

ESSAYS IN BEHAVIORAL AND PUBLIC ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Alexander Robert Rees-Jones

August 2013

© 2013 Alexander Robert Rees-Jones

ALL RIGHTS RESERVED

ESSAYS IN BEHAVIORAL AND PUBLIC ECONOMICS

Alexander Robert Rees-Jones, Ph.D.

Cornell University 2013

This dissertation presents two lines of research, each aimed at developing and assessing psychologically-motivated economics research in the realm of public policy.

In the first chapter I present a theory of tax sheltering activities motivated by prospect theory (Kahneman and Tversky, 1979), where a loss-averse citizen frames a refund as a gain and a tax payment as a loss. A unique implication of this theory is a discrete drop in the marginal benefit of tax sheltering once crossing the threshold into the gain domain. This drives excess tax sheltering among individuals owing money on tax day, and an excess mass of individuals to shelter precisely to the gain/loss threshold. I investigate these implications in 1979-1990 IRS panel of individual returns and find strong support for loss aversion. A mixture-modeling approach is developed to estimate model parameters and conduct policy simulations. Estimates suggest that psychologically-motivated framing effects can have substantial impact on tax revenue. I discuss the implications of these results for the detection and deterrence of tax evasion, the implementation of tax-incentivized public programs, and forecasting behavioral response to tax policy changes.

The second and third chapters assess current uses of happiness or subjective well-being (SWB) data in economic settings. Economists and policy makers often estimate the tradeoffs individuals accept and forecast the choices they will make. An increasingly-used approach to this exercise uses survey responses

to SWB questions as a direct measure of economists' notion of utility. The research presented here directly assesses these practices across a variety of settings. Chapter 2 reports the results of three surveys eliciting choice and SWB over alternatives in a battery of hypothetical scenarios. Chapter 3 reports the results of a field study of medical residency choice, allowing the side-by-side comparison of choice-based and SWB-based tradeoff estimates. Across these studies, we find that while choice and SWB rankings are often reasonably well aligned, systematic differences exist, and are particularly problematic for inference on marginal rates of substitution. We discuss the implications of our results for the use of SWB measures in economic applications and the comparative performance of different SWB-based approaches.

BIOGRAPHICAL SKETCH

Alexander Rees-Jones was born in New Jersey, USA in October, 1985, and relocated to Colorado at age five. He attended Cornell University for his undergraduate studies, where by chance he enrolled in a course in behavioral economics taught by Professor Ted O'Donoghue. He developed an immediate strong interest in this field, and became determined to pursue psychologically-motivated economic research as a career. After completing undergraduate majors in Economics and Mathematics, Alex continued into the economics PhD program, which he ultimately completed in 2013 under Professor O'Donoghue's supervision.

Alex resides in Cambridge, MA with his wife Elizabeth. After completing his PhD, he will work as a postdoctoral fellow at the National Bureau of Economic Research.

ACKNOWLEDGEMENTS

This dissertation was written under the supervision of Ted O'Donoghue, Dan Benjamin, and Francesca Molinari, all of whom were exceptionally generous with their time and support. Their positive impact on my research, and my life, cannot be overstated. Many other professors also contributed greatly to my experience as a student: Greg Besharov, Larry Blume, Aaron Bodoh-Creed, Ori Heffetz, and Miles Kimball are all worthy of special note. I am grateful for my fellow PhD students Kristen Cooper, Teevrat Garg, Mike Strain, and Ken Whelan; they all taught me a lot and have been great friends. I thank my parents, Rob and Trish Rees-Jones, for being supportive throughout every step of my education and my life. And I thank my wife, Liz, for everything. Meeting her was the most important thing I did in my time at Cornell.

The second and third chapter of this dissertation are the result of collaboration with Dan Benjamin, Ori Heffetz, and Miles Kimball, all of whom have my lasting respect and gratitude. I couldn't ask for better coauthors, and the amount I learned from our interactions spans far beyond the research presented here.

Research contained in this dissertation has been generously supported by the Cornell Institute for the Social Sciences, NIH/NIA grants R01-AG040787 and R01-AG020717-07 to the University of Michigan, and NIA grant T32-AG00186 to the NBER.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
1 Loss Aversion Motivates Tax Sheltering: Evidence from U.S. Tax Returns	1
1.1 Introduction	2
1.2 Theoretical framework	7
1.3 Data	15
1.4 Evidence of loss aversion	18
1.5 Alternative theories and robustness concerns	32
1.6 Structural estimation and policy experiments	37
1.7 Conclusion	45
1.8 Works cited	47
1.9 Acknowledgements	51
2 What Do You Think Would Make You Happier? What Do You Think You Would Choose?	52
2.1 Introduction	53
2.2 Survey design	59
2.2.1 Populations and studies	61
2.2.2 Scenarios	61
2.2.3 Main questions	64
2.3 Do people respond to the choice and SWB questions in the same way?	67
2.3.1 Within-subject results	67
2.3.2 Between-subjects results	73
2.3.3 Measurement error	75
2.4 Do other factors help predict choice, and by how much?	76
2.4.1 Response distributions	78
2.4.2 Explaining the variation in choice	78
2.4.3 Comparing the coefficients	82
2.4.4 Measurement error	83
2.5 Robustness	84
2.6 Heterogeneity in choice-SWB concordance	87
2.6.1 Comparing SWB measures	87
2.6.2 Comparing scenarios	88
2.6.3 Comparing respondents	92
2.7 Discussion	92
2.8 Works cited	96
2.9 Acknowledgements	100

3	Can Marginal Rates of Substitution Be Inferred From Happiness Data?	
	Evidence from Residency Choices	102
3.1	Introduction	103
3.2	Choice setting: the National Resident Matching Program (NRMP)	109
	3.2.1 Background	109
	3.2.2 Key features for our study	110
3.3	Sample and survey design	113
	3.3.1 Sample	113
	3.3.2 Survey design	115
3.4	Main analysis and results	121
	3.4.1 Single SWB measures	121
	3.4.2 Robustness	130
	3.4.3 Multi-question SWB indices	133
3.5	From slopes to orderings: predicting choice ranking from anticipated-SWB ranking	139
3.6	Concluding remarks	146
3.7	Works cited	149
3.8	Acknowledgements	153
A	Appendix to “Loss Aversion Motivates Tax Sheltering”	154
	A.1 Proofs	154
	A.2 Indirect tests of tax evasion	156
	A.3 Supplemental tables and figures	160
B	Appendix to “What Do You Think Would Make You Happier? What Do You Think You Would Choose?”	173

LIST OF TABLES

1.1	Summary statistics	16
1.2	Estimates of excess mass at zero balance due	22
1.3	Estimates of AGI shocks at zero balance due	26
1.4	Estimates of AGI shocks at zero balance due interacted with in- come source	28
1.5	Sheltering-relevant behaviors at zero balance due	31
1.6	Results of structural calculations	43
2.1	Study-specific information	62
2.2	Choice and SWB responses across studies and scenarios (within- subject data)	68
2.3	Regressions of choice on aspects of life	81
2.4	OLS regressions of choice on all aspects of life, by scenario	89
3.1	Main SWB and residency attribute survey questions	119
3.2	Rank-ordered logit estimates: choice vs. anticipated SWB	123
3.3	Tradeoff estimates: choice vs. anticipated SWB	127
3.4	Main SWB and residency attribute survey questions	136
3.5	Predicting binary choice from anticipated-SWB and attribute questions	140
A.1	Sample size across SSN codes and years	161
A.2	Structural parameter estimates	162
A.3	Structural parameter estimates (cont.)	163

LIST OF FIGURES

1.1	Implications of loss-averse utility for sheltering behavior	12
1.2	Distribution of balance due	19
1.3	Distribution of balance due by year-specific AGI quartile	21
1.4	Distribution of balance due in vicinity of zero	23
2.1	Raw response distributions (choice and aspects of life)	79
3.1	Survey response timeline	116
3.2	Distributions of variables by program rank	121
3.3	Tradeoff estimates: choice vs. anticipated SWB	129
3.4	Implications of iso-utility and iso-SWB curves for ordinal prediction	144
A.1	Graphs of fit of structural estimates by year and AGI level: 1979-1980	164
A.2	Graphs of fit of structural estimates by year and AGI level: 1981-1982	165
A.3	Graphs of fit of structural estimates by year and AGI level: 1983-1984	166
A.4	Graphs of fit of structural estimates by year and AGI level: 1985-1986	167
A.5	Graphs of fit of structural estimates by year and AGI level: 1987-1988	168
A.6	Graphs of fit of structural estimates by year and AGI level: 1989-1990	169
A.7	Charitable giving in the vicinity of zero balance due	170
A.8	Feldman and Slemrod (2007) evasion proxy values in the vicinity of zero balance due	171
A.9	Slemrod (1985) evasion proxy values in the vicinity of zero balance due	172

CHAPTER 1
LOSS AVERSION MOTIVATES TAX SHELTERING: EVIDENCE FROM
U.S. TAX RETURNS

Alex Rees-Jones

Abstract: This paper presents evidence that tax avoidance and evasion are influenced by loss aversion. I present a reference-dependent theory of individual response to taxation in which refunds are framed as gains and payments are framed as losses. The primary implication of this theory is a discrete drop in the marginal benefit of tax sheltering once crossing the threshold into the gain domain, leading to bunching in the distribution of balance due at the reference point. I investigate these implications in the 1979-1990 IRS panel of individual returns. The distribution of the balance exchanged with the IRS exhibits a point mass precisely at zero, and the distribution for those owing a tax payment is shifted in a manner consistent with higher pursuit of sheltering. This behavior is shown to be particularly prevalent among high-income filers and driven by those experiencing positive income shocks which are not mechanically withheld. Additionally, a number of proxies for the pursuit and employment of tax shelters are discontinuously prevalent precisely at the referent. A structural model of this behavior is developed to quantify the resulting excess tax sheltering, as well as the fraction of tax filers exhibiting loss-averse sheltering behavior. I discuss the application of these results to the detection and deterrence of tax evasion, as well as the forecasting of behavioral response to tax policy changes.

1.1 Introduction

Economists routinely model how individuals value money, as captured by the tradeoffs or risks they will tolerate in exchange for a marginal dollar. Our most common models offer tractable, principled, and “rational” approaches to forecasting such decisions, but abstract from many of the subtle psychological considerations which are often at work. The last 30 years have seen a flurry of interest in augmenting standard models with insights from psychology, with the most well-studied and famous example arguably being “prospect theory” of Kahneman and Tversky (1979). In a recent review of this work and the substantial literature it has inspired, Barberis (2013) notes that prospect theory and its modern variants are widely accepted as leading models of decision making in experimental settings. This positive assessment is tempered by the acknowledgment that “there are relatively few well-known and broadly accepted applications of prospect theory in economics,” which could permit the interpretation that this model may be “less relevant outside the laboratory.”¹

In this paper I present evidence that the insights of prospect theory are directly relevant to our understanding of a large-scale field setting of unambiguous economic importance: the manner in which individuals perceive and react to income taxes. The rationale for how gain/loss framing might affect a taxpayer is straightforward. Throughout the year, a taxpayer earns taxable income, takes actions which might be tax advantaged, and makes tax payments based on a forecast of the tax liability that will ultimately be owed. In preparation

¹Examples of existing field tests of loss aversion include the study of taxi-driver labor supply (Camerer, Babcock, Loewenstein, and Thaler, 1997; Farber, 2005; Farber, 2008; Crawford and Meng, 2011), housing prices (Genesove and Mayer, 1999), the putting behavior of professional golfers (Pope and Schweitzer, 2010), the effect of alternative policies to reduce shopping bag use (Homonoff, 2013), and behavior in financial markets (reviewed in Barberis, 2012).

for tax day, these activities must be precisely documented and reported to the IRS, and the “balance due”—the difference between the total taxes owed and the tax payments made—must be settled. If the balance due is positive, the tax filer must send payment to the IRS, and thus incur a loss in a very literal way. If the balance due is negative, a refund is due to the taxpayer, yielding a literal gain. A loss-averse citizen would react to this framing by experiencing a discrete increase in the marginal disutility of a dollar taxed as balance due crosses the gain/loss threshold. This disutility can impact observable behavior by influencing the taxpayer’s pursuit of income adjustments, deductions, and credits, or by influencing decisions on the amount of taxes to illegally evade. Identifying and quantifying the effect of loss aversion on these tax sheltering decisions will be the central focus of this paper.

The possibility of reference dependence affecting perceptions of the income tax has been the topic of a considerable amount of research. Several papers have presented theoretical treatments of loss-averse taxpayers, and have shown that loss aversion can rationalize a variety of features of the tax code, such as over-withholding and the rate of voluntary compliance.² The presence of this type of tax-framing effect has also seen support in a number of small-scale surveys and lab experiments.³ Despite encouraging results from this line of research, direct study of this phenomenon in the field has been limited, presumably due to data constraints and the difficulty of identification.⁴ This paper contributes

²See, for example, Elffers and Hessing (1997), Yaniv (2001), Bernasconi and Zanardi (2004), Kanbur, Pirttilä, and Tuomala (2008), and Dhami and al Nowaihi (2007, 2010).

³See, for example, Chang, Nichols, and Schultz (1987), Copeland and Cuccia (2002), Kirchler and Maciejovsky (2001), Robben et. al. (1990), Robben, Webley, Elffers, and Hessing (1990), or Schepanski and Shearer (1995). Schadewald (1989) presents experimental results where manipulations of reference points did not have significant effects.

⁴Notable related field studies include Feldman (2010), which utilized a change in tax withholding law to study the role mental accounting plays in retirement savings. Engström, Nordblom, Ohlsson, and Persson (2011) demonstrate a higher take-up rate of a specific deduction in Swedish tax code among individuals who are underwithheld and discuss the implications of

to the literature by presenting observable implications of loss aversion resulting from broadly-defined sheltering activities, which permits both the detection of this mechanism in the field and the estimation of its aggregate impact to tax revenues.

