scandinavian Journal of Economics 113.1 (2011): 212-227.

Koszegi, Botond, Matthew Rabin. 2006. A model of reference-dependent preferences. Quarterly Journal of Economics 121(4) 1133–1165.

Koszegi, Botond, Matthew Rabin. 2007. Reference-dependent risk attitudes. The American Economic Review 97(4) 1047–1073.

Koszegi, Botond, Matthew Rabin. 2009. Reference-dependent consumption plans. The American Economic Review 99(3) 909–936.

Ledger, D., & McCaffrey, D. (2014). Endeavour Partners Report: Inside Wearables: How the Science of Human Behavior Change Offers the Secret to Long-term Engagement. Endeavour website.

Lehmann, R., Denissen, J. J., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. Developmental Psychology, 49(2), 365.

Lewis, Z. H., Lyons, E. J., Jarvis, J. M., & Baillargeon, J. (2015). Using an electronic activity monitor system as an intervention modality: A systematic review. BMC public health, 15(1), 585.

Liao, P., Klasnja, P., Tewari, A., & Murphy, S. A. (2015). Micro-Randomized Trials in mHealth. arXiv preprint arXiv:1504.00238.

Loewenstein, G., Weber, R., Flory, J., Manuck, S., & Muldoon, M. (2001, November). Dimensions of time discounting. In Conference on survey research on household expectations and preferences (Vol. 31).

Matthey, Astrid, and Nadja Dwenger. "Don't aim too high: the potential costs of high aspirations." Jena Economic Research Paper 2007-097 (2007).

Milkman, K. L., Minson, J. A., & Volpp, K. G. (2013). Holding the Hunger Games hostage at the gym: An evaluation of temptation bundling. Management Science, 60(2), 283-299.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., ... & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. Proceedings of the National Academy of Sciences,108(7), 2693-2698.

Ogden, Cynthia L., et al. "Prevalence of childhood and adult obesity in the United States, 2011-2012." Jama 311.8 (2014): 806-814.

O'Gorman, J. G., & Baxter, E. (2002). Self-control as a personality measure.Personality and individual differences, 32(3), 533-539.

Patel, M. S., Asch, D. A., & Volpp, K. G. (2015). Wearable devices as facilitators, not drivers, of health behavior change. Jama, 313(5), 459-460.

Patel, S. R., & Hu, F. B. (2008). Short sleep duration and weight gain: a systematic review. Obesity, 16(3), 643-653.

Patel, S. R., Malhotra, A., White, D. P., Gottlieb, D. J., & Hu, F. B. (2006). Association between reduced sleep and weight gain in women. American journal of epidemiology, 164(10), 947-954.

Pope, Devin G., and Maurice E. Schweitzer. "Is Tiger Woods loss averse? Persistent bias in the face of experience, competition, and high stakes." The American Economic Review

101.1 (2011): 129-157.

Porter, C. E., & Donthu, N. (2006). Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. Journal of business research, 59(9), 999-1007.

Rai, A., Chen, L., Pye, J., & Baird, A. (2013). Understanding determinants of consumer mobile health usage intentions, assimilation, and channel preferences. Journal of medical Internet research, 15(8).

Royer, H., Stehr, M. F., & Sydnor, J. R. (2012). Incentives, commitments and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company (No. w18580). National Bureau of Economic Research.

Sapienza, P., Zingales, L., & Maestripieri, D. (2009). Gender differences in financial risk aversion and career choices are affected by testosterone.Proceedings of the National Academy of Sciences, 106(36), 15268-15273.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. Journal of personality and social psychology, 94(1), 168.

Shim, U., Kim, H. N., Roh, S. J., Cho, N. H., Shin, C., Ryu, S., ... & Kim, H. L. (2014). Personality traits and body mass index in a Korean population. PloS one, 9(3), e90516.

Smith, A. (2015). US Smartphone Use in 2015. Pew Research Center.

Smith, A. (2014). Older adults and technology use. Pew Research Center.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. Journal of personality and social psychology, 100(2), 330.

Sutin, A. R., Ferrucci, L., Zonderman, A. B., & Terracciano, A. (2011). Personality and obesity across the adult life span. Journal of personality and social psychology, 101(3), 579.

Suvorov, Anton, and Jeroen Van de Ven. "Goal setting as a self-regulation mechanism." Available at SSRN 1286029 (2008).

Teixeira, Pedro J., et al. "Exercise, physical activity, and self-determination theory: a systematic review." Int J Behav Nutr Phys Act 9.1 (2012): 78.

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., ... & Costa Jr, P. T. (2009). Facets of personality linked to underweight and overweight. Psychosomatic medicine, 71(6), 682.

Thorndike, A. N., Mills, S., Sonnenberg, L., Palakshappa, D., Gao, T., Pau, C. T., & Regan, S. (2014). Activity monitor intervention to promote physical activity of physicians-in-training: randomized controlled trial.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4), 297-323.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. The quarterly journal of economics, 1039-1061.

United States Census Bureau (2012). Current Population Survey. Annual Social and Economic Supplement.

Van Drongelen, A., Boot, C. R., Hlobil, H., Twisk, J. W., Smid, T., & van der Beek, A. J. (2014). Evaluation of an mHealth intervention aiming to improve health-related behavior and sleep and reduce fatigue among airline pilots.Scand J Work Environ Health, 40(6), 557-68.

van Reedt Dortland, A. K., Giltay, E. J., Van Veen, T., Zitman, F. G., & Penninx, B. W. (2012). Personality traits and childhood trauma as correlates of metabolic risk factors: the Netherlands Study of Depression and Anxiety (NESDA).Progress in Neuro-Psychopharmacology and Biological Psychiatry, 36(1), 85-91.

Wang, J. B., Cadmus-Bertram, L. A., Natarajan, L., White, M. M., Madanat, H., Nichols, J. F., ... & Pierce, J. P. Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized

Controlled Trial. Telemedicine and e-Health.

1.12 Figures

	(1) steps	(2) steps	(3) steps
treatment	173.1***	170.7***	181.9***
	(48.03)	(47.42)	(47.75)
age		-28.33***	
		(1.703)	
bmi		-134.9***	
		(4.097)	
male		-87.91	
		(45.32)	
State FE	No	No	Yes
r2	0.000247	0.0296	0.0141
N	52628	52423	52628

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.1: Step Intervention #1, Thanksgiving 2013. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects.

	(1) steps
treatment	199.6**
	(68.01)
male	-50.47
	(77.61)
treatment X male	-55.89
	(95.69)
bmi (std)	-122.4***
	(6.907)
treatment X bmi (std)	-19.30*
	(8.580)
age (std)	-29.10***
	(2.909)
treatment X age (std)	1.205
	(3.591)
system days (std)	0.351
	(0.333)
treatment X system days (std)	-0.567
	(0.410)
r2	0.0298
N	52423

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.2: Step Intervention #1, Thanksgiving 2013. Estimates from regression of daily steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) steps
treatment	123.7**
	(39.27)
recent lagged steps (std)	2953.5***
	(32.08)
treatment X recent lagged steps (std)	92.00*
	(39.40)
r2	0.333
N	52539

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.3: Step Intervention #1, Thanksgiving 2013. Estimates from regression of daily steps on the initial day of the intervention on treatment status, average steps on the preceding 6 days (standardized so it has mean 0 and variance 1), and an interaction term.

	(1) steps_5	(2) steps_25	(3) steps_50	(4) steps_75	(5) steps_95
main					
treatment	47.00	99.00*	202.0**	136.0*	297.0*
	(41.36)	(48.10)	(71.84)	(64.61)	(133.5)
r2					
N	52628	52628	52628	52628	52628

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.4: Step Intervention #1, Thanksgiving 2013. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles.

	(1) steps	(2) steps	(3) steps
treatment12	45.93	45.88	51.18
	(56.48)	(55.71)	(56.15)
age		-27.98***	
		(2.113)	
bmi		-141.6***	
		(5.114)	
male		-109.4	
		(56.13)	
State FE	No	No	Yes
r2	0.0000191	0.0310	0.0155
N	34640	34504	34640

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.5: Step Intervention #1, Thanksgiving 2013. Comparison of the two treatments. Treatment12 is an indicator that takes on a value of 0 if the notification included the person's step average ("Stay at flock's front by meeting your X step average.") and a value of 1 if the message did not ("Stay at flock's front by maintaining your average today."). Column (1) shows mean difference in steps between the treatments. The dependent variable is total steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects.

	(1) accepted	(2) accepted	(3) declined	(4) declined	(5) ignored	(6) ignored
age	-0.00299***	-0.00348***	-0.000404***	-0.000478***	0.00340***	0.00396***
	(0.000140)	(0.000142)	(0.0000611)	(0.0000622)	(0.000149)	(0.000151)
male	-0.0499***	-0.0467***	-0.000384	0.000101	0.0503***	0.0466***
	(0.00371)	(0.00370)	(0.00162)	(0.00162)	(0.00395)	(0.00393)
bmi	0.000814*	0.00104**	0.000000264	0.0000337	-0.000814*	-0.00107**
	(0.000338)	(0.000336)	(0.000148)	(0.000148)	(0.000359)	(0.000358)
system days	0.0000564***	-0.000102***	-0.000000436	-0.0000243**	-0.0000559***	0.000127***
	(0.0000159)	(0.0000181)	(0.00000693)	(0.00000793)	(0.0000169)	(0.0000192)
Total Days	No	Yes	No	Yes	No	Yes
r2	0.0186	0.0278	0.00130	0.00242	0.0199	0.0306
N	34504	34504	34504	34504	34504	34504

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.6: Step Intervention #1, Thanksgiving 2013. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the notifications (mutually exclusive categories). Note that the "ignore" category includes those who look at the app and ignore the notification and those who simply don't look at the app. Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects.

	(1) steps
accepted	2235.3***
	(84.37)
declined	-811.8***
	(188.0)
ignored	-138.8**
	(49.17)
r2	0.0169
N	52628

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.7: Step Intervention #1, Thanksgiving 2013. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally.

	(1) threshold_default	(2) threshold_default	(3) threshold_default
treatment	0.0101*	0.00998*	0.0107*
	(0.00426)	(0.00423)	(0.00425)
age		-0.00147***	
		(0.000152)	
male		-0.0170***	
		(0.00405)	
bmi		-0.00965***	
		(0.000366)	
State FE	No	No	Yes
r2	0.000106	0.0177	0.00811
N	52628	52423	52628

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.8: Step Intervention #1, Thanksgiving 2013. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects.

	(1) threshold_default
treatment	0.0139*
	(0.00607)
male	-0.0121
	(0.00693)
treatment X male	-0.00776
	(0.00854)
bmi (std)	-0.00891***
	(0.000617)
treatment X bmi (std)	-0.00112
	(0.000766)
age (std)	-0.00129***
	(0.000260)
treatment X age (std)	-0.000280
	(0.000321)
system days (std)	0.0000348
	(0.0000297)
treatment X system days (std)	-0.0000345
	(0.0000366)
r2	0.0178
N	52423

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.9: Step Intervention #1, Thanksgiving 2013. Estimates from regression of an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) threshold_goal	(2) threshold_goal	(3) threshold_goal
treatment	0.0176***	0.0176***	0.0185***
	(0.00455)	(0.00456)	(0.00453)
age		-0.000291	
		(0.000164)	
male		-0.00333	
		(0.00436)	
bmi		-0.00287***	
		(0.000394)	
State FE	No	No	Yes
r2	0.000285	0.00151	0.0120
N	52628	52423	52628

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.10: Step Intervention #1, Thanksgiving 2013. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the notification-based goal, only presented to the treatment group (average daily steps). Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects.

	(1) steps	(2) steps	(3) steps	(4) steps
treatment	161.7***	155.4***	159.1***	127.4**
	(45.09)	(44.79)	(45.04)	(45.36)
age		-28.44***		
		(1.829)		
bmi		-101.6***		
		(3.888)		
male		177.5***		
		(44.98)		
State FE	No	No	Yes	No
Date FE	No	No	No	Yes
r2	0.000269	0.0236	0.00637	0.00410
N	47856	47250	47856	47856

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.11: Step Intervention #2, fourth day in the system. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects.

	(1) steps
treatment	102.4
	(63.96)
male	125.2*
	(63.22)
treatment X male	104.7
	(89.96)
bmi (std)	-99.87***
	(5.504)
treatment X bmi (std)	-3.573
	(7.777)
age (std)	-32.17***
	(2.573)
treatment X age (std)	7.516*
	(3.658)
r2	0.0237
N	47250

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.12: Step Intervention #2, fourth day in the system. Estimates from regression of daily steps on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) steps	(2) steps	(3) steps	(4) steps	(5) steps	(6) steps	(7) steps
treatment	136.8	-155.6	19.84	133.7	274.8*	36.98	402.1**
	(123.7)	(121.3)	(119.0)	(109.7)	(117.9)	(116.2)	(125.9)
r2	0.000176	0.000277	0.00000456	0.000195	0.000825	0.0000140	0.00137
N	6950	5937	6102	7606	6583	7245	7433

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.13: Step Intervention #2, fourth day in the system. Estimates from regression of daily steps on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday. . . , Column (7) is Saturday.

	(1) steps_5	(2) steps_25	(3) steps_50	(4) steps_75	(5) steps_95
main					
treatment	151.0*	129.0**	186.0***	164.0**	-59.00
	(67.79)	(40.30)	(53.72)	(59.62)	(140.7)
r2					
N	47856	47856	47856	47856	47856

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.14: Step Intervention #2, fourth day in the system. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles.

	(1) accepted	(2) accepted	(3) declined	(4) declined	(5) ignored	(6) ignored
age	-0.000616**	-0.00137***	0.00100***	0.000831***	-0.000393	0.000531*
	(0.000234)	(0.000237)	(0.000132)	(0.000135)	(0.000249)	(0.000252)
male	-0.0838***	-0.0825***	0.00181	0.00210	0.0817***	0.0802***
	(0.00575)	(0.00572)	(0.00325)	(0.00325)	(0.00614)	(0.00609)
bmi	0.00592***	0.00607***	-0.000256	-0.000222	-0.00570***	-0.00588***
	(0.000493)	(0.000491)	(0.000279)	(0.000279)	(0.000527)	(0.000523)
Total Days	No	Yes	No	Yes	No	Yes
r2	0.0140	0.0252	0.00248	0.00434	0.0124	0.0272
N	23330	23330	23330	23330	23330	23330

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.15: Step Intervention #2, fourth day in the system. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects.

	(1) steps
accepted	859.0***
	(70.02)
declined	-150.1
	(128.7)
ignored	-81.06
	(50.24)
r2	0.00379
N	47856

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.16: Step Intervention #2, fourth day in the system. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored initial notification. These estimates should not be interpreted causally.

	(1) threshold_default	(2) threshold_default	(3) threshold_default	(4) threshold_default
treatment	0.0119**	0.0115**	0.0116**	0.0101*
	(0.00430)	(0.00429)	(0.00430)	(0.00433)
age		-0.00174***		
		(0.000175)		
male		0.00951*		
		(0.00431)		
bmi		-0.00809***		
		(0.000372)		
State FE	No	No	Yes	No
Date FE	No	No	No	Yes
r2	0.000159	0.0143	0.00467	0.00234
N	47856	47250	47856	47856

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.17: Step Intervention #2, fourth day in the system. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects.

	(1) threshold_default
treatment	0.00491
	(0.00613)
male	0.00305
	(0.00605)
treatment X male	0.0130
	(0.00862)
bmi (std)	-0.00766***
	(0.000527)
treatment X bmi (std)	-0.000867
	(0.000745)
age (std)	-0.00207***
	(0.000246)
treatment X age (std)	0.000674
	(0.000350)
r2	0.0144
N	47250

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.18: Step Intervention #2, fourth day in the system. Estimates from regression of an indicator for whether the person met the default goal of 10,000 steps on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) threshold_goal	(2) threshold_goal	(3) threshold_goal	(4) threshold_goal
treatment	0.0291***	0.0291***	0.0291***	0.0260***
	(0.00453)	(0.00456)	(0.00454)	(0.00456)
age		0.000211		
		(0.000186)		
male		0.00470		
		(0.00458)		
bmi		-0.000991*		
		(0.000396)		
State FE	No	No	Yes	No
Date FE	No	No	No	Yes
r2	0.000860	0.00102	0.00239	0.00555
N	47856	47250	47856	47856

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.19: Step Intervention #2, fourth day in the system. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the notification-based goal, only presented to the treatment group (average daily steps over the preceding 2 days + 500 steps). Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects.

	(1) threshold_goal
treatment	0.0317***
	(0.00651)
male	0.00718
	(0.00644)
treatment X male	-0.00528
	(0.00916)
bmi (std)	-0.00132*
	(0.000560)
treatment X bmi (std)	0.000654
	(0.000792)
age (std)	-0.000368
	(0.000262)
treatment X age (std)	0.00117**
	(0.000372)
r2	0.00128
N	47250

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.20: Step Intervention #2, fourth day in the system. Estimates from regression of an indicator for whether the person met the notification-based goal (average steps over the preceding 2 days + 500 steps) on the initial day of the intervention on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) 1d-Lag	(2) 2d-Lag	(3) 3d-Lag	(4) 4d-Lag	(5) 5d-Lag	(6) 6d-Lag	(7) 7d-Lag	(8) 8d-Lag	(9) 9d-Lag	(10) 10d-Lag	(11) 11d-Lag
treatment	0.0125**	0.0122**	0.00976*	0.0104*	0.0128**	0.00704	0.00414	0.00355	0.00548	0.00588	0.00582
	(0.00453)	(0.00453)	(0.00454)	(0.00454)	(0.00453)	(0.00454)	(0.00456)	(0.00456)	(0.00456)	(0.00457)	(0.00457)
r2	0.000159	0.000150	0.0000967	0.000110	0.000167	0.0000501	0.0000173	0.0000127	0.0000301	0.0000346	0.0000339
N	47856	47856	47856	47856	47856	47856	47856	47856	47856	47856	47856

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.21: Step Intervention #2, fourth day in the system. Estimates from regression of lagged indicator for whether the person met the notification-based goal (average steps over the preceding 2 days + 500 steps) on treatment status, for lags of 1 to 11 days. Regressions additionally controlling for state fixed effects, time fixed effects, or age, gender, BMI, and days since becoming a user (not shown) are essentially identical in terms of lagged treatment effect magnitudes and significances.

	(1) steps	(2) steps	(3) steps	(4) steps
treatment	12.35	15.86	12.06	15.03
	(15.47)	(15.39)	(15.44)	(15.36)
age		-29.46***		
		(0.478)		
bmi		-101.3***		
		(1.117)		
male		137.2***		
		(12.39)		
State FE	No	No	Yes	No
Date FE	No	No	No	Yes
r2	0.000000918	0.0205	0.00421	0.0158
N	694724	686587	694724	694724

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.22: Step Intervention #3, after a particularly sedentary week. Column (1) shows mean difference in steps between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total steps on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects.

	(1) accepted	(2) accepted	(3) declined	(4) declined	(5) ignored	(6) ignored
age	-0.00179***	-0.00225***	-0.0000528**	-0.0000941***	0.00184***	0.00235***
	(0.0000336)	(0.0000343)	(0.0000195)	(0.0000199)	(0.0000376)	(0.0000383)
male	-0.0634***	-0.0629***	-0.00995***	-0.00990***	0.0733***	0.0728***
	(0.000869)	(0.000866)	(0.000503)	(0.000503)	(0.000971)	(0.000967)
bmi	0.00466***	0.00486***	0.000934***	0.000952***	-0.00559***	-0.00581***
	(0.0000783)	(0.0000780)	(0.0000453)	(0.0000453)	(0.0000875)	(0.0000872)
system days	0.0000806***	-0.0000533***	-0.0000510***	-0.0000630***	-0.0000296***	0.000116***
	(0.00000325)	(0.00000386)	(0.00000188)	(0.00000224)	(0.00000363)	(0.00000431)
Total Days	No	Yes	No	Yes	No	Yes
r2	0.0184	0.0256	0.00265	0.00283	0.0183	0.0252
N	549596	549596	549596	549596	549596	549596

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.23: Step Intervention #3, after a particularly sedentary week. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects.

	(1) steps
accepted	1106.0***
	(24.37)
declined	401.5***
	(39.02)
ignored	-156.3***
	(15.71)
r2	0.00514
N	694724

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.24: Step Intervention #3, after a particularly sedentary week. Column (1) shows the mean difference in steps between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally.

	(1) mins	(2) bedtime	(3) waketime	(4) mins	(5) mins	(6) mins
treatment	4.943***	-5.150***	0.184	4.866***	4.925***	4.904***
	(0.841)	(0.849)	(1.069)	(0.835)	(0.841)	(0.838)
age				-0.373***		
				(0.0250)		
bmi				-0.971***		
				(0.0585)		
male				-16.31***		
				(0.669)		
State FE	No	No	No	No	Yes	No
Date FE	No	No	No	No	No	Yes
r2	0.000438	0.000466	0.000000376	0.0168	0.00187	0.00629
N	78833	78833	78833	78517	78833	78833

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.25: Sleep Intervention, after a particularly restless few days. Column (1) shows mean difference in minutes of sleep between treatment and control groups. The independent variable "treatment" is treatment status T_i. The dependent variable is total minutes of sleep on the day of the initial notification. Columns (2) and (3) consider bed and wake times as the outcome variable (in minutes relative to 7pm the evening of the notification). Column (4) includes age, BMI, and an indicator for male gender as covariates. Column (5) includes state fixed effects. Column (6) includes date fixed effects.

	(1) mins
treatment	5.064***
	(1.161)
male	-16.07***
	(1.510)
treatment X male	-0.415
	(1.685)
bmi (std)	-0.971***
	(0.131)
treatment X bmi (std)	0.00343
	(0.146)
age (std)	-0.332***
	(0.0566)
treatment X age (std)	-0.0570
	(0.0632)
system days (std)	0.00387
	(0.00601)
treatment X system days (std)	0.00415
	(0.00670)
r2	0.0169
N	78517

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.26: Sleep Intervention, after a particularly restless few days. Estimates from regression of daily minutes of sleep on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) mins	(2) mins	(3) mins	(4) mins	(5) mins
treatment	6.933***	2.756	7.102***	4.253**	2.737
	(1.797)	(2.494)	(1.942)	(1.627)	(1.803)
r2	0.000816	0.000132	0.000923	0.000348	0.000133
N	18229	9220	14482	19609	17293

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.27: Sleep Intervention, after a particularly restless few days. Estimates from regression of minutes of sleep on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday... Column (5) is Thursday. Note this intervention was *not* fielded on Friday nor Saturday nights.

	(1) mins_5	(2) mins_25	(3) mins_50	(4) mins_75	(5) mins_95
main					
treatment	7.000*	4.500***	5.000***	5.550***	3.000
	(3.551)	(1.093)	(0.720)	(0.671)	(1.535)
r2					
N	78833	78833	78833	78833	78833

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.28: Sleep Intervention, after a particularly restless few days. Columns indicate quantile effects for the 5th, 25th, 50th, 75th, and 95th percentiles.

	(1) accepted	(2) accepted	(3) declined	(4) declined	(5) ignored	(6) ignored
age	-0.00585***	-0.00641***	-0.000987***	-0.00105***	0.00684***	0.00747***
	(0.000123)	(0.000125)	(0.0000627)	(0.0000638)	(0.000131)	(0.000132)
male	-0.0707***	-0.0684***	0.000325	0.000594	0.0703***	0.0678***
	(0.00329)	(0.00327)	(0.00167)	(0.00167)	(0.00349)	(0.00347)
bmi	0.000704*	0.00110***	0.000607***	0.000655***	-0.00131***	-0.00176***
	(0.000287)	(0.000287)	(0.000146)	(0.000147)	(0.000305)	(0.000304)
system days	0.000109***	-0.0000759***	-0.0000460***	-0.0000684***	-0.0000634***	0.000144***
	(0.0000130)	(0.0000150)	(0.00000662)	(0.00000768)	(0.0000138)	(0.0000159)
Total Days	No	Yes	No	Yes	No	Yes
r2	0.0424	0.0512	0.00497	0.00550	0.0480	0.0579
N	63106	63106	63106	63106	63106	63106

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.29: Sleep Intervention, after a particularly restless few days. Linear probability model estimates for the impact of anthropomorphics on accepting, declining, or ignoring the initial notification (exclusive categories). Note the "ignore" category excludes users who did not open the app that day (as does the control group). Even numbered columns additionally include a covariate for total days the user ended up remaining in the system for, meant to control for selection effects.

	(1) mins
accepted	24.38***
	(1.082)
declined	-11.65***
	(1.891)
ignored	-0.0183
	(0.866)
r2	0.0110
N	78833

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.30: Sleep Intervention, after a particularly restless few days. Column (1) shows the mean difference in minutes of sleep between the omitted control group and those who either accepted, declined, or ignored the initial notification. These estimates should not be interpreted causally.

	(1) 1d-Lag	(2) 2d-Lag	(3) 3d-Lag	(4) 4d-Lag	(5) 5d-Lag	(6) 6d-Lag	(7) 14d-Lag
treatment	3.328***	2.123*	3.553***	4.035***	1.622	2.446**	1.882*
	(0.914)	(0.974)	(0.983)	(0.950)	(0.954)	(0.938)	(0.929)
r2	0.000197	0.0000741	0.000207	0.000286	0.0000460	0.000109	0.0000694
N	67170	64069	62970	63050	62771	62620	59164

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.31: Sleep Intervention, after a particularly restless few days. Estimates from regression of lagged minutes of sleep on treatment status, for lags of 1, 2, 3, 4, 5, 6, and 14 days. Regressions additionally controlling for state fixed effects or age, gender, BMI, and days since becoming a user (not shown) are essentially identical in terms of lagged treatment effect magnitudes and significances.

	(1) mins per evening	(2) mins per evening	(3) mins per evening
treatment	1.189**	1.138**	1.163**
	(0.394)	(0.380)	(0.393)
male		-15.43***	
		(0.305)	
bmi		-0.986***	
		(0.0266)	
age		-0.365***	
		(0.0114)	
system days		0.00411***	
		(0.00121)	
State FE	No	No	Yes
r2	0.000116	0.0707	0.00431
N	78833	78517	78833

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.32: Sleep Intervention, after a particularly restless few days. Column (1) shows estimates from a regression of minutes of sleep per day averaged across all days the user remains active in the system (including the day of the initial notification and after) on first-day treatment status. Columns (2) and (3) additionally control for state fixed effects and gender, BMI, age, and days since becoming a user, respectively.

	(1) threshold_default	(2) threshold_default	(3) threshold_default	(4) threshold_default
treatment	0.0154***	0.0151***	0.0154***	0.0153***
	(0.00326)	(0.00325)	(0.00326)	(0.00326)
age		-0.00174***		
		(0.0000975)		
male		-0.0514***		
		(0.00260)		
bmi		-0.00151***		
		(0.000228)		
State FE	No	No	Yes	No
Date FE	No	No	No	Yes
r2	0.000284	0.0110	0.00172	0.00547
N	78833	78517	78833	78833

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.33: Sleep Intervention, after a particularly restless few days. The independent variable "treatment" is treatment status T_i. The dependent variable is an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification. Column (2) includes age, BMI, and an indicator for male gender as covariates. Column (3) includes state fixed effects. Column (4) includes date fixed effects.

	(1) threshold_default
treatment	0.0159***
	(0.00452)
male	-0.0502***
	(0.00588)
treatment X male	-0.00162
	(0.00656)
bmi (std)	-0.00133**
	(0.000509)
treatment X bmi (std)	-0.000221
	(0.000569)
age (std)	-0.00149***
	(0.000220)
treatment X age (std)	-0.000312
	(0.000246)
system days (std)	-0.0000124
	(0.0000234)
treatment X system days (std)	0.0000231
	(0.0000261)
r2	0.0110
N	78517

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.34: Sleep Intervention, after a particularly restless few days. Estimates from regression of an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification on treatment status, an indicator for male gender, BMI, age, system days (how many days someone has had their wearable band for), and interaction terms between each and treatment status. BMI, age, and system days are standardized.

	(1) threshold_default	(2) threshold_default	(3) threshold_default	(4) threshold_default	(5) threshold_default
treatment	0.0285***	0.000530	0.0266***	0.00806	0.00746
	(0.00730)	(0.00988)	(0.00730)	(0.00611)	(0.00696)
r2	0.000838	0.000000312	0.000915	0.0000888	0.0000664
N	18229	9220	14482	19609	17293

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.35: Sleep Intervention, after a particularly restless few days. Estimates from regression of an indicator for whether the person met the default goal of 8 hours of sleep on the day of the initial notification on treatment status by day of the week. Column (1) is Sunday, Column (2) is Monday... Column (5) is Thursday. Note this intervention was *not* fielded on Friday nor Saturday nights.

1.13 Appendix: Demographic Mediators of Heterogeneity

It is plausible that openness to experience, conscientiousness (a more specific domain of time-preference), comfort with technology, and risk aversion are concepts related to accepting and responding to smartphone notifications. I lack direct measures of these broader concepts, but they each differ on average by demographics such as gender, age, and BMI. I use these generally accepted associations between demographics and those broader concepts to form predictions about demographic heterogeneity in responsiveness.

1.13.1 Openness to Experience and Conscientiousness

Costa and McCrae (1992) describe high openness to experience (or simply high openness) as being associated with keen imagination, intellectual curiosity, behavior flexibility, and attitudes and beliefs that tend to be fluid and nondogmatic. I expect these traits were associated with being influenced by, and accepting, smartphone-based walking or sleep interventions.

Because of the multi-faceted nature of the more general economic domain of time preference[14], I hone in on the personality trait of conscientiousness to form my hypotheses. Costa and McCrae (1992) describe high conscientiousness as being marked by self-discipline and diligence, hence related to the direction of impulses, and notions of self-control (O'Gorman and Baxster, 2002). Self-control intuitively predicts physical health (in addition to substance abuse, crime, and personal finances, see Moffitt et al., 2010). Indeed, I expect more conscientious people to be less impulsive and hence were more likely to accept, and be influenced by, my smartphone interventions.

There are significant average differences in personality traits between the genders (Del Giudice et al., 2012, N $>$ 10,000 Americans). Western men tend to be more open to experience, while western women tend to be more conscientious (Lehman et al., 2013, N $>$ 19,000 Germans; and Schmitt et al., 2008, N $>$ 17,000 people from 55 international cultures). To the extent that women are more conscientious on average, I expect them to be more likely to have accepted a smartphone notification. They might also have been

[14]Conscientiousness is a specific example of the economic concept of time preference. Frederick et al. (2002) provides an excellent review of time preference, the details of which are beyond the scope of this paper. They conclude that time preferences are difficult to measure because of the particularly strong influence of framing effects and domain-specificity. They also highlight that longitudinal studies evaluating the consistency of measures of time preference are lacking, and their correlation with plausibly related real world activities are modest at best (e.g. smoking, credit card use, seat belt use, exercise, dental checkups, vaccination). Indeed, splitting up "time preference" into more specific categories, e.g. impulsivity, compulsivity, inhibition, is likely the best path forward (Lowenstein et al., 2001; Khwaja et al., 2007; Borghans and Golsteyn, 2006; Ikeda et al., 2010).

more likely to act on the notification's suggestion and hence be impacted more by the intervention itself. Male openness to experience, however, would tend to predict the same for men. Results are consistent with gender only mattering along the extensive margin (i.e. acceptance) consistent with female conscientiousness, but not male openness. Gender did not appear to matter along the intensive margin (i.e. treatment effect heterogeneity).

On average there are also differences in personality with respect to age. Lehman et al. (2013) finds average neuroticism and extraversion are negatively related to age, agreeableness and conscientiousness are positively related to age, and openness to experience exhibits a curvilinear relationship with age (highest at midlife). Soto et al. (2011) uses a cross-sectional sample of over 1.2 million people, finding negative trends for openness from late childhood into adolescence, but positive trends for both genders into adulthood and middle age (with men being somewhat more open throughout adulthood, consistent with the above hypotheses focused solely on gender). Evidence is consistent with the openness at least increasing from early adulthood to middle age (the age groups most highly represented in my sample), implying that becoming "set in ones ways" may only operate much later in life. Given our sample is concentrated slightly below middle age, through the channel of openness, older people therefore ought to be more likely to accept, and be impacted by, the notifications. With respect to conscientiousness, Soto et al. 2011 finds a strong negative trend from late childhood into adolescence, but a strong positive trend from adolescence into emerging adulthood and beyond (with women showing slightly more conscientiousness overall, again consistent with above hypotheses focused solely on gender). This reinforces my predictions based on openness, namely that older people should have been more likely to accept, and be impacted by, the notifications; indeed, I find they were slightly more impacted by "Step Intervention #2." However, older people were robustly less likely to accept all notifications, likely driven by the dominant influence of lack of comfort or familiarity with technology.

The majority of evidence points to there being a modest negative relationship between BMI and openness, implying we might have expected higher BMI's to be associated with less acceptance of the notifications (and more declining and ignoring) and smaller treatment effects (Brummett et al., 2006; Van Reedt Dortland et al., 2012). However, many studies find zero or mixed results in measuring the relationship between BMI and openness (Chapman et al. 2009, Terracciano et al. 2009; Sutin et al. 2011; Shim et al., 2014). Evidence points to high BMI being associated with lower conscientiousness (Brummett et al., 2006; Shim et al., 2014), potentially exacerbating low acceptance and smaller treatment effects induced by openness. Of course, we sampling from the upper end of the motivation distribution. This is likely particularly relevant for BMI; I expect these people to be

particularly motivated and this effect to dominate. In fact, higher BMI people were more likely to accept the notifications. However, on Thanksgiving, higher BMI users were less responsive, although high BMI users still accepted the notification in greater numbers on Thanksgiving.