Section 1.2 presents a theoretical framework which underlies the identification strategy. Motivated by recent applications of “bunching” identification strategies in public finance,⁵ I explore the unique implications of loss aversion on the structure of the distribution of final balance due. This approach embraces the fact that the balance due reported to the IRS is not exogenously assigned; rather, it is manipulable by the taxpayer through the sheltering decisions described above. If the taxpayer’s assessment of the value of a marginal dollar is discontinuous and discretely higher in the loss domain, then the resulting distribution of manipulated balance due will itself be discontinuous. Individuals in the loss domain pursue excess tax sheltering activities relative to individuals in the gain domain. Moreover, a discrete fraction of taxpayers will be willing to shelter to precisely the gain/loss threshold, then discontinue pursuit of additional shelters due to the sudden drop in their marginal return.

Section 1.3 describes the data used to test these predictions, the IRS Statistics of Income 1979-1990 panel of individual returns. This dataset follows a panel of randomly-drawn taxpayers and contains information on most of the individual components of a tax calculation, allowing a detailed look at the sources of income and the nature of adjustments, deductions, and credits claimed. Repeated observations of the same individual over time allows direct observation of the manner in which taxpayers have modified their behavior, permitting the ob-

their results for loss aversion.

⁵For example, Saez (2010), Kleven et. al. (2011), and Chetty, Friedman, Olsen, and Pistaferri (2011).

servation of new avoidance activities and of changes in reported income. My primary sample consists of 229,116 tax returns filed by 53,177 taxpayers.

Section 1.4 turns to this data and tests the identifying distributional implications of loss aversion. As predicted by the model, significant excess mass is seen precisely at zero balance due. While present across all income quartiles, this excess mass is especially pronounced among high-income tax filers, particularly those with large income shocks to non-mechanically-withheld income sources. A variety of behaviors indicative of excess tax sheltering are associated with this bunching behavior. Taken together, these results are exactly in line with the theoretical predictions of the loss-averse model, and are difficult to rationalize with a non-reference-dependent baseline.

Section 1.5 considers a variety of alternative theories and possible confounds, and demonstrates their inability to explain the patterns observed in the data. In particular, this behavior can not be plausibly generated from fixed costs of being in the loss domain,⁶ avoidance of underwithholding penalties, liquidity constraints, discontinuities in the tax schedule, or hyperaccurate tax forecasting. Additional theoretical concerns, such as the interaction of loss aversion with paid tax preparers or the choice of the reference point, are discussed and ruled out as confounding factors.

Section 1.6 develops and estimates a structural model to quantify the implied excess sheltering. The distribution of balance due is modeled as a mixture of the distributions generated from loss-averse and “standard” types, with heterogeneity in model parameters and distributions permitted as a function of observables. These estimates suggest that approximately 29% of tax filers be-

⁶Examples include a belief in a jump in audit probability or the annoyance of having to write a check.

have in a manner consistent with loss averse sheltering, with this behavior again shown to be driven by high-income filers. Estimates of the total excess sheltering motivated by loss aversion amount to approximately 278 million dollars per year, expressed in 2013 dollars. Policy experiments on the revenue effects of changes in withholding laws are conducted, and shown to be potentially highly cost-effective means of revenue generation.

Section 1.7 concludes.

The analysis and results contained in this paper contribute to the literature in three primary ways. First, this paper constitutes one of the largest field tests of loss aversion to date, both in terms of the number of individuals affected and the magnitude of the economic consequences. Second, this paper informs a puzzle in public finance, seen by contrasting Chetty, Friedman, Olsen, and Pistaferri (2011) and Saez (2010): why don't middle- and high-income American taxpayers bunch at kinks in the tax schedule, as would be predicted by standard theory, and as has been seen in other countries? While many factors are likely involved, the analysis presented here suggests that psychological gain/loss framing, combined with the wide availability of tax shelters, motivates some taxpayers to bunch along an alternative dimension. American taxpayers *do* bunch at kink points, but the kink generating the most dramatic response is psychologically motivated. Finally, this framework enables the estimation of the aggregate impact of loss-averse tax sheltering, allowing a precise study of the resulting revenue costs and policy implications which was unavailable in the past literature.

1.2 Theoretical framework

In this section, I present a theoretical framework for understanding the role of loss aversion in tax sheltering decisions, and formally characterize the resulting distributional features which allow the detection of this psychological mechanism. In particular, I develop a model of the decisions of citizens who are in the process of filing their annual tax returns.

In preparation for tax day, taxpayers must complete and submit form 1040 or one of its variants, formally documenting their tax-relevant information for the year.⁷ Completing this form involves identifying oneself, documenting all taxable income, claiming credits or deductions to that taxable income due to participation in tax-incentivized behaviors, calculating the total taxes owed, and finally comparing these taxes owed to taxes already paid.⁸ These components are summed to the “balance due,” the amount of money that must be exchanged between the taxpayer and the IRS.

In the processes of filing form 1040, taxpayers have the opportunity to manipulate their balance due through legal or illegal tax sheltering. Legal means of tax sheltering typically entail reporting some tax-advantaged behavior which entitles the taxpayer to a reduction in tax or taxable income.⁹ Illegal channels for tax sheltering take the form of underreporting taxable income or overreporting

⁷Individuals with particularly simple taxable behavior may fill out shortened and simplified versions of this form, 1040A or 1040EZ.

⁸The US tax system is “pay as you go.” As citizens earn income throughout the year, they should periodically be paying portions of their tax for the year to the government, either through direct withholdings or estimated tax payments. To incentivize these payments, sufficient underwithholding results in a penalty. However, forecasting tax liability to-the-dollar is often a difficult task due to the uncertainty in most individuals’ future income and behaviors. As a result, such a level of accuracy is not expected, required, or incentivized.

⁹Concrete examples of legal means to reduce taxes include reporting sales of losing investments to realize capital losses, donating money to charity, investing money in a tax-preferred savings account, or writing off business expenses.

of tax-advantaged behaviors.

Finding and employing tax shelters is costly. For cases of legal tax avoidance, these costs include the effort necessary to find tax benefits for which the taxpayer qualifies, as well as the time and effort needed to document and claim those tax benefits. For the case of illegal evasion, these costs can include literal accounting effort as well as the expected future penalties that will be incurred if the evasion is detected. Any psychological stigma costs incurred by either evasion or avoidance can similarly be incorporated.

This framework introduces a tradeoff between the value of reducing tax payments and the cost which must be incurred to do so. Sheltering decisions made in light of this tradeoff can be represented by the utility maximization problem:

$$\max_{s \in \mathbb{R}^+} u(b^{PM}, s, \theta) = \underbrace{m(b^{PM} - s|\theta)}_{\text{utility over money}} - \underbrace{c(s)}_{\text{cost of sheltering}} \quad (1.1)$$

In the equation above, b^{PM} denotes the “pre-manipulation” value of balance due, and is modelled as a realization of a continuous random variable. s denotes the additional sheltering pursued in an attempt to manipulate balance due. θ captures additional included variables, such as wealth in the final-wealth-dependent case and the reference point in the loss-averse case. $m(\cdot)$ denotes the utility from money, while $c(\cdot)$ denotes the disutility generated from the costly pursuit of sheltering. Assume that $c(\cdot)$ is increasing and twice continuously differentiable. Further assume that $c(\cdot)$ is convex, which equates to assuming that the taxpayer pursues shelters with the lowest marginal cost first. The actual balance due reported to the government is the final, post-manipulation amount $b = b^{PM} - s$. If this value is positive, the IRS is owed money; if this value is negative, a refund is due to the taxpayer.

Below we will compare alternative sets of assumptions on the structure of utility over money in this model, and consider the implications of these structures on observed balance due.

To model a non-reference-dependent “baseline” case, assume that utility over money depends on weakly-concave, smooth preferences over final wealth: $m(b^{PM} - s|\theta) = m^{FWD}(w - b^{PM} + s)$. The primary distributional implication of this model is the continuity in balance due which it generates, expressed formally in proposition 1 below.

Proposition 1. *In the final-wealth-dependent sheltering model, if $m^{FWD}(\cdot)$ is twice continuously differentiable and the PDF of b^{PM} is continuous, then the PDF of $b = b^{PM} - s$ is continuous.*

Proof. See appendix A.1.

Put simply, smooth preferences combined with a smoothly-distributed pre-manipulation balance due generate a smoothly-distributed final balance due. As an illustrative example, consider the case where $m(\cdot)$ is linear with slope β , thus abstracting from the diminishing marginal utility of wealth. Assuming that some positive amount of sheltering is desirable,¹⁰ optimal sheltering would be determined by $s^* = c'^{-1}(\beta)$. $c'^{-1}(\cdot)$ denotes the inverse function of the derivative of $c(\cdot)$, which is guaranteed to exist and be increasing due to the assumed monotonicity and convexity of $c(\cdot)$. The final balance due is $b = b^{PM} - c'^{-1}(\beta)$, and the distribution of b corresponds to the distribution of b^{PM} shifted downward by the constant value $c'^{-1}(\beta)$, thus preserving the continuity of the distribution.

Now consider instead a loss-averse taxpayer, who has $m(b^{PM} - s|\theta) = -b^{PM} +$

¹⁰Formally, assuming that $c'(0) < m'(w - b^{PM}) = \beta$.

$s + \phi(-b^{PM} + s - r)$. The sum $-b^{PM} + s$ reflects the value of an avoided dollar of taxes in the same manner as in the linear case above, while the $\phi(\cdot)$ component allows for the influence of reference dependence. To capture loss aversion, ϕ is specified according to a piecewise-linear version of the prospect theory value function (Kahneman and Tversky, 1979):

$$\phi(x) = \begin{cases} \eta x & \text{if } x \geq 0 \\ \eta \lambda x & \text{if } x < 0 \end{cases} \quad (1.2)$$

λ is the coefficient of loss aversion, assumed to be greater than 1. η captures the weight on the loss-averse utility component relative to the direct utility component. r is assumed to be an exogenously determined reference value of balance due. In the sections to follow we will assume that the reference point is zero balance due. This decision is motivated by the intuitive appeal of that referent in this context, as well as the supporting results of experimental studies mentioned in the introduction. Furthermore, Carroll (1992) presents data from journals of tax-related thoughts in which subjects directly report considering taxes in terms of out-of-pocket gains or losses at the time of filing. Alternative theories of reference points, and their implications for the results of this paper, will be discussed in section 1.5.

To rule out corner solutions where sheltering is not pursued, assume $c'(0) < 1 + \eta$. The implied optimal sheltering behavior resulting from this model is expressed by the piecewise solution:

$$s^*(b^{PM}) = \begin{cases} c'^{-1}(1 + \eta \lambda) & \text{if } b^{PM} > c'^{-1}(1 + \eta \lambda) - r \\ b^{PM} & \text{if } b^{PM} \in [c'^{-1}(1 + \eta) - r, c'^{-1}(1 + \eta \lambda) - r] \\ c'^{-1}(1 + \eta) & \text{if } b^{PM} < c'^{-1}(1 + \eta) - r \end{cases} \quad (1.3)$$

In words, a sufficiently large pre-manipulation balance due results in high level of sheltering, $c'^{-1}(1 + \eta \lambda)$. A sufficiently small pre-manipulation balance due

results in low level of sheltering, $c'^{-1}(1 + \eta)$. For an intermediate range of pre-manipulation balance due, the level of sheltering chosen will be exactly the amount necessary to offset the pre-manipulation tax bill and reach the gain/loss threshold.

The fundamental difficulty of identifying this behavior is that pre-manipulation balance due is unobservable. To proceed, these results must be translated into distributional implications for the observed balance due, after the additional sheltering has taken place. Denoting the PDF of pre-manipulation balance due with f_b^{PM} and the distribution of final reported balance due as f_b , equation 1.3 implies that:

$$f_b(x) = \begin{cases} f_b^{PM}(x + c'^{-1}(1 + \eta)) & \text{if } x < r \\ F_b^{PM}(c'^{-1}(1 + \eta\lambda)) - F_b^{PM}(c'^{-1}(1 + \eta)) & \text{if } x = r \\ f_b^{PM}(x + c'^{-1}(1 + \eta\lambda)) & \text{if } x > r \end{cases} \quad (1.4)$$

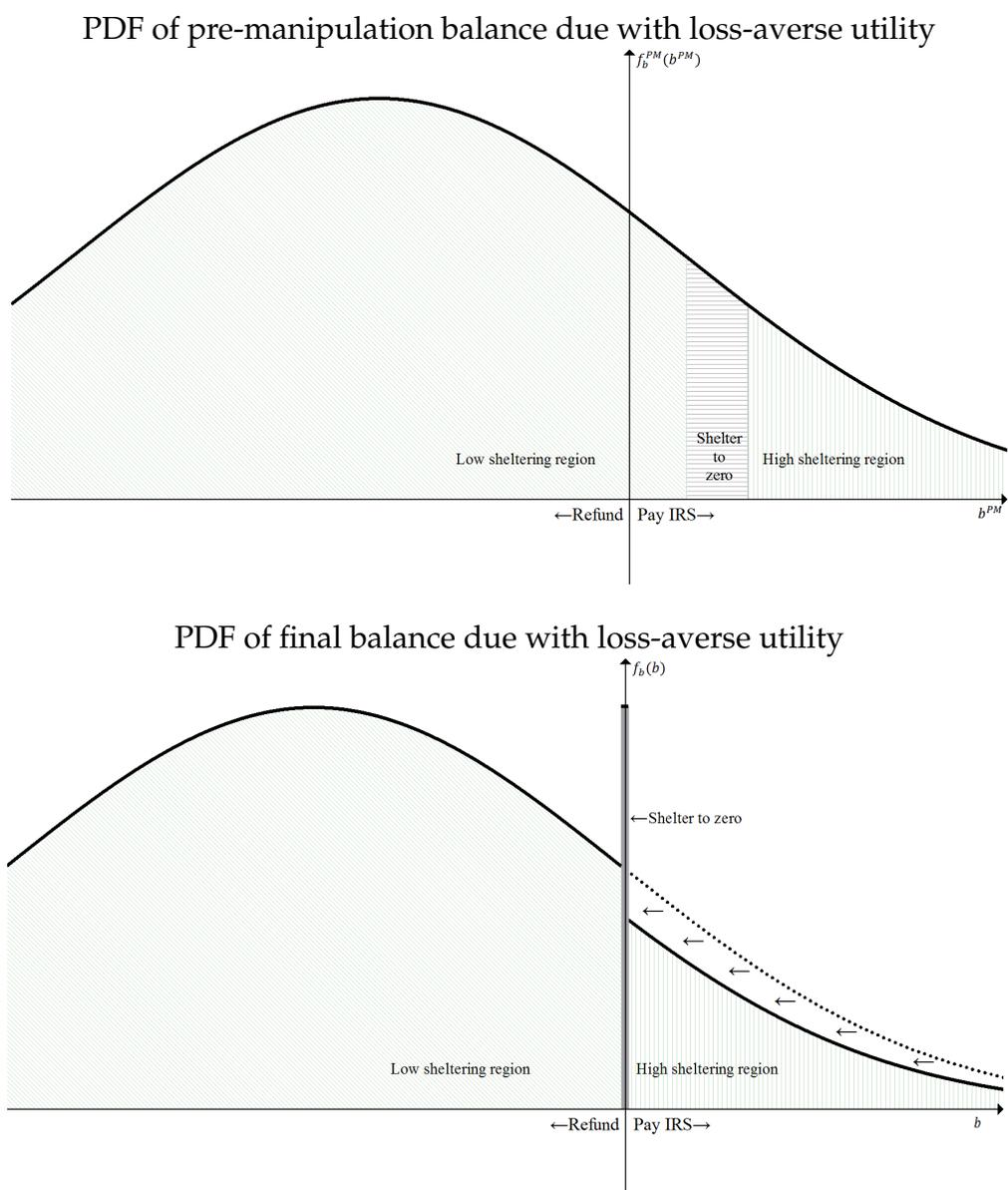
A graphical representation of this solution, and the relationship between the pre- and post-manipulation distributions, is presented in figure 1.1. The qualitative distributional implications which contrast with the final-wealth-dependent case are summarized in propositions 2 and 3 below.

Proposition 2. *Consider the loss-averse sheltering model. If r is in the support of the balance due distribution, then the PDF of balance due is discontinuous at r due to the point mass of individuals who shelter to reach that point.*

Proposition 3. *Consider the loss-averse sheltering model and assume r is in the support of the balance due distribution. For all $x \neq r$, $f_b(x)$ can be expressed as a horizontal shift of $f_b^{PM}(x)$, with a greater amount of shifting present when the balance due implies a loss ($x > r$).*

Both propositions follow immediately from the implied balance due distri-

Figure 1.1: Implications of loss-averse utility for sheltering behavior



Notes: Implications of loss-averse tax sheltering for the distribution of balance due. The first panel presents a hypothetical distribution of “pre-manipulation” balance due and indicates the optimal sheltering behavior from equation 1.3. The second panel indicates the final balance due that would be observed after loss-averse sheltering. The entire distribution is shifted to the left, with a fixed, larger shift for taxpayers with positive balance due. The darkly shaded region of taxpayers all shelter until reaching zero balance due, leading to a point mass in the observed distribution.

bution in equation 1.4. Proposition 3 is expressed relative to the unobserved distribution f_b^{PM} . However, additional assumptions on the structure of f_b^{PM} yield additional testable predictions. For example, if f_b^{PM} is single peaked, and if that peak p maps to the gain domain (consistent with the general phenomenon of overwithholding), then the implied shifting will result in an appearance of too little mass across the loss domain.

The propositions thus far have characterized the distribution of balance due induced by “standard” versus loss-averse agents. In reality, if loss aversion is present, it is likely that this psychological concern would only influence some fraction of tax filers. By assuming a mixture of these two types, additional testable hypotheses may be generated regarding the behaviors and characteristics of loss-averse shelterers. Intuitively, due to the excess accumulation of loss-averse taxpayers at the reference point, the relative frequency of this type will be especially high at this unique value of balance due. Examining the discontinuities in individual characteristics which occur at this point will be informative as to the traits of loss-averse individuals, particularly those who are able to shelter precisely to the referent.

To formalize this idea, assume that taxpayers make sheltering decisions according to equation 1.1, as above. Let fraction $p \in (0, 1)$ of taxpayers act according to an increasing, concave, and twice continuously differentiable final-wealth-dependent utility function $m^{FWD}(w - b^{PM} + s|\theta^{FWD}) - c(s|\theta^{FWD})$. Fraction $(1-p)$ act according to the loss-averse utility function $-b^{PM} + s + \phi(-b^{PM} + s - r|\theta^{LA}) - c(s|\theta^{LA})$. As above, assume $c(\cdot)$ is increasing, convex, and twice-continuously differentiable. Let (w, θ^{FWD}) and $(\eta, \lambda, \theta^{LA})$ be vectors of heterogeneous model parameters for each type. Assume that $c'(0|\theta^{LA}) < 1 + \eta\lambda$ for all $(\eta, \lambda, \theta^{LA})$, ruling

out the possibility of the degenerate $s^* = 0$ sheltering solution for loss-averse types. Regardless of their motivations, the behavior of individuals for whom $c'(0|\theta^{LA}) \geq 1 + \eta\lambda$ does not depend on positioning relative to a referent, and can be rationalized by a non-reference-dependent utility. As a result, these individuals are best grouped with the final-wealth-dependent types.

Refer to this joint set of assumptions as the “mixed-type sheltering model.”

Proposition 4. *Consider the mixed-type sheltering model. If r is in the support of the balance due distribution for the loss-averse type, then there exists a threshold value c such that $E[s^{*LA}(\lambda, \eta, \theta^{LA}|b^{PM} = 0)] - E[s^{*FDW}(w, \theta^{FDW}|b^{PM} = 0)] > c > 0$ implies:*

- a) $E[b^{PM}|b = r] > \lim_{b \rightarrow r^+} E[b^{PM}|b]$ and $E[s|b = r] > \lim_{b \rightarrow r^-} E[s|b]$; and
- b) $E[s|b = r] > \lim_{b \rightarrow r^+} E[s|b]$ and $E[b^{PM}|b = r] > \lim_{b \rightarrow r^-} E[b^{PM}|b]$.

Proof. See appendix A.1.

In words, proposition 4 says that if the average amount of sheltering among loss-averse types at zero balance due is sufficiently large relative to the the average amount of sheltering among final-wealth-dependent types reporting zero balance due, then the conditional expectation of both pre-manipulation balance due and sheltering should be discontinuously high at precisely the reference value. This result may be unsurprising, as it is essentially an immediate result of the excess mass of loss-averse taxpayers reporting the reference value. However, it is worthy of emphasis, because it offers additional testable predictions of this model. It is reasonable to assume that individuals pursuing loss-averse tax sheltering are pursuing a relatively high level of sheltering compared to the

general population, particularly since the population includes many individuals who pursue no sheltering at all. If that is the case, then we should expect proxies for both sheltering behavior and shocks to pre-manipulation balance due to exhibit a positive spike in prevalence at the candidate reference point. This offers an additional means of confirming the theory above, and a method for detecting the precise tax provisions used to manipulate final balance due to the referent.

This theory as a whole formalizes the notion that loss aversion will motivate individuals in the loss domain to work harder to reduce taxes, and that an excess mass of tax filers will cease their sheltering efforts at the referent due to the sudden drop in perceived payoff. Using the identifying features of such a model proposed above, the empirical results in section 1.4 will confirm these basic intuitions. Of course, decisions on how taxes are determined and how payments are made are quite complex, and this theory leaves a number of components unmodelled, such as the determination of initial withholding. Possible theoretical complications will be revisited in section 1.5, so we may jointly consider their effect on the theory and the ability of such concerns to generate the observed patterns in the data.

1.3 Data

The data considered in this study come from the the 1979-1990 IRS Statistics of Income (SOI) Panel of Individual Returns. The SOI Panel of Individual Returns is an unbalanced panel which follows a random sample of tax filers. Randomization occurred over social security numbers: five four-digit numbers were

Table 1.1: Summary statistics

	Mean	Standard Deviation	p5	p25	p50	p75	p95
Balance Due	-392	2509	-2393	-867	-377	-43	1477
Total Income	25178	21262	5022	11180	19560	32440	63110
Adjusted Gross Income	24698	20698	5000	11050	19285	31820	61747
Credits	82	496	0	0	0	26	398
Payments	3843	5029	337	1118	2458	4848	11229
Itemized Deductions	0.38						
Filed 1040	0.66						
Filed 1040A	0.26						
Filed 1040EZ	0.08						
Observations	229116						
Unique Taxpayers	53177						

Notes: The first section of the table presents the mean, standard deviation, and certain quantiles of several important variables. All values are measured in dollars. The second section of the table presents the fraction of respondents in certain categories.

drawn, and tax filers whose last four SSN digits matched one of these codes were included in the sample. Not all five codes were sampled in all years; appendix table A.1 illustrates the sampling pattern over time. These data contain many line items from the 1040 tax form and the relevant supplemental schedules, allowing direct observation of balance due and many steps of its calculation.

In the process of preparing the dataset, I exclude data according to several criteria. First, I restrict my sample to taxpayers in one of the 50 states or the District of Columbia. Second, I remove a small number of observations which were sampled under a different, stratified sampling regime.¹¹ Finally, I drop any data for filing years before 1979.¹² These exclusions remove 3,051 observations from the raw data, and yield a sample size of 291,275 person-years for 64,027 tax filers.

For most of the analyses surrounding loss aversion, I will further restrict the data to only individuals with non-zero tax liability (before the application of tax credits) as well as non-zero tax prepayments. This restriction excludes 62,159 observations from the data. The theory above assumes some degree of taxable activity has been pursued, and that some tax prepayments have taken place. The inclusion of individuals for whom this does not hold is problematic. For example, individuals not in the work force do not fit well with the theory laid out

¹¹As mentioned above, this sample was generated by randomization based on social security numbers, but not all numbers were sampled in all years of the panel. In years where a given social security number was excluded from this panel, it was not excluded from other IRS sampling frames. As a result, a small number of those taxpayers were randomly sampled to be part of that year's IRS tax model file. These observations were subsequently included in this panel, but flagged. Since these employed a different, stratified sampling technique, I exclude them to preserve the sampling structure of the primary panel.

¹²Tardy tax returns filed during the sampling period appear in the raw data, and thus a small number of tax returns for years before the sample began appear in the raw data. These observations are excluded.

in section 1.2, and will have balance due in the immediate region around zero for mechanical reasons. I refer to the dataset with these individuals excluded as the “primary sample,” in contrast to the full sample above. This sample consists of 229,116 tax returns filed by 53,177 taxpayers, and basic summary statistics are presented in table 1.1.

1.4 Evidence of loss aversion

In this section, I will test the predictions of the loss-averse sheltering model presented in section 1.2.

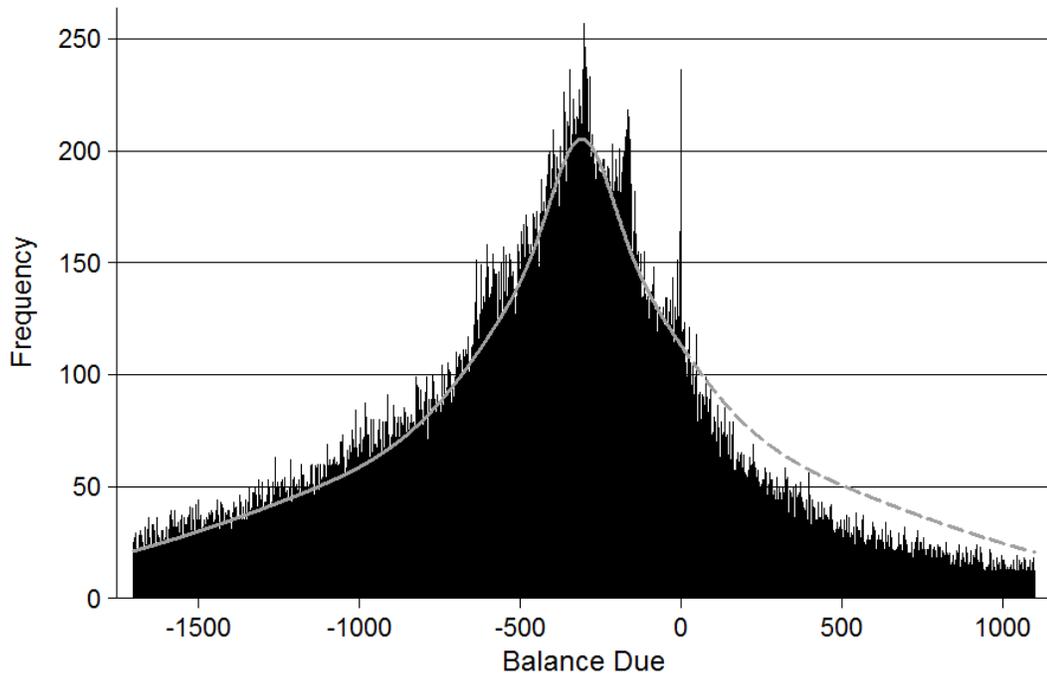
To begin, we will first examine the raw distribution of balance due among individuals in the primary sample. Figure 1.2 presents a frequency histogram with a bin size of \$1. As is visually apparent, this distribution is reasonably smooth and bell-shaped, although clearly with higher kurtosis than a standard normal distribution. Consistent with proposition 2, a sharp point mass is seen precisely at zero, and consistent with proposition 3 the distribution of balance due exhibits an apparant shifting for those with positive balance due.