1.13.2 Comfort with Adopting New Technology

Comfort with adopting new technology is a particularly relevant sub-domain of openness to experience. We might expect that above and beyond the predictions of general openness, those who are particularly at ease with new technologies might have been more responsive to our interventions.

There is little research on anxiety or comfort in specifically adopting wearable or app-based activity trackers. Rai et al. (2013) is one notable exception, surveying 1,132 Americans about their mHealth use, finding older people were less likely to adopt mHealth platforms (as might have been guessed based on the above discussion about openness). They also found women were more likely to use mHealth as a complement to traditional doctor visits (in contrast with male openness to experience), but no other relevant heterogeneity with respect to gender. They do not investigate BMI. I therefore employ older, better-researched technologies surrounding comfort with Internet use and adoption to ascertain whom we should expect to be most at ease accepting and responding to smartphone interventions.

Findings with respect to gender are somewhat mixed, though many studies report that women are on average more anxious about adopting new technologies (Broos, 2005; Huffman et al., 2013 noting that the common finding of increased technology self-efficacy for men is driven in-part by gender roles rather than biological sex alone). He and Freeman (2009) highlight that past findings are mixed with respect to gender; while many find women are more anxious about technology adoption and less likely to adopt, many also report no gender gap. Their study digs deeper by evaluating how gender plays a roll in forming beliefs about information technology. If we accept the hypothesis that women are on average less keen to adopt new technologies, it implies women should have been less likely to accept, and be impacted by, my smartphone interventions. In fact, while I observe no impact in terms of gender on the intensive margin of the interventions, women were indeed *more* likely to accept the challenges proffered by the interventions, calling into question the relevance of predictions based on comfort with technology with respect to gender.

Smith (2014) reports results from a large Pew Institute Survey, finding that 59% of

seniors report using the Internet (as opposed to 86% of all US adults). Czaja et al. (2006) confirms these findings, noting further that that relationship between age and adoption of technology is mediated by cognitive ability (fluid and crystallized intelligence). Porter and Donthu (2006) develop an advanced version of the technology acceptance model (TAM) to explain, in part, the age gap in technology adoption, finding that age (as well as income, education, and race) are associated with certain beliefs about the Internet, and these beliefs ultimately drive usage. This is consistent with Smith (2014), which points out that two distinct groups of seniors make up Internet users. Younger, more wealthy, and more highly educated seniors have adoption rates almost as high as the broader adult population, while older, less wealthy, and less healthy seniors tend to be completely offline and more skeptical of its benefits. Our sample is likely to be drawn primarily from the former. However, given the novelty of an mHealth intervention delivered through a smartphone app, paired with a wearable fitness tracker, it still seems plausible that older individuals would be less likely to accept, and be influenced by, my smartphone interventions (Rai et al., 2013). With respect to acceptance, this intuition turns out to be true: older people were less likely to accept the challenges. Along the intensive margin, however, these predictions fell flat. Only "Step Intervention #2" showed heterogeneity along the intensive margin, indicating that older people were slightly more responsive. This is consistent with predictions based on conscientiousness, but less so openness or comfort with technology.

1.13.3 Risk Aversion

Risk aversion is also plausibly related to willingness to accept, and respond to, smartphone interventions aimed at increasing physical activity and sleep. It is well established that risk aversion is associated with a broad range of risky-taking activities, including smoking, drinking, being overweight, seat belt use, having insurance, and investing (Anderson and Mellor, 2008; Barsky et al., 1997). If we deem responding to a smartphone notification a preventative, pro-health behavior (like wearing a seatbelt, investing in insurance, or not smoking or drinking), we should expect more risk averse people to accept, and be influenced by, my smartphone interventions. However, given that risk aversion tends to explain a small percentage of the variation in a range of risky outcomes (Barsky et al. 1997), we should be hesitant to expect predictions based on demographic differences in risk aversion to be particularly informative as to acceptance of, or responsiveness to, my interventions.

Men are generally more risk tolerant than women (Barsky et al 1997, Eckel and Grossman 2008, Borghans et al. 2009, Kimball et al. 2008, Sapienza et al. 2009). Therefore, with respect to risk aversion, we should expect women to more likely to accept, and re-

spond to, the smartphone interventions. While this turned out to be true with respect to acceptance, I find no evidence of heterogeneity with respect to gender and the intensive margin.

Older people are generally less risk tolerant (Sahm 2013 using within-person variation in gambling responses, citing several other papers finding the same thing using alternative methods). Therefore, with respect to risk aversion, we should expect older people to more likely to accept, and respond to, the smartphone interventions. This was indeed true with respect to responsiveness to "Step Intervention #2." However, with respect to acceptance, predictions based on risk aversion and age were incorrect.

Higher risk aversion is also intuitively associated with decreased likelihood of having a high BMI (Anderson and Mellor, 2008). I therefore expect high BMI individuals to be less responsive to the interventions. With respect to accepting, this turned out to be incorrect. However, with respect to "Step Intervention #1," this prediction held.

1.13.4 Results

My findings about challenge acceptance were constant across all interventions. Users who accepted the notification tended to be female, younger, and higher BMI. These results are consistent with predictions involving greater female conscientiousness and risk aversion (but not greater male openness nor comfort with technology). They are also consistent with older people being less comfortable with technology and more risk averse (but not the tendency of older people to be more open and conscientious). However, they are inconsistent with predictions involving BMI, likely driven by our self-selected sample of particularly motivated users.

In "Step Intervention #1," the only covariate that showed evidence of mediating treatment effects was BMI: the treatment was most effective for those with lower BMI, on the order of 20 steps per standard deviation of BMI. Despite the fact that we are drawing from a disproportionately motivated sample, the direction of this effect is consistent with the above predictions of higher BMI people being less openness to experience, less conscientious, and less risk averse.

In "Step Intervention #2," the treatment was slightly more effective for older users. This is consistent with predictions regarding older users being more conscientious, open, and risk averse. However, this is inconsistent with my prediction regarding older people being less comfortable with, and hence less responsive to, new technology.

"Step Intervention #3" and "Sleep Intervention" show no evidence of treatment response heterogeneiety with respect to gender, age, and BMI.

CHAPTER 2

Reconsidering Risk Aversion *

* with Daniel Benjamin and Miles Kimball

2.1 Abstract

Policymakers, economists, and the popular media have long been worried that Americans may not be investing appropriately in preparation for retirement. Setting default options associated with investing for retirement requires knowing at least the average level of risk aversion in the population. However, quantitative measures of an individual's risk tolerance vary, depending on a variety of factors that should not matter according to standard normative axioms, notably how the problem is framed. Our aim is to develop a surveying procedure, based on the philosophical tradition of deliberative thinking and logical reconciliation among contradictions, which attempts to overcome framing biases in measuring risk aversion. In moral philosophy, and more recently in the field of decision analysis (e.g. Raiffa, 1968), there is a long history of having individuals attempt to resolve their internal inconsistencies and thereby discover their preferences through reasoning. However, we are not aware of work that has developed systematic procedures for helping people think through their preference in important economic contexts such as investing for retirement. Using a sample of 628 subjects, the vast majority of whom are Cornell students, we first elicit "untutored" risk preferences among 5 different investment plans, which involve gambles over "how much you have to spend each year during retirement, from age 65 on." We ask questions of each subject using 7 different ways of framing the decision (e.g. varying whether irrelevant information is provided, whether independent choices are made sequentially or simultaneously, whether subjects are able directly to choose their favorite choice or can only choose between many pairs of pre-determined plans, and whether lotteries are already reduced). "Untutored" preferences are those measured before any sort of guidance or explicit effort on our part to engender reconciliation of logical contradic-

tions among frames. We then lead subjects through a reconciliation phase, allowing them to update their choices where they have made inconsistent or intransitive decisions (contradictions according to normative axioms of rational choice), finally reaching "reasoned" preferences, which we define as those measured after two rounds of inconsistency and intransitivity resolutions. We also allow subjects to update already-consistent preferences to test whether our procedure's results are driven by experimenter demand. While subjects readily update toward consistency (and transitivity), they rarely move in the other direction. However, few subjects completely endorse all the normative axioms implied by differences between similar frames. There is also substantial heterogeneity across frames in both original consistencies and propensity to update. Initially, subjects are particularly apt toward consistency in frames in which they could directly choose investment plans (or pieces of those plans), as opposed to frames in which they were forced to compare every potential pair of investment plans. They are particularly unlikely initially to endorse Reduction of Compound Lotteries, although this axiom also sees the most resolution toward endorsement. A subset of participants was invited back for a second wave of the experiment 2-4 weeks later. We observe some persistence in preferences across waves as well as further movements toward consistency across frames. Finally, using maximum likelihood estimation, we quantify risk aversion and incidence of decision errors among frames and over the course of our survey. Despite our large sample, analyses of some frames are still statistically underpowered. Among well-powered analyses, across the course of our survey, we see limited convergence in risk aversion among frames and an overall reduction in decision errors. While our study constitutes significant progress, we require a bigger sample to make concrete default recommendations. We are weighing several further tweaks to our procedure, including (1) designing a more heavy-handed reconciliation procedure (to facilitate even more updating toward consistency), (2) direct elicitation of second-, third-, and fourth-favorite choices (to allow for greater comparability across frames and boost the statistical power of our MLE procedure), (3) bringing subjects back into the lab for third or even fourth waves (to test the extent to which our procedure "sticks" over the course of more than a few weeks), and (4) collecting data from more nationally representative sample (given concerns about external validity).

2.2 Background and Motivation

Policymakers, economists, and the popular media have long been worried that Americans may not be investing appropriately in preparation for retirement (e.g., Poterba et al, 2005). Americans face the dilemma of complex financial and health decisions combined with a

high incidence of cognitive decline with age. Since default options matter enormously for actual investment choices (Beshears et al., 2008a), policymakers have turned to defaults as a policy tool for influencing individuals' asset allocations (e.g., the Pension Protection Act of 2006). But what default asset allocation is appropriate?

This project focuses on the practical question of how to measure risk preferences for calibrating long-term retirement savings, especially for setting fund contribution defaults, but also for personal investing. The difficulty is that knowing what a "good choice" is for an individual depends crucially on knowing her preferences. The key preference parameter in the simple benchmark theory of optimal portfolio choice (relevant for retirement savings as well as investment choices) is risk aversion, while the optimal saving rate depends on time preference and the elasticity of intertemporal substitution (EIS) as well as risk aversion. However, the very fact that individuals tend to stick to arbitrary default asset allocations and savings rates in defined contribution plans strongly suggests that, for many Americans, actual retirement saving and investment decisions do not reflect their preferences (Beshears et al., 2008b). Setting appropriate defaults requires at least knowing the average level of risk aversion, time preference, and EIS in the population, and ideally knowing each individual's parameters. These questions pose serious challenges despite economists' and decision analysts' extensive work on this issue over many years (e.g., Arrow, 1983). For example, quantitative measures of an individual's risk tolerance vary considerably, depending on a variety of factors that should not matter according to standard normative axioms, namely how the problem is framed (e.g. Kahneman and Tversky, 1979; Tversky and Kahneman, 1981). For determining optimal portfolio choice, it is particularly troublesome that risk aversion depends on whether the risks are larger-stakes or smaller-stakes[1]. For example, among Health and Retirement Study (HRS) respondents, we calculate that the mean estimated coefficient of relative risk aversion is 9.3 for gambles that could double permanent income, compared with 106.1 for gambles that could increase permanent income by 20%. Relative risk aversion of 106.1 strikes many economists as implausibly high, but risk aversion appears even greater at smaller stakes. In fact, individuals' choices in risky choice problems are inconsistent across stake sizes *regardless* of the functional form of the utility function, and hence these choices violate very basic assumptions of normative economic models (Rabin, 2000). A particular question of this project is whether individuals might be making an error in either their smaller-stakes choices or in their larger-stakes choices.

There are some economic models that can accommodate seemingly-inconsistent risk

[1]Evidence on time preference suggests equally important inconsistencies. For example, individuals are more patient between two future dates than between today and a future date (Strotz, 1955; Thaler, 1981). Estimates of the elasticity of intertemporal substitution range widely from about zero (Hall, 1988; Dynan, 1993) to well over 1.0 (Mulligan, 2002; Gruber, 2006). Our focus here is on risk aversion.

aversion[2] over smaller stakes and risk tolerance over larger stakes (e.g., Koszegi and Rabin, 2006). Although these models that accommodate anomalous behavior are written as non-standard preference specifications, most are explicitly meant to be descriptive (rather than derived from normative axioms). Hence these "preference-based" theories admit the interpretation that non-standard behaviors represent decision errors, rather than features of actual preferences (Beshears et al., 2008b). Ideally, saving and investment defaults would be set based on measures of risk preferences, time preference, and the EIS that are uncontaminated by decision errors.

When a person's decisions are inconsistent with normative axioms, the problem arises of distinguishing choices due to decision errors and choices due to preferences. Individuals with greater cognitive ability are generally less risk-averse (Dohmen et al. 2008). Moreover, experimentally reducing available cognitive resources via "cognitive load" (a cognitively-demanding task performed concurrently) increases risk-averse behavior (Benjamin et al., 2006). These findings suggest that high degrees of risk aversion may be driven at least in part by cognitive limitations. In moral philosophy, and more recently in the field of decision analysis (e.g., Raiffa, 1968), there is a long history of having individuals attempt to resolve their internal inconsistencies and thereby discover their own preferences through reasoning. Guided decision-making is taught widely in business schools today and is integrated in existing financial planning software, such as Lawrence Kotlikoff's ESPlanner. Indeed, many financial advisors currently implement a version of this idea, albeit usually with simple measures of risk tolerance and rules of thumb about optimal behavior, rather than with the results of economic research (e.g., TIAA-CREF, 2008). Few papers have measured reasoned preferences in a similar spirit as we propose here. For example, Loewenstein and Sicherman (1991) found that subjects were more likely to maximize the expected discounted value of cash flows after they were provided subjects with arguments for and against doing so. Druckman (2001) focuses on the popular "Asian disease problem" and suggests using the "both" frame as an appropriate baseline for ascertaining true preferences. McNeil et al. (1988) turns to the realm of medical decisions, suggesting eliciting preferences by asking a positive frame, a negative frame, and a combination of the two as a sort of sensitivity analysis. However, we are not aware of work that has developed systematic procedures for helping people think through their preferences in important economic context like investing for retirement.

Our central contribution is a procedure for identifying decision errors in standard savings and investment decision problems. After answering an initial set of investment ques-

[2]Similarly, there are models that can accommodate inconsistency in time preference over various horizons, such as quasi-hyperbolic discounting (Laibson, 1997).

tions, we provide subjects with an opportunity to update their choices (both when there are actual inconsistencies between similar frames, as well as "placebo" inconsistencies, allowing subjects to update choices that are already consistent). Our premise is that if an individual himself acknowledges having made a mistake after having updated his choices, then we have identified a decision error. Furthermore, once we have vetted an individual's choices in this way using a variety of checks for consistency (and transitivity), we tentatively accept the final choices as being closer to the individual's preferences. The basic idea is to allow individuals to revise their choices when they have exhibited inconsistent or intransitive behavior without being heavy-handed or suggesting that we want them to update their choices. Of course, it would be extremely difficult (and probably pointless) to try to explain to a typical experimental subject why an entire combination of choices is considered normatively problematic. A crucial feature of our experimental procedure is that no subject was ever asked to understand this whole chain of reasoning. Instead, depending on the subject's particular pattern of normatively inconsistent choices, we confronted the subject with individual links of the chain of logic, each of which is relatively easy to understand.

Let us call the initial set of choices *untutored preferences*, and the final set of choices a person makes after our reasoning process *reasoned preferences*. Our proposed procedure for eliciting reasoned preferences is a means toward better understanding the risk preferences on which appropriate retirement saving and investment depend. We detail which normative axioms people initially endorse, where they are most likely to correct where they have said they have made a mistake, and toward which frames people revise. We also focus on how estimates of risk aversion vary by frame across untutored and reasoned preferences. By better estimating the values of average reasoned risk preferences, we hope to reduce the range of uncertainty about risk aversion and therefore help identify optimal default asset allocations. Our more general aim is to develop techniques for identifying "reasoned preferences" for the even larger set of preferences that matter for difficult decisions that Americans face, even beyond the retirement saving decision. While our initial rounds of surveying have yielded significant progress in terms of reducing this range of uncertainty for risk aversion, we have not observed enough convergence in risk aversion among different frames to make exact default recommendations.

2.3 Importance of Setting Portfolio Defaults and Savings Rates Appropriately

If an individual is defaulted into a saving rate and asset allocation that is optimal for a *different* level of preference parameters than her own, she could suffer large expected welfare losses. Consider the continuous-time, 35-year-long investment-saving problem faced by a household with annual discount rate ρ and initial wealth $w_0 > 0$ (and no labor income), as in Merton (1969), except with Kreps-Porteus preferences that separate the EIS s from relative risk aversion γ. In each instant, the household chooses what fraction of wealth to consume, and what fraction to invest in a risky asset whose excess return is stochastic, with mean μ and standard deviation (per square-root of time) σ as opposed to a safe asset with rate of return r. However, suppose the household's allocation is optimal with respect to the (possibly *wrong*) risk aversion, time preference, and EIS parameters: $\hat{\gamma}$, $\hat{\rho}$, and $\hat{s}$. Let $u(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)$ denote the expected discounted utility, calculated using the household's true preference parameters γ, ρ, s of the state-contingent saving-investment path implied by solving the household's problem with risk aversion $\hat{\gamma}$, $\hat{\rho}$, and $\hat{s}$. Then let the real-valued function $c(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)$ denote the certain, constant level of consumption that, if consumed at every moment from now on, would give expected discounted utility equal to $u(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_o)$. We measure the welfare loss from choosing the suboptimal investment-saving path by:

$$\frac{c(\gamma, \hat{\gamma}, \rho, \hat{\rho}, s, \hat{s}; w_0)}{c(\gamma, \gamma, \rho, \rho, s, s; w_0)} = \left(\frac{1 - e^{-aT}}{a}\right)^{\frac{1}{1-s}} \left(\frac{1 - e^{-\hat{m}T}}{\hat{m}}\right)^{\frac{-s}{1-s}} \left(\frac{1 - e^{-\hat{a}T}}{\hat{a}}\right)^{-1} \tag{2.1}$$

$$a = (1 - s)(r + \frac{\mu}{2\gamma^2\sigma^2}) + s\rho \tag{2.2}$$

$$\hat{a} = (1 - \hat{s})(r + \frac{\mu}{2\hat{\gamma}^2\sigma^2}) + \hat{s}\hat{\rho} \tag{2.3}$$

$$\hat{m} = \rho + (1/s - 1)(r + \frac{\mu}{2\hat{\gamma}^2\sigma^2}(2\hat{\gamma} - \gamma) - \hat{a}) \tag{2.4}$$

This ratio equals 1 if the household's portfolio choice is optimal for its preferences but will be less than 1 otherwise. Given the assumptions on preferences, the ratio is independent of w_0, and it can be interpreted as the proportional reduction in initial wealth (and hence consumption in all states) that is equivalent in terms of welfare to the mistaken

	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$	$\gamma = 8$
$\hat{\gamma} = 1$	0.00	0.14	0.56	0.94
$\hat{\gamma} = 2$	0.06	0.00	0.07	0.29
$\hat{\gamma} = 4$	0.13	0.03	0.00	0.03
$\hat{\gamma} = 8$	0.17	0.07	0.02	0.00

Table 2.1: Welfare cost of investing with "wrong" preferences. Wealth loss (in percent) that is equivalent to mistaken behavior. True risk aversion, γ, is along the columns. Risk aversion used for portfolio choice, $\hat{\gamma}$, is along the rows.

behavior. Table 2.1 presents illustrative calculations, where we set $T = 35$, $r = 0.03$, $\mu = 0.03$, $\sigma = 0.15$, $s = \hat{s} = 0.375$, $\rho = \hat{\rho} = 0.075$, and we let γ and $\hat{\gamma}$ take on each of 1, 2, 4, and 8.

Even though existing literature has focused on the mistake of non-participation in financial markets (e.g., Haliassos and Bertaut, 1995), Table 2.1 shows that the more severe mistake is behaving less risk-aversely than one actually is (because the extra risk is extremely costly). If a household with risk aversion $\gamma = 1$ behaves like $\hat{\gamma} = 2$, the welfare loss is equivalent to 6% of wealth. By contrast, if $\gamma = 2$ and $\hat{\gamma} = 1$ (the household takes on too much risk) then the welfare loss is 14%. If $\gamma = 8$ and $\hat{\gamma} = 1$, then the loss is 94%! These calculations imply it is important to choose a default allocation appropriately; a one-size-fits-all default may be particularly harmful for individuals who are more risk-averse than the chosen default.

2.4 Frames and Axioms

We first elicit "untutored" risk preferences among 5 different investment plans. The 5 investment plans are labeled: A, BCE, BDF, BDE, and BDF, ordered weakly from safest to riskiest. All questions involve hypothetical payoffs in "how much you have to spend each year during retirement, from age 65 on." We ask questions of each subject using 7 different ways of framing the decision (e.g. varying whether irrelevant information is provided, whether independent choices are made sequentially or simultaneously, whether subjects are able directly to choose their favorite choice or can only choose between many pairs of pre-determined plans, and whether lotteries are already reduced). We refer to each of these 7 sets of questions as a frame. The number of questions used in each frame varies from only a single question to 10 questions. We divide these 7 frames into 2 broader categories: 4 Nodewise Action Choice Frames and 3 Pairwise Strategy Choice Frames. Similar frames only differ by axiomatic "baby steps," making things as simple as possible for subjects and allowing us isolate specific axiomatic expected utility violations as distinct from mistakes.

The simplest frame that includes all possible plans consists of only one question involving a two-period investment horizon over 20+ years. We call it the "Complete Contingent Action Plan" see Figure 2.1. It is also one of the Nodewise Action Choice Frames. It asks subjects to make 3 binary choices simultaneously: A v. B, C v. D, and E v. F. The latter two choices are unnecessary if a person chooses A. In later rounds of experimenting[3], if a subject chose A, we followed-up by asking her to pretend she could not choose A, forcing her also to choose between each of C v. D and E v. F.

The remaining 6 frames are derived from the Complete Contingent Action Plan. The simplest Nodewise Action Choice Frame, which we call "Single Action in Isolation," consists of 2 questions (see Figure 2.2). Each is simply one of the individual questions asked in the Complete Contingent Action Plan conditional on answering B $>$ A, sans information about any other choices. In other words, C v. D and E v. F are each asked separately without additional context.

Adding a layer of complexity, the next Nodewise Action Choice Frame, which we call "Single Action with Backdrop," again asks participants to separately choose in each of C v. D and E v. F (i.e. 2 questions, see Figure 2.3). However, unlike Single Action in Isolation, the additional context available in the Complete Contingent Action Plan frame is grayed out.

We form the final Nodewise Action Choice Frame, "Two Contingent Actions with Backdrop," by simply allowing the subject to choose C v. D and E v. F simultaneously (i.e. a single question, see Figure 2.4). Note this is essentially combining the two questions from the Single Action with Backdrop frame into a single question. Or, viewed another way, it is the same as the Complete Contingent Action Plan frame but without allowing a decision between A and B.

The Pairwise Strategy Choice Frames are a bit more complicated, each consisting of 10 questions (i.e. 5 potential investment plans choose 2 = 10 questions). The most complicated, "Pairwise Choices Between Complete Strategies," asks a subject to choose between all possible pairs of already filled-out Complete Contingent Action Plans (see Figure 2.5).

Pairwise Choices Between Complete Strategies forms the basis of the remaining 2 Pairwise Strategy Choice frames. The first, "Pairwise Choices Between Compound Lotteries," simplifies things by stripping the figures of grayed out information and de-emphasizing the inter-temporal aspect of learning about probabilistic outcomes at age 50, but does not reduce the compound lotteries (see Figure 2.6).

The second, "Pairwise Choices Between Reduced Simple Lotteries" simply reduces the

[3]Namely, during the second round of surveying covering 311 subjects (as well as the second wave covering 264 subjects).

compound lotteries from Pairwise Choices Between Compound Lotteries (see Figure 2.7).

Of course, as already mentioned, it would be extremely difficult (and probably pointless) to try to explain to a typical experimental subject why an entire set of choices among all frames is considered normatively problematic. Instead, we systematically confront subjects *only* with instances where their risk preferences violate normative axioms *between similar frames* (what we refer to as a "step"). See Table 2.2 for the names of the steps (denoting the axiom associated with that step), as well as the names of the two frames associated with each. Moving forward, we therefore use the words "step" and "axiom" interchangeably. Note that the adaptive software only walks the subject through whatever set of simple steps was relevant for *that subject's* choices.

Step/Axiom	**Frame 1**	**Frame 2**
Irrelevance of Background Counterfactuals	Single Action in Isolation	Single Action with Backdrop
Simple Actions = State-Contingent Actions	Single Action with Backdrop	Two Contingent Actions with Backdrop
Irrelevance of Counterfactual Choices	Two Contingent Actions with Backdrop	Complete Contingent Action Plan
Fusion + Shift from Nodewise to Pairwise	Complete Contingent Action Plan	Pairwise Choices between Complete Strategies
Complete Strategies = Implied Lotteries	Pairwise Choices between Complete Strategies	Pairwise Choices Between Compound Lotteries
Reduction of Compound Lotteries	Pairwise Choices Between Compound Lotteries	Pairwise Choices Between Reduced Simple Lotteries

Table 2.2: Name of each axiomatic baby step (or simply "step") and the two frames associated with each.

2.5 Experiment Design and Methods

The experiment has 6 parts: (1) Pre-Test, (2) Training Batteries, (3) Main Body Part 1: Elicitation of Untutored Preferences, (4) Psychological and Cognitive Batteries, (5) Main Body Part 2: Elicitation of Reasoned Preferences, and (6) Follow-Up and Demographic Batteries.

There are a variety of challenges to implementing the kind of experiment we propose, such as avoiding "experimenter demand effect" in which a subject agrees with a normative

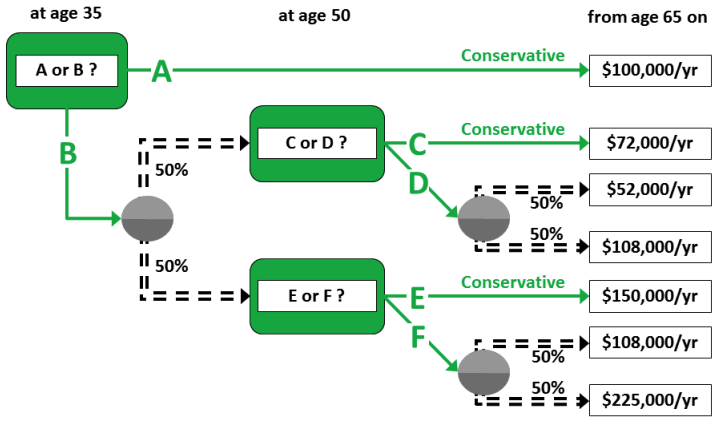

Figure 2.1: Complete Contingent Action Plan (1 question). Subjects are prompted with instructions: ”Imagine you are currently 35 years old. You need to make three decisions. The decision between A and B takes place now. If you choose B, you also need to make two decisions that will lock in how you will invest at age 50. If you choose A, you do not need to make any more decisions. All decisions will affect how much money you will be able to spend each year during retirement (from age 65 on). In each decision, you will choose between two strategies: risky or conservative. Each conservative strategy will guarantee you a fixed amount to spend each year during retirement. Each risky strategy allows for possibly higher amounts. You do not need to choose the same kind of strategy in each decision.”

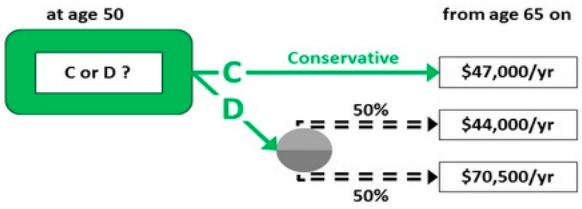

Figure 2.2: Single Action in Isolation (2 questions). Subjects are prompted with instructions: "You will be asked to make decisions that will affect how much you will be able to spend each year during retirement (from age 65 on), imagining that you are currently 50 years old. You will choose between two strategies: risky or conservative. The conservative strategy will guarantee you a fixed amount to spend each year during retirement. Under the risky strategy, higher amounts are possible."

axiom in order to please the experimenter (or in order to appear "rational" to herself or others) rather than because she has been genuinely persuaded or genuinely holds those preferences. We return to these issues after describing the experimental procedure in greater detail.

2.5.1 Design Considerations

Informal conversations with subjects after our initial pilot study (see "Appendix: Pilot Study") indicated that many found it hard to understand how two frames represented the same decision problem. Because we want to make sure that subjects *do* understand how certain combinations of choices violate normative axioms, we decided to take several precautions to minimize subject confusion in our main experiment. First, we study much simpler risky decision problems, where the equivalences follow from small steps of logical reasoning[4]. Second, we quiz subjects on their comprehension of basic tenants of prob-

[4]Our pilot data also made us realize that when pilot subjects behaved inconsistently, it was impossible to pinpoint which of many possible decision errors might account for subjects' inconsistent behavior because there were many differences in the pilot across the annual and long-term framing of returns. For example, the annual framing requires individuals to compound the returns (which is difficult), and the chance that stocks lose money is much higher in the annual framing (and individuals may be loss-averse). We focused on studying the inconsistency between relatively high risk aversion over smaller-stakes choices and relatively low

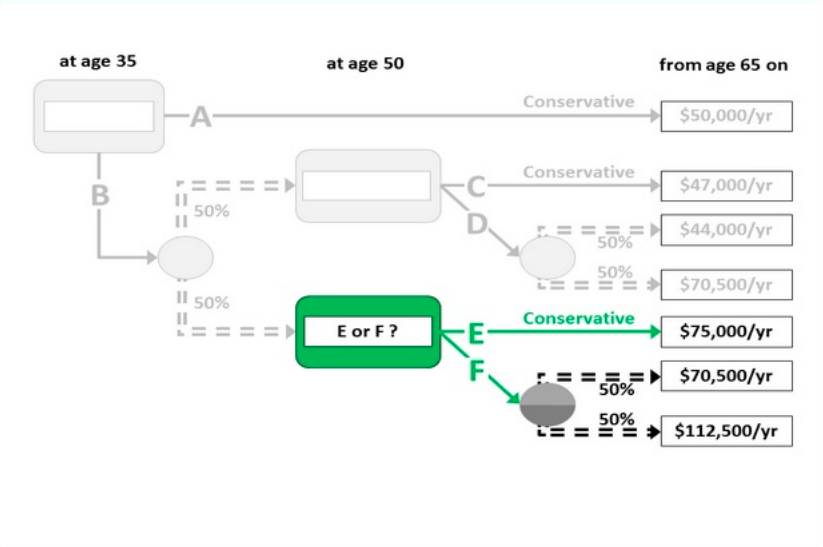

Figure 2.3: Single Action with Backdrop (2 questions). Subjects are prompted with instructions: "You will be asked to make decisions that will affect how much you will be able to spend each year during retirement (from age 65 on), imagining that you are currently 50 years old. You will also be given information about how decisions you made when you were 35 turned out, that are beyond your control at this point. These grayed-out parts of the picture are things that could have happened, but you know for sure did not happen. You will choose between two strategies: risky or conservative. The conservative strategy will guarantee you a fixed amount to spend each year during retirement. Under the risky strategy, higher amounts are possible."

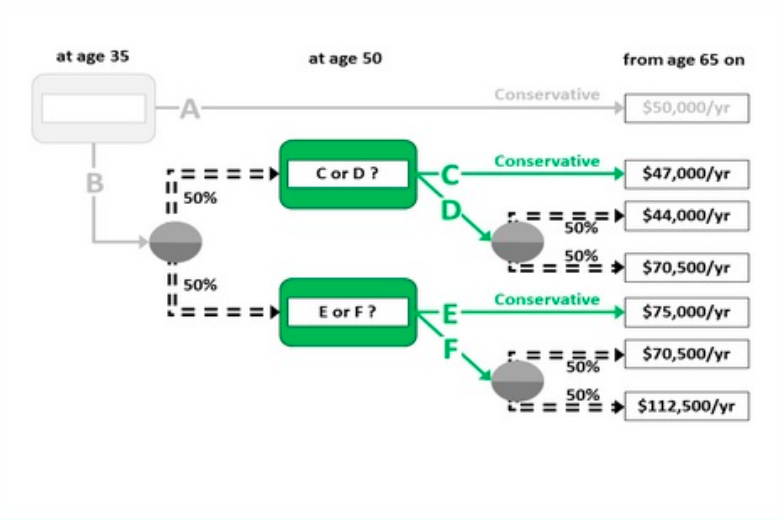

Figure 2.4: Two Contingent Actions with Backdrop (1 question). Subjects are prompted with instructions: "Imagine you are currently 35 years old, and have chosen risky decision B. You do not yet know how this decision has turned out. So, you need to make two decisions that will lock-in how you will invest at age 50. These decisions will affect how much money you will be able to spend each year during retirement (from age 65 on). In each decision, you will choose between two strategies: risky or conservative. Each conservative strategy will guarantee you a fixed amount to spend each year during retirement. Each risky strategy allows for possibly higher amounts. You do not need to choose the same kind of strategy in each decision."

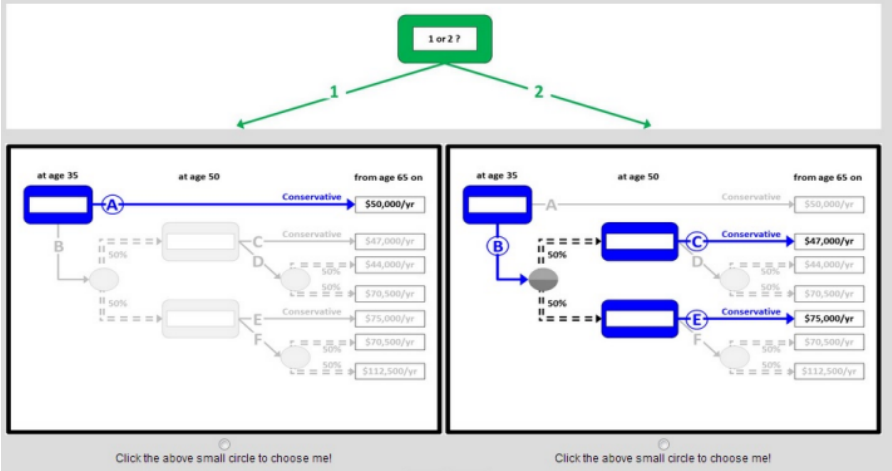

Figure 2.5: Pairwise Choices Between Complete Strategies (10 questions). Subjects are prompted with instructions: "In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2. Each investment plan has a set of choices locked in along the way (at age 35 and age 50), shown by circled letters, that lead to possible levels of yearly spending during retirement (from age 65 on). Grayed-out parts are used to show things that can't happen if you choose that investment plan. Spinners show the chances of different outcomes. From a spinner, the chance of taking each fork is shown next to that fork. Each fork can lead either to a locked in choice, or directly to a level of yearly spending during retirement. Note that a path with a 50% chance at one fork, followed by another 50% chance at a later fork, means that there is only 25% chance of getting all the way to the end of that path. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page)."

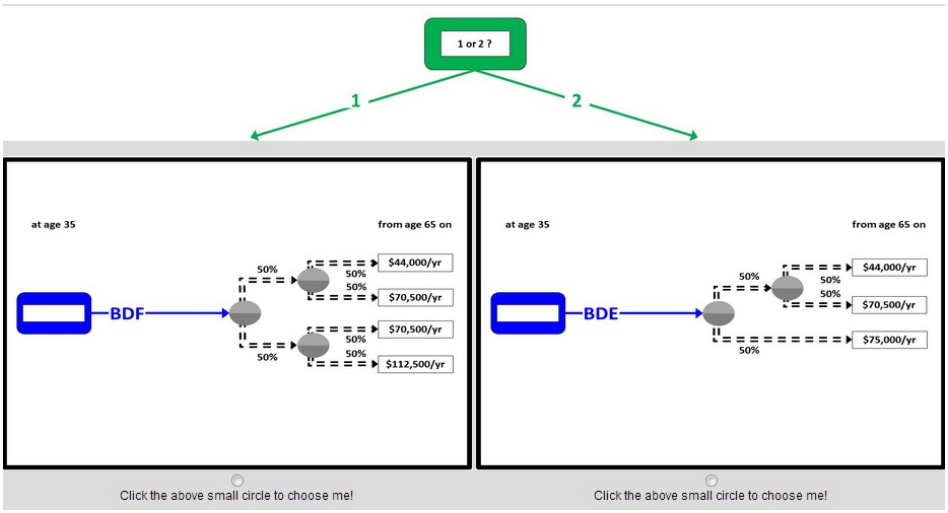

Figure 2.6: Pairwise Choices Between Compound Lotteries (10 questions). Subjects are prompted with instructions: "In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2, imagining that you are currently 35 years old. Each investment plan has different possible outcomes for how much you will be able to spend each year during retirement (from age 65 on). Note that some investment plans are named by one letter, "A"; other investment plans are named by more than one letter, like "BCE". Spinners show the chances of different outcomes. From a spinner, the chance of taking each fork is shown next to that fork. Each fork can lead either to another spinner, or directly to a level of yearly spending during retirement. Note that a path with a 50% chance at one fork, followed by another 50% chance at a later fork, means that there is only a 25% chance of getting all the way to the end of that path. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page)."

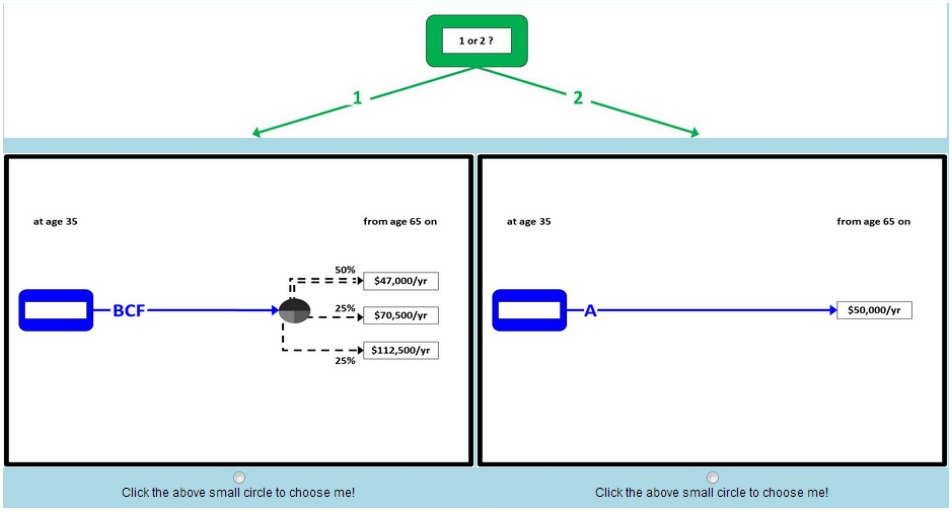

Figure 2.7: Pairwise Choices Between Reduced Simple Lotteries (10 questions). Subjects are prompted with instructions: ”In each question in this section you need to make a choice between two investment plans, Option 1 and Option 2, imagining that you are currently 35 years old. Each investment plan has different possible outcomes for how much you will be able to spend each year during retirement (from age 65 on). Note that some investment plans are named by one letter, like ”A” ; other investment plans are named by more than one letter, like ”BCE”. The picture shows the chance of each outcome happening next to the outcome (if the chance is less than 100%). Note that a 50% chance is twice as likely to happen as a 25% chance. To select your choice for each question, you have to click the small gray circle under the plan (not shown on this page). ”

ability, the various symbols and graphs used throughout the experiment to describe their decisions, and the assumptions we ask them to make, but only after leading them through an extensive training section. We verified through extensive piloting that these training sections are clear and our presentation of choice problems readily understandable. Third, the decision problems in our experiment exclusively use probabilities of 25% and 50%, which are more familiar to most people than the probabilities often used in tests of expected utility. Fourth, in order to make sure that subjects are motivated, we conduct the experiment in a laboratory setting where we have better control over subjects' attention. Finally, we use lotteries over yearly income during retirement as a means of simplifying the decisions and reducing cognitive burden.

2.5.2 Subject Population

We collected 2 rounds of data at Cornell's LEEDR and Business Simulation laboratories on 628 people during 2013-2014 (almost all undergraduates) using adaptive survey software we designed especially for this experiment (using the Rand Corporation's Multimode Interviewing Capability technology). Sessions were scheduled for 2 hours each, but mean completion time was 68 minutes (not including initial introductions and interactions which last about 10 minutes). Subjects were paid $40 for 2 hours. This lack of direct incentivization is often worrying in survey research. However, our focus is on saving for retirement; fully incentivizing subjects with a large fraction of lifetime resources is well beyond our budget constraints (and would be even be well beyond our budget constraint if implemented in a developing country)[5]. The first round consisted of 317 subjects (July, November, and December 2013). The second round consisted of 311 subjects (April 2014). We invited all 311 subjects from the second round back to the laboratory for a follow-up session 2-4 weeks after the initial experiment. Almost 85% (264 subjects) returned for the second wave. During this second wave, subjects only receive the Pre-Test, an abridged Training

risk aversion over larger-stakes choices because this discrepancy is a major stumbling block to determining an appropriate retirement portfolio default allocation. Rather than trying to explain why large-stakes risk tolerance is inconsistent with small-stakes risk aversion, we focus instead on identifying at *which* of the intermediate steps of the argument the subject is inconsistent. This approach has the advantage that we can pinpoint which decision errors subjects are making, as well as which normative axioms the subjects themselves repudiate.

[5]Waiting many decades for payouts to realize is also more of an ambitious experiment than we hoped to undertake. Fortunately, our Cornell student population, although not representative, is perhaps non-representative in a way that makes up for the lack of direct, financial incentive. Cornell students are particularly motivated to perform well on difficult tasks, so much so that this (non-representative) internal motivation might be sufficient to overcome lack of direct, financial incentive. That being said, other non-representative characteristics of the particular student subject pool might sully the external validity of our findings, although we cannot do much more than caveat this potential difficultly.

section, and both parts of the Main Body. Our motivation was to see how their decisions, risk aversion, consistency, and transitivity change over the course of several weeks, and to what extent any learning that took place during the first wave "stuck" with participants over the short-term. We should note each participant who returns for a second wave sees a re-randomized version of the survey—this mitigates concern over participants simply remembering and repeating exactly what they did in the previous wave.

2.5.3 Pre-Test

The Pre-Test elicited preliminary risk aversion estimates for each subject using simple binary choices between hypothetical safe and risky assets. We present here a synopsis of our findings using these data. For more information on the questions asked, and for the exact details and outputs of these analyses, see "Appendix: Pre-Test." After calculating cardinal measures of risk aversion from subjects' categorical responses (using the procedure developed by Kimball et al., 2008), we found (1) higher cognitive function was associated with lower risk aversion in both waves and (2) male subjects were generally less risk averse. These align with our intuition and past studies on risk aversion. Interestingly, among subjects surveyed in two waves, higher cognitive function was associated with lower risk aversion to a greater extent in the first wave than in the second wave. This is consistent with our procedure lessening the impact of cognition on measured risk aversion over the course of several weeks. Finally, we implemented a randomization for all 311 subjects surveyed in the second round: approximately half were given the Pre-Test at the end of the survey (instead of at the beginning). In both waves 1 and 2, being given the Pre-Test at the end of the survey was associated with lower imputed log risk aversion. Again, if we think of risk aversion as being associated with cognitive bias, our procedure (at least temporarily) appears to alleviate this bias.

2.5.4 Training Batteries

The training sections reviewed basic facets of probability theory (through examples using coin tosses and dice rolls), taught subjects to interpret the symbols and figures used to convey the choices they would be faced with in the main body of the survey, and reviewed central assumptions we wanted subjects to hold while making their decisions. Each training module was followed by a short quiz meant to test subjects' understanding. Subjects could not continue to the next part of the survey until they got nearly all quiz questions correct.

The assumptions participants were instructed to make were as follows: (1) "The government provides free medical insurance, and you are in good health," (2) "The government

no longer provides social security (i.e. monthly checks)," (3) "There is no inflation," (4) "Imagine that your friends and extended family outside of your household do not need financial help from you, and you cannot ask them for money," (5) "When you retire at age 65, you plan to move into rental housing that will have a monthly payment," and (6) "You have no other resources beyond the amounts specified by your decisions. For example, any money you get from selling your existing home has already been figured into the yearly spending you can afford." See the Survey Walk-Through Appendix for more details on all the training sections.

2.5.5 Main Body Part 1: Elicitation of Untutored Preferences

Here we elicited subjects' untutored preferences among the same 5 different investment plans using 7 different ways of framing the decision. This section, which stops after eliciting untutored risk preferences, without any experimental intervention, is comparable to most existing studies of risk preferences and is of interest in its own right. Again, all questions involve hypothetical payoffs in "amount you have to spend each year during retirement, from age 65 on."

We randomized across participants which set of monetary amounts[6] was used as payoffs (among 6 different sets)[7]. To be clear, each subject in each wave only ever saw one set of monetary amounts (i.e. monetary amounts are held fixed for a given subject and wave). Subjects who returned for a second wave received a re-randomized version of the survey, thereby potentially receiving a different set of monetary amounts. However, within a wave, a participant was always asked questions drawing from the same monetary amounts. The 6 possible sets of monetary differ in how attractive the risky options is (see Table 2.3). Each monetary amount is associated with a different level of constant relative risk aversion (CRRA): 1.576, 2.958, 4.865, 7.184, 12.113, or 17.967. The 3rd, 5th, and 6th amounts shown in Table 2.3 are always 100k, 150k, and 225k, respectively. The 2nd amount is the

[6]Our goal in constructing these dollar amounts was to provide gambles that would maximize initial inconsistencies. Subjects are faced with three fundamental choices: A v. B, C v. D, and E v. F (where the latter two are only relevant if "B" is initially chosen, resulting in 5 potential investment plans). However, according to expected utility and CRRA, and given our monetary amounts, choosing the safer option in one pair usually implies one should also choose the safer option in the other pairs (and vice-versa for the choosing the risky option). We did not choose these amounts to maximize power in measuring cardinal risk aversion (see below MLE).

[7]Unfortunately, our first round of data collection (317 students in summer 2013) were found to have suffered from a systematic programming error. Inadvertently, our software was only randomizing between 2 sets of monetary levels (those amounts that made riskier investment plans more attractive). Therefore, during the second round of data collection, we only randomized among the 3 monetary levels that made the riskier investment plans look the least attractive. One monetary level was never used. This approach does not substantively differ from having simply randomized over the 5 monetary levels throughout the entire data collection process.

dollar value such that an expected utility maximizer with CRRA preferences (given the constant coefficient of relative risk aversion associated with that set of monetary amounts) would be indifferent between 100k for sure and a 50-50 chance of 150k and the unknown amount. Finally, the 1st amount is the 2nd amount squared and divided by 100.

1	**2**	**3**	**4**	**5**	**6**
52K	64K	74K	81K	88K	92K
72K	80K	86K	90K	94K	96K
100K	100K	100K	100K	100K	100K
108K	120K	129K	135K	141K	144K
150K	150K	150K	150K	150K	150K
225K	225K	225K	225K	225K	225K

Table 2.3: Columns represent 6 potential sets of monetary amounts, associated with constant coefficients of relative risk aversion of 1.576, 2.958, 4.865, 7.184, 12.113, or 17.967, respectively. Descriptions of how these amounts are calculated can be found in the main text.

Approximately half of subjects were further randomized and given monetary amounts that are exactly half of those appearing in Table 2.3. According to CRRA this should not make a difference.

We also randomized the order that the frames were initially asked during Main Body Part 1. This order was also used to determine the order in which subjects were allowed to update their choices in Main Body Part 2. For approximately one-third of subjects, the order was: Single Action in Isolation, Single Action with Backdrop, Two Contingent Actions with Backdrop, Complete Contingent Action Plan, Pairwise Choices Between Complete Strategies, Pairwise Choices Between Compound Lotteries, and Pairwise Choices Between Reduced Simple Lotteries. For another one-third, the order was reversed. A final one-third was given a random ordering. This provides a means of empirically testing whether reasoned preferences are more relevant for optimal policy than untutored preferences. If reasoned preferences truly reflect underlying preferences, then reasoned preferences should not depend on the order of reasoning. Unfortunately, we are not yet statistically well powered enough to analyze this randomization using our MLE procedure (see below). Further planned data collection, however, will facilitate this analysis.

Approximately half of subjects received all of the survey questions oriented exactly as appear in the figures presented in this section; remaining subjects were randomly assigned to a "down" randomization that flipped the orientation of all figures such that option "A" appeared on the bottom rather than at the top. This was meant to mitigate the tendency of people to choose things toward the top of surveys and figures. Fortunately, this did not

seem to matter for subjects' choices. Also, throughout the initial elicitation of preferences, the ordering of questions was randomized within a frame (for frames with more than one question). Moreover, for relevant questions, the orientation of which option appeared on the left and which option appeared on the right was also randomized. This was again to mitigate the tendency of option orientation on the screen to impact choices.

2.5.6 Psychological and Cognitive Batteries

Interspersed between frames in Main Body Part 1, we measured several psychological and cognitive covariates. These included: cognitive reflection task battery (Frederick, 2005), number series battery (adopted from the CogUSA, McArdle et al., 2007-2009), abbreviated Big-Five personality battery (Gosling et al., 2003), probabilistic sophistication battery (developed by Miles Kimball, see the Survey Walk-Through Appendix for details), need for cognition battery (Cacioppo et al., 1984), and elicitation of SAT scores. Batteries appeared in random order in between the elicitation of frames in Main Body Part 1. In other words, subjects alternated between answering all of the questions in a particular frame and these psychological and cognitive batteries.

2.5.7 Main Body Part 2: Elicitation of Reasoned Preferences

During Main Body Part 2 of the experiment we elicited reasoned preferences by providing subjects the opportunity to systematically review and update their choices from each of the 7 frames answered in Part 1. The adaptive software systematically confronted subjects with one round of inconsistency checks (providing them the opportunity to update inconsistencies between similar frames, as well as the opportunity to update a random subset of already-consistent choices), one round of intransitivity checks (allowing them to explicitly rank options within a frame among which cyclical preferences are detected), another round of inconsistency checks (using an identical algorithm as the first round), and another round of intransitivity checks (also using an identical algorithm as the first round). We consider a participant's choices after these checks to be their reasoned preferences. The purpose of this experimental intervention is to cause subjects to reason more fully through their risk preferences. By comparing their reasoned preferences with their untutored preferences, we can learn where untutored preferences exhibited decision errors that advisors and policymakers could help individuals overcome. In designing this part of the experiment we tried to avoid being paternalistic, but instead strove to be deferentially light-handed. We note that while our focus is on risk preferences, this approach could be adopted for any sort of preferences (e.g. time preferences).

2.5.8 Inconsistency Checks

The procedure for a particular inconsistency worked as follows: (1) when two choices subjects made violated some axiom in Table 2.2, they were asked whether they think those two choices should be the same ; (2) if they said their choices should be the same, they were given the opportunity to revise, and then were asked why they revised as they did; (3) if they said their choices should not be the same, they were asked why they should not be the same. During the first round of surveying (covering 317 subjects), responses to these "why" questions were open-ended. We used these open-ended responses to generate multiple-choice responses that were used in the second round (covering 311 subjects). For details on the exact wording of these questions, see "Appendix: Flow of Inconsistency Checks." In the vast majority of instances where a subject revised, they said it was because they either originally made a mistake or learned something new about their preferences in answering the questions. Experimenter demand was not often cited as a reason or switching. Those who did not revise most often cited that they believed the situations were different enough to merit different answers.

We also presented participants with "placebo inconsistencies" to test the influence of experimenter demand. These checks were randomly interspersed among the standard inconsistency checks. Subjects were given the opportunity to change already-consistent choices, using nearly identical text and questions as the normal inconsistency checks. Of course, the questions necessarily differed slightly: when subjects said their choices should be the same, they were asked "why" without being given the opportunity to update; when subjects said their choices should be different, they were given the opportunity to update, and then asked why they did so. Subjects were given a random number of these placebo checks. For each pair of questions in a step that could potentially trigger an inconsistency, but did not, a random one-third were pushed to the subject as placebo checks. Fortunately, we find that people updated already-consistent choices less than 2% of the time. In contrast, people revised actual inconsistencies more than 40% of the time. Indeed, our results do not appear to be driven by experimenter demand.

2.5.9 Intransitivity Checks

We also confronted participants with intransitivities (i.e. instances of cyclical preferences) among their choices in each of the Pairwise Strategy Choice Frames. Subjects were presented with, and asked to explicitly rank, the choices among which they had an intransitivity (or note that they could not rank them). Their pairwise choices in that frame were then adjusted accordingly. See Figure 2.8for an example of one such resolution. Given 10 ques-

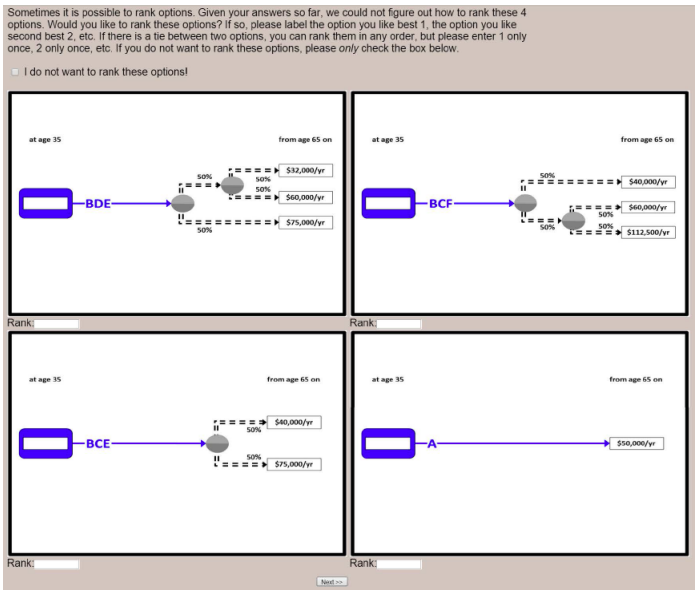

Figure 2.8: Example of an intransitivity resolution.

tions were necessary to elicit all pairwise preferences among 5 potential investment plans (i.e. 5 choose 2 = 10), intransitivities could come in a variety of flavors: 5-way, 4-way, or 3-way.

If a subject selected "I do not want to rank these options," they were given the opportunity to answer a follow-up question, "Why couldn't you rank the options on the previous slide?" Their options were: (1) "I couldn't rank the options because they are all equally good to me," (2) "I couldn't rank the options because I don't know which option I prefer," (3) "I should be able to rank the options, but it's extremely hard," or (4) "I couldn't rank the options for another reason."

2.5.10 Follow-Up and Demographic Batteries

This section first asked subjects how much they agreed that the survey evoked each of the following emotions on a 1-6 scale: enjoyment, annoyance, stress, and frustration (Fagerlin et al., 2007). Responses are shown in Table 2.4. Subjects somewhat enjoyed the experience (44.7% respond 4/6 or higher); were somewhat annoyed by it (54.2% respond 4/6 or higher); were not overly stressed by it (66.3% respond 3/6 or lower), and were slightly frustrated by it (36.2% respond 4/6 or higher).

Finally, we administered a demographics survey. Mean subject age was 20.9 years. Approximately 90% of subjects were between the ages of 18 and 22 (i.e. college students). Subjects were disproportionately female, with only 35.4% being male. Further details and questions can be found in the Survey Walk-Through Appendix.

Emotion	1 (least)	2	3	4	5	6 (most)
Enjoyment	11.8%	18.1%	25.4%	32.5%	9.0%	3.2%
Annoyance	4.6%	17.0%	24.2%	23.8%	17.8%	12.6%
Stress	23.8%	22.2%	20.3%	19.5%	10.2%	3.8%
Frustration	15.2%	21.1%	22.2%	24.6%	10.2%	6.6%

Table 2.4: Percentage of respondents answering on a scale of 1-6, "To what extent do you agree with the following statement": "I enjoyed thinking through these choices," "Thinking through these choices was annoying," "Thinking through these choices made me feel stressed," "Thinking through these choices was frustrating."

2.6 Results

We now describe which normative axioms subjects initially endorsed, where they were most likely to update their choices, and toward which frames they actually revised. Overall, for any given axiom, most people endorsed consistency, although there was substantial heterogeneity across axioms. There were always people who did not endorse consistency for a given axiom, and there were relatively few people who were completely consistent across axioms. However, subjects rarely moved from consistency to inconsistency.

2.6.1 Inconsistency Checks

Among all subjects' first wave responses, out of 30 potential inconsistencies, subjects averaged approximately 5.6 inconsistencies in their untutored preferences and approximately 2.6 inconsistencies in their reasoned preferences[8].

In order hone in on resolution behavior at each axiom, we derive "inconsistency rates" for each step for subjects' untutored and reasoned preferences. Steps differ in the number of inconsistencies possible between the two associated frames. Inconsistency rates control for these differences, facilitating cross-axiom comparisons of inconsistency. For a given step, the inconsistency rate is simply equal to the total inconsistencies for that axiom for a given subject, divided by the total potential number of inconsistencies in that step, averaged across all subjects.

See Table 2.5 for results from all 628 subjects (first wave only). We also report two-sided p-values associated with testing differences in inconsistency rates between untutored and reasoned preference for each axiom (using differences in proportions tests).

[8]Among participants who were surveyed during the second round (the vast majority of whom returned for a second wave), people average approximately 6 inconsistencies in their untutored preferences during the first wave, approximately 3 inconsistencies in their reasoned preferences at the end of the first wave, approximately 4 inconsistencies in their untutored preferences during the second wave, and approximately 2 inconsistencies in their reasoned preferences at the end of the second wave.

See Table 2.6 for results using only subjects who participated in the second round of surveying (covering 311 subjects in wave 1 and 264 subjects in wave 2). Here, we report inconsistency rates at 4 points in time (untutored and reasoned for each wave), and 3 sets of proportion tests (between untutored and reasoned preferences in the first wave, between wave 1 reasoned preferences and wave 2 untutored preferences, and between untutored and reasoned preferences in the second wave).

Axiom	**Total**	**Untutored**	**Reasoned**	**P-Value**
Irrelevance of Background Counterfactuals	2	0.118	0.055	$<$0.0005
Simple Actions = State-Contingent Actions	2	0.116	0.055	0.002
Irrelevance of Counterfactual Choices	2	0.101	0.107	0.602
Fusion + Shift from Nodewise to Pairwise	4	0.215	0.115	$<$0.0005
Complete Strategies = Implied Lotteries	10	0.189	0.079	$<$0.0005
Reduction of Compound Lotteries	10	0.221	0.080	$<$0.0005

Table 2.5: Average inconsistency rates for untutored and reasoned preferences, and two-sided tests for differences in proportions (i.e. differences in inconsistency rates). "Total" denotes total potential inconsistencies for a given axiom. Inconsistency rates are calculated by averaging across subjects: total inconsistencies divided by total potential inconsistencies, by axiom.

Results are similar whether we look at first wave responses from all subjects or restrict attention to first wave responses from the second round of surveying. Initially, subjects were particularly apt to endorse axioms involving the Nodewise Action Choices Frames relative to axioms involving the Pairwise Strategy Choices Frames. They were especially unlikely initially to endorse: (1) Shift Between Nodewise and Pairwise and (2) Reduction of Compound Lotteries. However, Reduction of Compound Lotteries also saw the most movement toward consistency.

The second wave of surveying (which was completed by 264 of 311 subjects invited to return 2-4 weeks after their first wave surveys) allows us to analysis the short-term impact of our procedure. Two axioms exhibited significant retention in reduced inconsistencies, as is evidenced by (1) significant differences in inconsistency rates between wave 1 untutored preferences and wave 1 reasoned preferences, and (2) insignificant differences in inconsistency rates between wave 1 reasoned preference and wave 2 untutored preferences: (1) Irrelevance of Background Counterfactuals and (2) Fusion + Shift from Nodewise to

Axiom	**Untut. Wave 1**	**Reas. Wave 1**	**Untut. Wave 2**	**Reas. Wave 2**	**P-Val. Wave 1**	**B/w P-Val.**	**P-Val. Wave 2**
Irrelevance of Background Counterfactuals	0.113	0.056	0.079	0.035	<0.0005	0.1137	0.0010
Simple Actions = State-Contingent Actions	0.117	0.084	0.058	0.059	0.0477	0.0768	0.9040
Irrelevance of Counterfactual Choices	0.119	0.138	0.084	0.111	0.3090	0.0021	0.1038
Fusion + Shift from Nodewise to Pairwise	0.207	0.126	0.150	0.096	<0.0005	0.0917	<0.0005
Complete Strategies = Implied Lotteries	0.202	0.089	0.129	0.075	<0.0005	<0.0005	<0.0005
Reduction of Compound Lotteries	0.236	0.092	0.161	0.083	<0.0005	<0.0005	<0.0005

Table 2.6: Average inconsistency rates for untutored and reasoned preferences in waves 1 and 2, respectively, and two-sided tests for differences in proportions (i.e. differences in inconsistency rates) within wave 1 between untutored and reasoned preferences, between wave 1 reasoned preferences and wave 2 untutored preferences, and within wave 2 between untutored and reasoned preferences. Inconsistency rates are calculated by averaging across subjects: total inconsistencies divided by total potential inconsistencies, by axiom.

Pairwise.

Irrelevance of Counterfactual Choices was the only step in which subjects exhibited essentially no updating. One possible theory for why there was such little updating between these questions is because subjects forgot to treat frames as independent. They might have been trying to diversify their answers across frames in a misplaced attempt to adopt one of the central tenants of investing[9]. A more testable theory is whether subjects who choose "A" in the Complete Contingent Action Plan are driving the results. Recall for subjects surveyed during the second round, if a person chose "A" in the Complete Contingent Action Plan, they were prompted for their second-favorite choice as though they could not choose B (distinct from the Two Contingent Actions with Backdrop decision, which does not even show A v. B). These people might have cared far less about their second-favorite options, which were used to determine whether they had an inconsistency in this step, and therefore are unusually, highly inconsistent. Considering only participants who were *not* given the opportunity to report their second-favorite choice conditional on choosing "A" in the Complete Contingent Action Plan, the inconsistency rate decreases slightly, but the drop is not statistically significant. Considering participants who *could* report a second-favorite choice, both (1) subjects who did *not* choose "A" (and therefore whose inconsistencies are driven by their favorite choice) and (2) subjects who *did* in fact choose "A" (and therefore whose inconsistencies are driven by their second-favorite choice), show no statistically significant change in the inconsistency rate in the first wave. However, both groups show a small, significant increase in the inconsistency rate in the second wave. These findings suggest that people choosing "A" are *not* driving the lack of increased consistency in Irrelevance of Counterfactual Choices.

Aside from this anomaly, it appears that subjects readily updated toward consistency. However, there are other measures of consistency that paint a less rosy picture. See Table 2.7for the percentages of respondents with any inconsistency among all possible axioms, restricting attention to subjects in the second round of surveying who were invited back for a second wave. Although approximately 87% of subjects initially had at least one inconsistency, this number only declined to approximately 69% by the end of the first wave. Wave 2 saw the percentage with at least one inconsistency go from 74% to 58% from untutored to reasoned preferences. So, while there were significant improvements toward consistency in most individual axioms, most subjects still retain at least one inconsistency after two full sessions.

[9]Unfortunately this is the sole step in which the follow-up questions about why a subject wanted to retain an inconsistency were inadvertently not implemented. It is plausible they would have responded that the frames differed significantly enough that they wanted to make different choices (as in all other steps), though we cannot say for sure.

Survey Stage	Mean	Std. Dev.
Wave 1 Untutored	0.871	0.335
Wave 1 Reasoned	0.688	0.464
Wave 2 Untutored	0.736	0.441
Wave 2 Reasoned	0.576	0.495

Table 2.7: Percentage of respondents with any inconsistency.

Table 2.8 shows, for each axiom, conditional on being inconsistent, the percentage of the time subjects chose to update (or not). For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2. Consistent with our findings above, subjects are least likely to update when faced with inconsistencies in Irrelevance of Counterfactual Choices. For the axiom with the most updating toward consistency, Reduction of Compound Lotteries, we see subjects tend to update toward their choices in the non-reduced lotteries (i.e. from their initial decisions in Pairwise Choices Between Reduced Simple Lotteries toward their decisions in Pairwise Choices Between Compound Lotteries).

Table 2.9 details why subjects said they did not revise (only among participants surveyed during the second round, i.e. those whose answers come from multiple-choice questions). Results are similar across the axioms. Overall, 57% of the time subjects said the situations were sufficiently different as to merit different answers; 25% of the time they said they were indifferent[10]; and the remainder of the time subjects were either deferring to apparent experimenter demand, did not know what their preferences were, were confused, or "Other."

Table 2.10 details why subjects said they revised one of their answers, again only among participants surveyed during the second round. Results are again broadly similar across the axioms. Overall, 45% of the time subjects cite that they previously made a mistake in making their choices; 35% of the time that they learned something new about their preferences in answering questions; and 12% of the time that they are indifferent. The remainder of the time, they cite some combination of experimenter demand, not knowing their own preferences, confusion, or "Other." The axiom Simple Actions = State-Contingent Actions is a bit of an outlier: subjects are only 37% likely to say they initially made a mistake, 37% likely to say they learned something new, 17% likely to cite indifference (with literally nobody citing experimenter demand).

For more details on how subjects update, particularly whether they tend to update toward riskier or safer options, see "Appendix: Updating Conditional on Riskiness of Choices." It appears there is a weak preference toward updating toward riskier choices,

[10]Note that in these cases, we still consider subjects to be "inconsistent" despite specifically saying they were indifferent.

but this is by no means ubiquitous across axioms.

Axiom	Toward Frame 1	Toward Frame 2	No Update	Swap	Total Choices
Irrelevance of Background Counterfactuals	23%	24%	49%	3%	197
Simple Actions = State-Contingent Actions	20%	12%	62%	5%	196
Irrelevance of Counterfactual Choices	7%	3%	88%	0%	246
Fusion + Shift from Nodewise to Pairwise	21%	17%	55%	6%	449
Complete Strategies = Implied Lotteries	21%	25%	48%	4%	1504
Reduction of Compound Lotteries	27%	20%	45%	5%	2005

Table 2.8: Percentage of the time that subjects updated toward each frame, did not update, or swapped their choices for normal inconsistency checks.

Tables 2.11, 2.12, and 2.13 show these percentages for the placebo inconsistency checks. These results stand in stark contrast with those from the regular inconsistency checks. Subjects only very rarely update results that are already consistent. When asked why they do not update, more than 89% of the time subjects cite that the situations are similar enough to merit the same answer, 7% of the time subjects cite indifference, and the remainder of the time they cite some combination of experimenter demand (only 0.4%), that they don't know their own preferences, confusion, or "Other." Results are similar across axioms. In the rare occasions where subjects do switch an already-consistent answer, they tend to cite having learned something new about their preferences, indifference, having made a mistake before, and confusion. There is heterogeneity among axioms, but these instances happen so infrequently as not to merit further analysis.