To help formalize these notions, I fit a parametric distribution to the histogram of negative values of balance due (the gain domain). I then extrapolate predicted frequencies into the region of positive balance due (the loss domain). Specifically, I model the conditional distribution of negative balance due as

$$f(b|b < 0) = \frac{\sum_{i=1}^3 p_i \phi\left(\frac{b-\mu}{\sigma_i}\right)}{\sum_{i=1}^3 p_i \left(\Phi\left(\frac{-\mu}{\sigma_i}\right) - \Phi\left(\frac{b-\mu}{\sigma_i}\right)\right)} \quad (1.5)$$

This equation defines a three-component mixture of normal distributions, with the normal PDF and CDF denoted with ϕ and Φ , respectively. While a single

Figure 1.2: Distribution of balance due



Notes: Histogram of balance due in \$1 bins. The graph is centered on -300 with range restricted to $[-1700, 1100]$. For details of the calculation of the fitted distribution, see equation 1.5 in section 1.4.

normal distribution exhibits somewhat poor fit due to the high kurtosis discussed above, a mixture of several normals closely fits the sharp peak observed. A common mean is assumed to preserve symmetry. p_i denotes the mixing probabilities. The denominator ensures that this conditional distribution integrates to 1 on its restricted range.¹³ I estimate parameters for this model via maximum likelihood and overlay the predicted frequencies in figure 1.2. As is visually clear, the frequency distribution of balance due exhibits substantial excess mass precisely at zero, and frequencies in the loss domain are substantially lower than would be forecasted from the remainder of the distribution.

¹³ \underline{b} is the lowest value of balance due considered, set to -1700 in figure 1.2.

Due to difficulties about to be discussed, this fitted distribution is not meant to be interpreted as a structural estimate of the true counterfactual distribution without loss aversion. Rather, it is intended as a rough guide to help demonstrate the unique regime change occurring precisely at zero balance due.

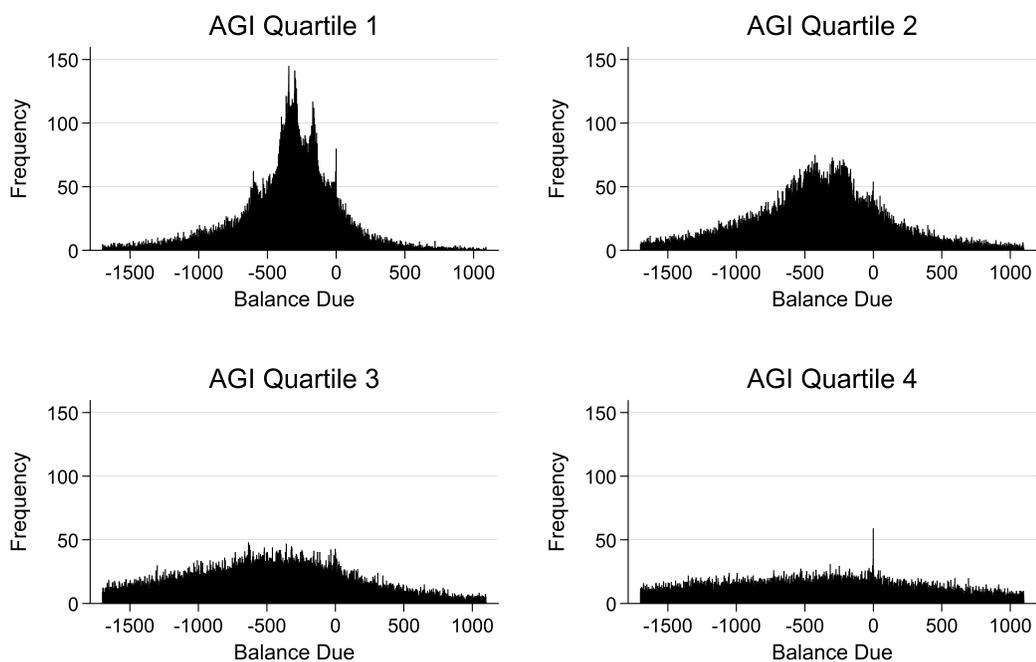
A possible concern with the interpretation of figure 1.2 is that it ignores cross-group heterogeneity. Such heterogeneity is clearly an important determinant of this distribution; for example, differences are expected between high-income and low-income individuals since they have different income sources and available tax shelters, and since greater earnings allow for greater under- or over- withholding. Figure 1.3 plots the histogram of balance due separating observations into year-specific adjusted gross income quartiles. Clear differences can be seen in the distributions: while it is generally quite smooth and dispersed for high AGI individuals, it is more irregular and concentrated for those with low AGI. However, across these quite different distributions, the predictions of propositions 2 and 3 still hold.

An additional concern with the interpretation of figure 1.2 is that the fitted distribution relies on an assumption of symmetry. However, the detection of missing mass in the loss domain does not rely on this assumption. To illustrate, figure 1.4 presents a histogram of balance due “zoomed in” on zero, restricting the range to $[-100, 100]$. To estimate a fitted distribution, I proceed in a manner similar to that in Chetty, Friedman, Olsen, and Pistaferri (2011) and model balance due as a seventh-order polynomial, allowing for a discontinuity at zero and “dummying out” the spike itself. Formally, I estimate

$$C_j = \alpha + \sum_{i=1}^n \beta_i \cdot b^i + \gamma \cdot I(b_j = 0) + \delta \cdot I(b_j > 0) + \epsilon_j \quad (1.6)$$

j indexes each dollar amount of balance due b from -100 to 100, with corre-

Figure 1.3: Distribution of balance due by year-specific AGI quartile



Notes: Histogram of balance due in \$1 bins, graphed separately for different year-specific AGI quartiles. Each graph is centered on -300 with range restricted to $[-1700, 1100]$.

sponding counts C_j . The polynomial approximates the smooth distribution of b , although it allows for a discontinuity at zero through the inclusion of δ . These estimates are reported in column 1 of table 1.2, and the predicted distribution is graphed over the histogram in figure 1.4. The negative and statistically significant estimate of δ indicates a downward shift in the distribution when crossing the gain/loss threshold.

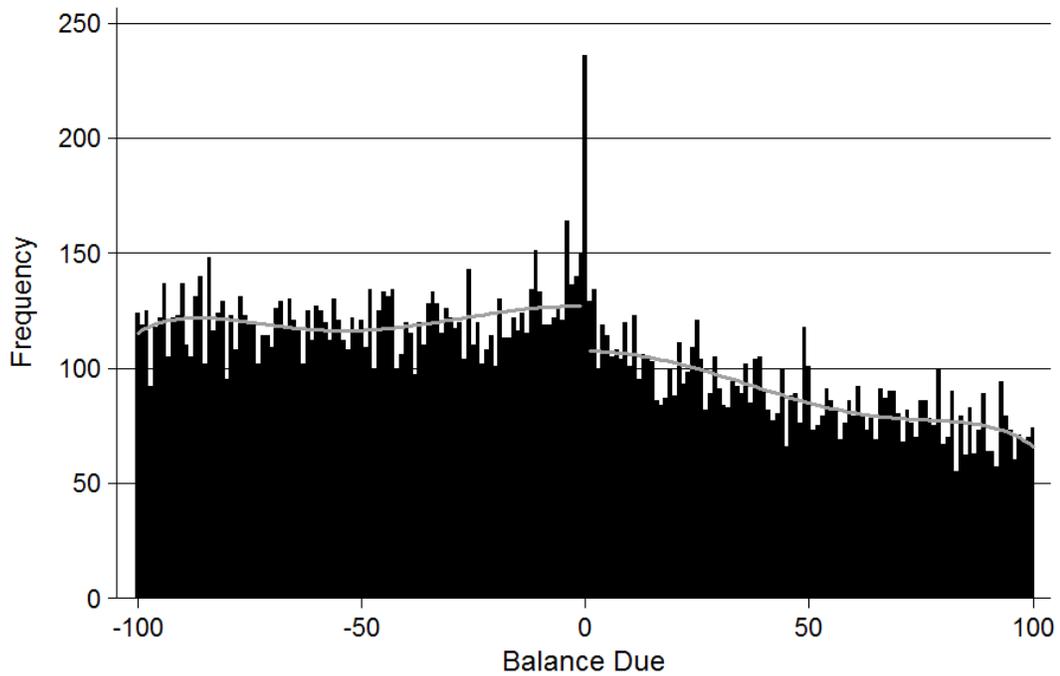
This framework additionally allows for quantification of the excess mass seen at zero balance due. The ratio $\frac{\gamma}{\alpha}$ measures this mass at zero in percentage terms, relative to the mass predicted from the distribution in the gain domain. To clarify the interpretation of this ratio, note that an estimate of 0 would im-

Table 1.2: Estimates of excess mass at zero balance due

	(1)	(2)	(3)	(4)	(5)
	All AGI groups				
	1st AGI Quartile	2nd AGI Quartile	3rd AGI Quartile	4th AGI Quartile	4th AGI Quartile
$\gamma: I(\text{balance due} = 0)$	109.21*** (12.48)	37.73*** (6.68)	19.64** (6.06)	14.36** (5.22)	37.48*** (4.62)
$\delta: I(\text{balance due} > 0)$	-19.20** (6.36)	-5.73 (3.41)	-2.78 (3.09)	-4.92 (2.66)	-5.78* (2.35)
$\alpha: \text{Constant}$	126.79*** (3.69)	42.27*** (1.97)	34.36*** (1.79)	28.64*** (1.54)	21.52*** (1.36)
Balance due polynomial	X	X	X	X	X
$\frac{\gamma}{\alpha}$: Excess mass estimate	0.86*** (0.11)	0.89*** (0.17)	0.57** (0.19)	0.50** (0.19)	1.74*** (0.27)
N	201	201	201	201	201
R^2	0.76	0.71	0.44	0.33	0.43

Notes: Standard errors in parentheses. Standard errors for excess mass estimates calculated with the delta method. Regression sample limited to observations with balance due $\in [-100, 100]$. For details of regressions, see equation 1.6 in section 1.4. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1.4: Distribution of balance due in vicinity of zero



Notes: Histogram of balance due in \$1 bins. Range restricted to $[-100, 100]$. For details of the calculation of the fitted distribution, see equation 1.6 in section 1.4.

ply that the frequency observed at the referent is precisely that predicted by the gain-domain polynomial. An estimate of 1 would indicate 100% excess mass at this point; the difference between the actual frequency and the predicted frequency (γ) is the same as the predicted frequency itself (α). The estimates from column 1 imply that there is 86% excess mass at zero balance due (95% confidence interval: [65%, 108%]).

One might question whether this constitutes a large or small amount of excess mass. For the sake of comparison, consider the amount of excess mass generated by other features of the tax system studied with comparable bunching identification strategies. Chetty, Friedman, Olsen, and Pistaferri (2011) estimated the elasticity of labor supply based off of the excess mass generated

by a discontinuity in the marginal tax rate faced by Danish workers. In their full-sample analysis, they found an excess mass (argued to be due to behavioral response to this discontinuity) which amounted to 81% of the height of the smoothed distribution. Saez (2010) looked for bunching at a variety of discontinuities in the tax schedule, both due to jumps in tax rates as well as the two kinks in the EITC schedule. Bunching was not observed at most kinks, although excess mass was seen at the first EITC kink and the first kink in the tax schedule (where the marginal tax rate jumps from zero to positive). Overall, Saez's final conclusion is that we actually see very little bunching at all. Furthermore, unlike the bunching seen in these papers, which is diffuse around the predicted point, this bunch is a specific pointmass at the precise location theoretically predicted. Relative to other applications of bunching identification strategies, these results are notably sharp.

The remaining columns of table 1.2 repeat this exercise while restricting the data to different AGI quartiles. While a significant amount of excess mass is seen at zero across all four quartiles, it is clear that this bunching behavior is markedly more pronounced among high-AGI tax filers. The distribution of the top year-specific AGI quartile exhibits 174% excess mass at zero (95% confidence interval: [121%, 227%]), substantially higher than the excess mass in the first three quartiles (89%, 57%, and 50% respectively). By and large, this phenomenon appears to be primarily driven by high-income individuals.

Having confirmed the basic distributional predictions of loss-aversion, we now turn to the prediction implied by proposition 4. If loss-averse shelterers pursue a relatively high level of sheltering relative to the general population, we should expect individuals reporting zero balance due to have unusually

high values of pre-manipulation balance due and unusually high pursuit of tax shelters. While neither of these variables are formally observed, proxies associated with their levels may be constructed through the examination of different components of the total balance due calculation.

Inference on the pre-manipulation value of balance due may be conducted by examining individual-level income shocks. Tax withholdings are determined based on estimates of your taxable income for the year. To the extent that these estimates were forecasted from prior years' taxable income, individuals with stable income streams are insulated from large shocks to balance due. In contrast, individuals experiencing a large and positive shock to taxable income will face a relatively large balance due unless their income shock was fully anticipated. Such individuals are likely to face a loss.

To explore the income shocks experienced by individuals at zero balance due, I make use of the panel nature of these data. In tables 1.3 and 1.4 I report estimates from models of the form

$$\Delta AGI = \alpha + \beta I(b = 0) + \mathbf{CF} + \epsilon \quad (1.7)$$

where \mathbf{CF} represents the relevant included controls. In all such regressions, standard errors are clustered by taxpayer. If the individuals at zero balance due are, on average, recipients of unduly-large income shocks, we would expect the coefficient on $I(b = 0)$ to be positive.

Table 1.3: Estimates of AGI shocks at zero balance due

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable : Δ AGI						
Balance due = 0	3766** (1373)	4133** (1375)	5460*** (1451)	3752* (1696)	4922** (1717)	4971** (1643)
Balance due > 0		-16 (102)	693*** (99)		624*** (130)	1213*** (114)
Filing-year fixed effects	X	X	X	X	X	X
Balance due polynomial		X	X		X	X
Lagged AGI polynomial			X			X
Taxpayer fixed effects				X	X	X
<i>N</i>	148325	148325	148325	148325	148325	148325
<i>R</i> ²	0.00	0.15	0.22	0.00	0.22	0.41

Notes: Standard errors, clustered by taxpayer, in parentheses. Xs indicate the presence of filing-year or taxpayer fixed effects, a third-order polynomial in lagged AGI, or a third-order polynomial in balance due interacted with $I(\text{balance due} > 0)$ to allow for discontinuity at zero. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

The first column of table 1.3 indicates that relative to all other taxpayers, individuals precisely at zero balance due report an additional \$3,766 of income growth, on average. In spirit similar to the excess mass estimates of the previous section, column 2 estimates a model where ΔAGI is a smooth function of balance due but allowing for a discontinuity at zero. The third column additionally allows the amount of AGI growth to depend on a third-order polynomial in last-year's AGI, which can capture the notion that a larger absolute amount of year-to-year income growth is expected among higher-income individuals. The fourth through sixth columns repeat these exercises with the inclusion of taxpayer-specific fixed effects. Across these regressions the estimated excess amount ranges from \$3,752 to \$5,460. It is worthy of note that despite the large overall sample size, these estimates are still reasonably imprecise, as they are identified solely from the individuals precisely at the referent. While this imprecision means these reported magnitudes come with a fair degree of uncertainty, it is clear that all estimates are positive in a statistically significant manner and their confidence intervals suggest an economically significant magnitude, giving strong evidence of relatively large income shocks being experienced by these bunching individuals.

The ability of an increase in income to generate a shock to balance due depends on the source of income. For example, if a taxpayer's earnings are primarily from a fixed wage or salary, an increase to that salary can relatively easily be accounted for in the employer's withholding calculations. On the other hand, income from non-wage and non-salary sources often have more complex requirements for accurate tax withholdings. For example, a small-business owner, whose income is reported through schedule C, must forecast his earnings at the beginning of the tax year. If this filer subsequently outperforms his predicted

Table 1.4: Estimates of AGI shocks at zero balance due interacted with income source

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable : Δ AGI						
Balance due = 0	564 (650)	577 (651)	1018 (680)	113 (815)	280 (863)	774 (801)
Income from Schedule C-F	90 (52)	-1570*** (66)	-959*** (61)	252* (105)	-824*** (107)	-606*** (109)
Balance due = 0 × Income from Schedule C-F	7654* (3136)	9308** (3137)	11019*** (3260)	8229* (3654)	10652** (3649)	9596** (3511)
Balance due > 0		439*** (99)	1119*** (99)		424*** (129)	1217*** (120)
Balance due > 0 × Income from Schedule C-F		120 (124)	-311* (123)		523** (170)	75 (159)
Filing-year fixed effects	X	X	X	X	X	X
Balance due polynomial		X	X		X	X
Lagged AGI polynomial			X			X
Taxpayer fixed effects				X	X	X
<i>N</i>	148325	148325	148325	148325	148325	148325
<i>R</i> ²	0.00	0.16	0.22	0.00	0.22	0.41

Notes: Standard errors, clustered by taxpayer, in parentheses. Xs indicate the presence of filing-year or taxpayer fixed effects, a third-order polynomial in lagged AGI, or a third-order polynomial in balance due interacted with $I(\text{balance due} > 0)$ to allow for discontinuity at zero. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

earnings, that shock will carry through and result in underwithholding.

To investigate the relationship between income sources and the income shock at zero I calculate a dummy variable indicating the filing of schedule C (business income), schedule D (capital gains and losses), schedule E (royalties, partnerships, S corporations, rental real estate, etc.), or schedule F (farm income). Income from these sources is likely unpredictable at the beginning of the tax year, allowing large shocks to these income categories to translate into shocks to pre-manipulation balance due. The structure of table 1.4 exactly follows table 1.3, but includes a dummy variable indicating that the taxpayer filed any of the above-noted schedules, as well as an interaction variable between this dummy and $I(b = 0)$. Among individuals without income from these sources, no statistically significant excess income growth is detected. However, as inferred from the interaction term, individuals who file one of this group of schedules and report a balance due of zero are experiencing exceptionally high income growth, with estimates of the excess growth ranging from \$7,654 to \$11,019.

The results of tables 1.3 and 1.4 serve as a confirmation of the first prediction detailed in proposition 4. The second prediction of this proposition suggests that sheltering measures must also be discretely higher at zero balance due.

In table 1.5, I present estimates of the relationship between the employment of tax-reducing provisions and the value of balance due. After the calculation of total income on form 1040, the tax filer may subsequently reduce this value by reporting adjustments to income and by claiming deductions. After the resulting tax is calculated, it may be further reduced by claiming tax credits. For these three categories of activities, I examine if—conditional on year, balance due, and

the prior year's reported AGI—individuals reporting zero balance due are more likely to pursue a tax shelter in that category. I similarly examine if the raw amount of sheltering in that category is markedly increased. The estimates of columns 1, 3, and 5 indicate that pursuit of non-zero amounts of all three sheltering categories are higher precisely at zero, although the effect is only statistically significant for the pursuit of adjustments to income. The estimates of columns 2, 4, and 6 indicate that the raw amount of all three sheltering categories are higher precisely at zero, although the effect is statistically insignificant for the pursuit of credits. Overall, the evidence supports the prediction that the individuals bunching at zero, and the behavioral type they represent, are pursuing tax-reducing activities to a greater degree than relevant comparison groups.

Unaudited tax returns, like these, are not ideal for direct inference on the rate of illegal tax evasion. However, methods for conducting indirect inference on the degree of evasion have been suggested by Slemrod (1985) and Feldman and Slemrod (2007). While these methods are not ideally suited to this setting, for completeness I use each to generate proxies for tax evasion and estimate its prevalence in the vicinity of zero balance due. These analyses are reported in appendix section A.2, and briefly summarized here.

The approach proposed by Slemrod (1985) is based on the incentives introduced by the bin thresholds in tax tables. Since tax tables are not used by the higher-income individuals driving the primary bunching result, responsiveness to these thresholds is not observed to be discontinuously prevalent at zero balance due, constituting an (understandable) null result.

The approach proposed by Feldman and Slemrod (2007) is designed to estimate the rate of underreporting of different sources of income based on the rate

Table 1.5: Sheltering-relevant behaviors at zero balance due

	(1)	(2)	(3)	(4)	(5)	(6)
	Adjustments		Itemized Deduction		Credits	
	> 0	Amount	> 0	Amount	> 0	Amount
Balance due = 0	0.10**	516**	0.02	911*	0.02	179
	(0.03)	(172)	(0.03)	(436)	(0.03)	(144)
Balance due > 0	0.05***	138***	0.00	293***	-0.01*	27*
	(0.00)	(20)	(0.00)	(48)	(0.00)	(10)
Filing-year effects	X	X	X	X	X	X
Balance due polynomial	X	X	X	X	X	X
Lagged AGI polynomial	X	X	X	X	X	X
<i>N</i>	148325	148325	148325	148325	148325	148325
<i>R</i> ²	0.19	0.10	0.38	0.53	0.47	0.05

Notes: OLS regressions with standard errors clustered at the individual level. Xs indicate the presence of filing-year fixed effects, a third-order polynomial in lagged AGI, or a third-order polynomial in balance due interacted with $I(\text{balance due} > 0)$ to allow for discontinuity at zero. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

of charitable giving. By assuming that the income elasticity of charitable giving is invariant to income source, the rate of income underreporting can be estimated based on cross-income-category rates of charitable giving.¹⁴ I implement

¹⁴The authors assume that the rate of income reporting on schedules C-F is potentially less than 100%, whereas the rate of reporting on other incomes sources is 100% due to factors such as the effectiveness of income reports. These assumptions are supported by evidence from IRS estimates of voluntary reporting ratings, which show that the 1988 voluntary reporting percentages (VRPs) for income sources associated with the above schedules range from 18.6% for informal supplies income to 92.5% for partnership and S corporation income. This contrasts with the 99.1% VRP for wages and salaries. The primary explanation for the differences across income sources are the degree of “visibility” of the income to the IRS, and thus the scope for undetected evasion.

a variant of their estimation strategy in my data and use it to form a metric of total underreporting. This metric is shown to be sharply spiked at zero balance due, consistent with higher income underreporting at that point.

To summarize, this section has documented a clear excess mass of tax filers reporting zero balance due. This excess mass is primarily driven by high-income filers and is strongly associated with experiencing a large, positive income shock from a source of income with nonmechanical withholding. A variety of correlates and proxies for sheltering are notably spiked at zero. These results are precisely in line with the predictions of the loss-averse theory presented in section 1.2, and suggest that psychologically-motivated gain/loss framing impacts the filing of taxes.

1.5 Alternative theories and robustness concerns

In this section, I consider several candidate alternative theories and their ability to generate the behavior observed in the previous section. While previous literature and intuitive considerations bring several options to mind, ultimately they are not well suited to explaining the patterns in the data. I will additionally consider several possible complications with the theoretical model presented in section 1.2, and their relevance in explaining the highlighted results.

Fixed costs incurred in the loss domain: If a substantial fixed cost were assigned to individuals as they crossed the threshold into positive balance due, this could naturally generate bunching at zero through incentivizing additional tax sheltering. Possible sources of such a fixed cost include the annoyance of having to write out a check (which is unnecessary if balance due is non-positive), or

beliefs of a discretely different and higher audit probability among individuals with positive balance due. While these fixed costs are clearly not large in objective terms, a large subjective belief in their value might occur. Such a fixed cost would determine a threshold \bar{b} such that the individual would shelter to zero if pre-manipulation balance due fell within $[0, \bar{b}]$. Put simply, if the cost of writing a check to the IRS were such that taxpayers were willing to incur the costs associated with up to \$100 of additional sheltering, taxpayers would shelter to zero if dealt a baseline balance due between \$0 and \$100. If baseline balance due did not fall within this region, marginal incentives would not be affected, and the distribution would be no different than in the case with no fixed cost. Thus, these incentives can generate bunching at zero, but have different implications for the distribution over positive balance due amounts. This theory would suggest a region with no probability mass immediately to the right of zero, but farther in the tail the distribution would be consistent with the no-fixed-cost baseline. This pattern does not conform with the distribution observed in figure 1.2, suggesting that perceptions of a substantial fixed cost being incurred while passing zero are not driving this behavior.

Avoidance of the underwithholding penalty: The underwithholding penalty, and the discontinuity in the tax schedule it induces, can drive bunching behavior in tax sheltering activity without the presence of loss aversion. However, it is important to note that the threshold for underwithholding is not at zero (or indeed very close). Across the years in the sample, there is a grace window fixed in percentage terms (typically 10% of total tax due, a substantial amount for the high-income filers driving the observed bunching behavior), with the size of the window bounded below at a minimum of several hundred dollars. The bunching behavior we see precisely at zero is not at the threshold for the

underwithholding penalty, and cannot be rationalized in this manner.

Liquidity constraints: Taxpayers who are liquidity constrained must eliminate their balance due or face a payment they cannot make, which can naturally lead to higher rates of tax sheltering for individuals with positive pre-manipulation balance due. However, if liquidity constraints covered a range of small but non-zero asset values, it would imply a diffusion of excess mass to the right of zero in figure 1.2. Such a diffusion is clearly not present. This fact, combined with the general implausibility of the lack of literally any liquid assets among such high-income filers, suggests that liquidity cannot be the primary driver of this bunching behavior.

Bunching due to discontinuities in the tax schedule or extremely accurate forecasting: As demonstrated in Chetty, Friedman, Olsen, and Pistaferri (2011), discontinuities in the tax schedule, formed by the discrete jumps in the marginal tax rate as one crosses into a higher tax bracket, can naturally generate bunching in the distribution of total taxes owed. This bunching is naturally somewhat diffused by things such as adjustment costs, as explored in that work. Saez (2010) reports not finding evidence of this behavior for higher-income taxfilers in US data, which limits the concern that this could be a confound. However, if some population of taxfilers were attempting to bunch at these thresholds, had the ability to do so precisely, and planned their tax prepayments assuming their final tax liability would precisely land on this threshold, this would naturally generate point mass in the distribution of b^{PM} at zero. However, this would require these taxpayers to have literally zero imprecision in their forecasts of their taxable behavior for the year and in tax prepayments. Given that some degree of forecasting error is an unambiguous reality of this decision en-

vironment, we would instead expect the excess mass induced by this type of bunching to be diffuse around zero. That no such diffusion is observed in figure 1.2, combined with the results of Saez (2010), suggests that this explanation is not plausible. Furthermore, the primary drivers of this bunching behavior are individuals with non-wage non-salary income (which is inherently difficult to forecast) who are having dramatically atypical, high-income years, making to-the-dollar accuracy of forecasting implausible.

Interaction with tax preparers: A substantial fraction of taxpayers do not handle their own tax returns, but instead relegate this task to a paid tax preparer. In principle this could complicate the manner in which sheltering decisions are made. A simple reformulation of the baseline model in section 1.2, accounting for this complication, would be:

$$\max_s \underbrace{u_P(m_C(-b^{PM}) + c_C(s))}_{\text{utility of client}} - \underbrace{c_P(s)}_{\text{costs to tax preparer}} \quad (1.8)$$

The subscripts P and C refer to the tax preparer and client, respectively. In this formulation, the tax preparer balances the utility gains from customer satisfaction (capturing concerns such as the increased probability of return service and good recommendations) and the costs of pursuing additional sheltering. If u_P is an increasing, concave, twice-differentiable function, the qualitative results of section 1.2 above hold without modification—if the client has smooth utility, the distribution of balance due will be smooth, and if the client is loss averse, the distribution will bunch at zero due to higher sheltering in the loss domain.

Interestingly, the use of tax preparers is particularly prevalent among individuals reporting zero balance due, even when controlling for prior AGI. An estimate analogous to those reported in table 1.5 indicates an increase in the probability of using a paid tax preparer of 15 percentage points (standard error:

6.3 percentage points).