2.6.2 Intransitivity Checks

The results from the intransitivity checks (only applicable to the Pairwise Strategy Choice Frames) are also interesting. Let us focus on participants who are surveyed during the second round of surveying (and hence are invited back for a second wave)[11].

[11]Focusing on all subjects' first wave responses yields similar results. Overall, subjects' untutored preferences averaged 1.2 intransitivities among all Pairwise Strategy Choice Frames. After the first round of

Axiom	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**Total**
Irrelevance of Background Counterfactuals	54.6%	21.6%	3.1%	9.3%	9.3%	2.1%	97
Simple Actions = State-Contingent Actions	73.2%	17.9%	2.4%	1.6%	4.1%	0.8%	123
Irrelevance of Counterfactual Choices	55.0%	25.8%	1.7%	9.2%	3.3%	5.0%	120
Fusion + Shift from Nodewise to Pairwise	62.7%	20.5%	2.0%	7.2%	4.0%	3.6%	249
Complete Strategies = Implied Lotteries	55.9%	24.3%	3.7%	7.3%	3.2%	5.7%	725
Reduction of Compound Lotteries	54.5%	27.4%	3.0%	5.1%	4.1%	5.9%	920
Total	56.9%	24.8%	3.0%	6.3%	4.0%	5.1%	2234

Table 2.9: Percentage of the time that subjects gave each of the following responses to "Why do you want to make different choices in these two situations?": (1) "The two situations are different enough that I want different choices", (2) "Some of the options are equally good to me, so it doesn't matter which one I choose", (3) "I chose how I thought the experimenters wanted me to chose", (4) "I don't know which options I prefer", (5) "I don't know or am confused", or (6) "Other."

Axiom	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Total
Irrelevance of Background Counterfactuals	47.9%	35.1%	11.7%	1.1%	1.1%	1.1%	2.1%	94
Simple Actions = State-Contingent Actions	36.9%	36.9%	16.9%	0.0%	4.6%	1.5%	3.1%	65
Fusion + Shift from Nodewise to Pairwise	47.1%	31.0%	13.2%	1.1%	4.6%	2.3%	0.6%	174
Complete Strategies = Implied Lotteries	45.7%	34.7%	10.9%	1.0%	4.7%	1.3%	1.8%	709
Reduction of Compound Lotteries	45.1%	36.4%	12.0%	1.1%	2.3%	1.4%	1.6%	979
Total	45.4%	35.3%	11.8%	1.0%	3.4%	1.4%	1.7%	2021

Table 2.10: Percentage of the time that subjects gave each of the following responses to "Why did you want to change your choices as you did?": (1) "I made a mistake when I first chose", (2) "Answering all of these questions made me change what I want", (3) "Some of the options are equally good to me, so it doesn't matter which one I choose", (4) "I chose how I thought the experimenters wanted me to chose", (5) "I don't know which options I prefer", (6) "I don't know or am confused", or (7) "Other."

Axiom	Update Frame 1	Update Frame 2	No Update	Update Both	Total
Irrelevance of Background Counterfactuals	0%	0%	100%	0%	506
Simple Actions = State-Contingent Actions	0%	0%	100%	0%	449
Irrelevance of Counterfactual Choices	0%	0%	99%	0%	172
Fusion + Shift from Nodewise to Pairwise	1%	3%	95%	0%	355
Complete Strategies = Implied Lotteries	0%	0%	97%	0%	2107
Reduction of Compound Lotteries	1%	0%	97%	0%	2249

Table 2.11: Percentage of the time that subjects updated each frame, did not update, or updated both their choices for placebo inconsistency checks.

Axiom	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**Total**
Irrelevance of Background Counterfactuals	90.2%	7.6%	0.2%	0.8%	0.6%	0.6%	490
Simple Actions = State-Contingent Actions	85.2%	11.1%	0.2%	1.9%	0.9%	0.7%	432
Irrelevance of Counterfactual Choices	85.5%	9.3%	0.0%	0.0%	1.7%	3.5%	172
Fusion + Shift from Nodewise to Pairwise	91.4%	7.1%	0.0%	0.6%	0.6%	0.3%	338
Complete Strategies = Implied Lotteries	88.9%	7.0%	0.5%	1.2%	1.1%	1.3%	2063
Reduction of Compound Lotteries	90.6%	6.7%	0.4%	0.8%	1.0%	0.6%	2191
Total	89.4%	7.3%	0.4%	1.0%	1.0%	0.9%	5686

Table 2.12: Percentage of the time that subjects gave each of the following responses to "Why do you want to make the same choices in these two situations?": (1) "The two situations are similar enough that I want to make the same choices", (2) "Some of the options are equally good to me, so it doesn't matter which one I choose", (3) "I chose how I thought the experimenters wanted me to chose", (4) "I don't know which options I prefer", (5) "I don't know or am confused", or (6) "Other."

Axiom	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**Total**
Irrelevance of Background Counterfactuals	16.7%	50.0%	25.0%	0.0%	0.0%	8.3%	0.0%	12
Simple Actions = State-Contingent Actions	23.1%	7.7%	15.4%	7.7%	0.0%	23.1%	23.1%	13
Fusion + Shift from Nodewise to Pairwise	26.7%	13.3%	20.0%	6.7%	6.7%	20.0%	6.7%	15
Complete Strategies = Implied Lotteries	13.3%	40.0%	23.3%	0.0%	3.3%	13.3%	6.7%	30
Reduction of Compound Lotteries	23.5%	26.5%	29.4%	0.0%	0.0%	17.6%	2.9%	34
Total	20.2%	28.8%	24.0%	1.9%	1.9%	16.3%	6.7%	104

Table 2.13: Percentage of the time that subjects gave each of the following responses to "Why did you want to change your choices as you did?": (1) "I made a mistake when I first chose", (2) "Answering all of these questions made me change what I want", (3) "Some of the options are equally good to me, so it doesn't matter which one I choose", (4) "I chose how I thought the experimenters wanted me to chose", (5) "I don't know which options I prefer", (6) "I don't know or am confused", or (7) "Other."

During the first wave, initially, subjects averaged 1.3 intransitivities in their untutored preferences. During this wave, subjects were confronted with a total of 433 intransitivities; 20% of the time subjects did not update, instead checking the box: "I do not want to rank these options!" These subjects were given the opportunity to answer a follow-up question, "Why couldn't you rank the options on the previous slide?" Their options were: (1) "I couldn't rank the options because they are all equally good to me," (2) "I couldn't rank the options because I don't know which option I prefer," (3) "I should be able to rank the options, but it's extremely hard," or (4) "I couldn't rank the options for another reason." 59% of subjects who could not rank answered that they did not know which option they preferred; 27% answered that it was extremely hard; the remainder were evenly split between the other two options. After the first wave was over (i.e. after 2 rounds of inconsistency checks and 2 rounds of intransitivity checks), subjects' reasoned preferences only averaged 0.4 intransitivities.

In the second wave, subjects were confronted with a total of 313 intransitivities, averaging 1.1 and 0.6 intransitivities in their untutored and reasoned preferences, respectively. There thus appeared to be mild "stick" to our procedure with respect to transitivity, but less so than with respect to consistency. Not only did people have approximately the same number of intransitivities at the beginning of the second wave as at the beginning of the first wave, but by the end of the second wave they appeared to have even *more* intransitivities than at the end of the first wave. It thus appears that sometimes, as people become more consistent, they become less transitive (something very much at odds with traditional decision theory). 39% of the time during the second wave, subjects did not update, citing that they did not want to rank the options among which they had an intransitivity. 43% of these subjects followed-up that they did not know which option they preferred; 20% answered that it was extremely hard; 24% responded that the options were all equally good; and the rest cited "another reason."

There was little heterogeneity in terms of incidence of intransitivities among the Pairwise Strategy Choice Frames. One exception was the fact that both waves' untutored preferences in Pairwise Choices Between Reduced Simple Lotteries were slightly more likely to be intransitive. However, in both waves, this difference disappears by the time reasoned preferences were reached. With respect to the type of intransitivities: 3-way intransitivities

inconsistency checks, average intransitivities dropped to 1.0, indicating that as people initially became more consistent, they also become more transitive. The initial round of intransitivity checks further lowered this average to 0.4. The second round of inconsistency checks slightly raised the average number of intransitivities to 0.5, only to see it drop to 0.3 after the second round of intransitivity checks. Untutored preferences in Pairwise Choices Between Reduced Simple Lotteries were slightly more likely to be intransitive. However, this difference disappeared by the time reasoned preferences were reached.

were the most common (52% overall), followed by 4-way intransitivities (32% overall), and finally 5-way (16%). These percentages were roughly homogenous across frames and over time.

2.7 Maximum Likelihood Estimation

Let us now turn to quantifying risk aversion and incidence of decision errors using maximum likelihood estimation (MLE). In so doing, we wish to explore how estimates of risk aversion and incidence of decision errors vary by frame, as well as how they change between untutored and reasoned preferences[12]. Estimation is carried out using the Berndt-Hall-Hall-Hausman (BHHH) numerical optimization algorithm, which uses the outer product of scores to approximate the hessian (see Berndt et al., 1974). We use a MATLAB implementation programmed by Thomas Jørgensen[13], which also enables the computation of robust standard errors. Although orders of magnitude faster, point estimates using BHHH match those found using global search optimization.

We assume subjects are expected utility maximizers with utility functions that obey constant relative risk aversion (CRRA) of the canonical form, with (directly unobserved) risk aversion parameter γ_{ift}, for individual i, frame f, and snapshot t. We use the term "snapshot" to refer to a particular point in the survey. We denote snapshot 1 as after elicitation of initial, untutored preferences; snapshot 2 as after the initial round of inconsistency checks; snapshot 3 as after the initial round of intransitivity checks; snapshot 4 as after the second round of inconsistency checks; and finally snapshot 5 as after the second round of intransitivity checks (i.e. when subjects have reached their first set of reasoned preferences). Snapshots 6-10 refer to the same points in the survey for the second wave. Because of individual response error, ϵ_{ift}, which is assumed to be the same across questions within a snapshot and frame, we can only observe η_{ift}. We also make the following distributional assumptions:

$$x_{ift} = ln(\gamma_{ift}) \tag{2.5}$$

[12]To provide further evidence on whether reasoning ability plays a role in driving a wedge between untutored and reasoned preferences, we will eventually, once we have more data, explore how cognitive traits relate to untutored preferences, reasoned preferences, and the frequency and type of decision errors we uncover. Are reasoned preferences less correlated with cognition and more homogeneous across individuals than untutored preferences? How does untutored risk aversion vary with gender? How about reasoned risk aversion? Do reasoned preferences depend on the order in which individuals reason through their violations of normative axioms?

[13]http://www.econ.ku.dk/phdstudent/jorgensen/code.htm

$$\eta_{ift} = x_{ift} + \epsilon_{ift} \tag{2.6}$$

$$\epsilon_{ift} \sim N(0, \sigma^2_{\epsilon_{ft}}) \tag{2.7}$$

$$x_{ift} \sim N(\mu_{ft}, \sigma^2_{x_{ft}}) \tag{2.8}$$

Our MLE therefore estimates the vector of parameters: (μ_{ft}, $\sigma_{x_{ft}}$, $\sigma_{\epsilon_{ft}}$). μ_{ft} is the mean of the distribution of log risk aversion for a particular frame and snapshot; $\sigma_{x_{ft}}$ is the standard deviation of the distribution of log risk aversion for a particular frame and snapshot. $\sigma_{\epsilon_{ft}}$ represents within-frame decision error for a particular frame and snapshot.

Next, we define CE_k as the log certainty equivalent of choice k (where k can be one of: A, BCE, BCF, BDE, or BDF) conditional on some known coefficient of relative risk aversion λ, as well as the (also known) monetary amounts on which the subject's payoffs are based (see Table 2.3 where ML_r denotes the monetary amount in row r for the specific column randomly assigned to a given subject):

$$CE_A = ln[ML_3] \tag{2.9}$$

$$CE_{BCE} = ln\left[((1-\lambda)(\frac{1}{2}\frac{{ML_2}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{ML_5}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right] \tag{2.10}$$

$$CE_{BCF} = ln\left[((1-\lambda)(\frac{1}{2}\frac{{ML_2}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{ML_4}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{ML_6}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right] \tag{2.11}$$

$$CE_{BDE} = ln\left[((1-\lambda)(\frac{1}{4}\frac{{ML_1}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{ML_4}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{ML_5}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right] \tag{2.12}$$

$$CE_{BDF} = ln\left[((1-\lambda)(\frac{1}{4}\frac{{ML_1}^{1-\lambda}}{1-\lambda} + \frac{1}{2}\frac{{ML_4}^{1-\lambda}}{1-\lambda} + \frac{1}{4}\frac{{ML_6}^{1-\lambda}}{1-\lambda}))^{\frac{1}{1-\lambda}}\right] \tag{2.13}$$

2.7.1 Nodewise Action Choice Frames

As described above, in each of the Nodewise Action Choice Frames, respondents choose among either 4 or 5 different investment plans: A, BCE, BCF, BDE, or BDF. "A" is an

option only in the Complete Contingent Action Plan. Let CE_k denote the chosen plan's conditional log certainty equivalent. Let CE_{-k} denote the vector of other plans' conditional log certainty equivalents. Utilizing a multinomial logit discrete choice functional form, the MLE problem can be written as follows, where m either iterates to 4 or 5:

$$\max_{\mu,\sigma_x,\sigma_\epsilon} \sum_{i=1}^{n} ln \left(\int_{-\infty}^{\infty} \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma_x^2}} f\left[\frac{CE_k}{\sigma_\epsilon}, \frac{CE_{-k}}{\sigma_\epsilon} \right] \mathrm{d}x \right) \tag{2.14}$$

$$f(\frac{CE_k}{\sigma_\epsilon}, \frac{CE_{-k}}{\sigma_\epsilon}) = \frac{1}{\sum_m e^{\frac{CE_m - CE_k}{\sigma_\epsilon}}} \tag{2.15}$$

For Complete Contingent Action Plan, we conduct two separate analyses, treating them as two separate frames. "Complete-1" uses all available data for each snapshot, including both initial response data, and, where available, follow-up data on subjects' second-favorite choices (conditional on having initially chosen "A"). This specification requires a slight alteration of the likelihood function relative to what is written above. The inner likelihood function with $m = 5$ (representing their first-favorite choice versus the other 4 options) remains the same. However, when the person originally chose "A" and indeed was prompted to provide their second-favorite choice, we multiply the inner likelihood function with another nearly identical likelihood function that has $m = 4$ (representing their second-favorite choice versus the remaining 3 options).

On the other hand, "Complete-4" only considers 4 possible investment plans for each snapshot (all those except "A"). For subjects who initially chose "A," we consider their second-favorite choice among the other 4 options, if that information is available; otherwise, the subject is not included in the estimate. For subjects who initially did not choose "A," we only consider their favorite choice among the 4 options sans "A."

Note that all available data are used for each snapshot and frame, meaning the first 5 snapshots are better powered than the latter 5 (given all 627 subjects have data on the first wave, i.e. the first 5 snapshots, but only 264 of a possible 311 subjects returned for a second wave, i.e. the latter 5 snapshots). Unfortunately, our analyses in this section are statistically underpowered across all snapshots (but especially for snapshots 6-10), resulting in likelihood functions that break down and behave erratically. This is driven by the fact that subjects in these frames are only asked for their favorite investment plan (or sometimes second-favorite investment plan) among the 5 possible options. This is in contrast with the Pairwise Complete Strategy Choice Frames, which are indeed well-powered, and which elicit a full ranking of preferences among the 5 possible investment plans (i.e. we have information not only on a subject's favorite investment plan, but their second favorite, third favorite, and fourth favorite.). Initial results by frame and snapshot can be seen in Figure

2.9. Note that Two Contingent Actions with Backdrop behaves erratically, with enormous standard errors even in early snapshots 2 and 3 (note the large scale of the y-axes). We conduct diagnostic tests to get a sense of the behavior of the likelihood function when our standard errors are large. We do this by manually varying one of the three parameters while fully optimizing over the remaining two. We next plot the log likelihood as a function of the manually varied parameter, noting its behavior. We repeat this separately for each of the three parameters. For all snapshots, Two Contingent Actions with Backdrop's likelihood functions indeed also behave erratically. These plots and more details of these tests can be found in "Appendix: MLE Diagnostic Tests."

We therefore ignore Two Contingent Actions with Backdrop and re-plot the MLE results (Figure 2.10). Still, two sets of results are problematic: estimates from snapshots 6-10 as well as both Complete Contingent Action Plan frames. The diagnostic procedure yields similar conclusions for these sets of estimates: their likelihood functions break down and act erratically. We therefore try individually dropping, in turn, (1) both Complete-1 and Complete-4 and (2) snapshots 6-10.

Results after dropping both Complete Contingent Action Plan frames can be seen in Figure 2.11, leaving only Single Action in Isolation and Single Action with Backdrop. Standard errors are again fairly large across all parameters. The most that can be concluded from these results is that the two frames are indistinguishable, which is consistent with out conclusions above. The diagnostic tests in the appendix still indicate a slightly erratic likelihood function.

Results after dropping snapshots 6-10 can be seen in Figure 2.12. These are similarly uninformative. Most frames are indistinguishable. The only apparent result is that "Complete-1" has a higher error response variance. This is intuitive: considering choices that involve both an initial choice among 5 investment plans, as well as a second-best choice among 4 investment plans (conditional on having chosen "A" initially), is more difficult than simply making a single choice among 4 investment plans (as is the case in all other frames in the figure).

2.7.2 Pairwise Complete Strategy Frames

The MLE for the Pairwise Complete Strategy Frames differs only slightly from the Node-wise Action Choice Frames in terms of setup, but its results are far more conclusive. Diagnostic tests described above are repeated and do not indicate unstable likelihood functions nor lack of statistical power (see "Appendix: MLE Diagnostic Tests"). Recall here that subjects must choose between each of 10 pairs of options. Let q index each of these 10

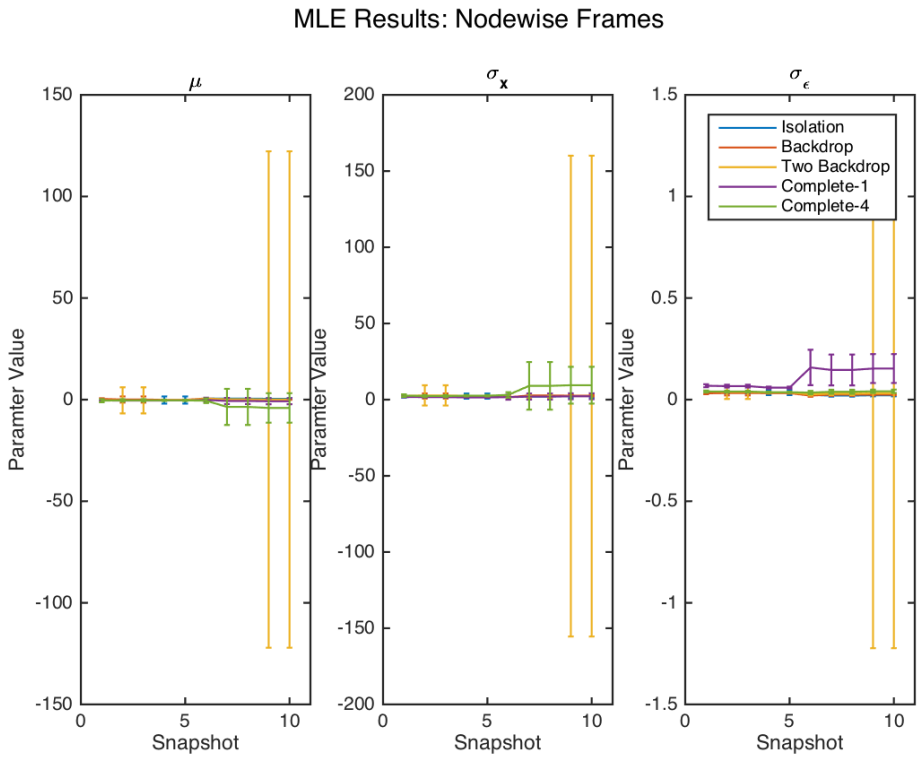

Figure 2.9: MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. "Isolation" refers to Single Action in Isolation; "Backdrop" refers to Single Action with Backdrop; "Two Backdrop" refers to Two Contingent Actions with Backdrop; "Complete-1" refers to Complete Contingent Action Plan with all available data; and "Complete-4" refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame.

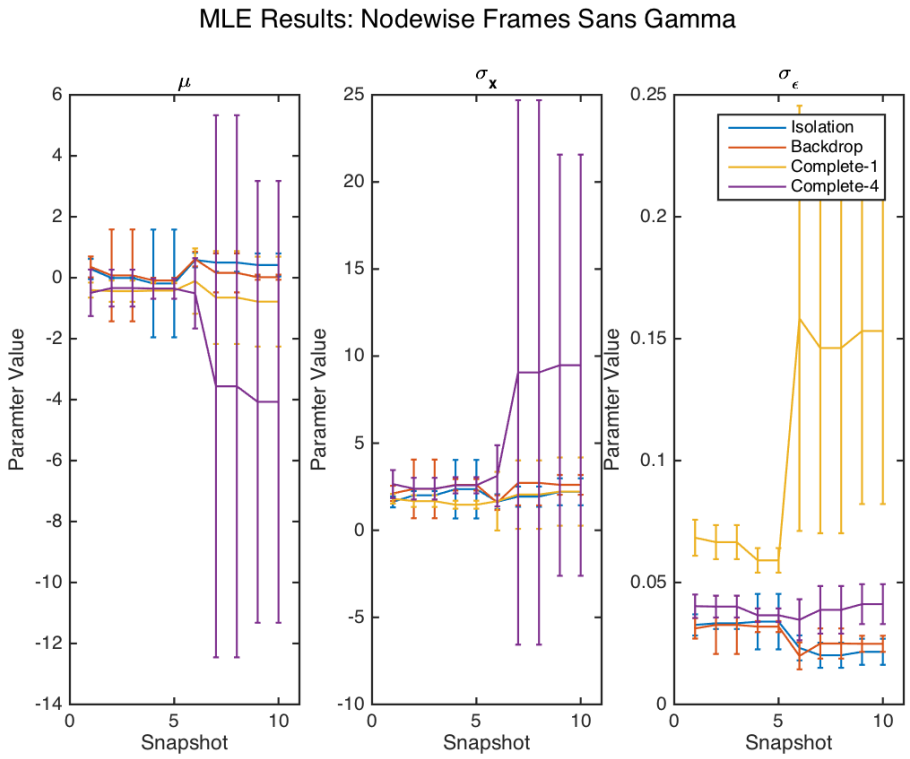

Figure 2.10: MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. "Isolation" refers to Single Action in Isolation; "Backdrop" refers to Single Action with Backdrop; "Complete-1" refers to Complete Contingent Action Plan with all available data; and "Complete-4" refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame.

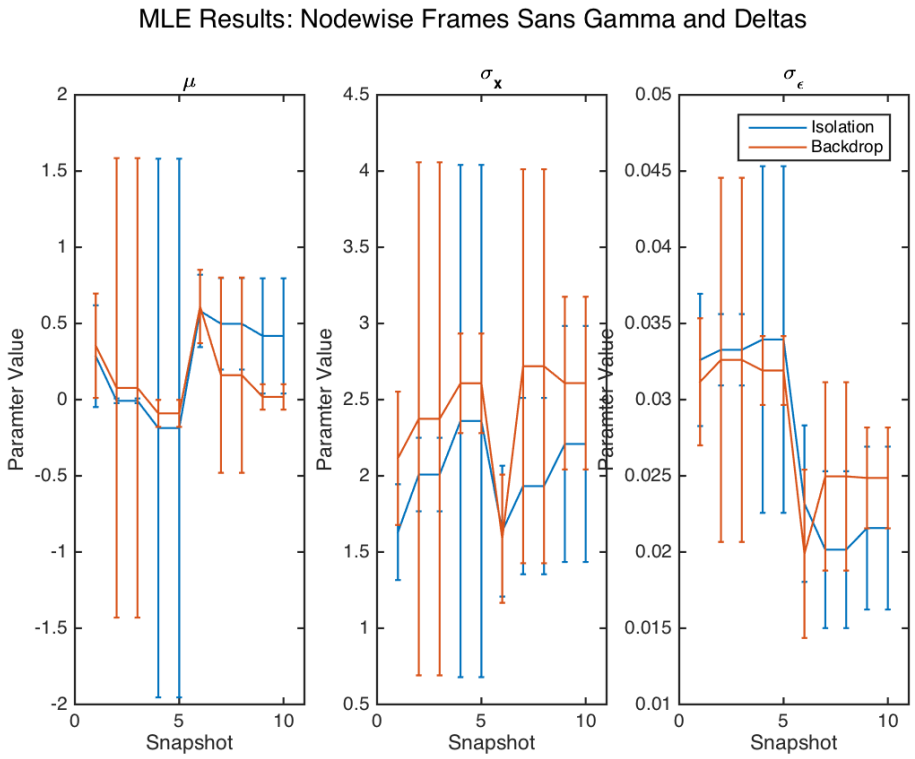

Figure 2.11: MLE results for Nodewise Action Choices Frames, snapshots 1-10. Error bars denote standard errors. "Isolation" refers to Single Action in Isolation; and "Backdrop" refers to Single Action with Backdrop. All available data are used for each snapshot and frame.

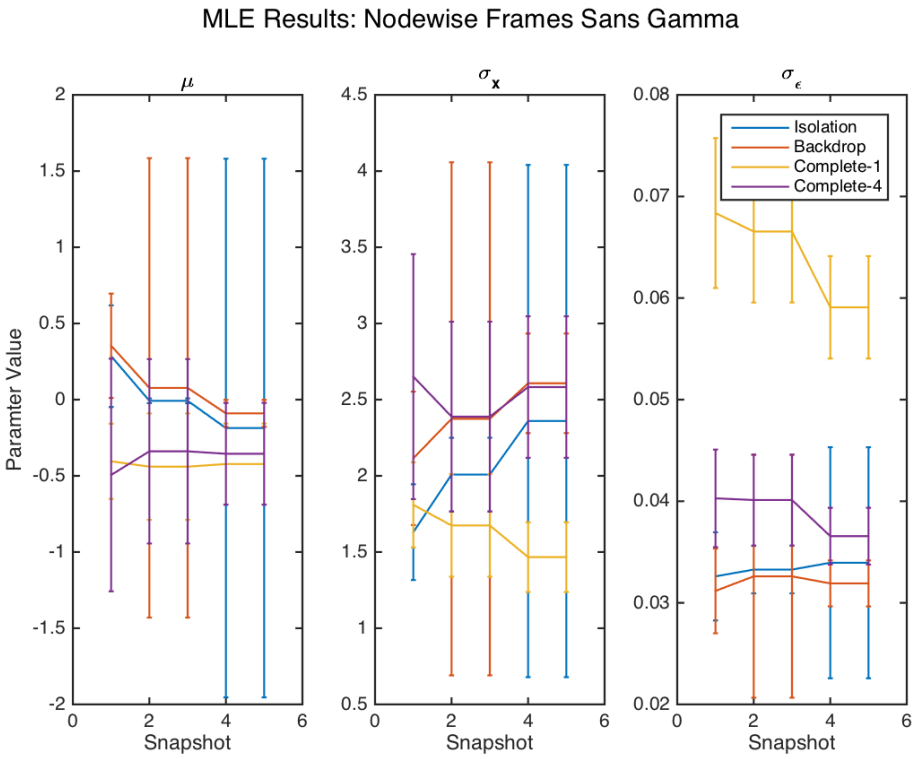

Figure 2.12: MLE results for Nodewise Action Choices Frames, snapshots 1-5. Error bars denote standard errors. "Isolation" refers to Single Action in Isolation; "Backdrop" refers to Single Action with Backdrop; "Complete-1" refers to Complete Contingent Action Plan with all available data; and "Complete-4" refers to Complete Contingent Action Plan but only using data on choices BCE, BCF, BDE, and BDF (not A). Unless otherwise indicated, all available data are used for each snapshot and frame.

pairs, and furthermore assume that for a given q, a subject chooses option $q1$ over $q2$ when the former certainty equivalent is greater than the latter (after including error on each term so a gamble is evaluated as $CE_k + \epsilon_k$, where response errors are distributed normal with zero mean and σ_ϵ). Utilizing 10 different multinomial logit discrete choice functions (each between 2 options), the MLE can be written as follows:

$$\max_{\mu,\sigma_x,\sigma_\epsilon} \sum_{i=1}^{n} ln \left(\int_{-\infty}^{\infty} \frac{1}{\sigma_x \sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma_x^2}} \prod_q f \left[\frac{CE_{q1}}{\sigma_\epsilon}, \frac{CE_{q2}}{\sigma_\epsilon} \right] \mathrm{d}x \right) \tag{2.16}$$

$$f(\frac{CE_{q1}}{\sigma_\epsilon}, \frac{CE_{q2}}{\sigma_\epsilon}) = \frac{1}{1 + e^{\frac{CE_{q2} - CE_{q1}}{\sigma_\epsilon}}} \tag{2.17}$$

Results by frame and snapshot can be seen in Figure 2.13. Note that all available data are used for each snapshot and frame, meaning the first 5 snapshots are again better powered than the latter 5. Results are broadly consistent with our non-MLE analyses above.

First focusing attention on the leftmost pane: there is limited convergence in mean risk aversion across the frames. Untutored preferences for Pairwise Choices Between Reduced Simple Lotteries start with relatively high mean risk aversion, declining over the course of the survey. Untutored preferences for both Pairwise Choices Between Complete Strategies and Pairwise Choices Between Compound Lotteries start with relatively low mean risk aversion, increasing over the course of the survey. Focusing next on the middle pane: Pairwise Choices Between Reduced Simple Lotteries are associated with overall lower variance in risk aversion compared to the other two pairwise frames. Across all frames, this variance slightly increases over the course of the survey. Finally, focusing on the rightmost pane, all pairwise frames see a significant overall decrease in the error response variance, except between waves, when it jumps up slightly before again declining. Pairwise Choices Between Reduced Simple Lotteries are, overall, associated with lower error, as might be expected, given these questions are already reduced (and hence simpler).

2.8 Conclusions and Plans for Further Research

People generally revise toward consistency during our procedure, although few subjects are completely consistent across all 30 potential inconsistencies. There is substantial heterogeneity across frames in both initial consistencies and propensity to update. Initially, subjects are particularly apt toward consistency in Nodewise Action Choice Frames (as opposed to Pairwise Strategy Choice Frames). They are particularly unlikely initially to endorse Reduction of Compound Lotteries, although this axiom also sees the most resolu-

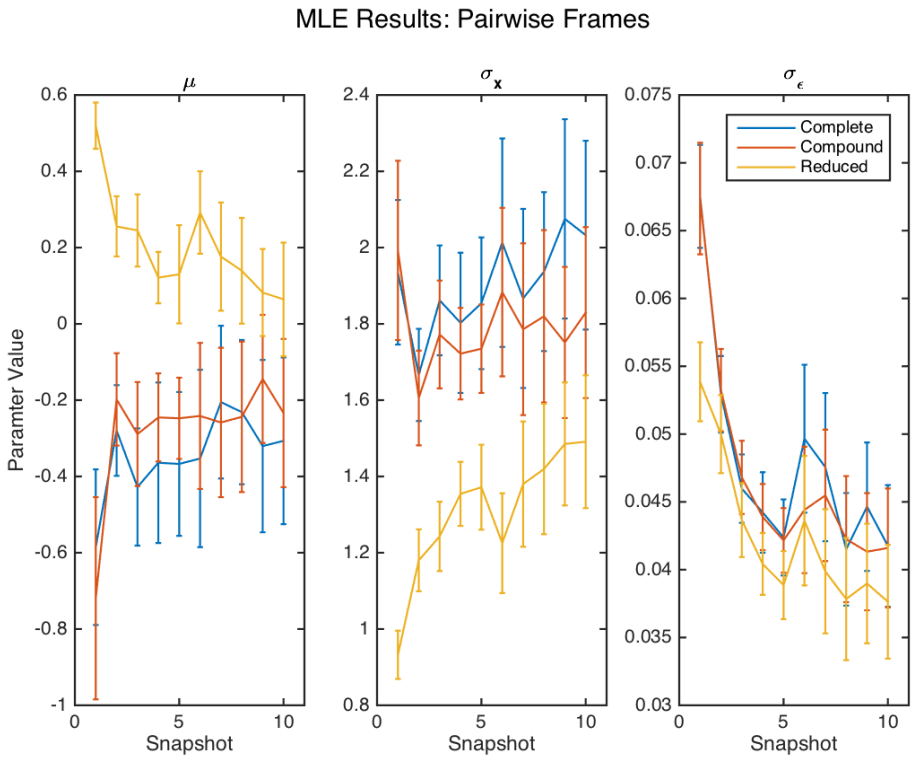

Figure 2.13: MLE results for Pairwise Strategy Choices Frames, snapshots 1-10. All available data are used for each snapshot and frame. Error bars denote standard errors. "Complete" refers to Pairwise Choices Between Complete Strategies; "Compound" refers to Pairwise Choices Between Compound Lotteries; and "Reduced" refers to Pairwise Choices Between Reduced Simple Lotteries.

tion toward endorsement after sufficient reasoning. Moreover, when brought back into the laboratory for a second wave 2-4 weeks later, we find evidence of persistence: people are far more consistent than they were relative to their wave 1 untutored preferences, but still not quite as consistent as they were relative to their wave 1 reasoned preferences. While subjects also tend to reduce their intransitivities among Pairwise Strategy Choice Frames, there is less evidence of persistence between waves.