Tax returns filled out by a paid preparer are completed by an individual with much greater knowledge and command of the tax code, leading to a richer variety of tax shelters that could be expected to be brought to bear. Furthermore, using audited returns from 1979, Erard (1993) demonstrated that tax noncompliance is dramatically higher among individuals with CPA- or lawyer-prepared returns, as opposed to self-prepared returns.¹⁵ Under this interpretation, the use of paid tax preparers may be serving as a proxy for high pursuit of sheltering, in which case its discontinuity at zero serves as further evidence in support of the loss-averse model.

Choice of the reference point: The theoretical discussion of loss aversion has assumed a common framing of a zero balance due referent, consistent with past experimental studies and with Kahneman and Tversky's original presentation. This modeling decision seems plausible in the framing of balance due, given the literal gain or loss that will occur relative to that point.

An alternative approach prevalent in recent literature is the expectations-based reference-dependent model of Koszegi and Rabin (2006, 2007). Variants of this model have been successfully employed in several recent papers (see, e.g., Crawford and Meng (2011), Ericson and Fuster (2011), or Gill and Prowse (2012)), although it has also been challenged by difficulties in explaining the endowment effect (Heffetz and List, 2012). The approach of the papers listed above, as well as other prevalent reference-dependent work (Camerer, Babcock, Loewenstein, and Thaler, 1997; Genesove and Mayer, 2001; Kahneman, Knetsch,

¹⁵Self-prepared returns understated their income 39.2% of the time, with a mean level of non-compliance of \$244. CPA or lawyer-prepared returns understated their income 63% of the time, with a mean level of noncompliance of \$1,786.

and Thaler, 1990; Pope and Schweitzer, 2010) is to pursue a particular model of the reference point and generally “assume away” model heterogeneity in the reference point formation process. Given the sum of this literature, it seems reasonable to imagine that across different individuals and contexts, different factors might determine the reference point.

The bunching observed at zero balance due suggests that a referent of zero, in the “narrow frame” of balance due amount, is indeed considered by some positive fraction of the population; however, it is possible that some additional fraction is reference dependent relative to an alternative model not studied in this paper. The exploration of alternative reference points in this process is empirically challenging, but a potentially fruitful avenue for future research.

1.6 Structural estimation and policy experiments

Given the results of the previous sections, it appears clear that the observed excess sheltering and bunching behavior is driven by loss aversion. In this section, I will develop and estimate a structural model of this behavior to answer several remaining questions.

While the evidence of the preceding sections suggests that some individuals are loss averse, it is not informative as to the precise fraction of individuals exhibiting evidence of these psychological considerations. Knowing the fraction of individuals behaving in a manner consistent with loss aversion is of significant interest to both behavioral economists (to determine the prevalence of this widely-studied behavior) and to tax policy makers (to accurately forecast their behavior), the structural model I develop will explicitly allow a mixtures

of “standard” types and loss averse types and estimate the fraction of each, permitting inference on the number of individuals driving the observed bunching behavior.

It would additionally be desirable to explicitly estimate the preference parameters of loss-averse individuals, such as the coefficient of loss aversion. Unfortunately, these preference parameters cannot be separately identified from the parameters of the cost function, but this does not hinder the use of this model for quantifying the effects of loss aversion.¹⁶ The amount of additional sheltering pursued in the loss domain relative to the gain domain *is* identified.

I will use these estimates to conduct a counterfactual policy experiment, considering the implications of shifting the distribution of baseline balance due, as could be achieved by, e.g., more lenient withholding laws. Such a shift would move a range of individuals into the loss domain, and thus motivate additional tax sheltering for this group. Quantifying this effect is helpful in understanding the magnitude of effects this behavior would generate when manipulations to tax law are made.

The structural model I estimate is an extension of the theory laid out in section 1.2. There are two types, standard agents and loss-averse agents. As before, the distribution of balance due of loss-averse agents is given by equation 1.4, which is reproduced here after a change-in-variables helpful for understanding identification:

$$f_b^{LA}(x|\mu, \nu, f) = \begin{cases} f(x - \mu) & \text{if } x < 0 \\ F(-\mu + \nu) - F_\epsilon(-\mu) & \text{if } x = 0 \\ f(x - \mu + \nu) & \text{if } x > 0 \end{cases} \quad (1.9)$$

¹⁶The lack of separate identification can be immediately inferred from equation 1.3.

Define $\mu = E[b^{PM}] - c'^{-1}(1 + \eta)$ and define $\nu = c'^{-1}(1 + \eta\lambda) - c'^{-1}(1 + \eta)$. μ can be interpreted as the mean of a counterfactual distribution in which the taxpayer always pursues the low sheltering amount. ν measures the difference between the high and the low sheltering amount. The constant shift induced by ν is clearly identified, and its estimate permits inference on the difference in sheltering across gain or loss framing.

Denote the fraction of loss-averse taxpayers with p_{LA} . The remainder are standard agents, and do not pursue different sheltering strategies based on the realization of baseline balance due. The distribution of their balance due is thus determined as

$$f_b^{SA}(x|\mu, \psi, f) = f(x - \mu + \psi) \quad (1.10)$$

Standard agents will likely have a different baseline sheltering behavior, or potentially even a different mean level of baseline balance due (capturing differences in withholding behavior). The inclusion of the term ψ allows for these differences.

As in section 1.2, the solutions presented have assumed some positive amount of sheltering for loss-averse taxpayers, at least in the loss domain. If an individual is reference dependent, but only has access to tax sheltering technologies that are perceived to be so costly that they go unused, no asymmetry in sheltering behavior can be observed, because no sheltering behavior exists. Since these individuals' behavior can be rationalized by a non-reference-dependent utility function, they are grouped with the standard agents. Estimates of the fraction of loss-averse shelterers sheltering thus exclude individuals with loss-averse motivations but insufficient access to sheltering technologies.

The presence of these two types of taxpayers generates a mixture distribution, parameterized by p_{LA} , f , μ , ν , and ψ . Formally,

$$f_b(x|\mu, \nu, \psi, p_{LA}, f) = p_{LA} \cdot f_b^{LA}(x|\mu, \nu, f) + (1 - p_{LA}) \cdot f_b^{SA}(x|\mu, \psi, f) \quad (1.11)$$

f is of unknown form, and will itself be modeled as a mixture of normal distributions, $f(x) = \sum_{i=1}^2 q_i \phi\left(\frac{x}{\sigma_i}\right)$, in order to capture the high-kurtosis bell shape observed in figure 1.2. This introduces the additional parameters q_1 , σ_1 , and σ_2 which must be estimated.

As emphasized in section 1.4, heterogeneity in model components across both years and income levels is to be expected. Furthermore, the results of previous sections suggest that high income filers are those driving the observed loss-averse behavior, suggesting that the probability of being a loss-averse type should be allowed to vary by AGI level. To model this heterogeneity, I allow all model parameters to be year specific, and to be a function of AGI (normalized within year). For unrestricted parameters, I model this as a linear relationship. For parameters constrained to be positive, I model this as a log-linear relation-

ship. Probabilities take logit form. Formally,

$$\mu_t = c_t^\mu + \beta_t^\mu \cdot std(AGI) \quad (1.12)$$

$$\psi_t = c_t^\psi + \beta_t^\psi \cdot std(AGI) \quad (1.13)$$

$$\ln(v_t) = c_t^\nu + \beta_t^\nu \cdot std(AGI) \quad (1.14)$$

$$\ln(\sigma_{1t}) = c_t^{\sigma_1} + \beta_t^{\sigma_1} \cdot std(AGI) \quad (1.15)$$

$$\ln(\sigma_{2t}) = c_t^{\sigma_2} + \beta_t^{\sigma_2} \cdot std(AGI) \quad (1.16)$$

$$q_{1t} = \frac{e^{\theta_{1t}}}{e^{\theta_{1t}} + 1} \quad (1.17)$$

$$\theta_{1t} = c_{1t}^\theta + \beta_{1t}^\theta \cdot std(AGI) \quad (1.18)$$

$$p_{LA_t} = \frac{e^{\gamma_{1t}}}{e^{\gamma_{1t}} + 1} \quad (1.19)$$

$$\gamma_{1t} = c_{1t}^\gamma + \beta_{1t}^\gamma \cdot std(AGI) \quad (1.20)$$

I estimate the 168 parameters¹⁷ of this system of equations via maximum likelihood. Numerical optimization is conducted through alternating application of Newton-Raftson and the Berndt-Hall-Hall-Hausman algorithms. Parameter estimates are reported in appendix tables A.2 and A.3, and the predicted fit of the mixtures is graphed in appendix figures A.1 - A.6.

Examining the raw estimates, two results become immediately clear. First, consistent with prior results in this paper, loss-averse behavior is estimated to be rare among low-income filers, but is prevelant among high income filers. As reported in figures A.1 - A.6, the estimated probability of reference dependence for an individual with AGI 1 standard deviation below the year-specific mean is low, ranging from 0% to 9%. In contrast, the estimated probability of loss aversion for an individual with AGI 1 standard deviation above the year-specific mean ranges from 43% to 94%. As reported in table 1.6, the estimated fraction

¹⁷7 constants and 7 AGI coefficients for each of the 12 panel years.

of the population exhibiting loss aversion, unconditional on income and averaged across years, is 29%.

I will now turn to calculating the total excess sheltering these estimates imply, as well as the counterfactual exercise described above. These calculations require two additional sources of data. First, to make dollar amounts comparable across years, I use the consumer price index calculator of the BLS to translate year-specific dollar amounts into 2013 dollars. Additionally, calculations of total sheltering require multiplying per-capita estimates by the total number of tax filers in each year, which I obtained from estimates published by the IRS.

To calculate an estimate of the per capita additional sheltering motivated by loss aversion, I multiply the posterior probability that an individual is loss averse¹⁸ by the predicted amount of excess sheltering the model would forecast for a loss-averse type (conditional on year, AGI, and reported balance due). This excess sheltering is zero for all individuals in the gain domain, v for individuals in the loss domain, and the necessary amount to reach zero for individuals at the referent. Relevant estimates are reported in rows 2-4 of table 1.6. Overall, expressed in 2013 dollars, the estimated model implies a yearly average of 278 million dollars less tax revenue collected due to tax sheltering motivated by loss aversion.

While this raw total is interesting, it is not the ideal number for policy considerations. This number captures the total amount of additional tax revenue

¹⁸Formally, the posterior probability that an observation was generated by the reference-dependent component, conditional on the parameter estimates and the observed value of balance due x , is calculated with a simple application of Bayes rule:

$$\frac{f_b^{LA}(x|\mu_t(AGI), \nu_t(AGI), \sigma_{1t}(AGI), \sigma_{2t}(AGI), q_{1t}(AGI))}{f_b(x|\mu_t(AGI), \nu_t(AGI), \psi_t(AGI), \sigma_{1t}(AGI), \sigma_{2t}(AGI), q_{1t}(AGI), p_{LA_t}(AGI))} \quad (1.21)$$

Table 1.6: Results of structural calculations

	79	80	81	82	83	84
Average probability of loss-averse type	0.39	0.38	0.25	0.22	0.23	0.28
Mean per capita extra sheltering (in dollars)	0.34	0.033	1.48	0.77	0.34	1.21
Total extra sheltering (in millions of dollars)	31.6	3.06	141.6	73.1	32.3	120.5
Total extra sheltering (in millions of 2013 dollars)	101.5	8.62	362.4	176.2	75.5	270.0
Mean per capita effect of withholding change (in dollars)	0.11	0.012	0.12	0.17	0.31	0.26
Total effect of withholding change (in millions of dollars)	10.1	1.09	11.7	16.5	29.8	25.7
Total effect of withholding change (in millions of 2013 dollars)	32.5	3.07	30.0	39.8	69.8	57.7

	85	86	87	88	89	90	Average
Average ref. dep. probability	0.23	0.24	0.28	0.35	0.31	0.29	0.29
Mean per capita extra sheltering (in dollars)	0.48	3.27	2.36	2.01	0.90	1.89	
Total extra sheltering (in millions of dollars)	48.5	337.4	252.8	220.9	101.1	215.3	
Total extra sheltering (in millions of 2013 dollars)	104.7	715.4	518.2	435.1	190.1	383.3	278.4
Mean per capita effect of withholding change (in dollars)	0.17	0.38	0.26	0.37	0.24	0.37	
Total effect of withholding change (in millions of dollars)	17.1	39.6	28.1	40.8	26.7	41.5	
Total effect of withholding change (in millions of 2013 dollars)	37.0	84.0	57.5	80.5	50.2	73.9	51.3

Notes: Calculations from structural exercises described in section 1.6.

which could be captured if loss-averse filers could be induced to always behave as if they were in the gain domain. It is unclear what type of policy could actually accomplish this. A more practical consideration focuses on what would happen if specific steps were taken to move the gain/loss threshold. The primary policy lever available to manipulate loss-averse sheltering is changing the framing of the taxes owed at tax day. For example, changing withholding laws to yield a different average baseline balance due would, for a range of taxpayers, affect whether they are in the gain or loss domain, and thus affect their sheltering behavior.

This is directly relevant to the phenomenon of overwithholding (for related behavioral study of overwithholding, see Jones (2012)). Most tax filers get refunds, and in my full sample, the average balance due is -\$324. This overwithholding is often discussed as undesirable for taxpayers, as these excess tax prepayments constitute a (small) interest-free loan to the government. As discussed in previous literature, this overwithholding is beneficial in the presence of reference dependence, as it reduces the motives for additional tax sheltering. With the structural estimates provided, we may precisely characterize the magnitude of such an effect.

To do so, I calculate the additional loss-averse sheltering that would be expected as a result of shifting the balance due distribution by \$300 to offset overwithholding. Examining figure 1, this is conceptually recentering the distribution such that zero balance due falls closer to the peak, rather than in the right tail of the distribution. Estimates of the change in sheltering are calculated in the precise same manner as the totals previously considered, and the increases in sheltered tax revenue are reported in rows 5-7 of table 1.6. These estimates

imply that such a shift would result in 51 million dollars of new sheltering, expressed in 2013 dollars and averaged across years. These results suggest that attention to these psychological considerations while designing and implementing tax changes can have substantial effects on collected revenue, and, to the extent that such manipulations are “cheap,” may prove highly cost effective.

1.7 Conclusion

Taken together, the results of this paper suggest that loss aversion has a significant role in the economic psychology of taxation, and affects the behavior of tax filers in substantial ways. The observed sharp bunching at zero and the concurrent shifting of the balance due distribution for positive values are immediate implications of loss aversion, and the distribution they generate is reasonably implausible under existing alternative theories.

These results contribute to the growing literature demonstrating the importance of loss aversion in the field (e.g. Camerer, Babcock, Loewenstein, and Thaler, 1997; Genesove and Mayer, 1999; Farber 2005, 2008; Pope and Schweitzer, 2010; Crawford and Meng, 2011), and demonstrate the applicability of this model for improving our understanding of tax behavior. Estimates of the population rate of loss-averse sheltering suggest that this behavioral phenomenon is remarkably prevalent in population, at least for this particular setting. Furthermore, the estimates presented in section 1.6 suggest that substantial amounts of sheltering are being motivated by these simple and predictable psychological concerns.

Several immediate policy recommendations can be drawn from the results

of this study. First, this model implies that individuals reporting near-zero balance due are especially likely to be employing tax shelters. To the extent that illegal means are used to achieve this sheltering, targeted auditing of these individuals would assist in enforcing tax compliance. Second, this study suggests that the phenomenon of overwithholding is even more beneficial to the IRS than previously expected, as it ensures that a reasonably small fraction of tax filers face the additional sheltering motivations associated with underwithholding in their baseline balance due. Consideration of the psychological consequences, and the increase in sheltering they would motivate, should be taken into account when considering manipulating withholding laws. Finally, this study suggests that consideration of psychological motivations may be essential to understanding bunching in taxable behavior, a topic of substantial recent interest in labor economics and public finance. The relative lack of bunching at tax kinks seen in Saez (2010) is puzzling from the view of previous theory, but easy to explain when considering the psychological motivations present: the salient points where individuals are drawn to bunch on tax day are different from those generated from simple examination of the tax schedule. Even if individuals did attempt to bunch at these tax kink points through their labor decisions, their psychologically-motivated sheltering decisions could “smooth out” the excess mass this would generate. However, the lack of bunching at tax kink points could also be due to a variety of other considerations. Identifying the degree to which psychological considerations are responsible for these puzzling findings, and the implications of this confound for the estimation of labor supply elasticity based on administrative tax data, are fruitful avenues for future research.

1.8 Works cited

- Bakija, J. & Heim, B.** 2011. "How Does Charitable Giving Respond to Incentives and Income? New Estimates from Panel Data." *National Tax Journal*, 64(2): 615-650.
- Barberis, N.** 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment." *Journal of Economics Perspectives*, 27: 173-196.
- Bernasconi, M. & Zanardi, A.** 2004. "Tax Evasion, Tax Rates, and Reference Dependence." *FinanzArchiv*, 60: 422-445.
- Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R.** 1997. "Labor Supply of New York City Cabdrivers: One Day at a Time." *The Quarterly Journal of Economics*, 112(2), 407-41.
- Carrol, J. S.** 1992. "How Taxpayers Think About Their Taxes: Frames and Values." In *Why People Pay Taxes: Tax Compliance and Enforcement*, ed. Joel Slemrod, pp. 43-63.
- Chang, O. H., Nichols, D. R., & Schultz Jr, J. J.** 1987. "Taxpayer Attitudes Towards Tax Audit Risk." *Journal of Economic Psychology*, 8: 299-309.
- Chetty, R., Friedman, J., Olsen, T., & Pistaferri, L.** 2011. "Adjustment Costs, Firm Responses, and Micro vs Macro Labor Supply Elasticities: Evidence from Danish Tax Records" *The Quarterly Journal of Economics*, 126(2): 749-804.
- Copeland, P. & Cuccia, A.** 2002. "Multiple Determinants of Framing Referents in Tax Reporting and Compliance." *Organizational Behavior and Human Decision Processes*, 88(1): 499-526.

- Crawford, V. P. & Meng, J.** 2011. "New York City Cabdrivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income." *American Economic Review*, 101: 1912-1932.
- Dhmi, S. & al Nowaihi, A.** 2007. "Why Do People Pay Taxes? Prospect Theory versus Expected Utility Theory." *Journal of Economic Behavior & Organization*, 64(1): 171 - 192.
- Dhmi, S. & al Nowaihi, A.** 2010. "Optimal Income Taxation in the Presence of Tax Evasion: Expected Utility versus Prospect Theory." *Journal of Economic Behavior & Organization*, 75: 313-337.
- Elffers, H. & Hessing, D. J.** 1997. "Influencing the Prospects of Tax Evasion." *Journal of Economic Psychology*, 18(2-3), 289 - 304.
- Engström, P., Nordlom, K., Ohlsson, H., & Persson, A.** 2011. "Loss Evasion and Tax Aversion." Uppsala Center for Fiscal Studies, Working Paper 2011:11.
- Erard, B.** 1993. "Taxation with Representation: An Analysis of the Role of Tax Practitioners in Tax Compliance." *Journal of Public Economics*, 52(2): 163-97.
- Ericson, K. & Fuster, A.** 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments." *The Quarterly Journal of Economics*, 126(4): 1879-1907.
- Feldman, N. & Slemrod, J.** 2007. "Estimating Tax Noncompliance with Evidence from Unaudited Tax Returns" *The Economic Journal*, 117: 327-352.
- Feldman, N.** 2010. "Mental Accounting Effects of Income Tax Shifting" *Review of Economics and Statistics*, 92(1): 70-86.

- Genesove, D. & Mayer, C.** 2001. "Loss Aversion And Seller Behavior: Evidence From The Housing Market." *The Quarterly Journal of Economics*, 116(4): 1233 - 1260.
- Gill, D. & Victoria Prowse, V.** 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition." *American Economic Review*, 102(1): 468-503.
- Heffetz, O. & List, J.** 2012 "Is the Endowment Effect an Expectations Effect?" Working Paper.
- Homonoff, T.** 2013. "Can Small Incentives Have Large Effects? The Impact of Taxes versus Bonuses on Disposable Bag Use." Working paper.
- Jones, D.** 2012. "Inertia and Overwithholding: Explaining the Prevalence of Income Tax Refunds." *American Economic Journal: Economic Policy*, 4(1): 158-85
- Kahneman, D., Knetsch, J., & Thaler, R.** 1990 "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy*, 98(6): 1325-1348.
- Kahneman, D. & Tversky, A.** 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263-91.
- Kanbur, R., Pirttilä, J., & Tuomala, M.** 2008. "Moral Hazard, Income Taxation, and Prospect Theory." *The Scandinavian Journal of Economics*, 110(2): 321-337.
- Kirchler, E. & Maciejovsky, B.** 2001. "Tax Compliance Within the Context of Gain and Loss Situations, Expected and Current Asset Positions, and Profession." *The Journal of Economic Psychology*. 22: 173-194.

- Kleven, H., Knudsen, B., Kreiner, C., Pedersen, S., & Saez, E.** 2011. "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark." *Econometrica*, 79(3): 651-692.
- Koszegi, B. & Rabin, M.** 2006. "A Model of Reference-Dependent Preferences." *The Quarterly Journal of Economics*, 121(4), 1133-1166.
- Koszegi, B. & Rabin, M.** 2007. "Reference-Dependent Risk Attitudes." *American Economic Review*, 97(4): 1047-1073.
- Pope, D. & Schweitzer, M.** 2010. "Is Tiger Woods Loss Averse? Persistent Bias in the Face of Experience, Competition, and High Stakes." *American Economic Review*, 101(1): 129-157.
- Robben, H., Webley, P., Elffers, H., & Hessing, D.** 1990. "Decision Frames, Opportunity, and Tax Evasion: An Experimental Approach." *Journal of Economic Behavior and Organization*, 14: 353-361.
- Robben, H., Webley, P., Weigel, R., Wärneryd, K., Kinsey, K., Hessing, D., Martin, F. A., Elffers, H., Wahlund, R., Van Langenhove, L., Long, S., & Scholz, J.** 1990. "Decision Frames and Opportunity as Determinants of Tax Cheating." *Journal of Economic Psychology*, 11: 341-364.
- Saez, E.** 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2: 180-212.
- Schadewald, M.** 1989. "Reference Point Effects In Taxpayer Decision Making." *Journal of the American Taxation Association*. 89(10): 68-84.
- Schepanski, A. & Shearer, T.** 1995. "A Prospect Theory Account of the Income

Tax Withholding Phenomenon." *Organizational Behavior and Human Decision Processes*, 63(2): 174 - 186.

Slemrod, J. 2007. "Cheating Ourselves: The Economics of Tax Evasion." *Journal of Economic Perspectives*, 21(1): 25-48.

Yaniv, G. 1999. "Tax Compliance and Advance Tax Payments: A Prospect Theory Analysis." *National Tax Journal*, 52(4): 753-764.

1.9 Acknowledgements

I am grateful to John Bakija, Dan Benjamin, Greg Besharov, Steve Coate, Ori Heffetz, Ben Ho, David Laibson, Francesca Molinari, Ted O'Donoghue, Ken Whelan, and seminar audiences at Cornell and the NBER for comments and suggestions that improved the project. I thank the Cornell Institute for the Social Sciences, the National Bureau of Economic Research, and the National Institute on Aging (grant T32-AG00186) for generous research support, and Anthony Hawkins for research assistance.

CHAPTER 2

WHAT DO YOU THINK WOULD MAKE YOU HAPPIER? WHAT DO YOU THINK YOU WOULD CHOOSE?

Daniel J. Benjamin, Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones

Abstract: Would people choose what they think would maximize their subjective well-being (SWB)? We present survey respondents with hypothetical scenarios and elicit both choice and predicted SWB rankings of two alternatives. While choice and predicted SWB rankings usually coincide in our data, we find systematic reversals. We identify factors—such as predicted sense of purpose, control over one’s life, family happiness, and social status—that help explain hypothetical choice controlling for predicted SWB. We explore how our findings vary by SWB measure and by scenario. Our results have implications regarding the use of SWB survey questions as a proxy for utility.

All things considered, how satisfied are you with your life as a whole these days?

Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?

*Much of the time during the past week, you felt you were happy. Would you say yes or no?*¹

2.1 Introduction

Economists increasingly use survey-based measures of subjective well-being (SWB) as an empirical proxy for utility. In many applications, SWB data are used for testing or estimating preference models, or for conducting welfare evaluations, in situations where these are difficult to do credibly with choice-based revealed-preference methods. Examples include estimating the negative externality from neighbors' higher earnings (Erzo F.P. Luttmer, 2005), individuals' tradeoff between inflation and unemployment (Rafael Di Tella, Robert J. MacCulloch, and Andrew J. Oswald, 2003), and the effect of health status on the marginal utility of consumption (Amy Finkelstein, Luttmer, and Matthew J. Notowidigdo, 2008). Such work often points out that in addition to being readily available where choice-based methods might not be, SWB-based proxies avoid the concern that choices may reflect systematically biased beliefs about

¹The first of these three questions is from the World Values Survey; similar questions appear in the Euro-Barometer Survey, the European Social Survey, the German Socioeconomic Panel, and the Japanese Life in Nation survey. The second question is from the U.S. General Social Survey; similar questions appear in the Euro-Barometer survey, the National Survey of Families and Households, and the World Values Survey. The third question is from the University of Michigan's Survey of Consumers; similar questions appear in the Center of Epidemiologic Studies Depression Scale, the Health and Retirement Study, and the Gallup-Healthways Well-Being Index.

their consequences (e.g., George Loewenstein, Ted O'Donoghue, and Matthew Rabin, 2003; Daniel T. Gilbert, 2006). It hence interprets SWB data as revealing what people would choose if they were well-informed about the consequences of their choices for SWB, and uses SWB measures to proxy for utility under the assumption that people make the choices they think would maximize their SWB. This paper provides evidence for evaluating that assumption.

We pose a variety of hypothetical decision scenarios to three respondent populations: a convenience sample of 1,066 adults, a representative sample of 1,000 adult Americans, and 633 students. Each scenario has two alternatives. For example, one scenario describes a choice between a job that pays less but allows more sleep versus a job with higher pay and less sleep. We ask respondents which alternative they think they would choose. We also ask them under which alternative they anticipate greater SWB; we assess this “predicted SWB” using measures based on each of the three commonly-used SWB questions posed in the epigraph above. We test whether these two rankings coincide.² To the extent that they do not, we attempt to identify—by eliciting predictions about other consequences of the choice alternatives—what else besides predicted SWB explains respondents’ hypothetical choices, and to quantify the relative contribution of predicted SWB and other factors in explaining these choices.

In designing our surveys, we made two methodological decisions that merit discussion. First, while the purpose of our paper is to help relate choice behavior to SWB measures, those measures are based on reports of respondents’

²In the terminology of Daniel Kahneman, Peter P. Wakker, and Rakesh K. Sarin (1997), our work can be viewed as comparing “decision utility” (what people choose) with “predicted utility” (what people predict will make them happier). We avoid these terms, however, because our “decisions” are hypothetical; and because we ask respondents to predict their responses to common SWB survey questions, rather than the integral over time of their moment-by-moment affect.

general levels of realized SWB, whereas our survey questions elicit respondents' predictions comparing the SWB consequences of specific choices. To compare SWB rankings with choice rankings under the same information set and beliefs, however, we must measure predictions about SWB because it is only predictions that are available at the moment of choice. Moreover, to link SWB with choice, we must focus on the SWB consequences of specific choices.