We are able to conclude much more from the MLE of Pairwise Strategy Choice Frames than of Nodewise Action Choice Frames. This is because in the Nodewise Frames, we only ask participants their favorite choice (and sometimes their second-favorite choice in the Complete Contingent Action Plan), and hence our MLE analyses are statistically underpowered; these frames are mostly indistinguishable according to our MLE results. The Pairwise Frames are far better statistically powered because we ask subjects to choose between each potential pair of options among all 5 investment plans (meaning we can always ascertain subjects' second, third, and fourth favorite choices). Subjects tend to be more risk averse in their responses for Pairwise Choices Between Reduced Simple Lotteries as compared to the other Pairwise Frames. That being said, estimates of mean risk aversion mildly converge among all Pairwise Frames. As expected, decision errors are more prevalent in the non-reduced Pairwise Frames, and generally decline (but more so for the non-reduced frames).

Overall, we have demonstrated a method that makes progress in reducing the range of uncertainty surrounding risk aversion. However, the method is still imperfect. More data collection and further tweaks to our procedure are necessary for us to state what "the" level of risk aversion across frames is definitively (and provide default asset allocation recommendations). We have several ideas moving forward. First, our procedure focused on being as light-handed as possible in terms of encouraging resolution; we gave subjects ample opportunity to do what they wanted, and our results strongly reject experimenter demand as a driving force. However, heavy-handed promotion could help close the gap between measures of risk aversion derived from different frames (e.g. by changing the wording of the inconsistency checks to more strongly encourage resolution). Perhaps this extra push could facilitate even more updating and hence greater convergence in measures of risk aversion across frames. Second, as mentioned above, we would like to elicit second-, third-, and fourth-favorite choices among the Nodewise Action Choice Frames. This will help boost the statistical power of our MLE procedure and provide a more comprehensive picture of subjects' preferences. It would also facilitate greater comparability between Nodewise and Pairwise Frames (the latter already elicits a full set of preferences). Third, it would be interesting to bring subjects back into a lab for third or even fourth waves. We saw

substantial evidence of persistence of first wave reasoned preferences into the second wave; it would be interesting to evaluate how preferences persist over greater periods of time. Moreover, subjects had the greatest amount of consistency, and lowest within-frame error responses, after the second wave. It would be interesting see whether this trend continues into further waves (or to what extent there are declining marginal returns to further rounds of inconsistency and intransitivity checks). Fourth, it would be useful to start collecting data on a more nationally representative sample. External validity is a concern that we can only really address by broadening our subject pool.

Once we are confident we have validated a procedure for accurately assessing reasoned risk preferences (that transcend framing effects), our next step will be to find a short survey that successfully approximates the results of the more thorough procedure. Such a survey could be implemented to identify individual-specific optimal asset allocation (in addition to continuing to inform mean levels of population risk aversion and hence default allocations).

Since the question of determining optimal policy when choices violate normative axioms is central to behavioral welfare economics (Bernheim and Rangel, 2008), we believe that the idea of eliciting reasoned preferences could have quite general implications for economics and policy. We envision that applying the kind of procedure we develop here could help distinguish decision errors from reasoned preferences in a wide range of economic settings aside from investment decision making and for other kinds of preferences besides risk preferences (e.g. time preferences, elasticity of intertemporal substitution).

2.9 References and Works Consulted

Arrow, Kenneth J. (1983). "Behavior under uncertainty and its implications for policy." In Stigum, B.P., and F. Wenstøp (eds.), Foundations of Utility and Risk Theory with Applications. The Netherlands, Reidel.

Barsky, Robert B., F. Thomas Juster, Miles S. Kimball, and Matthew D. Shapiro (1997). "Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study." Quarterly Journal of Economics, 112(2), 537-579.

Benartzi, Shlomo, and Richard H. Thaler (1999). "Risk aversion or myopia? Choices in repeated gambles and retirement investments." Management Science, 45(3), 364-381.

Benjamin, Daniel J., Sebastian A. Brown, and Jesse M. Shapiro (2006). "Who is 'behavioral'? Cognitive ability and anomalous preferences." Harvard University mimeo, May.

Bernheim, B. Douglas, and Antonio Rangel. Beyond revealed preference: choice theoretic foundations for behavioral welfare economics. No. w13737. National Bureau of Economic Research, 2008.

Berndt, Ernst R., et al. "Estimation and inference in nonlinear structural models." Annals of Economic and Social Measurement, Volume 3, number 4. NBER, 1974. 653-665.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian (2008a). "The importance of default options for retirement saving outcomes: Evidence from the United States." In Kay, Stephen J., and Tapen Sinha (eds.), Lessons from Pension Reform in the Americas.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian (2008b). "How are preferences revealed?" Journal of Public Economics, 92, 1787-1794.

Beshears, John, James J. Choi, David I. Laibson, and Brigitte C. Madrian (2009). "Can psychological aggregation manipulations affect risk-taking? Evidence from a framed field experiment." Harvard University mimeo, February.

Burks, Stephen, Jeffrey Carpenter, Lorenz Goette, and Aldo Rustichini (2008). "Cognitive skills explain economic preferences, strategic behavior and job attachment." IZA Discussion Paper No. 3609, July.

Cacioppo, John T., and Richard E. Petty. "The need for cognition: Relationship to attitudinal processes." Social perception in clinical and counseling psychology 2 (1984): 113-140.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde (2008). "Are risk aversion and impatience related to cognitive ability?" IZA Discussion Paper No. 2735, April.

Druckman, James N. "Evaluating framing effects." Journal of Economic Psychology 22.1 (2001): 91-101.

Dynan, Karen E. (1993). "How Prudent Are Consumers?" Journal of Political Economy, 101(6), 1104-1113.

Fagerlin, Angela, Brian J. Zikmund-Fisher, Peter A. Ubel, Aleksandra Jankovic, Holly A. Derry, and Dylan M. Smith (2007). "Measuring numeracy without a math test: Development of the Subjective Numeracy Scale." Medical Decision Making, 27, 672-680.

Frederick, Shane. "Cognitive reflection and decision making." Journal of Economic perspectives (2005): 25-42.

Fuchs, Victor R. (1982). "Time preference and health: An exploratory study." In Victor R. Fuchs (ed.), Economic aspects of health, chapter 3, pp. 93-120. The University of Chicago Press, Chicago.

Fudenberg, Drew, and David Levine (2006). "A Dual Self Model of Impulse Control," American Economic Review, 96, 1449-1476.

Gosling, Samuel D., Peter J. Rentfrow, and William B. Swann. "A very brief measure of the Big-Five personality domains." Journal of Research in personality 37.6 (2003): 504-528.

Gruber, Jonathan (2006), "A Tax-Based Estimate of the Elasticity of Intertemporal Substitution," NBER Working Paper No. 11945.

Gul, Faruk, and Wolfgang Pesendorfer (2005). "The Revealed Preference Theory of Changing Tastes," Review of Economic Studies, 72, 429-448.

Haliassos, Michalis and Carol Bertaut (1995). "Why Do So Few Hold Stocks?" The Economic Journal. 195, 1110-29.

Hall, Robert (1988). "Intertemporal Substitution in Consumption," Journal of Political Economy, 96(2), 339-357.

Huck, Steffan, and Wieland Muller (2008). "Allais for all: Revisiting the paradox." University College London mimeo, November 3.

Kahneman, Daniel, and Amos Tversky (1979). "Prospect theory: An analysis of decision under risk." Econometrica, 47, 263-291.

Kimball, Miles S., Claudia R. Sahm, and Matthew D. Shapiro. "Imputing risk tolerance from survey responses." Journal of the American statistical Association 103.483 (2008): 1028-1038.

Koszegi, Botond, and Matthew Rabin (2006). "A model of reference-dependent preferences." Quarterly Journal of Economics, 121(4), 1133-1166.

Laibson, David I. (1997). "Golden eggs and hyperbolic discounting," Quarterly Journal of Economics, 112(2), 443-477.

Loewenstein, George, and Nachum Sicherman (1991). "Do workers prefer increasing wage profiles?" Journal of Labor Economics, 9(1), 67-84.

Loewenstein, George, and Drazen Prelec (1992). "Anomalies in intertemporal choice: Evidence and an interpretation," Quarterly Journal of Economics, 107(2), 573-597.

Loewenstein, George, and Ted O'Donoghue (2007). "The Heat of the Moment: Modeling Interactions Between Affect and Deliberation," Cornell University mimeo, June.

McArdle, John, Willard Rodgers, and Robert Willis. Cognition and Aging in the USA (CogUSA) 2007-2009. ICPSR36053-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-04-16.

McNeil, Barbara J., Stephen G. Pauker, and Amos Tversky. "On the framing of medical decisions." (1988).

Merton, Robert C. (1969). "Lifetime portfolio selection under uncertainty: the continuous time case." Review of Economics and Statistics, 51, 247-257.

Mulligan, Casey B. (2002). "Capital, Interest, and Aggregate Intertemporal Substitution," NBER Working Paper No. 9373.

Poterba, James M., Joshua Rauh, Steven F. Venti, and David A. Wise (2005). "Utility Evaluation of Risk in Retirement Saving Accounts," in Wise, David A. (ed.), Analyses in the Economics of Aging. Chicago: University of Chicago Press.

Rabin, Matthew (2000). "Risk aversion and expected-utility theory: A calibration theorem." Econometrica, 68(5),1281-92.

Raiffa, Howard (1968). Decision Analysis. Reading, MA: Addison-Wesley.

Strotz, Robert (1955). "Myopia and inconsistency in dynamic utility maximization," Review of Economic Studies, 23(3), 165-180.

Thaler, Richard H. (1981), "Some empirical evidence on dynamic inconsistency," Economic Letters, 8, 201-207.

Thaler, Richard H., and Hersh M. Shefrin (1981). "An economic theory of self-control," Journal of Political Economy, 89(2), 392-410.

TIAA-CREF website as of August 4, 2008.

Tversky, Amos, and Daniel Kahneman (1981). "The framing of decisions and the psychology of choice." Science, 211, 453-58.

Wiseman, David B., and Irwin P. Levin (1996). "Comparing risky decision making under conditions of real and hypothetical consequences." Organizational Behavior and Human Decision Processes, 66(3), 241-250.

2.10 Appendix: Pilot Study

We conducted a pilot experiment in Fall 2008 that partly motivates the design of the main experiment in this paper. Our pilot builds on Benartzi and Thaler's (1999) finding that individuals choose to invest 41% of their retirement portfolio in stocks when shown annual rates of return on stocks and bonds, but they invest 82% in stocks when shown 30-year rates of return (but see Beshears et al., 2009, for counterevidence). These two presentations should not differentially affect behavior because they provide essentially the same information. Our pilot study sought to assess which (if either) choice more accurately reflects individuals' risk preferences. We surveyed 54 adult subjects (mean age: 39.5 years) in an outdoor pedestrian mall. While Benartzi and Thaler showed half their subjects the annual returns and half the long-term returns, we showed all subjects both presentations. As

in Benartzi and Thaler's research, we found that subjects put a significantly higher fraction in stocks in the long-term presentation (63% v. 41%, $p < 0.0001$). Our key innovation was a follow-up question: we explained that the two presentations represented the same rates of return, and we asked subjects what they would prefer, now that they knew the two presentations were just two ways of framing the same information. Now subjects invested 55% in stocks, which is significantly larger than the 41% in the one-year framing ($p < 0.05$) and not-quite-significantly smaller than the 63% in the long-term framing ($p = 0.13$).

2.11 Appendix: Flow of Inconsistency Checks

To be more concrete, let us outline the exact flow of questions during a representative inconsistency check (with screenshots from an actual example from the survey).

First, a subject is asked: "In one question you chose X over Y but in another question you chose Y over X. Do you think the two situations are different enough that it makes sense to have different choices, or should they be the same?" The screen also displays the filled out choices. They may answer one of the following: (1) "It makes sense to have the same choices in both questions", or (2) "It makes sense to have different choices." See Figure 2.14 for a screenshot.

Let us assume they answered: "It makes sense to have different choices." They are then asked "Why do you want to make different choices in these two situations?" and given the following options (which were developed after extensive piloting): (1) "The two situations are different enough that I want different choices", (2) "Some of the options are equally good to me, so it doesn't matter which one I choose", (3) "I chose how I thought the experimenters wanted me to chose", (4) "I don't know which options I prefer", (5) "I don't know or am confused", or (6) "Other". See Figure 2.15 for a screenshot. After answering this question, they either move on to the next inconsistency, the next placebo inconsistency, or the next part of the survey. Note that we explicitly address experimenter demand in asking this question; this option is rarely chosen.

Instead, let us assume they answered: "It makes sense to have the same choice in both questions." They are then asked: "Please look at your choices from before. Which better represents what you want to do in both, X or Y?" The following choices are possible: (1) "Choice of X" (with image showing filled out choice), (2) "Choice of Y" (with image showing filled out choice), (3) "I changed my mind: I realized that it does make sense to have different choices in these two situations. I would like to change *both* of my choices", or (4) "I changed my mind: I realized that it does make sense to have different choices in these two situations. I would like to keep my current choices." See Figure 2.16 for a screenshot.

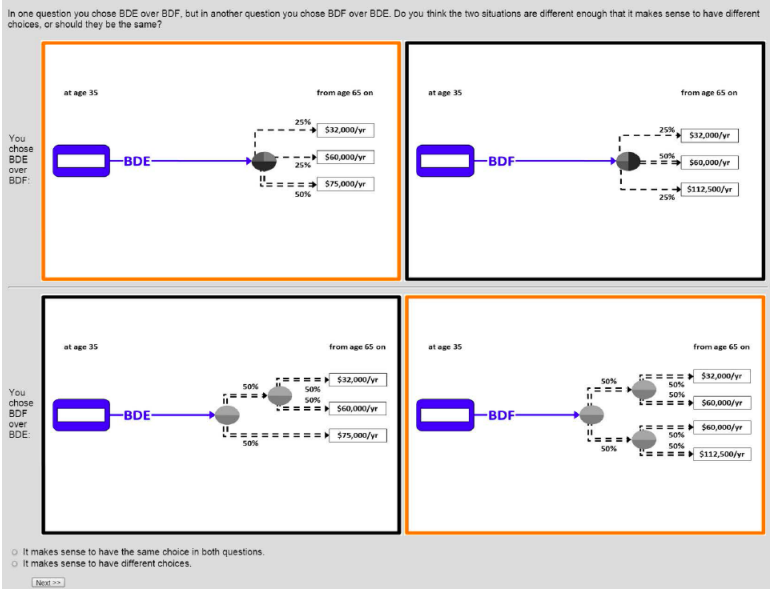

Figure 2.14: Example of an inconsistency check, initial question.

If they answer (4), they are brought to the screen in the prior paragraph. Otherwise, they are shown a new screen with their updated preferences filled out in the actual choice graphics, and told "Is this what you wanted your choices to be changed to? If so, click next. If not, click back and change your choices to what you want." See Figure 2.17 for a screenshot. After a subject has verified her new preferences, she is asked "Why did you want to change your choices as you did?" She is given the following choices (again developed after extensive piloting): (1) "I made a mistake when I first chose", (2) "Answering all of these questions made me change what I want", (3) "Some of the options are equally good to me, so it doesn't matter which one I choose", (4) "I chose how I thought the experimenters wanted me to chose", (5) "I don't know which options I prefer", (6) "I don't know or am confused", or (7) "Other". See Figure 2.18 for a screenshot. Note that we explicitly address experimenter demand in asking this question; this option is rarely chosen.

The flow of a placebo inconsistency is very similar, see Figures 2.19, 2.20 2.21, 2.22, 2.23, and 2.24.

2.12 Appendix: Pre-Test

The Pre-Test consisted of 4 rounds of 3 questions each, yielding 4 distinct measures of risk aversion. The introduction to this section of the survey read:

"We will now ask you to make choices in make-believe situations. These are serious make-believe situations involving long-run investing decisions you could face. Please focus

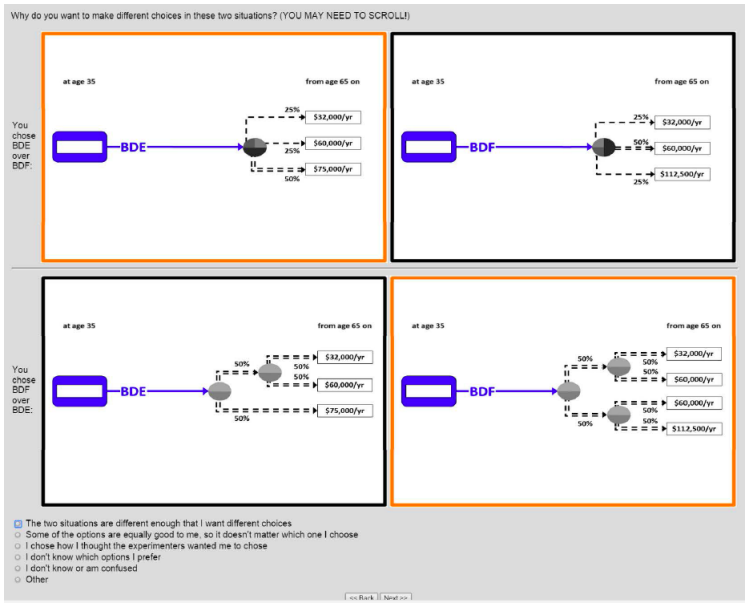

Figure 2.15: Example of an inconsistency check, conditional on answering ”It makes sense to have different choices.”

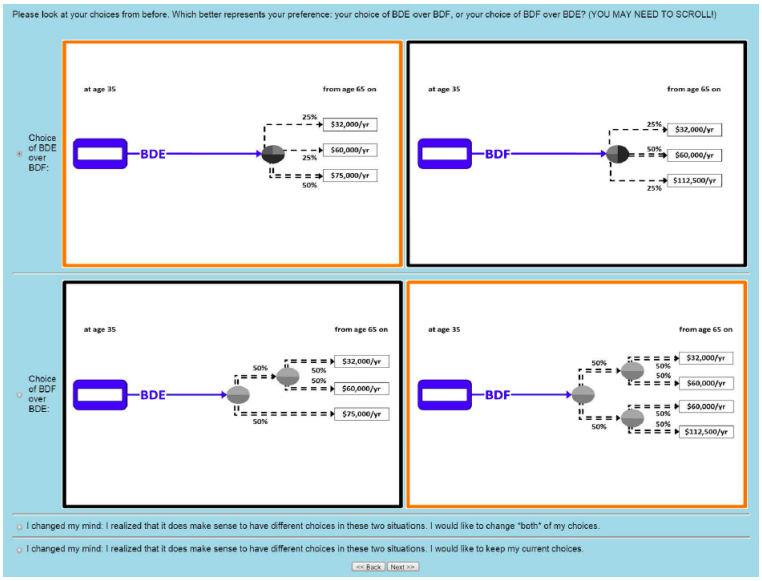

Figure 2.16: Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices.”

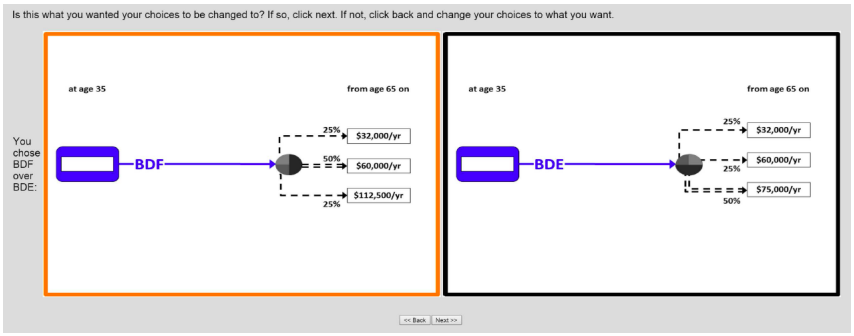

Figure 2.17: Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices” and now verifying updated preferences

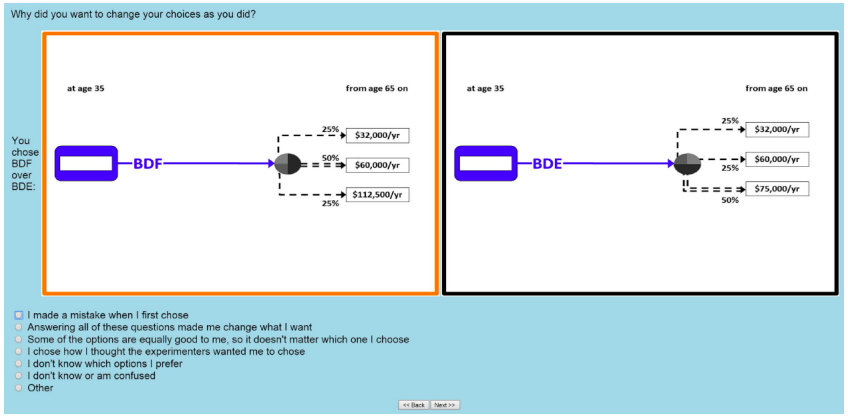

Figure 2.18: Example of an inconsistency check, conditional on answering ”It makes sense to have the same choices,” having verified updated preferences.

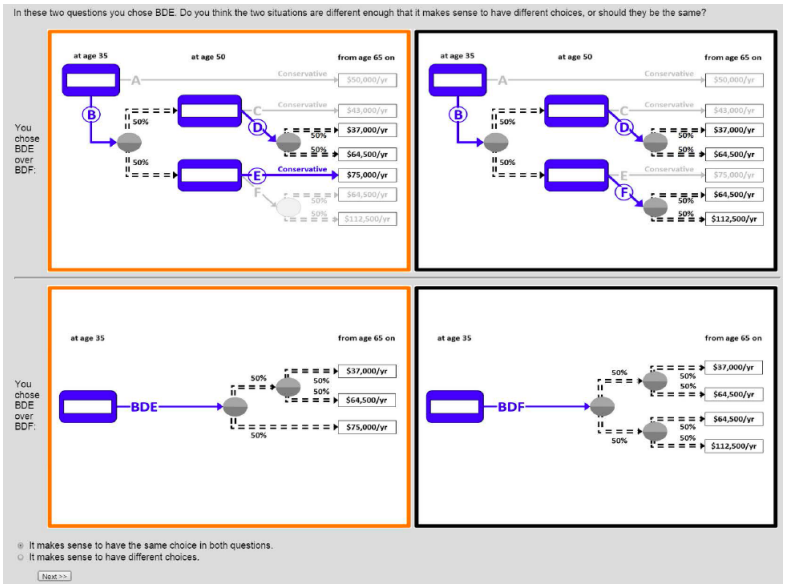

Figure 2.19: Example of an placebo inconsistency check, initial question.

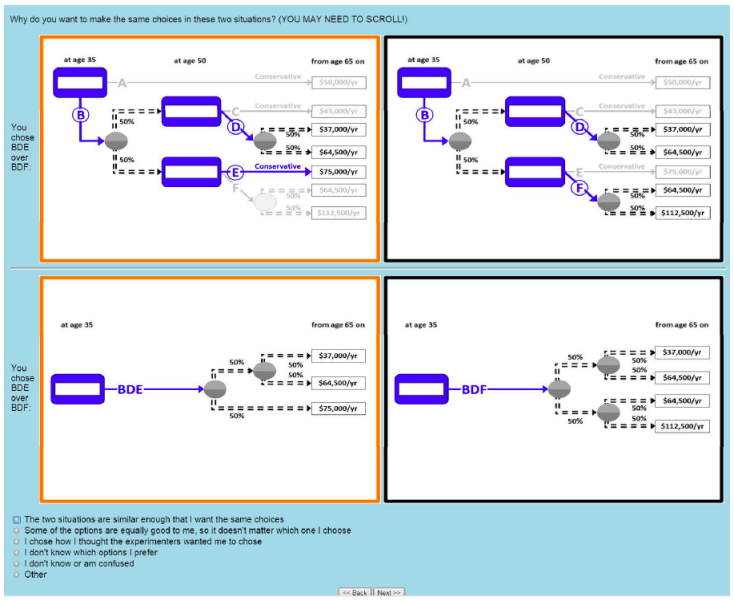

Figure 2.20: Example of an placebo inconsistency check, conditional on answering ”It makes sense to have the same choice...”

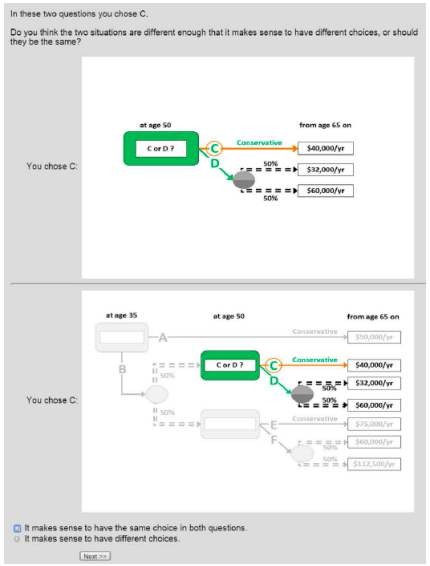

Figure 2.21: Example of an placebo inconsistency check, initial question.

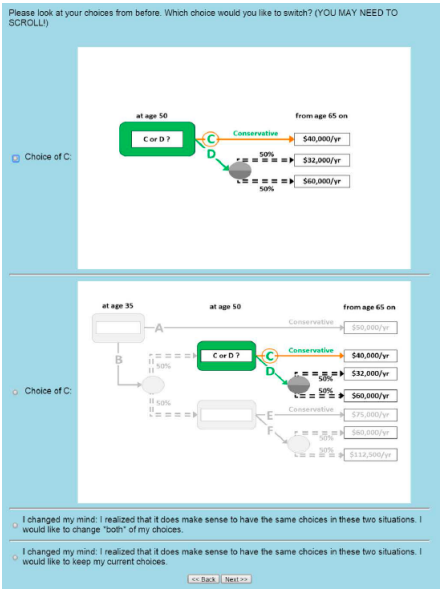

Figure 2.22: Example of an placebo inconsistency check, conditional on answering ”It makes sense to have different choices...”

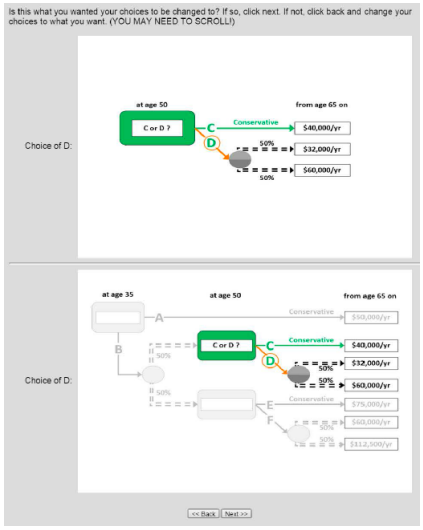

Figure 2.23: Example of an placebo inconsistency check, conditional on answering ”It makes sense to have different choices...” and now verifying updated preferences.

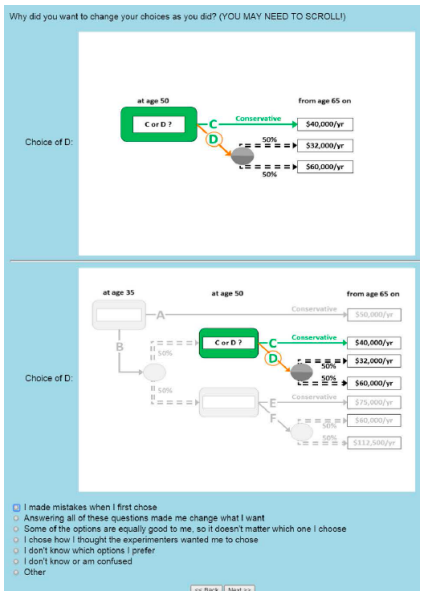

Figure 2.24: Example of an placebo inconsistency check, conditional on answering "It makes sense to have different choices..." and having verified updated preferences.

on the situations we give you, without bringing in any other opportunities you have in the real world. What happens to you depends on both your choices and luck."

The questions themselves were all of the following form: "Imagine that you must choose between two retirement plans. One guarantees you will receive X each year during your retirement, from age 65 on. The other is uncertain, and gives you a 50-50 chance of receiving either a higher amount every year during retirement, or a lower amount every year during retirement. Pretend that this is the only money you will be able to spend each year during retirement. Medical care and taxes have already been taken care of, but this money is all you have for rent, food, clothing, entertainment, etc. Which would you choose?"

Subjects could then either choose (1) X for sure to spend each year during retirement, or (2) 50-50 chance of Y or Z to spend each year during retirement.

X, Y, and Z varied from question-to-question. The 4 rounds were identical other than varying the upside of the risky choice. After answering these questions, we applied the procedure developed in Kimball et al. 2008 to impute a cardinal measure of risk aversion from subjects' categorical responses.

Subjects during wave 1 averaged an imputed log risk aversion of 1.99 with a standard deviation of 0.69 (and a median of 1.77). During wave 2, subjects averaged an imputed log risk aversion of 1.96 with a standard deviation of 0.75 (and a median of 1.62). The within-person correlation between these measures was 0.71. Among subjects who were surveyed in both waves, there is no statistically significant difference in these measures

between waves (p = 0.5350 of the hypothesis that the mean difference being not equal to 0). Of course, we do not necessarily need persistence across waves for our procedure to elicit a "purer" measure of risk aversion.

Using results from our demographic and psychological batteries, we could verify some common conceptions about risk aversion. First, higher cognitive function was associated with lower risk aversion in both waves, but more so in the first wave than in the second wave. Cognition was measured as the first principal component derived from a combination of several batteries: probabilistic sophistication battery, number series battery (CogUSA), number of statistics and economics classes taken, SAT math score, and a cognitive reflection task (Frederick, 2005). Note that all factor loadings for the cognition measure were in the intuitive direction. Among all subjects, regressing wave 1's imputed log risk aversion measure on the first principal component of cognition and gender yields a highly significant negative coefficient estimate for cognition ($p < 0.001$) with an R^2 of 0.1249. Restricting the sample to only those subjects who participated in 2 waves, regression wave 1's imputed log risk aversion measure on the first principal component of cognition and gender also yields a highly significant negative coefficient estimate for cognition ($p < 0.001$) with an R^2 of 0.1020. Doing the same for wave 2 also yields a significant negative coefficient for cognition (p = 0.007) with a much lower R^2 of 0.0602. This is consistent with our procedure lessening the impact of cognition of measured risk aversion. Given that risk aversion is associated with lower cognitive functioning, our results are therefore consistent with our procedure partially correcting for the biases introduced by low cognition on risk aversion.

Moreover, throughout these analyses, the coefficient on an indicator for male gender was always highly significant, and indicated (as expected) that male subjects are generally less risk averse.

Recall also there was a randomization involving the Pre-Test. Among participants brought in for both waves, approximately half of participants were given the Pre-Test at the end of the survey, rather than at the beginning; the hope was to observe how our experimental procedure impacted the very simple risk aversion questions asked in the Pre-Test. In both waves 1 and 2 among subjects who were surveyed in both waves, being given the Pre-Test at the end of the survey was associated with lower log risk aversion. In wave 1, regressing the measure of imputed log risk aversion on an indicator for having given the Pre-Test at the end of the survey (rather than at the beginning) yielded a coefficient estimate of -0.166 (p = 0.046). In wave 2, a similar analysis yielded a coefficient estimate of -0.202 (p = 0.030). Again, if we think of risk aversion as being associated with cognitive bias, our procedure (at least temporarily) alleviates said bias.

2.13 Appendix: Updating Conditional on Riskiness of Choices

Table 2.14 shows, for each axiom and inconsistency, the percentage of time each frame is associated with the riskier choice. Again, for each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2.

Axiom	Frame 1 Riskier	Frame 2 Riskier	Total
Irrelevance of Background Counterfactuals	.4	.59	197
Simple Actions = State-Contingent Actions	.52	.47	196
Irrelevance of Counterfactual Choices	.50	.50	246
Fusion + Shift from Nodewise to Pairwise	.38	.61	449
Complete Strategies = Implied Lotteries	.47	.52	1504
Reduction of Compound Lotteries	.51	.48	2005

Table 2.14: For each axiom and inconsistency, the percentage of time each frame is associated with the riskier choice. For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2.

Tables 2.15 and 2.16 show, conditional on either Frame 1 or Frame 2 being the riskier choice in an inconsistency, toward which frames people tend to update (or whether they choose not to update or to swap their choices). When Frame 1 is the riskier choice in an inconsistency, subjects tend to update toward that choice. However, when Frame 2 is the riskier choice, results are more mixed.

2.14 Appendix: MLE Diagnostic Tests

We conducted diagnostic tests of our MLE procedure to get a sense of the behavior of the likelihood function. We did this by manually varying one of the three parameters while fully optimizing over the remaining two. We next plotted the average log likelihood (across subjects) as a function of the manually varied parameter, noting its behavior. We repeated this separately for each of the three parameters, for each frame and snapshot. Moreover, to get a sense of precision using a measure that can be visualized in the same space as the average log likelihoods, we next drew horizontal lines representing confidence intervals. These horizontal lines were drawn at vertical distances from the global optimum equal to

Axiom	Toward Frame 1	Toward Frame 2	No Update	Swap Choices	Total
Irrelevance of Background Counterfactuals	.33	.22	.41	.02	80
Simple Actions = State-Contingent Actions	.22	.07	.66	.02	103
Irrelevance of Counterfactual Choices	.05	.01	.92	0	120
Fusion + Shift from Nodewise to Pairwise	.21	.17	.55	.04	174
Complete Strategies = Implied Lotteries	.23	.22	.48	.05	715
Reduction of Compound Lotteries	.31	.17	.45	.05	1023

Table 2.15: Conditional on Frame 1 being the riskier choice in an inconsistency, percentage of the time subjects' update toward each frame (or whether they choose not to update or to swap their choices). For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2.

Axiom	Toward Frame 1	Toward Frame 2	No Update	Swap Choices	Total
Irrelevance of Background Counterfactuals	.16	.25	.54	.03	117
Simple Actions = State-Contingent Actions	.18	.17	.56	.07	93
Irrelevance of Counterfactual Choices	.09	.05	.84	0	126
Fusion + Shift from Nodewise to Pairwise	.2	.17	.54	.06	275
Complete Strategies = Implied Lotteries	.19	.28	.47	.04	789
Reduction of Compound Lotteries	.23	.24	.46	.05	982

Table 2.16: Conditional on Frame 2 being the riskier choice in an inconsistency, percentage of the time subjects' update toward each frame (or whether they choose not to update or to swap their choices). For each axiom "Frame 1" and "Frame 2" refer to the frames as labeled in Table 2.2.

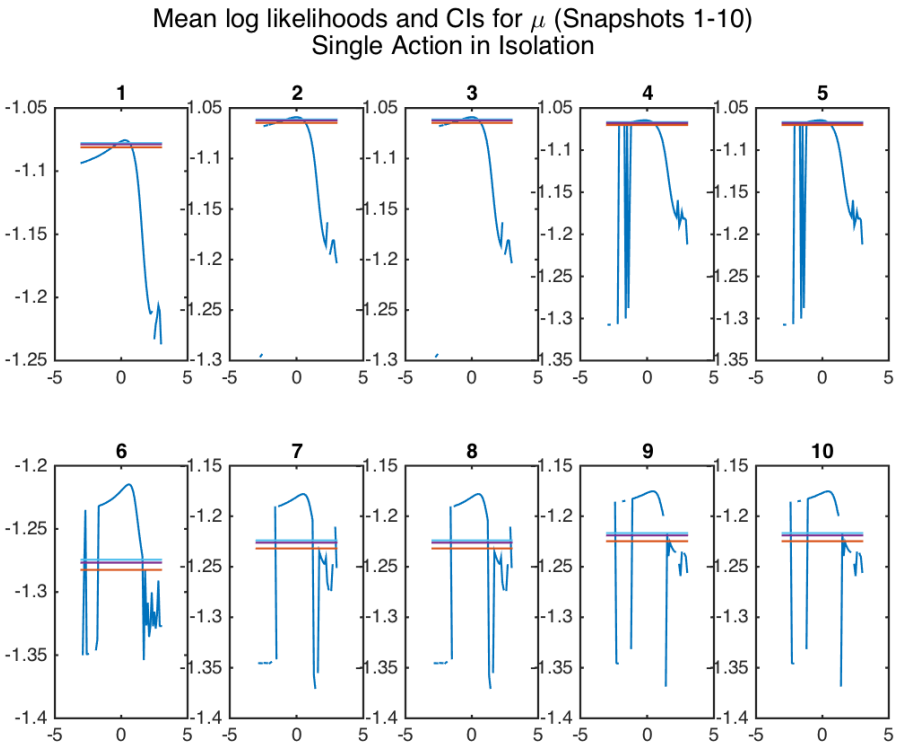

Figure 2.25: Diagnostic tests for μ, Single Action in Isolation, Snapshots 1-10.

the critical values from the likelihood ratio test. We conclude several things from these tests: (1) Pairwise Strategy Choice Frames are statistically well powered across all snapshots, (2) Nodewise Action Choice Frames for snapshots 1-5 are barely well powered, but snapshots 6-10 are certainly underpowered.

We first present results for the Nodewise Action Choice Frames for each parameter and all 10 snapshots (Figures 2.25-2.39). Horizontal lines from top to bottom always represent 90%, 95%, and 99% confidence intervals. As is obvious, the likelihood functions for many parameters (especially μ and often for σ_x) behave erratically and often have extremely wide confidence intervals. We appear to lack statistical power. Next, we show abbreviated results for the Pairwise Strategy Choice Frames, namely only for snapshot 1 and each parameter (Figures 2.40-2.48). The results for the other snapshots are extremely similar. Here, we are statistically well powered.

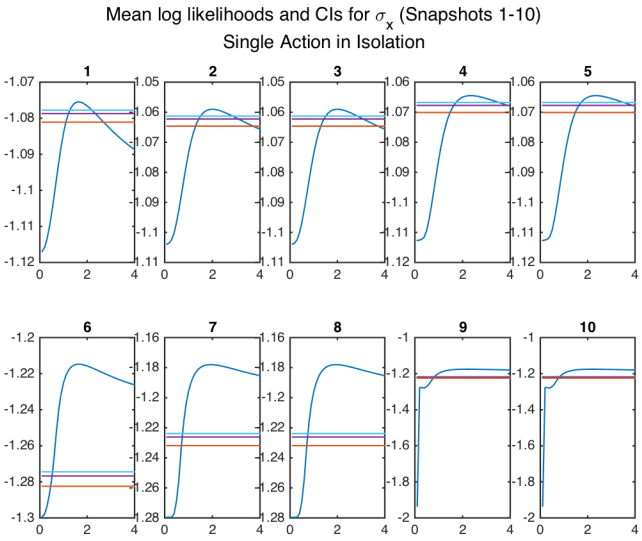

Figure 2.26: Diagnostic tests for σ_x, Single Action in Isolation, Snapshots 1-10.

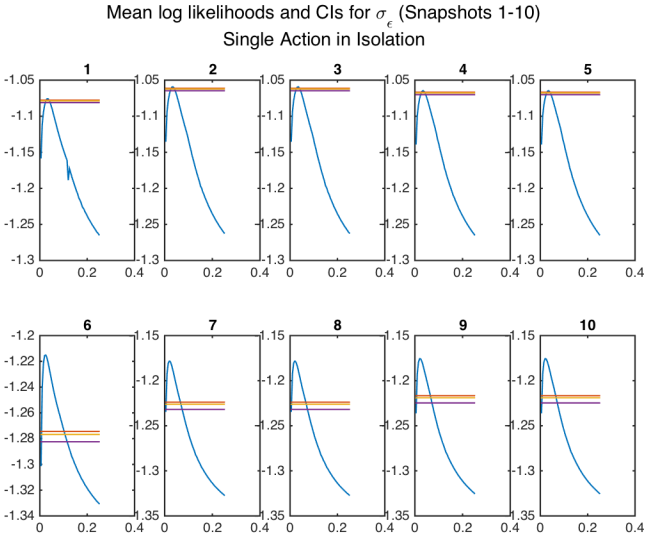

Figure 2.27: Diagnostic tests for σ_ϵ, Single Action in Isolation, Snapshots 1-10.

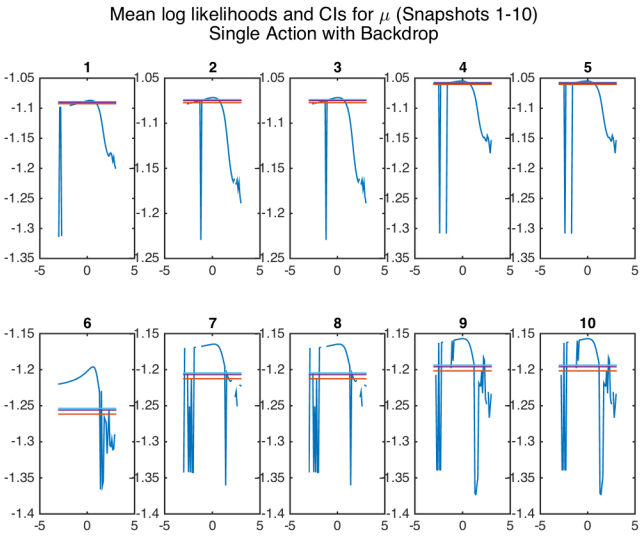

Figure 2.28: Diagnostic tests for μ, Single Action with Backdrop, Snapshots 1-10.

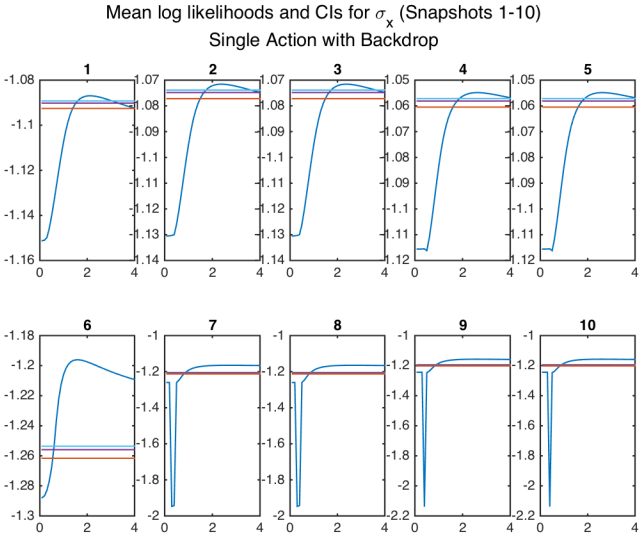

Figure 2.29: Diagnostic tests for σ_x, Single Action with Backdrop, Snapshots 1-10.

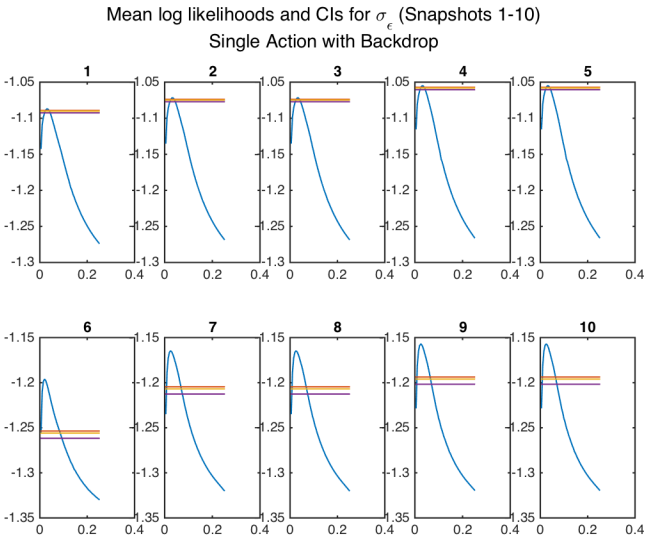

Figure 2.30: Diagnostic tests for σ_ϵ, Single Action with Backdrop, Snapshots 1-10.

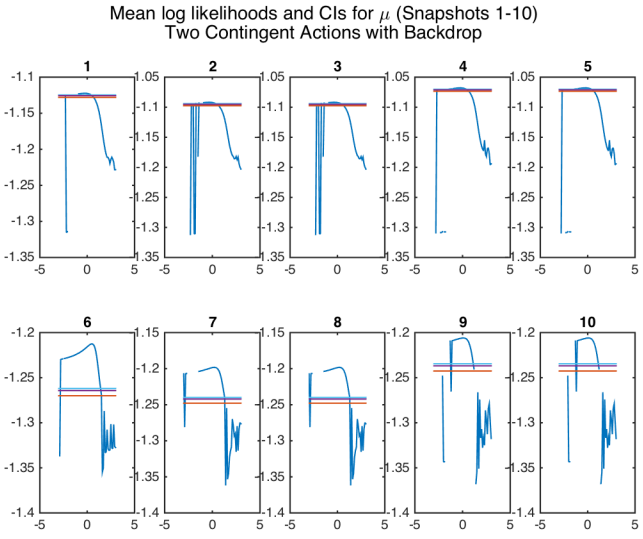

Figure 2.31: Diagnostic tests for μ, Two Contingent Actions with Backdrop, Snapshots 1-10.

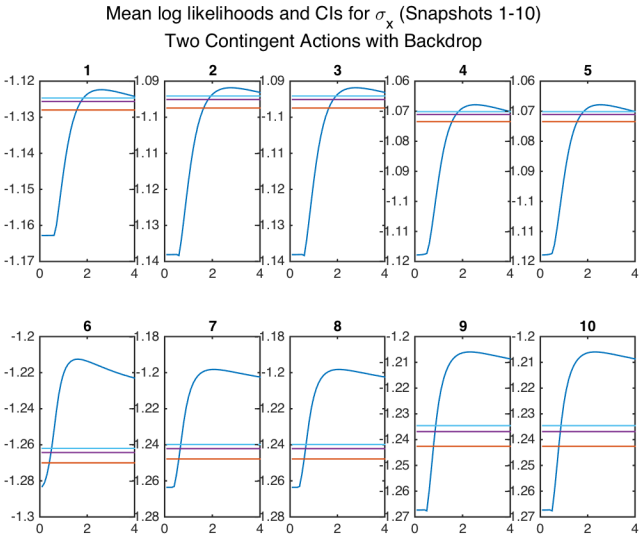

Figure 2.32: Diagnostic tests for σ_x, Two Contingent Actions with Backdrop, Snapshots 1-10.

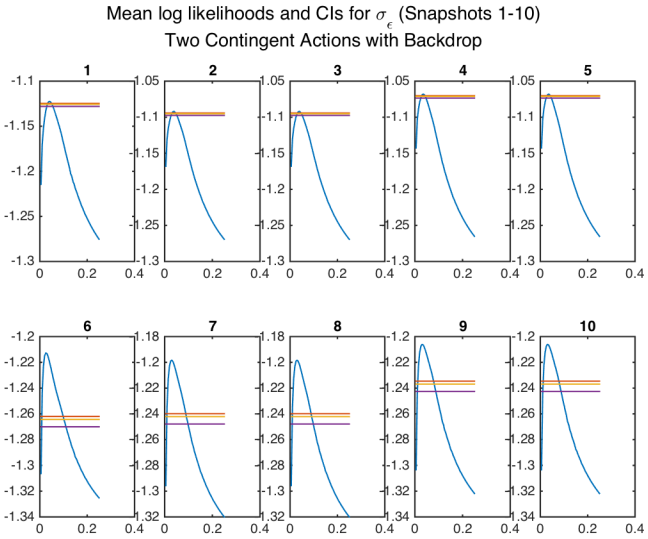

Figure 2.33: Diagnostic tests for σ_ϵ, Two Contingent Actions with Backdrop, Snapshots 1-10.

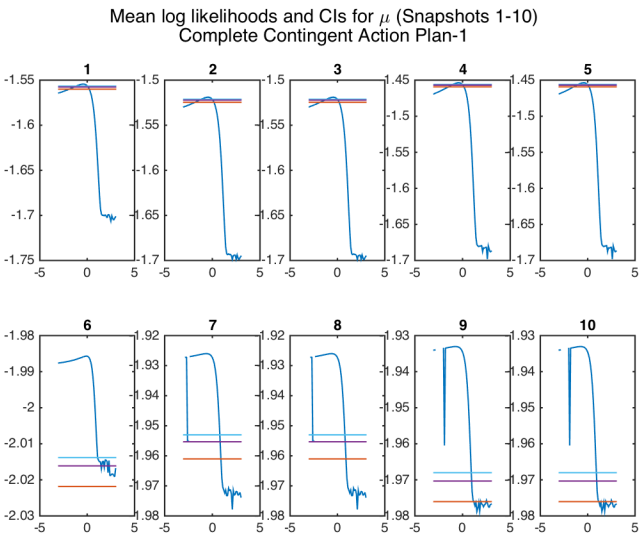

Figure 2.34: Diagnostic tests for μ, Complete Contingent Action Plan-1, Snapshots 1-10.

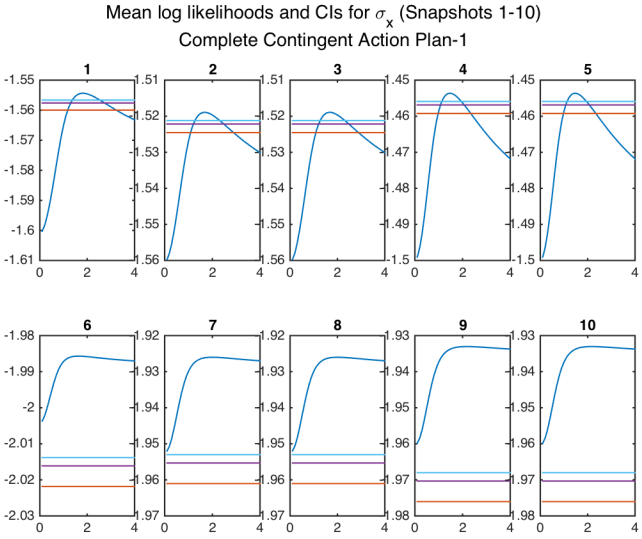

Figure 2.35: Diagnostic tests for σ_x, Complete Contingent Action Plan-1, Snapshots 1-10.

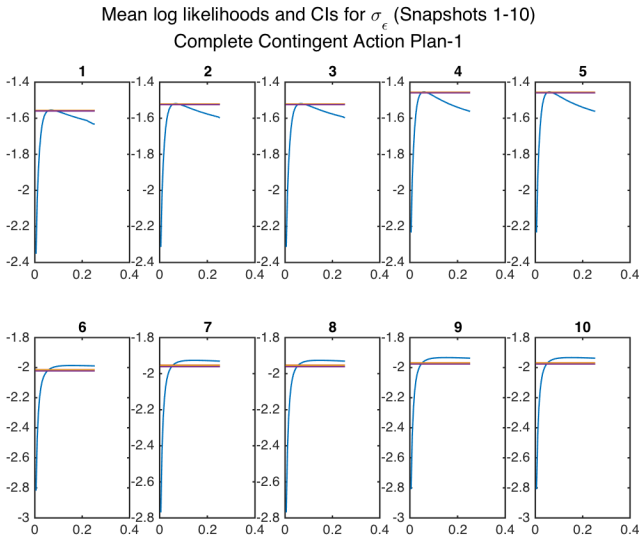

Figure 2.36: Diagnostic tests for σ_ϵ, Complete Contingent Action Plan-1, Snapshots 1-10.

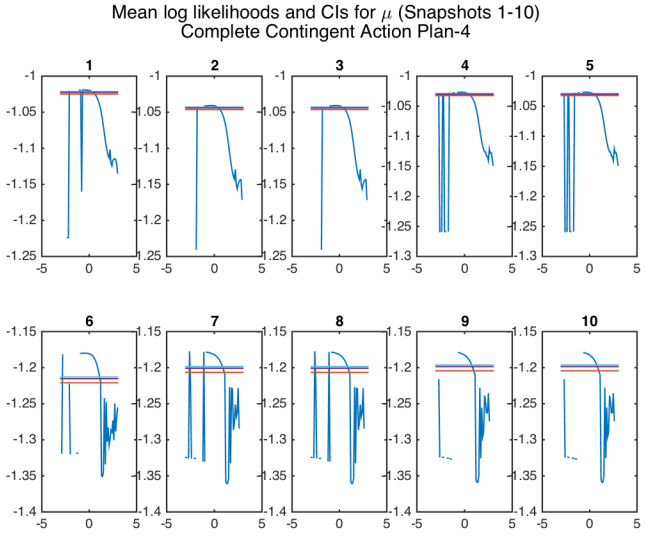

Figure 2.37: Diagnostic tests for μ, Complete Contingent Action Plan-4, Snapshots 1-10.

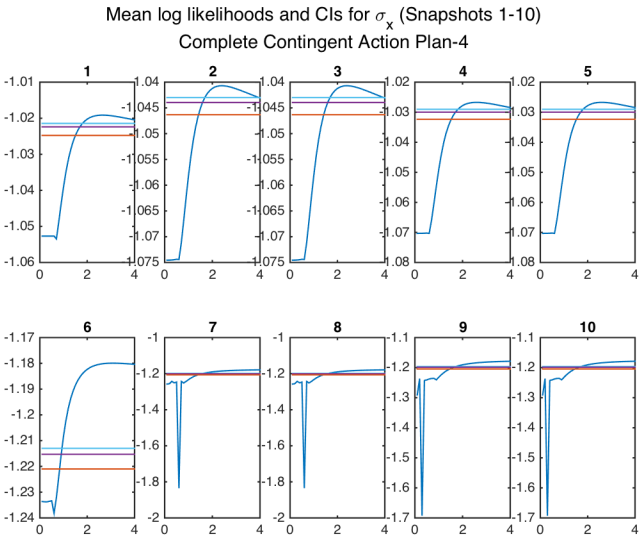

Figure 2.38: Diagnostic tests for σ_x, Complete Contingent Action Plan-4, Snapshots 1-10.

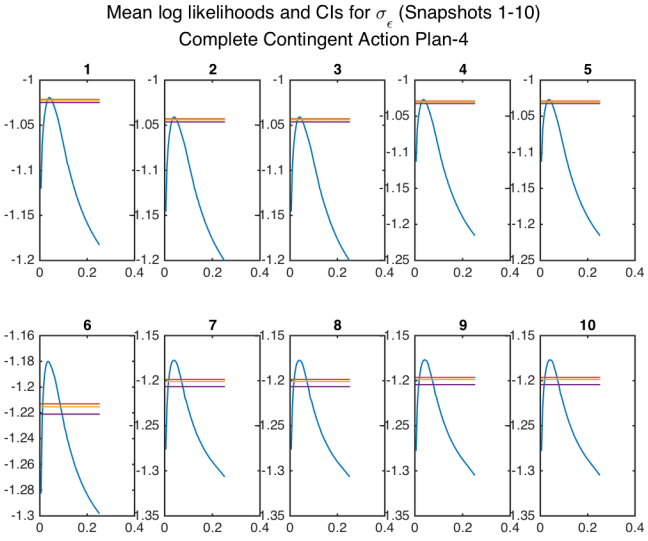

Figure 2.39: Diagnostic tests for σ_ϵ, Complete Contingent Action Plan-4, Snapshots 1-10.

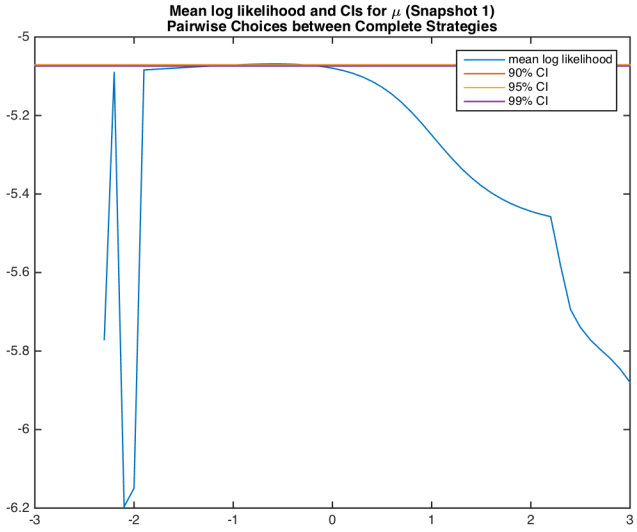

Figure 2.40: Diagnostic tests for μ, Pairwise Choices between Complete Strategies, Snapshot 1.

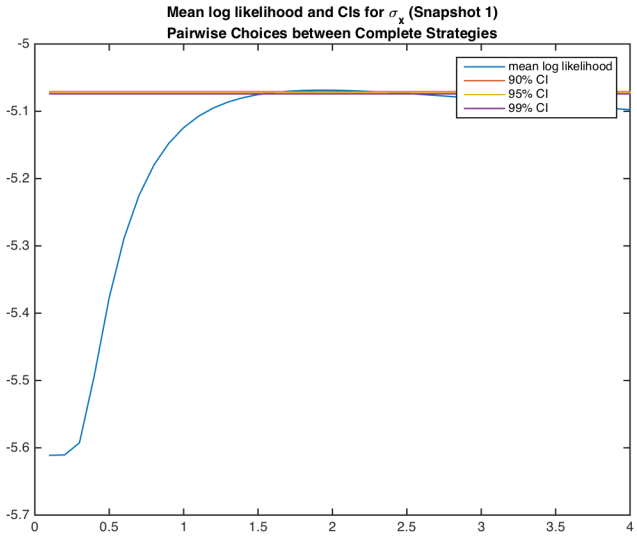

Figure 2.41: Diagnostic tests for σ_x, Pairwise Choices between Complete Strategies, Snapshot 1.

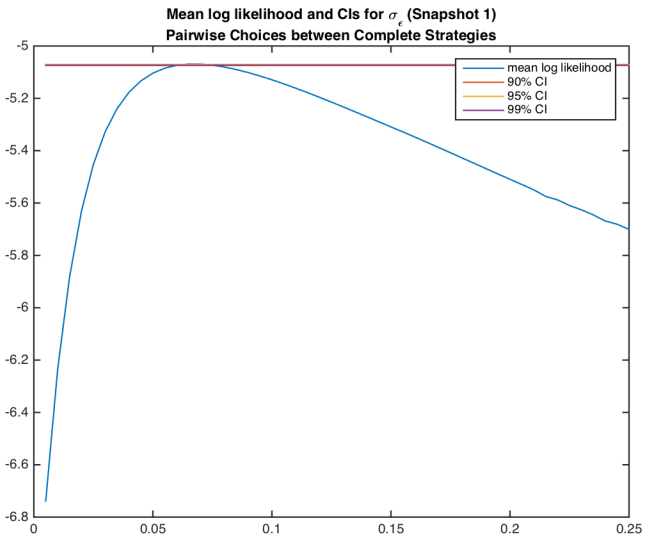

Figure 2.42: Diagnostic tests for σ_ϵ, Pairwise Choices between Complete Strategies, Snapshot 1.

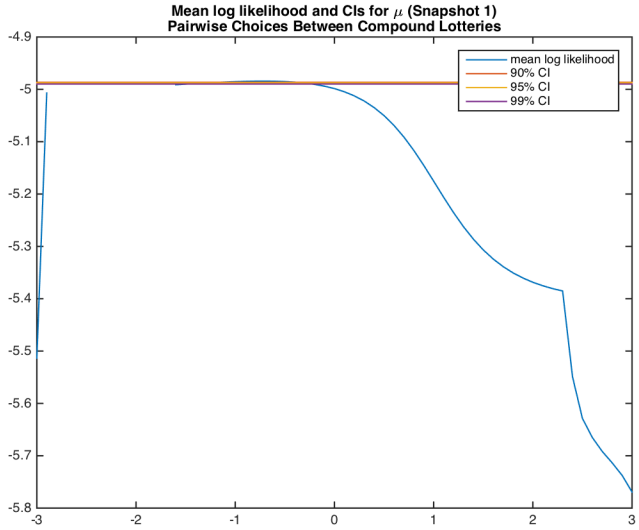

Figure 2.43: Diagnostic tests for μ, Pairwise Choices Between Compound Lotteries, Snapshot 1.

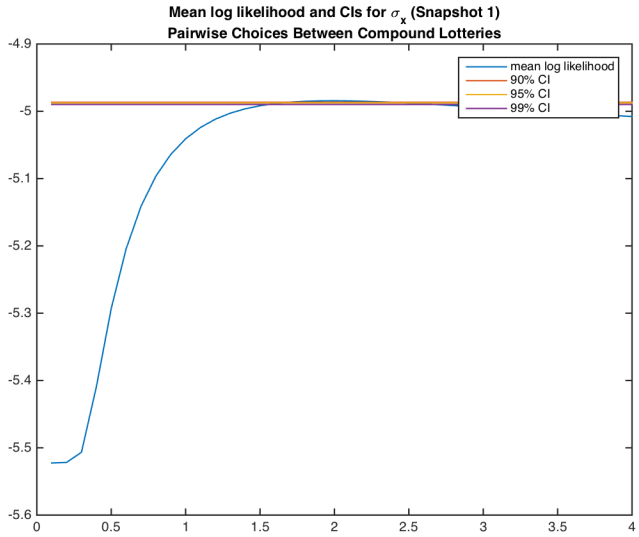

Figure 2.44: Diagnostic tests for σ_x, Pairwise Choices Between Compound Lotteries, Snapshot 1.

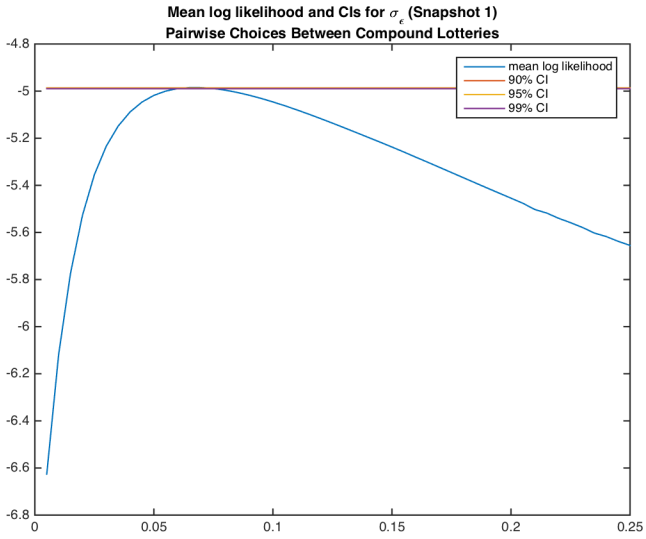

Figure 2.45: Diagnostic tests for σ_ϵ, Pairwise Choices Between Compound Lotteries, Snapshot 1.

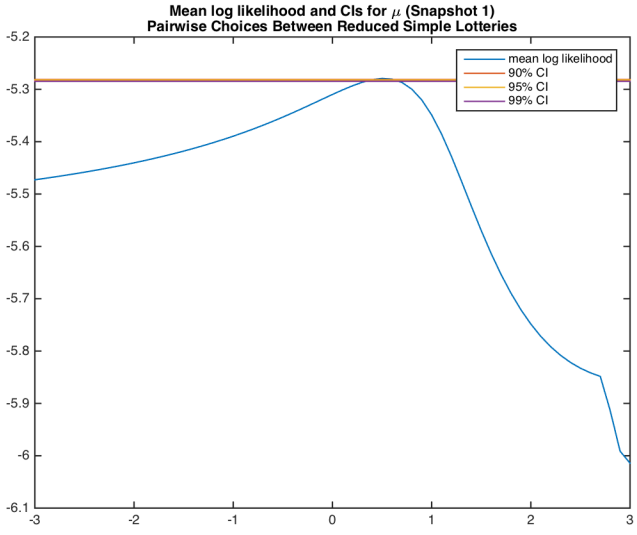

Figure 2.46: Diagnostic tests for μ, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1.

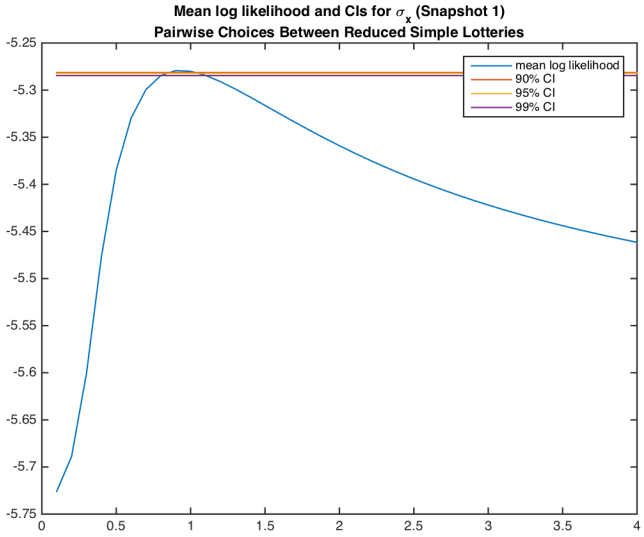

Figure 2.47: Diagnostic tests for σ_x, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1.

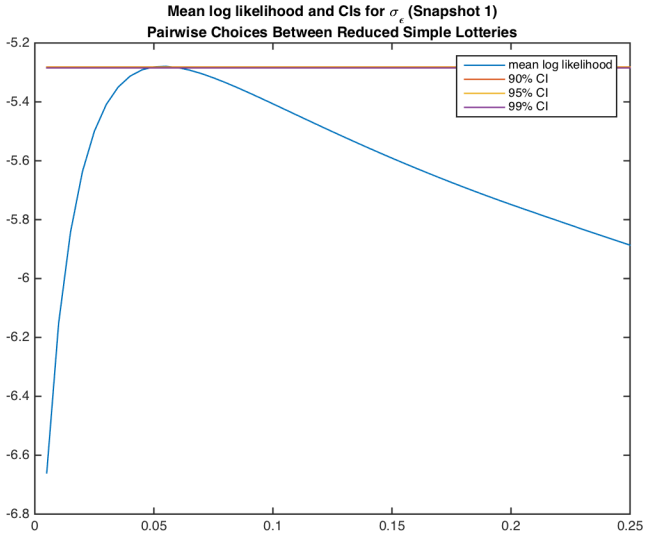

Figure 2.48: Diagnostic tests for σ_ϵ, Pairwise Choices Between Reduced Simple Lotteries, Snapshot 1.

CHAPTER 3

The Genetics of Cigarette Excise Tax Responsiveness

3.1 Abstract

Tobacco is among the leading causes of preventable death worldwide. While public policies and increased awareness of the health consequences of smoking have been effective at reducing its incidence in America, cigarette excise taxes do not appear to deter smoking across the entire distribution of smokers. Fletcher (2012), using data from the National Health and Nutrition Examination Survey (NHANES), looks toward genetics as a source of tax response heterogeneity. He tests for differential responsiveness to cigarette excise taxes based on individuals' genotypes on single nucleotide polymorphism (SNP) rs2304297, finding that only individuals with the GG genotype are responsive along both extensive and intensive margins, and individuals with other genotypes are unresponsive along both margins. However, better-powered genome-wide association (GWA) studies that attempt to uncover genetic variants associated with smoking do not implicate rs2304297 as being related to human smoking behavior (Tobacco and Genetics Consortium (TAG), 2010; Liu et al., 2010; Thorgeirsson et al., 2010). For behavioral phenotypes, incongruence between candidate gene studies and genome-wide association studies is fairly common, given lack of statistical power in the former driven by the fact that genetic variation in behavioral phenotypes is associated with small effect sizes for any particular SNP. Moreover, lack of extensive genetic data in NHANES only allowed Fletcher to control for self-reported ethnicity, not subtler within-population genetic variation (i.e. population stratification). I therefore have three goals. First, I repeat Fletcher's tax response heterogeneity analysis with respect to rs2304297 and state excise taxes, using the Health and Retirement Study (HRS) instead of NHANES. I do not replicate his main finding of genetic tax response heterogeneity along the extensive margin in the HRS. I am able partially to replicate his finding that individuals with the GG genotype on rs2304297 are more responsive to cigarette taxes

along the intensive margin. I find this effect in the HRS sample including subjects of all ethnicities, but not the sample including only subjects of European decent. Second, I show variation in rs2304297 is driven by ethnicity, i.e. population stratification. Third, I use estimates from two large, well-powered GWA studies to perform out-of-sample prediction on the HRS's European-decent subjects. These predictors, or polygenic scores, utilize the full gamut of SNPs available on the HRS and can be interpreted as measures of genetic predisposition. In particular, I construct measures of genetic predisposition to two outcomes plausibly related to smoking behavior: lifetime maximum cigarettes per day (CPD) and educational attainment (EA). Controlling for population stratification, I show these measures robustly predict cigarette consumption along both the extensive and intensive margins in the expected directions. However, when interacted with state tax rates, and additionally controlling for a host of relevant covariates, I find *no* evidence of tax response heterogeneity along either margin. It therefore seems appropriate to turn to other, non-genetic explanations for the fact that cigarette excise taxes do not have an impact across the entire distribution of smokers.

3.2 Motivation

Tobacco is among the leading causes of preventable death worldwide. The World Health Organization (WHO)[1] estimates that about 5.4 million people die every year from tobacco use (or about one person every 6 seconds). Almost 600,000 of these deaths are due to second-hand smoke. It's estimated that 100 million deaths were caused by tobacco in the 20th century. To get a sense of scale, 50 million people died during World War II (see Keegan 1989). In 2014, the U.S. Surgeon General reported that nearly 500,000 adult Americans would die prematurely because of smoking that year (see "The health consequences of smoking–50 years of progress"). The report also highlights that smoking costs America over $289 billion each year in terms of medical expenditures and productivity losses.

Federal and state governments have thus sought to discourage cigarette consumption with cigarette taxes, combined with bans on various forms of advertisement (e.g. to children, on television), restrictions on sales (e.g. to children), and limitations on where smokers may light up (e.g. workplaces, restaurants, on commercial and public transportation). While public policies and increased awareness of the health consequences of smoking have been effective at reducing its incidence in America, cigarette excise taxes do not appear to deter smoking behavior across the entire distribution of smokers (Fletcher, 2012). Us-

[1] Accessed 2/18/2015, see www.who.int/mediacentre/factsheets/fs339/en/ and www.who.int/tobacco/mpower/tobacco$_$facts/en/

ing data from the National Health and Nutrition Examination Survey (NHANES), Fletcher (2012) looks toward genetics as a novel source of tax response heterogeneity. He tests for differential responsiveness to cigarette excise taxes based on individuals' genotype on single nucleotide polymorphism (SNP) rs2304297, finding that only individuals with the GG genotype are responsive along both extensive and intensive margins, i.e. evidence of "policy-by-gene" interaction (a specific example of a "gene-by-environment" interaction). He finds individuals with other genotypes are unresponsive along both margins.

Analyzing variation in this particular SNP as a source of tax response heterogeneity, as opposed to other forms of genetic variation, was admittedly out of necessity. At the time, NHANES only contained information on 8 SNPs, among which rs2304297 was the only one plausibly related to smoking.

Human DNA is billions of units long; SNPs are the tens of millions of markers that commonly differ among us. SNPs are a more fundamental unit of genetic variations than genes, as genes themselves typically contain many SNPs. SNPs typically come one of two alleles, one inherited from each parent. By holding fixed the "reference allele" at a particular SNP to one of the two possible alleles, we can measure the genetic information at that SNP as the number of reference alleles (i.e. 0, 1, or 2).

Rs2304297 is located on chromosome 8, specifically within the CHRNA6 gene, which encodes an alpha subunit of neuronal nicotine acetylcholine receptors related to nicotine use in the brain (Mineur and Picciotto, 2008). A number of "candidate" gene studies have specifically linked CHRNA6 with tobacco use outcomes (Saccone et al., 2007; Zeiger et al., 2008; Greenbaum and Lerer, 2009; Hoft et al., 2009). However, better-powered genome-wide association (GWA) studies do not implicate rs2304297, nor other SNPs in CHRNA6, in human smoking behavior (Tobacco and Genetics Consortium (TAG), 2010; Liu et al., 2010; Thorgeirsson et al., 2010). While it is still likely that SNPs in CHRNA6 have some effect on smoking, the fact that these SNPs are not implicated in these better-powered GWA studies implies that their effect sizes are much smaller than those SNPs that are indeed implicated. These GWA studies implicate three genes for lifetime maximum cigarettes smoked per day: CHRNA3 on chromosome 15, coding for *another* member of the nicotinic acetylcholine receptor family of genes; LOC100188947 on chromosome 10, a non-coding RNA region; and EGLN2 (near CYP2A6) on chromosome 19, coding for a transcriptional complex involved in oxygen homeostasis. For smoking initiation, BDNF on chromosome 11 is implicated, while for smoking cessation a region near DBH on chromosome 9 is implicated. To get a sense of scale, note that in the TAG consortium, the three most significant individual SNPs accounted for 0.5%, 0.03%, and 0.19% of the variation in maximum CPD, initiation, and cessation, respectively.

Candidate gene studies were typically employed before technologies allowing measurement of large numbers of SNPs were readily available. For behavioral phenotypes or complex diseases, incongruence between older candidate gene studies and GWA studies is fairly common, given lack of statistical power in the former driven by the fact that genetic variation in behavioral phenotypes is generally driven by a large number of genetic variants, each with small effect sizes[2] (Benjamin 2012).

Lack of extensive genetic data in NHANES also only enabled control for self-reported ethnicity, not subtler within-population genetic variation (i.e. population stratification). Different ancestral groups even within a broad ethnic category tend to differ systematically in their allele frequencies (Price et al., 2009). Outcomes can also differ systematically for non-genetic reasons, causing spurious association. For example, if one tried to measure the genetic underpinnings of chopstick use, yet failed to control for population stratification, one would no doubt find all sorts of associations with SNPs related to Asian ancestry (Lander and Schork, 1994). However, these associations would only be based on happenstance cultural mechanisms, not true genetic associations. In conducting association tests, researchers therefore typically control for at least the first 10 principal components (PCs) derived from variation in *millions* of SNPs, capturing subtle ancestral differences among subjects. These continuous controls not only have a direct geographic interpretation (Novembre et al., 2008), but help avoid false positives driven by stratification[3]. While Fletcher acknowledges he could not rule out stratification as a confounding factor in his analysis, he was unable to test the extent to which it drove his results.

This paper has three goals. First, I repeat Fletcher's analysis of tax response heterogeneity with respect to rs2304297 and state excise taxes using the Health and Retirement Study (HRS). The HRS is a longitudinal study of older Americans with comprehensive genetic and smoking information, including data on variation in millions of SNPs. To my knowledge, the HRS is the largest available sample of Americans that offers such extensive individual genetic data and longitudinal smoking data. I show that Fletcher's main finding

[2]A notable exception involves Alzheimer's disease and the APOE gene. Plaques found in the brains of Alzheimer's patients contain apolipoproteins, which are produced by the APOE gene. SNPs coding for variation in APOE were originally discovered using the candidate gene approach and are among the most powerful common genetic variants that predict Alzheimer's (Strittmatter et al., 1993). However, the candidate approach is *not* this successful in a broader range of medical and behavioral phenotypes.

[3]Traditional measures of ethnicity and ancestral background also have flaws: (1) they are discrete and therefore coarse, in contrast with the (ever increasing) diversity of modern populations; and (2) they are subjectively measured via in-person surveys. Genetic counterparts to these metrics are (1) continuous, thereby allowing for a more nuanced appreciation of the effects of ethnicity on behavior; (2) objective, immune to participants misrepresenting their true identities and odd surveyor demand effects, and (3) increasingly cheap to measure. Whether conclusions derived these genetic measures match their survey-based cousins may help elucidate which notion of ethnic identity dominates in terms of its impact on economic behavior (a topic I do not delve further into here).

of genetic tax response heterogeneity along the extensive margin does not replicate in the HRS. I find doubling taxes is associated with $\sim$5% lower likelihood of smoking across all rs2304297 genotypes. I am able partially to replicate his finding that individuals with the GG genotype on rs2304297 are more responsive to cigarette taxes along the intensive margin: I find the effect in the HRS sample including subjects of all ethnicities, but not the sample including only subjects of European decent. In the full sample, I find doubling taxes is associated with $\sim$1.13 fewer CPD among individuals with the GG genotype, compared with $\sim$0.86 fewer CPD among individuals with other genotypes. Even my results for the full sample differ from Fletcher's results, which are consistent *only* with responsiveness for the GG genotype group. Among only Europeans-decent subjects, I find doubling taxes is associated with $\sim$1.08 fewer CPD across all rs2304297 genotypes.

Second, I evaluate the extent to which variation in rs2304297 is driven by population stratification. I find that the first 10 PCs derived from all SNPs explain $\sim$28% of the variation in rs2304297 among HRS participants. Genetically European participants are more likely to have an additional G reference allele, while genetically African participants are much less likely to have an additional G reference allele. I also find that variation in rs2304297 predicts state tax rate, reflecting the fact that different states tend to be composed of individuals from different ancestral backgrounds. These results indicate Fletcher's findings, as well as my limited successful replications, are likely to be driven by population stratification.

Third, I use estimates from two large, well-powered GWA studies to perform out-of-sample prediction on the HRS's European-decent subjects. These predictors, or polygenic scores, utilize the full gamut of SNPs available on the HRS and can be interpreted as measures of genetic predisposition to a particular phenotype or behavior. They capture much more genetic variation than any potential candidate SNP and are therefore better powered for predictive and gene-by-environment interaction analyses. Indeed, despite an inability to robustly identify individual SNPs that jointly account for the full heritability of a given trait ("missing heritability"), we are still able to use the joint predictive power of a large number of SNPs to investigate genetic predisposition (Benjamin et al., 2009). In particular, I construct measures of genetic predisposition to two phenotypes plausibly related to smoking behavior: high lifetime maximum cigarettes per day (CPD) and high educational attainment (EA), the latter proxying for genetic predisposition to high cognition (Okbay et al, 2015). I use a GWA study based on maximum CPD to construct my polygenic scores, rather than GWA studies associated with other smoking phenotypes (e.g. ever versus never smoker, smoking cessation, or age of first cigarette), given analyses using these other phenotypes found far fewer robust genetic associations.

Controlling for population stratification, I show these polygenic scores robustly predict cigarette consumption along both the extensive and intensive margins in the expected directions. However, when interacted with state tax rates, and additionally controlling for state economic conditions (per capita income and unemployment), self-reported individual health, state fixed effects, and wave fixed effects, I find *no* evidence of response heterogeneity along either margin. It therefore seems appropriate to turn to other, non-genetic potential explanations for the fact that cigarette excise taxes do not have an impact across the entire distribution of smokers.

More broadly speaking, insight into genetic heterogeneity may help tailor-fit addiction recovery methods (an example of "personalized medicine"). Indeed, identifying individuals who are most susceptible to different types of incentives could help improve the cost-effectiveness of recovery programs, isolating those who might respond to financial versus non-financial incentives or programs. The fact that my results constitute null findings disables me from forwarding any such cost-effectiveness strategy. However, this study still forwards the gene-by-environment interaction literature, which has historically been plagued by lack of replicability and low power, driven by utilization of candidate gene approaches (Duncan and Keller, 2011).

3.3 Related Literature

An extensive literature exists detailing the effects of taxes and prices on cigarette consumption. See Chaloupka and Warner (2000) for an excellent review of the literature to that point. Citing evidence from both aggregate and individual data, various authors find sizable elasticities along both the intensive and extensive margin (for individual data, roughly between 0 and -0.9, centered at approximately -0.7). Studies range in their care in addressing endogeneity and parsing the margins of consumption. Something of a consensus emerges that older people are less price responsive, but this is by no means ubiquitous (e.g. Wasserman et al., 1991 finding no difference between teenagers and adults but unstable estimates over time; DeCicca amd McLeod, 2008 finding sizable responsiveness among older adults, especially among less educated and low-income households). Maclean et al. (2015) also uses the HRS, finding extremely modest intensive margin tax elasticities between -0.02 and -0.04.

My focus on excise taxes rather than sales taxes is well supported. Evidence suggests that sales taxes on alcohol, which are only added to the price at the register, are far less salient than excise taxes; their associated elasticities are therefore much smaller (Chetty et al., 2009, using a field experiment to directly test for relevant demand effects). Goldin

and Homonoff (2013) extend this analysis to cigarettes, finding that sales taxes are more salient to poorer individuals. Relevant to our purposes, they verify that sales taxes are far less salient than excise taxes across income groups.

The budding subdiscipline of "genoeconomics" investigates the intersection of genetics and economics (Benjamin et al., 2012; Cesarini et al., 2009, 2010, 2012; Zyphur et al., 2009). However, little is known about biologically driven heterogeneity in individual responsiveness to taxes on addictive goods. Guo et al. (2010) explores sin good use with respect to both genotype and age-based legal status, i.e. the 21 year old drinking age, finding certain genes mitigate use to a greater extent at ages when said activity is illegal. Boardman (2009) uses twin pairs from the NHANES, finding daily smoking activity is highly heritable and that there is significant variation in the influence of genetics on smoking across states, with genetic influences lowest in states with high taxes on cigarettes. Using polygenic scores for detecting gene-by-environment interactions is also relatively unexplored. Okbay et al. (2015) uses an EA-based polygenic score and the Swedish Twin Registry, finding a decline in the explanatory power yielded by the score, as well as a decline in its associated regression coefficient, for younger birth cohorts. They argue that these declining genetic effects can be explained by a series of education reforms in Sweden during the middle of the 20th century. Belsky et al. (2013) constructs a risk score using the 3 SNPs robustly identified as being associated with smoking behavior by the TAG consortium to construct a polygenic score for a sample of New Zealanders. They find the score is unrelated to smoking initiation, but predicts (above and beyond familial background) conversion to daily smoking as a teenager (and conversion to heavy smoking), persistence of smoking, using smoking as a coping mechanism for stress, and failing to be able to quit.

3.4 Data

3.4.1 Health and Retirement Study

The HRS's biennial waves spanning 1992-2010 offer data on 2.5 million SNPs (imputed to over 20 million) for 12,595 older individuals, along with comprehensive socioeconomic and demographic information. Residence by state is available for each wave for most individuals (enabling matching to relevant cigarette excise tax rates). Also available is self-reported information on whether an individual has ever been a smoker, is a current smoker, and daily cigarette consumption (cigarettes per day).

3.4.2 Genome-Wide Association Studies

Past genetic studies, estimating effect sizes for each SNP, are necessary to calculate polygenic scores measuring genetic predisposition. I use the results of two large genetic studies that omit HRS to perform out-of-sample prediction on HRS and generate these polygenic scores. These studies are the Tobacco and Genetics Consortium (TAG, 2010), with n = 74,053 (for predisposition to high CPD[4]) and Okbay et al. (2015), with n = 285,072 (for predisposition to high EA). I focus on TAG's analysis of lifetime maximum CPD (as opposed to other smoking-related outcomes) given greater success in identifying highly robust SNPs associated with that outcome.

3.4.3 Cigarette Excise Taxes

Cigarette tax data at the federal and states levels are available from The Tax Burden on Tobacco (2012), compiled by Orzechowski and Walker. Federal cigarette taxes changed in 1993, 2000, 2002, and 2009. See Figure 3.3 for federal and average state taxes over time, expressed in cents per cigarette. Note that the most significant increase in 2009 was part of the State Children's Health Insurance Program (SCHIP). Given that the HRS is only administered every two years, I consider "the tax" for a specific wave and individual to be the average tax in her location over the prior year. I ignore local and other municipal excise taxes. While I conduct all below analyses with real tax rates (using the STATA's CPI package), results are almost identical using nominal tax rates. State unemployment rates and state real per capita income are from FRED and BEA, respectively.

3.5 Summary Statistics

See Tables 3.1 and 3.2 for summary statistics on static and time-dependent variables, respectively. On average, the sample was born in approximately 1938 and is disproportionately female. By design, the HRS over-samples minorities relative to the percentages of these groups in the overall American population. Almost 57% of the sample admits to having ever been a smoker, while 13% of the sample smokes in any given wave of the survey. Among all subjects and waves (including those who do not smoke), average CPD is almost 2 (among only smokers this number is almost 16). Moreover, among all subjects

[4]There are two other large cigarette consortia: Liu et al. (2010) using the Oxford-GlaxoSmithKline (OxGSK) consortium, n = 41,150; and Thorgeirsson et al. (2010) using the European Network of Genetic and Genomic Epidemiology (ENGAGE) consortium, n = 46,481. Unfortunately, only TAG's results are publicly available. However, TAG reports that their main findings replicate in a combined sample of the three consortia.

and waves, the average combined real (in 2000 dollars) tax from both federal and state governments is about $\$1.03$ per pack.

3.6 Replication of Fletcher (2012)

First, I repeat Fletcher (2012)'s main results using the HRS rather than NHANES. He regresses an indicator for tobacco use on the log state excise tax rate, an indicator for the GG genotype on rs2304297, an interaction term between the two, and additional covariates (see his Table 2). Exogeneity of the tax rate rests on an institutional feature of state budgetary regimes: all states except Vermont have balanced budget rules. In other words, they cannot issue and carry forward long-term debt as the federal government can. While the series of tax increases in the 1960s and 1970s were an endogenous response to the smoking and drinking culture of the era, tax increases in the 1990s and 2000s are more plausibly exogenous responses to countercyclical budgetary shortfalls, meant merely as a convenient way to raise necessary revenue (Maag and Merriman, 2003).

Fletcher's analysis clusters standard errors at the state level because variation in taxation is at the state level. While the NHANES contains cross-sectional data, the HRS contains longitudinal data, so I must additionally cluster standard errors by individual, lest end up with standard errors that are too tight. I therefore employ a 2-dimensional clustering procedure (Cameron et al., 2011; Thompson, 2011), implemented by the cluster2 package in STATA developed by Mitchell Petersen[5].

See Figure 3.4 for my extensive margin results. Column (1) includes only the tax rate; the coefficient is significant and negative, and thus similar to the first column in Fletcher's Table S1, although my effect size is larger in absolute value. Column (2) includes only the rs2304297 GG indicator; the covariate is significant and negative, and is thus similar to the second column in Fletcher's Table S2, although my effect size is smaller in absolute value. Column (3) mimics Fletcher's main specification in his Table 2. Fletcher's finding of genetic tax response heterogeneity along the extensive margin does not replicate in the HRS. He does not find a significant main tax effect, but significant negative effects for both the rs2304297 GG genotype indicator and its interaction with log taxes. In the HRS, I only find a significant main tax effect, namely that doubling taxes is associated with $\sim$4.5% lower likelihood of being a smoker across all rs2304297 genotypes (i.e. no significant interaction). Fletcher notes his findings are robust to analyzing only self-reported European-decent subjects, although he does not include additional covariates in this verifi-

[5] www.kellogg.northwestern.edu/faculty/petersen/htm/papers/se/se_programming.htm

cation (his Table S3). My Column (4) mimics this specification *with* additional covariates, finding that doubling taxes is associated with $\sim$5.2% lower likelihood of smoking across all rs2304297 genotypes among self-reported European-decent subjects (i.e. no significant interaction). Therefore, along the extensive margin, Fletcher's results based on the NHANES do not replicate in the HRS.

See Figure 3.5 for my intensive margin results. Column (1) regresses CPD on the log tax rate alone, finding a significant, negative association. Column (2) regresses CPD on an indicator for GG genotype on rs2304297, finding no significant association. Column (3) mimics the Fletcher's only intensive margin specification (his Table S2). My estimates imply that doubling taxes is associated with $\sim$1.13 fewer CPD among individuals with the GG genotype, but $\sim$0.86 fewer CPD among individuals with other genotypes. However, Fletcher's results indicate that *only* individuals with the GG genotype are responsive, and that a doubling of taxes is associated with GG genotype individuals reducing consumption by $\sim$0.64 CPD. Non-GG genotype subjects, according to Fletcher's analysis, do not respond to taxes at all. Therefore, although both our analyses are consistent with some degree of tax response heterogeneity with respect to the GG genotype on rs2304297, our results still differ. Column (4) restricts the sample to self-reported European-decent subjects; the coefficients on rs2304297 and the interaction lose significance, indicating that among all subjects, doubling taxes is associated with a reduction of $\sim$1.08 CPD.

In sum, along the extensive margin, I do not replicate Fletcher (2012)'s results in the HRS. Along the intensive margin, I find suggestive evidence of tax response heterogeneity with respect to the GG genotype on rs2304297 and the full HRS sample, but my findings still differ from Fletcher's. Differences between the HRS and NHANES samples are no doubt driving these divergent findings.

3.7 Population Stratification and rs2304297

Rs2304297 genotypes marginally predict individual state tax rates (Figure 3.6). Here, I use indicators for the three potential genotypes: GG, CG, and CC (omitted). This mimics the specification presented by Fletcher in his Table S4. My results indicate that variation in rs2304297 marginally predicts state tax rate, reflecting the fact that different states tend to be composed of individuals from different ancestral backgrounds. This is indirect evidence of population stratification driving variation in rs2304297.

I next directly test whether variation in rs2304297 is driven by population stratification, see Figure 3.7. Column (1) regresses an indicator for the GG genotype on indicators for ancestral background based on principal components (PCs) derived from all genotyped

SNPs (and all HRS participants) and HRS's quality control manual[6] for delineating whether someone is genetically of European, African, or Hispanic/Asian decent (with the latter as the omitted category). In particular, African decent individuals have PC1 $>= 0.008$, European-decent individuals have PC1 < 0.008 and PC2 < 0.006, and Hispanic/Asian decent individuals have PC2 $>= 0.006$. Coefficients on each indicator are highly significant. Column (2) directly regresses the GG indicator on the first 10 PCs, finding significant explanatory power of the first three PCs, and joint predictive power in terms of R^2 of $\sim$13.0%. Columns (3) and (4) repeat the analyses in the first two columns, only using a count variable for the dependent variable that enumerates how many G alleles an individual has (it can therefore take on the values of 0, 1, or 2). Results are consistent with the first two columns, indicating that the first 10 PCs derived from all SNPs and HRS participants explain $\sim$27.7% of the variation in the number of G alleles on rs2304297. Participants of European decent are more likely to have an additional G reference allele, while participants of African decent are much less likely to have an additional G reference allele.

These results indicate Fletcher's findings, as well as my limitedly successful replications along the intensive margin, are at least partially driven by population stratification. Indeed, given the fact that participants of Europeans decent are more likely to have a GG genotype on rs2304297, and Fletcher finds GG genotype people are the only ones responsive to taxes, it is likely that at least some of the gene-by-environment interaction is driven by ancestry. This result is consistent with MacClean (2015), finding greater intensive margin responsiveness to cigarette taxes among Europeans on the HRS. However, it is inconsistent with other studies using younger samples, consistently finding greater responsiveness among non-Europeans (Gruber and Zinman, 2001; DeCicca et al., 2000; Chaloupka and Pacula, 1999).

3.8 Polygenic Scores and Cigarette Tax Response Heterogeneity

I next use effect size estimates from two large, well-powered GWA studies (Tobacco and Genetics Consortium, 2010; Okbay et al., 2015) to perform out-of-sample prediction on the HRS's European-decent subjects. These predictors, or polygenic scores, utilize the full gamut of SNPs available on the HRS and measure genetic predisposition to a particular phenotype or behavior. They capture much more genetic variation than any single candidate SNP and are therefore better powered for predictive and gene-by-environment

[6] hrsonline.isr.umich.edu/sitedocs/genetics/HRS_QC_REPORT_MAR2012.pdf

interaction analyses. In particular, I construct measures for genetic predisposition to two phenotypes related to smoking behavior: lifetime maximum CPD and EA. I choose lifetime maximum CPD as my smoking phenotype because the relevant GWA study found many more robustly replicable SNPs than those based on ever versus never smoking status or smoking cessation. Okbay et al. (2015) provides compelling evidence that EA is a proxy-phenotype for cognition; studying EA is preferred given measuring years of schooling is much easier than measuring cognition, facilitating a much larger sample size (and more statistical power) for genetic discovery. I calculate these polygenic scores only for the European-decent HRS subsample because the GWA studies on which the coefficient estimates are based are also conducted on European-decent samples; coefficient estimates from GWA studies are generally not robust to different ancestral backgrounds. I therefore only conduct my heterogeneity analyses in this section on the HRS subjects of European decent.

These scores are simple to calculate. GWA studies, including those whose effect size estimates I utilize, posit millions of regressions[7] of the following form, one per SNP:

$$y_i = \mu + \beta_j x_{ij} + \delta z_i + \epsilon_i \tag{3.1}$$

y_i is the phenotype or behavior of interest for individual i. β_j is the effect of SNP x_{ij}. z_i is a vector of non-genetic controls (e.g. age, gender, and PCs); δ is the associated vector of coefficients; and ϵ_i is the residual. Note this additive specification assumes that the difference between having 0 and 1 reference alleles is the same as the difference between having 1 and 2 reference alleles. Gene-gene interactions are also ignored.

How exactly to weigh and combine these coefficients into predictive polygenic scores is a subject of active research, a full discussion of which is beyond the scope of this paper. If we define x_i as the stacked column vector of SNPs x_{ij} for individual i, and $\hat{\beta}$ as the row vector of related coefficient estimates from a particular GWA study, then we can calculate a simple polygenic score for individual i:

$$\hat{g_i} = \hat{\beta} x_i \tag{3.2}$$

While this method fails to take into account double counting and systematic spatial correlation among SNPs (i.e. linkage disequilibrium), in practice, it has shown to be successful in measuring genetic predisposition (e.g. for EA and HRS participants, see Okbay et al., 2015). It is important to note that GWA estimates used to construct my polygenic scores

[7]Multiple testing necessitates Bonferroni-adjusted significance levels. Qualifying for "genome-wide significance" requires a p-value of $5 * 10^{-8}$ or less.

come from younger birth cohorts than those included in HRS. If genetic associations differ among birth cohorts, my scores will tend to be worse predictors and thereby lessen our power to detect heterogeneity. This same logic applies spatially given the fact that not all cohorts used to derive the GWA estimates are American, while the HRS is indeed entirely American.

Figures 3.2 and 3.1 show histograms of the cigarette- and EA-based scores, respectively, for the HRS sample of European decent. Higher scores represent increased genetic predisposition to cigarette consumption and cognition. Note their raw correlation is -0.0837[8].

Both scores are significant predictors of smoking status and CPD in the expected directions, i.e. the cigarette score predicts higher incidence of smoking along both consumption margins, and the EA score predicts lower incidence and of smoking along both consumption margins (Figure 3.8 for smoking status and the cigarette score, Figure 3.9 for CPD and the cigarette score, Figure 3.10 for smoking status and the EA score, and Figure 3.11 for CPD and the EA score). Note that these analyses ignore the longitudinal nature of the data since all covariates are non-time varying. There is also no need for clustering at the state nor individual levels. Along the extensive margin, the cigarette score has incremental predictive power in terms of R^2 of less than 0.1% (comparing Columns 2 and 3). Along the intensive margin, this incremental predictive power is $\sim 0.1\%$. The EA scores have a much higher incremental predictive power: $\sim 0.7\%$ along both margins. This difference owes to the much larger sample size of the GWA study underlying EA.

I then repeat and expand on the tax heterogeneity analyses conducted above, separately substituting rs2304297 for each score. Aside from avoiding issues of population stratification that come with analyzing rs2304297 without PC controls, my analyses based on the polygenic scores are better powered than their candidate gene counterparts given the scores capture much more genetic variation than any single SNP. See Figure 3.12 for smoking status and the cigarette score, Figure 3.13 for CPD and the cigarette score, Figure 3.14 for smoking status and the EA score, Figure 3.15 for CPD and the EA score. These analyses use only the European sub-sample of the HRS as defined by the HRS's own genetic quality control guidelines. Again, this is because the GWA studies' estimates underlying the polygenic scores are only valid for European-decent individuals.

Column (1) in each figure uses the same covariates as Fletcher (2012) and my above replication attempts. Column (2) in each figure replaces self-reported indicators of ethnicity with 10 PCs (derived using *only* the European-decent subsample and all available

[8] I also estimate a genetic correlation of -0.284 (se: 0.064) between these traits using the LD score method developed by Bulik-Sullivan et al. (2015) and implemented in their LDSC python software package (with LD score data outlined in Finucane et al., 2015).

SNPs, thereby controlling for even more subtle population stratification among those with European decent). I then try a number of additional specifications by adding different covariates to Column (2). Column (3), continuing to use the European PCs, adds state economic conditions as controls (log per capita income and log unemployment, see Goldin and Homonoff, 2013; Ruhm, 2005). I control for state economic conditions because if states introduce excise taxes to overcome budgetary shortfalls, which themselves tend to happen in bad economic climates, excluding relevant proxies could introduce omitted variable bias if cigarette consumption is also correlated with the business cycle. Column (4), again using the PCs, instead controls for self-reported change in health, given the central role of health shocks in changing smoking behavior among older populations. Column (5) uses state fixed effects (preferred to controlling directly for time-invariant state-specific trends of cigarette consumption). Column (6) uses wave fixed effects, thereby controlling for unobserved temporal heterogeneity. Finally, Column (7) in each figure includes all these covariates together in a single specification.

In all specifications, for both scores and along both margins, I robustly find *no* evidence of genetic tax response heterogeneity among European-decent subjects, i.e. the interaction terms between the scores and log tax rates are always insignificant (even at the 10% level). While I continue to find in many specifications that taxes deter behavior along both margins, only the EA scores, not the cigarette scores, are significant predictors of behavior.

Figure 3.12 shows results for smoking status and the cigarette score. Among all specifications, there is no evidence of the cigarette score remaining predictive nor of any interaction effect with the log tax rate. However, taxes do have the expected, dissuasive impact on smoking status in most specifications. This effect ranges from a doubling of taxes being associated with a reduction in the likelihood of being a smoker by $\sim$6.5$\%$ to $\sim$9.6$\%$. When economic covariates or wave fixed effects are included, as in Columns (3), (6), and (7), taxes no longer significantly impact smoking status. This points to time-specific economic conditions, which happen to be correlated with increases in cigarette excise taxes thanks to states using these policies during bad economic times (Maag and Merriman, 2003), driving much of the observed tax effect. While broader questions involving the relationship among economic conditions, government policy, and cigarette consumption are interesting, they are beyond the scope of this paper. Toward my aims, it is sufficient to highlight that there is no evidence of tax response heterogeneity along the extensive margin with respect to genetic predisposition to heavy smoking.

Figure 3.13 displays results for CPD and the cigarette score, with broadly similar conclusions as the extensive margin. The tax effect, when it appears, does not display heterogeneity with respect to genetic predisposition to heavy daily smoking. A doubling of

the tax rate is associated with a reduction between ∼0.87 and ∼0.1.39 CPD depending on the specification. The score is again not directly predictive along the intensive margin, and there is no evidence of tax response heterogeneity.

Figure 3.14 presents results for smoking status and the EA score. Across all specifications, the EA score is predictive of smoking status in the expected direction: higher genetic predisposition to EA is associated with lower propensity to smoke. The tax effect again appears in most specifications, ranging from a doubling of taxes being associated with a reduction in the likelihood of being a smoker by ∼5.3% to ∼8.6%. This disappears in Columns (3), (6), and (7), when economic covariates or wave fixed effects are included. There is no evidence of genetic tax response heterogeneity along the extensive margin with respect to the EA score.

Figure 3.15 shows results for CPD and the EA score. Again, across all specifications, the EA score is predictive of CPD in the expected direction: higher genetic predisposition to EA is associated with fewer CPD. The tax effect again appears in most specifications. Unlike prior analyses, the tax effect does not disappear in Column (6) (with only the inclusion of wave fixed effects); however, the significant point estimate is much lower than in the other specifications (∼-0.46 versus between ∼-1.16 to ∼-1.70). This again points to time-specific economic conditions driving much of the observed tax effect. Again, the interaction term between the score and log tax rate is insignificant, indicating no genetic tax response heterogeneity along the intensive margin with respect to the EA score.

3.9 Conclusions

Genetic predisposition to heavy smoking and high cognition do not appear to mediate the impact of cigarette excises taxes on consumption among HRS participants. It therefore seems appropriate to turn to other, non-genetic explanations for why cigarette excise taxes do not have an impact across the entire distribution of smokers, and why elasticities appear to be falling over time. These explanations include internet-based purchases that evade state taxes (Goolsbee et al., 2010; Emery et al., 2002; Ribisl et al., 2001); substitution toward other, differentially taxed tobacco products or cigarettes with higher nicotine content (Obsfeldt et al., 1998, Delnevo et al., 2004; Evans and Farrelly, 1998; Obsfeldt and Boyle, 1997), especially e-cigarettes (Huang et al., 2014); the pool of smokers having dwindled to the most addicted, least responsive smokers; and responding to tax changes not by adjusting number of cigarettes smoked, but by the manner of smoking (e.g. consuming a particular cigarette for longer by taking deeper drags and smoking to the butt, Adda and Cornaglia, 2006, 2012; Abrevaya and Puzzello, 2012). Some of these explanations are less

relevant to the HRS's older sample of Americans, namely internet-based smuggling and substitution (especially toward e-cigarettes). The notion that the pool of smokers among the HRS sample has dwindled to the most addicted, least responsive, is likely the most relevant explanation.

That being said, further replication is warranted to verify these results beyond the HRS and the particular polygenic scores considered here. Although a null finding, this study forwards the gene-by-environment interaction literature, which has historically been plagued by lack of replicability and low power, driven by utilization of candidate gene approaches (Duncan and Keller, 2011). The editor of *Behavior Genetics* has responded to these challenges by requiring potentially publishable candidate gene studies to be well-powered, adjust for all sources of multiple testing, and replicate in at least one other sample (Hewitt, 2011). One particularly relevant limitation of my study is the fact that not many genetic variants associated with smoking have been robustly identified. Once more have been discovered, it is possible we will observe response heterogeneity with respect to those newly discovered variants. Using a polygenic score based on estimates from a better-powered GWA study of smoking, which capture the effects of variants with smaller effect sizes, might also uncover response heterogeneity. Replication with a younger sample would also be particularly useful, ensuring that my null results are not driven by the HRS's older sample. These analyses could also be extended to alcohol, coffee, fatty foods, sugary foods, or entire diets; any consumption that is impacted by genetic predisposition and depends on some price or policy might be analyzed for this sort of response heterogeneity.

3.10 Figures

Table 3.1: Summary Statistics, Static Variables

Variable	Mean	Std. Dev.	N
Birth Year	1938.391	10.575	12507
Female	0.591	0.492	12507
Black	0.133	0.34	12507
Hispanic	0.096	0.295	12506
Other Race	0.03	0.169	12507
Years of Education	12.582	3.164	12489
rs2304297 (GG)	0.512	0.5	12454
rs2304297 (G)	1.384	0.702	12454
European Decent by PCs	0.774	0.418	12419
African Decent by PCs	0.134	0.341	12419
Hispanic or Asian Decent by PCs	0.092	0.289	12419
CPD Polygenic Score	4.148	1	8651
EA Polygenic Score	3.528	1	8651

Table 3.2: Summary Statistics, Time-Varying

Variable	Mean	Std. Dev.	N
Ever Smoker	0.566	0.496	124510
Smoker	0.13	0.336	108592
CPD	1.944	6.547	107521
CPD (Smokers)	15.932	11.333	13119
Real Cigarette Tax (cents/pack)	103.589	59.179	94861
Married	0.551	0.497	125070
Income ($1000's)	14.18	38.452	96452

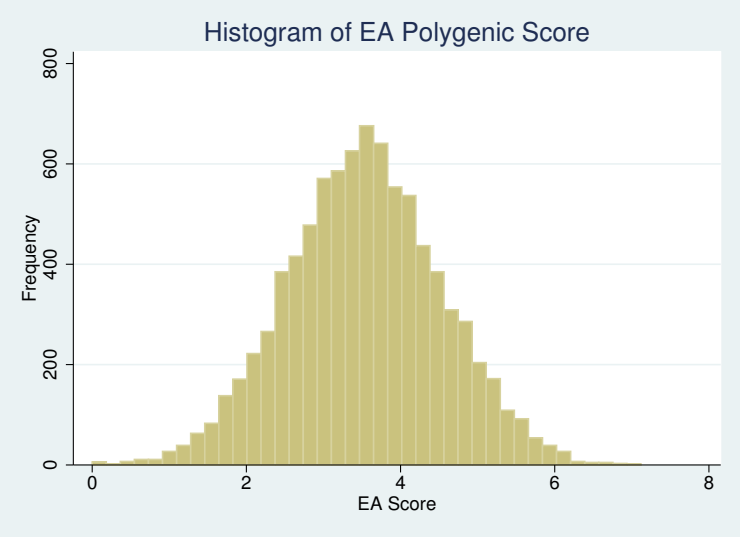

Figure 3.1: Histogram of polygenic scores based on GWA study estimates from Okbay et al. (2015), measuring genetic predisposition to high educational attainment (EA).

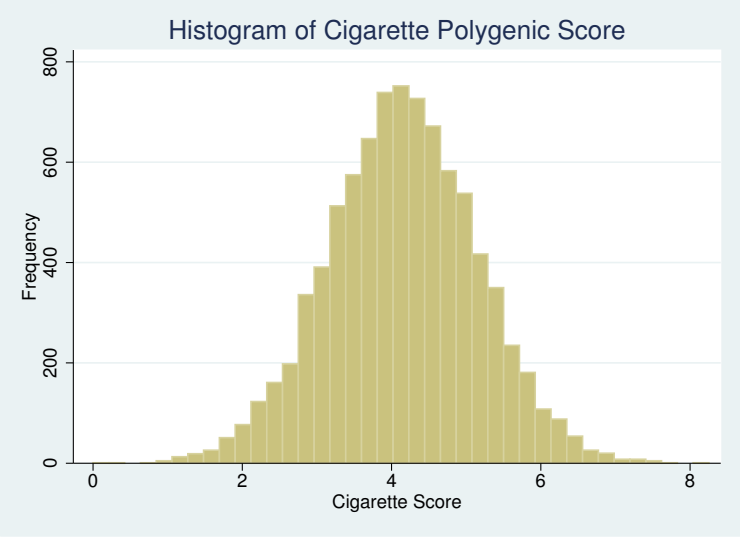

Figure 3.2: Histogram of polygenic scores based on GWA study estimates from the TAG Consortium (2010), measuring genetic predisposition to high daily cigarette consumption.

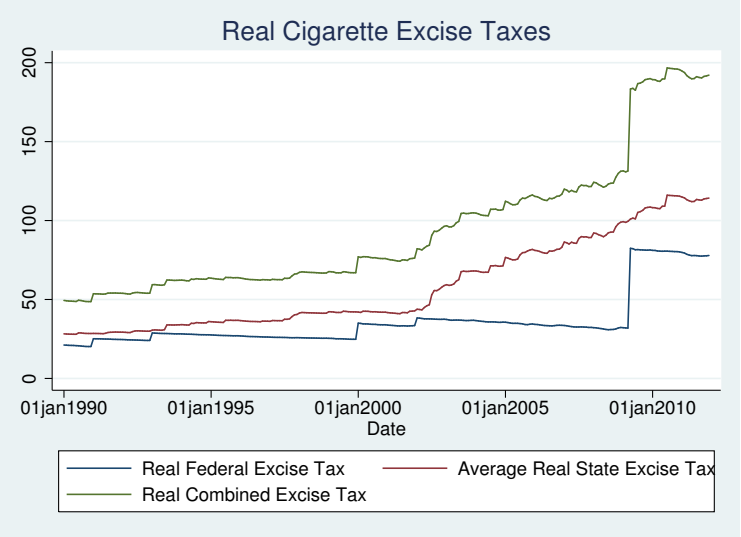

Figure 3.3: Federal, average state, and combined real (in 2000's dollars) cigarette excise taxes over time. The tax is levied per pack (20 cigarettes), expressed here in cents per pack. Note a dramatic increase in 2009 corresponding to provisions in the State Children's Health Insurance Program (SCHIP).

	(1) Smoke	(2) Smoke	(3) Smoke	(4) Smoke
ln(Tax)	-0.0472***		-0.0454***	-0.0524***
	(0.00520)		(0.00711)	(0.00878)
rs2304297(GG)		-0.00987*	0.0411	0.0118
		(0.00522)	(0.0341)	(0.0301)
ln(Tax) X GG			-0.00842	-0.00173
			(0.00733)	(0.00637)
Birth Year			0.0267	0.0794
			(0.0891)	(0.0951)
Birth Year Squr'd			-0.00000478	-0.0000185
			(0.0000231)	(0.0000246)
Female			-0.0412***	-0.0322***
			(0.00621)	(0.00631)
Black			-0.000273	
			(0.0127)	
Hispanic			-0.0823***	-0.0863***
			(0.0141)	(0.0165)
Other Race			-0.0304	
			(0.0190)	
EA			-0.0132***	-0.0145***
			(0.000962)	(0.00127)
Married			-0.0745***	-0.0787***
			(0.00734)	(0.00796)
Income (1000's)			-0.000297***	-0.000228***
			(0.0000934)	(0.0000754)
Observations	92305	108125	90501	76360
R^2	0.005	0.000	0.056	0.056

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.4: Fletcher (2012)'s specification along the extensive margin. OLS regression of smoking status on log cigarette excise tax rate, an indicator for having the GG genotype on rs2304297, an interaction term between log tax rate and the GG indicator, and covariates: birth year, birth year squared, an indicator for being female, indicators for self-reported race (with the omitted category being white), educational attainment (EA), being currently married, and income in thousands of dollars. Columns (1), (2), and (3) use the full HRS sample; Column (4) uses only self-reported whites. Standard errors are clustered by state and individual.

	(1) CPD	(2) CPD	(3) CPD	(4) CPD
ln(Tax)	-0.940***		-0.855***	-1.084***
	(0.0818)		(0.126)	(0.163)
rs2304297(GG)		0.0934	1.313*	0.422
		(0.0926)	(0.677)	(0.722)
ln(Tax) X GG			-0.274**	-0.0733
			(0.139)	(0.148)
Birth Year			1.269	1.621
			(1.579)	(1.575)
Birth Year Squr'd			-0.000293	-0.000383
			(0.000409)	(0.000408)
Female			-1.027***	-0.967***
			(0.113)	(0.115)
Black			-1.125***	
			(0.211)	
Hispanic			-2.260***	-2.452***
			(0.273)	(0.311)
Other Race			-0.494	
			(0.303)	
EA			-0.238***	-0.272***
			(0.0192)	(0.0279)
Married			-1.177***	-1.320***
			(0.109)	(0.131)
Income (1000's)			-0.00462***	-0.00397***
			(0.00166)	(0.00152)
Observations	91434	107060	89636	75608
R^2	0.005	0.000	0.043	0.045

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.5: Fletcher (2012)'s specification along the intensive margin. OLS regression of cigarettes per day (CPD) on log cigarette excise tax rate, an indicator for having the GG genotype on rs2304297, an interaction term between log tax rate and the GG indicator, and covariates: birth year, birth year squared, an indicator for being female, indicators for self-reported race (with the omitted category being white), educational attainment (EA), being currently married, and income in thousands of dollars. Columns (1), (2), and (3) use the full HRS sample; Column (4) uses only self-reported whites. Standard errors are clustered by state and individual.

	(1) ln(Tax)	(2) ln(Tax)
rs2304297(CG)	0.0636***	0.0219*
	(0.0118)	(0.0127)
rs2304297(GG)	0.0747***	0.0231*
	(0.0113)	(0.0129)
Birth Year		-0.478***
		(0.0938)
Birth Year Squr'd		0.000123***
		(0.0000242)
Black		-0.0612***
		(0.0132)
Hispanic		0.116***
		(0.0102)
Other Race		0.0620***
		(0.0195)
EA		0.00821***
		(0.00115)
Income (1000's)		0.000313***
		(0.000111)
Wave FE	No	Yes
Observations	94471	90871
R^2	0.002	0.404

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.6: Fletcher (2012)'s specification. OLS regression of log cigarette excise tax rate on indicators for genotype on rs2304297 (CC as the omitted category), birth year, birth year squared, indicators for self-reported race (with the omitted category being white), educational attainment (EA), and income in thousands of dollars. Column (2) additionally includes wave fixed effects. Standard errors are clustered by state and individual.

	(1) rs2304297(GG)	(2) rs2304297(GG)	(3) rs2304297(G)	(4) rs2304297(G)
European by PC's	0.0533***		0.0735***	
	(0.0146)		(0.0188)	
African by PC's	-0.475***		-0.997***	
	(0.0180)		(0.0231)	
PC1		-19.96***		-40.83***
		(0.468)		(0.598)
PC2		1.575***		4.595***
		(0.467)		(0.597)
PC3 - PC10	No	Yes	No	Yes
Observations	12367	12367	12367	12367
R^2	0.128	0.130	0.268	0.277

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.7: OLS regressions testing for the extent to which population stratification drives variation in rs2304297. Column (1) regresses an indicator for the GG genotype on indicators for race (determined by the first two principal components (PCs) derived from the full HRS sample and all genotyped SNPs, according to HRS's quality control manual, with the omitted category being "Hispanic/Asian"). Column (2) regresses an indicator for the GG genotype on the first 10 principal components (PCs) of genotyped SNPs from the full HRS sample. Columns (3) and (4) repeat these analyses with the *number* of G alleles on rs2304297 as the dependent variable.

	(1) Smoke	(2) Smoke	(3) Smoke
Cig. Score	0.0124***		0.00966**
	(0.00461)		(0.00463)
Birth Year		2.108***	2.101***
		(0.113)	(0.113)
Birth Year Squr'd		-0.000544***	-0.000542***
		(0.0000292)	(0.0000292)
Female		0.00517	0.00462
		(0.00929)	(0.00929)
PC's	No	Yes	Yes
Observations	6133	6133	6133
R^2	0.001	0.035	0.035

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.8: OLS regressions testing the predictive power of a polygenic score for smoking on smoking status. Column (1) regresses smoking status on the score. Column (2) regresses smoking status on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification.

	(1) CPD	(2) CPD	(3) CPD
Cig. Score	0.382***		0.323***
	(0.109)		(0.111)
Birth Year		41.06***	40.82***
		(2.684)	(2.684)
Birth Year Squr'd		-0.0106***	-0.0105***
		(0.000693)	(0.000693)
Female		-0.739***	-0.757***
		(0.239)	(0.239)
PC's	No	Yes	Yes
Observations	6133	6133	6133
R^2	0.002	0.027	0.028

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.9: OLS regressions testing the predictive power of a polygenic score for smoking on cigarettes per day (CPD). Column (1) regresses CPD on the score. Column (2) regresses CPD on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification.

	(1) Smoke	(2) Smoke	(3) Smoke
EA Score	-0.0334***		-0.0306***
	(0.00449)		(0.00445)
Birth Year		2.108***	2.058***
		(0.113)	(0.112)
Birth Year Squr'd		-0.000544***	-0.000531***
		(0.0000292)	(0.0000289)
Female		0.00517	0.00299
		(0.00929)	(0.00927)
PC's	No	Yes	Yes
Observations	6133	6133	6133
R^2	0.009	0.035	0.042

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.10: OLS regressions testing the predictive power of a polygenic score for educational attainment (EA) on smoking status. Column (1) regresses smoking status on the score. Column (2) regresses smoking status on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification.

	(1) CPD	(2) CPD	(3) CPD
EA Score	-0.771***		-0.721***
	(0.105)		(0.105)
Birth Year		41.06***	39.89***
		(2.684)	(2.653)
Birth Year Squr'd		-0.0106***	-0.0103***
		(0.000693)	(0.000685)
Female		-0.739***	-0.790***
		(0.239)	(0.239)
PC's	No	Yes	Yes
Observations	6133	6133	6133
R^2	0.008	0.027	0.034

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.11: OLS regressions testing the predictive power of a polygenic score for educational attainment (EA) on cigarettes per day (CPD). Column (1) regresses CPD on the score. Column (2) regresses CPD on birth year, birth year squared, an indicator for being female, and the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) combines these covariates into a single specification.

	(1) Smoke	(2) Smoke	(3) Smoke	(4) Smoke	(5) Smoke	(6) Smoke	(7) Smoke
ln(Tax)	-0.0648***	-0.0677***	-0.0238	-0.0674***	-0.0960***	-0.0258	0.000547
	(0.0232)	(0.0226)	(0.0239)	(0.0226)	(0.0211)	(0.0229)	(0.0225)
Cig. Score	-0.00506	-0.00653	-0.00521	-0.00410	-0.00252	-0.00446	-0.000777
	(0.0234)	(0.0235)	(0.0236)	(0.0235)	(0.0234)	(0.0235)	(0.0241)
ln(Tax) X Score	0.00232	0.00276	0.00254	0.00205	0.00198	0.00238	0.00146
	(0.00507)	(0.00500)	(0.00503)	(0.00497)	(0.00500)	(0.00498)	(0.00512)
Non-Race Cov.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Self-Rep. Race	Yes	No	No	No	No	No	No
PC's	No	Yes	Yes	Yes	Yes	Yes	Yes
Economic Cov.	No	No	Yes	No	No	No	Yes
Health Cov.	No	No	No	Yes	No	No	Yes
State FE	No	No	No	No	Yes	No	Yes
Wave FE	No	No	No	No	No	Yes	Yes
Observations	63819	63819	63208	60327	63819	63819	59737
R^2	0.061	0.062	0.066	0.061	0.072	0.069	0.078

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.12: OLS regressions testing for cigarette excise tax response heterogeneity along the extensive margin with respect to genetic variation in a polygenic score for cigarette smoking. All analyses include only genetically European-decent HRS participants. Column (1) regresses smoking status on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual.

	(1) CPD	(2) CPD	(3) CPD	(4) CPD	(5) CPD	(6) CPD	(7) CPD
ln(Tax)	-0.872*	-0.927*	-0.0670	-0.945*	-1.386***	-0.187	0.517
	(0.488)	(0.475)	(0.537)	(0.486)	(0.417)	(0.495)	(0.460)
Cig. Score	0.456	0.418	0.428	0.525	0.501	0.444	0.579
	(0.525)	(0.524)	(0.518)	(0.549)	(0.470)	(0.519)	(0.487)
ln(Tax) X Score	-0.0696	-0.0604	-0.0609	-0.0848	-0.0766	-0.0649	-0.0943
	(0.111)	(0.110)	(0.109)	(0.114)	(0.0992)	(0.109)	(0.102)
Non-Race Cov.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Self-Rep. Race	Yes	No	No	No	No	No	No
PC's	No	Yes	Yes	Yes	Yes	Yes	Yes
Economic Cov.	No	No	Yes	No	No	No	Yes
Health Cov.	No	No	No	Yes	No	No	Yes
State FE	No	No	No	No	Yes	No	Yes
Wave FE	No	No	No	No	No	Yes	Yes
Observations	63172	63172	62570	59843	63172	63172	59261
R^2	0.050	0.050	0.055	0.052	0.059	0.063	0.072

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.13: OLS regressions testing for cigarette excise tax response heterogeneity along the intensive margin with respect to genetic variation in a polygenic score for cigarette smoking. All analyses include only genetically European-decent HRS participants. Column (1) regresses cigarettes per day (CPD) on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual.

	(1) Smoke	(2) Smoke	(3) Smoke	(4) Smoke	(5) Smoke	(6) Smoke	(7) Smoke
ln(Tax)	-0.0531***	-0.0546***	-0.0115	-0.0581***	-0.0861***	-0.0142	0.00757
	(0.00745)	(0.00733)	(0.0121)	(0.00758)	(0.00589)	(0.0103)	(0.00842)
EA Score	-0.0159***	-0.0157***	-0.0158***	-0.0166***	-0.0151***	-0.0159***	-0.0163***
	(0.00380)	(0.00386)	(0.00392)	(0.00384)	(0.00506)	(0.00380)	(0.00517)
ln(Tax) X Score	-0.000405	-0.000262	-0.000303	-0.0000261	-0.000353	-0.000258	-0.000178
	(0.00106)	(0.00104)	(0.00105)	(0.00106)	(0.00109)	(0.00104)	(0.00110)
Non-Race Cov.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Self-Rep. Race	Yes	No	No	No	No	No	No
PC's	No	Yes	Yes	Yes	Yes	Yes	Yes
Economic Cov.	No	No	Yes	No	No	No	Yes
Health Cov.	No	No	No	Yes	No	No	Yes
State FE	No	No	No	No	Yes	No	Yes
Wave FE	No	No	No	No	No	Yes	Yes
Observations	63819	63819	63208	60327	63819	63819	59737
R^2	0.063	0.063	0.068	0.063	0.074	0.070	0.079

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.14: OLS regressions testing for cigarette excise tax response heterogeneity along the extensive margin with respect to genetic variation in a polygenic score for educational attainment (EA). All analyses include only genetically European-decent HRS participants. Column (1) regresses smoking status on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual.

	(1) CPD	(2) CPD	(3) CPD	(4) CPD	(5) CPD	(6) CPD	(7) CPD
ln(Tax)	-1.158***	-1.183***	-0.325	-1.323***	-1.703***	-0.463**	0.110
	(0.151)	(0.147)	(0.249)	(0.158)	(0.122)	(0.196)	(0.185)
EA Score	-0.325***	-0.329***	-0.329***	-0.347***	-0.317***	-0.331***	-0.337***
	(0.0844)	(0.0846)	(0.0870)	(0.0838)	(0.0984)	(0.0842)	(0.101)
ln(Tax) X Score	0.000180	0.00369	0.00347	0.00875	0.00193	0.00424	0.00596
	(0.0221)	(0.0215)	(0.0218)	(0.0217)	(0.0212)	(0.0215)	(0.0214)
Non-Race Cov.	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Self-Rep. Race	Yes	No	No	No	No	No	No
PC's	No	Yes	Yes	Yes	Yes	Yes	Yes
Economic Cov.	No	No	Yes	No	No	No	Yes
Health Cov.	No	No	No	Yes	No	No	Yes
State FE	No	No	No	No	Yes	No	Yes
Wave FE	No	No	No	No	No	Yes	Yes
Observations	63172	63172	62570	59843	63172	63172	59261
R^2	0.051	0.051	0.056	0.053	0.060	0.064	0.074

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 3.15: OLS regressions testing for cigarette excise tax response heterogeneity along the intensive margin with respect to genetic variation in a polygenic score for educational attainment (EA). All analyses include only genetically European-decent HRS participants. Column (1) regresses cigarettes per day (CPD) on the log cigarette excise tax rate, the score, their interaction, and covariates: birth year, birth year squared, educational attainment (EA), income in thousands of dollars, and indicators for self-reported race. Column (2) replaces indicators for self-reported race with the first 10 principal components (PCs) of genotyped SNPs from European-decent HRS subjects. Column (3) augments this specification with state economic covariates (log per capita income and log unemployment). Column (4) instead includes self-reported change in health as a covariate. Column (5) instead uses state fixed effects, and Column (6) wave fixed effects. Column (7) combines all covariates from Columns (3) through (6). Standard errors are clustered by state and individual.

3.11 References and Works Consulted

Abrevaya, Jason, and Laura Puzzello. 2012. "Taxes, Cigarette Consumption, and Smoking Intensity: Comment." American Economic Review, 102(4): 1751-63.

Adda, Jérôme, and Francesca Cornaglia. 2006. "Taxes, Cigarette Consumption, and Smoking Intensity." American Economic Review, 96(4): 1013-1028.

Adda, Jérôme and Francesca Cornaglia, 2013. "Taxes, Cigarette Consumption, and Smoking Intensity: Reply," American Economic Review, American Economic Association, vol. 103(7), pages 3102-14, December.

Ayyagari, P., P. Deb, J. Fletcher, W. Gallo and J.L.Sindelar (2013). Understanding Heterogeneity in Price Elasticities in the Demand for Alcohol for Older Individuals, Health Economics. 22: 89-105.

Belsky, Daniel W., et al. "Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study." JAMA psychiatry 70.5 (2013): 534-542.

Benjamin, D. J. et al. The promises and pitfalls of genoeconomics. Annu. Rev. Econom. 4, 627–662 (2012).

Boardman, Jason D. "State-level moderation of genetic tendencies to smoke."American Journal of Public Health 99.3 (2009): 480.

Bulik-Sullivan, B. et al. An Atlas of Genetic Correlations across Human Diseases and Traits. bioRxiv (Cold Spring Harbor Labs Journals, 2015).

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Robust inference with multiway clustering." Journal of Business Economic Statistics 29.2 (2011).

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., ... Poulton, R. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science, 301(5631), 386-389.

Cesarini D, Dawes CT, Johannesson M, Lichtenstein P, Wallace B. 2009. Genetic Variation in Preferences for Giving and Risk-Taking. Quarterly Journal of Economics. 124(2): 809-842.

Cesarini D, Johannesson M, Lichtenstein P, Sandewall O, Wallace B. 2010. Genetic Variation in Financial Decision Making. Journal of Finance. 65(5): 1725-1754.

Cesarini D, M Johannesson, PKE Magnusson, B Wallace. 2012. The Behavioral Genetics of Behavioral Anomalies. Management Science, 58: 21–34.

Chaloupka, Frank J., and Rosalie Liccardo Pacula. "Sex and race differences in young people's responsiveness to price and tobacco control policies." Tobacco Control 8.4 (1999): 373-377.

Chaloupka FJ, Warner KE. The economics of smoking. In: Culyer AJ, Newhouse JP, eds. Handbook of health economics. Amsterdam: North-Holland, 1539-1627, 2000.

Chetty, Raj, and Adam Looney and Kory Kroft. Salience and Taxation: Theory and Evidence. American Economic Review 99(4): 1145-1177, Sep. 2009.

Dave, D. and Saffer, H. (2008). Alcohol demand and risk preference. Journal of Economic Psychology, 29, 810-31.

DeCicca, Philip, Donald Kenkel, and Alan Mathios. "Racial difference in the determinants of smoking onset." Journal of Risk and Uncertainty 21.2-3 (2000): 311-340.

DeCicca, Philip and Logan McLeod (2008). "Cigarette taxes and older adult smoking: Evidence from recent large tax increases", Journal of Health Economics, 27(4): 918-929.

Delnevo, C. D., et al. "Cigar use before and after a cigarette excise tax increase in New Jersey." Addictive behaviors 29.9 (2004): 1799-1807.

Duncan, L. Keller, M. A critical review of the first 10 years of candidate gene by environment interaction research in psychiatry. Am. J. Psych. 168, 1041–1049 (2011).

Emery, Sherry, et al. "Was there significant tax evasion after the 1999 50 cent per pack

cigarette tax increase in California?." Tobacco Control 11.2 (2002): 130-134.

Evans, William N., and Matthew C. Farrelly. "The compensating behavior of smokers: taxes, tar, and nicotine." The Rand journal of economics (1998): 578-595.

Finucane, H. K. et al. Partitioning heritability by functional category using GWAS summary statistics. bioRxiv (Cold Spring Harbor Labs Journals, 2015).

Fletcher J.M. (2012) Why Have Tobacco Control Policies Stalled? Using Genetic Moderation to Examine Policy Impacts. PLoS ONE 7(12): e50576. doi:10.1371/journal.pone.0050576

Goldin, Jacob, and Tatiana Homonoff. 2013. "Smoke Gets in Your Eyes: Cigarette Tax Salience and Regressivity." American Economic Journal: Economic Policy, 5(1): 302-36.

Goolsbee, A., et al (2009). Playing With Fire: Cigarettes, Taxes and Competition From The Internet. The National Bureau on Economic Research. Retrieved from: http://www.nber.org/papers/w1561

Greenbaum, L and B Lerer. (2009). Differential contribution of genetic variation in multiple brain nicotinic cholinergic receptors to nicotine dependence: recent progress and emerging open questions. Molecular Psychiatry, 14: 912-945.

Gruber, Jonathan, and Jonathan Zinman. "Youth smoking in the United States: evidence and implications." Risky behavior among youths: An economic analysis. University of Chicago Press, 2001. 69-120.

Hällfors J, Loukola A, Pitkäniemi J, Broms U, Männistö S, Salomaa V, Heliövaara M, Lehtimäki T, Raitakari O, Madden PA, Heath AC, Montgomery GW, Martin NG, Korhonen T, Kaprio J. Scrutiny of the CHRNA5-CHRNA3-CHRNB4 smoking behavior locus reveals a novel association with alcohol use in a Finnish population based study. Int J Mol Epidemiol Genet. 2013 Jun 25;4(2):109-19.

Hewitt, J. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. Behav. Genet. 42, 1–2 (2011).

Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivières, S., Jahanshad, N., ... Sämann, P. G. (2015). Common genetic variants influence human subcortical brain

structures. Nature.

Hoft, Nicole R, Robin P Corley, Matthew B McQueen, Isabel R Schlaepfer, David Huizinga and Marissa A Ehringer. Genetic Association of the CHRNA6 and CHRNB3 Genes with Tobacco Dependence in a Nationally Representative Sample. Neuropsychopharmacology (2009) 34, 698–706.

Huang, Jidong, John Tauras, and Frank J. Chaloupka. "The impact of price and tobacco control policies on the demand for electronic nicotine delivery systems."Tobacco control 23.suppl 3 (2014): iii41-iii47.

Joslyn G1, Brush G, Robertson M, Smith TL, Kalmijn J, Schuckit M, White RL. Chromosome 15q25.1 genetic markers associated with level of response to alcohol in humans. Proc Natl Acad Sci U S A. 2008 Dec 23;105(51):20368-73. doi: 10.1073/pnas.0810970105. Epub 2008 Dec 8.

Keegan, J. (Ed.). (1989). The Times Atlas of the Second World War. Times Books.

Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832-8 (2010).

Liu, J. et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. 42, 436–440 (2010).

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... Kristiansson, K. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature, 518(7538), 197-206.

Maag, Elaine, and David Merriman. Tax Policy Responses to Revenue Shortfalls. Urban Institute. 2003.

Maclean, J. C., Kessler, A. S., Kenkel, D. S. (2015). Cigarette Taxes and Older Adult Smoking: Evidence from the Health and Retirement Study. Health economics.

Manning, WG, Blumberg L, and Moulton L. The demand for alcohol: The differential response to price. Journal of Health Economics, 14:123-148, 1995.

Mineur YS, Picciotto MR (2008). Genetics of nicotinic acetylcholine receptors: relevance to nicotine addiction. Biochem Pharmacol 75: 323–333.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98–101. doi: 10.1038/nature07331

Ohsfeldt, Robert L., and Raymond G. Boyle. "Tobacco excise taxes and rates of smokeless tobacco use in the US: an exploratory ecological analysis." Tobacco Control 3.4 (1994): 316.

Ohsfeldt, Robert L., Raymond G. Boyle, and Eli Capilouto. "Letter: Effects of tobacco excise taxes on the use of smokeless tobacco products in the USA."Health economics 6.5 (1997): 525-531.

Okbay, A. et al. Education-associated SNPs are enriched for brain function and disorders. Working paper (2015).

Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904-909.

Purcell, S.M. et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-52 (2009).

Ribisl, Kurt M., Annice E. Kim, and Rebecca S. Williams. "Web sites selling cigarettes: how many are there in the USA and what are their sales practices?."Tobacco Control 10.4 (2001): 352-359.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., ... McMahon, G. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science, 340(6139), 1467-1471.

Rosenquist, J. N. et al. Cohort of birth modifies the association between FTO genotype and BMI. Proc. Nat. Acad. Sci. U. S. A. 112, 354–359 (2015).

Saccone SF, Hinrichs AL, Saccone NL, Chase GA, Konvicka K, Madden PA et al. Cholin-

ergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Hum Mol Genet 2007; 16: 36–49.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421-427.

Slemrod, Joel and Kopczuk, Wojciech, 2002. "The optimal elasticity of taxable income," Journal of Public Economics, Elsevier, vol. 84(1), pages 91-112, April.

Strittmatter, Warren J., et al. "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease." Proceedings of the National Academy of Sciences 90.5 (1993): 1977-1981.

Tauras, John A. "Differential impact of state tobacco control policies among race and ethnic groups." Addiction 102.s2 (2007): 95-103.

Thompson, Samuel B. "Simple formulas for standard errors that cluster by both firm and time." Journal of Financial Economics 99.1 (2011): 1-10.

Thorgeirsson, T. et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nat. Genet. 42, 448–453 (2010).

Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat. Genet. 42, 441–447 (2010).

US Department of Health and Human Services. (2014). The health consequences of smoking—50 years of progress. A report of the Surgeon General.

Wagenaar AC, Salois MJ, Komro KA. Effects of beverage alcohol price and tax levels on drinking: A meta-analysis of 1003 estimates from 112 studies. Addiction. 2009;104(2):179-90.

Wasserman J, Manning WG, Newhouse JP, Winkler JD. The effects of excise taxes and regulations on cigarette smoking. Journal of Health Economics 1991;10(1):43-64.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., ... Lim, U.

(2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics, 46(11), 1173-1186.

Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, Chasman DI, Rose LM, Thorleifsson G, Steinthorsdottir V, Mägi R, et al. (2012). FTO genotype is associated with phenotypic variability of body mass index. Nature 490, 267-272.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Madden, P. A., Heath, A. C., ... Genetic Investigation of ANthropometric Traits (GIANT) Consortium. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics, 44(4), 369-375.

Young DJ, Bielinska-Kwapisz A (2003), Alcohol Consumption, Beverage Prices and Measurement Error. Journal of Studies on Alcohol, 64(2): 235-8.

Zeiger, Joanna S., Brett C. Haberstick, Isabel Schlaepfer, Allan C. Collins, Robin P. Corley, Thomas J. Crowley, John K. Hewitt, Christian J. Hopfer, Jeffrey Lessem, Matthew B. McQueen, Soo Hyun Rhee, and Marissa A. Ehringer. The neuronal nicotinic receptor subunit genes (CHRNA6 and CHRNB3) are associated with subjective responses to tobacco. Hum. Mol. Genet. (2008) 17(5): 724-734.

Zyphur M, Narayanan J, Arvey R, Alexander G. 2009. The Genetics of Economic Risk Preferences. Journal of Behavioral Decision Making. 22(4): 367-377.