Second, while economists generally prefer data on incentivized choices, our choice data consist of responses to questions about predicted choice in hypothetical scenarios. This is a limitation of our approach because the two may not be the same.³ However, using hypothetical scenarios allows us to address a much wider variety of relevant real-world choice situations. It also allows us to have closely comparable survey measures of choice and SWB.⁴ For brevity, hereafter we will sometimes omit the modifiers "predicted" and "hypothetical" when the context makes it clear that by "choice" and "SWB" we refer to our survey questions.

We have two main results. First, we find that overall, respondents' SWB predictions are a powerful predictor of their choices. On average, SWB and choice coincide 83 percent of the time in our data. We find that the strength of this relationship varies across choice situations, subject populations, survey methods, questionnaire structure variations, and measures of SWB, with coincidence ranging from well below 50 percent to above 95 percent.

³Although economists generally prefer data on incentivized choices, in some situations self-reports may be more informative about preferences, e.g., when temptation, social pressure, or family bargaining might distort real-world choices away from preferences. (As we mention below, our data are silent on which method best elicits preferences.)

⁴The advantage in having closely comparable (survey-based) measures is that when we find discrepancies between choice responses and SWB responses, these discrepancies can be attributed wholly to differences in question content rather than at least partially to differences in how respondents react to the perceived realness of the consequences of their response.

Our second main result is that discrepancies between choice and SWB rankings are systematic. Moreover, we can indeed identify other factors that help explain respondents' choices. As mentioned above, in addition to eliciting participants' choices and predicted SWB, in some surveys we also elicit their predictions regarding particular aspects of life other than their own SWB. The aspects that systematically contribute most to explaining choice, controlling for own SWB, are sense of purpose, control over life, family happiness, and social status. At the same time, and in line with our first main result above, when we compare the predictive power of own SWB to that of the other factors we measure, we find that across our scenarios, populations, and methods, it is by far the single best predictor of choice.

We use a variety of survey versions and empirical approaches in order to test the robustness of our main results to alternative interpretations. For example, while most of our data are gathered by eliciting both choice and predicted SWB rankings from each respondent, in some of our survey variations we elicit the two rankings far apart in the survey, or we elicit only choice rankings from some participants and only SWB rankings from others. As another example, we assess the impact of measurement error by administering the same survey twice (weeks or months apart) to some of our respondents. While these different approaches affect our point estimates and hence the relative importance of our two main results, both results appear to be robust.

As steps toward providing practical, measure-specific and situation-specific guidance to empirical researchers as to when the assumption that people's choices maximize their predicted SWB is a better or worse approximation, we analyze how our results differ across SWB measures and across scenarios. Com-

paring SWB measures, we find that in our data, a “life satisfaction” measure (modeled after the first question in the epigraph) is a better predictor of choice than either of two “happiness” measures (modeled after the second and third questions in the epigraph) that perform similarly to each other. Comparing scenarios, we find that in scenarios constructed to resemble what our student respondents judge as representative of important decisions in their lives, predicted SWB coincides least often with choice, and other factors add relatively more explanatory power. We also find that in scenarios where one alternative offers more money, respondents are systematically more likely to choose the money alternative than they are likely to predict it will yield higher SWB. Under some conditions, this last finding suggests that the increasingly common method of valuing non-market goods by comparing the coefficients from a regression of SWB on income and on the amount of a good⁵ systematically estimates a higher value than incentivized-choice-based methods of eliciting willingness-to-pay (since the weight of money in predicted SWB understates its weight in choice).

Much previous research has studied the relationship between choice and happiness.⁶ Our work is most closely related to experiments reported in Amos Tversky and Dale Griffin (1991), Christopher H. Hsee (1999), and Hsee, Jiao

⁵Recent examples have valued deaths in one’s family (Angus Deaton, Jane Fortson and Robert Tortora, 2010), the social costs of terrorism (Bruno S. Frey, Simon Luechinger, and Alois Stutzer, 2009), and the social cost of floods (Luechinger and Paul A. Raschky, 2009).

⁶In a spirit similar to ours, Gary S. Becker and Luis Rayo (2008) propose (but do not pursue) empirical tests of whether things other than happiness matter for preferences in empirically-relevant choice situations. Relatedly, Ricardo Perez-Truglia (2010) tests empirically whether the utility function inferred from consumption choices is distinguishable from the estimated happiness function over consumption. In contrast to our approach, these tests and their interpretation are affected by whether individuals correctly predict the SWB consequences of their choices.

Our work is also related to a literature in philosophy that poses thought experiments in hypothetical scenarios in order to demonstrate that people’s preferences encompass more than their own happiness (e.g., Robert Nozick, 1974, pp. 42-45), but that literature focuses on extreme situations, such as being hooked up to a machine that guarantees happiness, and focuses on an abstract conception of happiness that is broader than empirical measures.

Zhang, Fang Yu, and Yiheng Xi (2003) that use methods similar to some of ours.⁷ However, because our goal is to provide guidance for interpreting results from the empirical economics literature, our paper differs from these prior papers in two fundamental ways. First, both our scenarios and our SWB measures are tailored to be closely relevant to the economics literature. Thus, rather than primarily focusing on narrow affective reactions to specific consumption experiences (e.g., the “enjoyment” of a sound system), as in Hsee (1999) and Hsee et al. (2003), we purposefully model our measures on the SWB questions asked in large-scale social surveys, and we focus on a range of scenarios that we designed to be relevant to empirical work in economics as well as scenarios that are judged by our respondents to represent important decisions in their lives. Second, crucially, we elicit predictions about other valued aspects of the choice alternatives. Indeed, it has often been observed that factors beyond one’s own happiness (in the narrow sense measured by standard survey measures) may matter for choice.⁸ As far as we are aware, however, our work is the first to quantitatively estimate the relative contribution of predicted SWB and these other factors in explaining choice.

The rest of the paper is organized as follows. Section 2.2 discusses the survey design and subject populations. Section 2.3 asks whether participants choose the alternative in our decision scenarios that they predict will generate greater

⁷These papers find discrepancies between choice and predicted affective reactions, in hypothetical scenarios carefully designed to test theories about why the two may differ. Tversky and Griffin (1991) theorize that payoff levels are weighted more heavily in choice, while contrasts between payoffs and a reference point are weighted more heavily in happiness judgments. Hsee (1999) and Hsee et al. (2003) theorize that when making choices, individuals engage in “lay rationalism,” i.e., they mistakenly put too little weight on anticipated affect and too much weight on “rationalistic” factors that include payoff levels as well as quantitatively-measured attributes. Our finding that factors other than SWB help predict choice provides a different possible perspective on the evidence from these earlier papers.

⁸For a few recent examples, see Ed Diener and Christie Scollon (2003), Loewenstein and Peter A. Ubel (2008, pp. 1801-1804), Hsee, Reid Hastie, and Jingqui Chen (2008, p. 239), and Marc Fleurbaey (2009).

SWB. Section 2.4 asks whether aspects of life other than SWB help predict choice, controlling for SWB, and compares the relative predictive power of the factors that matter for choice. Section 2.5 presents robustness analyses. Section 2.6 characterizes the heterogeneity in choice-SWB concordance across SWB measures, scenarios, and respondent characteristics. Section 2.7 concludes and discusses other possible applications of our methodology and implications of our findings. For example, while our paper focuses on testing measures that are based on existing SWB survey questions, our methodology can be used to explore whether alternative, novel questions could better explain choice. And while our data cannot inform us regarding the best way to elicit preferences, if one assumes that hypothetical choices reveal preferences, then our findings may imply that individuals do not exclusively seek to maximize SWB as currently measured. The appendix lists our decision scenarios. For longer discussions, as well as detailed information on all survey instruments, pilots, robustness analyses, and additional results, see our working paper, Daniel J. Benjamin, Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones (2010) with its Web Appendix (hereafter BHKR).

2.2 Survey design

While our main evidence is based on 29 different survey versions, they all share a similar core that consists of a sequence of hypothetical pairwise-choice scenarios. To illustrate, our “Scenario 1” highlights a tradeoff between sleep and income. Followed by its SWB and choice questions, it appears on one of our questionnaires as follows:

Say you have to decide between two new jobs. The jobs are exactly the same in almost every way, but have different work hours and pay different amounts.

Option 1: A job paying \$80,000 per year. The hours for this job are reasonable, and you would be able to get about 7.5 hours of sleep on the average work night.

Option 2: A job paying \$140,000 per year. However, this job requires you to go to work at unusual hours, and you would only be able to sleep around 6 hours on the average work night.

Between these two options, taking all things together, which do you think would give you a happier life as a whole?

Option 1: Sleep more but earn less			Option 2: Sleep less but earn more		
<i>definitely happier</i>	<i>probably happier</i>	<i>possibly happier</i>	<i>possibly happier</i>	<i>probably happier</i>	<i>definitely happier</i>
X	X	X	X	X	X
Please circle one X in the line above					

If you were limited to these two options, which do you think you would choose?

Option 1: Sleep more but earn less			Option 2: Sleep less but earn more		
<i>definitely choose</i>	<i>probably choose</i>	<i>possibly choose</i>	<i>possibly choose</i>	<i>probably choose</i>	<i>definitely choose</i>
X	X	X	X	X	X
Please circle one X in the line above					

In within-subject questionnaires, respondents are asked both the SWB question and the choice question above. In between-subjects questionnaires, respondents are asked only one of the two questions.

2.2.1 Populations and studies

We conducted surveys among 2,699 respondents from three populations: 1,066 patients at a doctor’s waiting room in Denver who participated voluntarily; 1,000 adults who participated by telephone in the 2009 Cornell National Social Survey (CNSS) and form a nationally representative sample;⁹ and 633 Cornell students who were recruited on campus and participated for pay or for course credit. The Denver and Cornell studies include both within-subject and between-subjects survey variants, while the CNSS study is exclusively within-subject.

Table 2.1 summarizes the design details of these studies. It lists each study’s respondent population, sample size, scenarios used (see 2.2.2 below), types of questions asked (see 2.2.3 below), and other details such as response scales, scenario order, and question order.¹⁰ The rest of this section explains the details summarized in the table.

2.2.2 Scenarios

Our full set of 13 scenarios is given in the Appendix. Table 2.1 reports which scenarios are used in which studies, and in what order they appear on different questionnaires. As detailed in the Appendix, some scenarios are asked in different versions (e.g., different wording, different quantities of money, etc.) and some scenarios are tailored to different respondent populations (e.g., while we

⁹The CNSS is an annual survey conducted by Cornell University’s Survey Research Institute. For details: <https://sri.cornell.edu/SRI/cnss.cfm>.

¹⁰The median age in our Denver, CNSS, and Cornell samples is, respectively, 47, 49, and 21; the share of female respondents is 76, 53, and 60 percent. For summary statistics, see BHKR table A3.

Table 2.1: Study-specific information

Study Location	Denver		CNSS	Cornell	
Choice vs. SWB: Within- or Between-Subjects	Within	Between	Within	Within	Between
Sample Population	Volunteers at a doctor's waiting room		Nationally representative	Cornell students	
Observations	497	569	1000	432 ^a	201
Scenarios used	1, 3, 4, 11, 12, 13	1, 2, 3, 4 (version 2), 12 (v2), 13	1	1-10 (with v2 for scenario 4)	
SWB Question Format	Observations for each SWB question format				
(i) Life Satisfaction					
<i>(Isolated)</i>	164	569			
(ii) Happiness with Life as a Whole					
<i>(Isolated)</i>	162		1000		
(iii) Felt Happiness					
<i>(Isolated)</i>	171				
(iv) Own Happiness with Life as a Whole					
<i>Isolated</i>				107	201
<i>First/Last In Series</i>				107	
(v) Immediately Felt Own Happiness					
<i>Isolated</i>				110	
<i>First/Last In Series</i>				108	
SWB Response Scale	6-point		Binary	7-point	
Choice Response Scale	6-point		Binary	6-point	
Meta-Choice Question?	Yes	No	No	Yes	No
Order variations					
Scenario order	4-1-11-12-13-3	1-2-12-13-3-4	1	1-2- ... -9-10	
	3-13-12-11-1-4	3-13-12-2-1-4 ^b			
Question order	Choice-Meta-SWB SWB-Choice-Meta		SWB-Choice	Choice-SWB	
Aspects of life order	Two opposite orderings of aspects				
Summary: number of questionnaire versions	12	4	1	8	4

Notes: See section I for the framing of the choice, SWB, and meta-choice questions. See the Appendix for a full description of each scenario. The scenarios corresponding to the scenario-numbers above are: (1) sleep vs. income, (2) concert vs. birthday, (3) absolute income vs. relative income, (4) legacy vs. income, (5) apple vs. orange, (6) money vs. time, (7) socialize vs. sleep, (8) family vs. money, (9) education vs. social life, (10) interest vs. career, (11) concert vs. duty, (12) low rent vs. short commute, (13) friends vs. income.

a Of these, 230 were surveyed twice, allowing us to conduct measurement-error-corrected estimation.

b Scenario 4 is always presented last because it is followed by both a choice and a SWB question. In order to have a clean between-subjects design, we did not want subjects to know we were interested in both choice and SWB until after subjects were done with the rest of the scenarios. We also note that this scenario is presented in four different order-versions, so strictly speaking, the Denver between-subjects study includes the four questionnaire versions reported in the table's bottom row, times four (sixteen versions in total).

ask students about school, we ask older respondents about work). In constructing the scenarios, we were guided by four considerations.

First, we chose scenarios that highlight tradeoffs between options that the literature suggests might be important determinants of SWB. Hence, respondents face choices between jobs and housing options that are more attractive financially versus ones that allow for: in Scenario 1, more sleep (Kahneman et al., 2004; William E. Kelly, 2004); in Scenario 12, a shorter commute (Stutzer and Frey, 2008); in 13, being around friends (Kahneman et al., 2004); and in 3, making more money relative to others (Luttmer, 2005; see Heffetz and Robert H. Frank, 2011, for a survey).

Second, since some of us were initially unsure we would find any divergences between predicted choice and SWB, in our earlier surveys we focused on choice situations where one's SWB may not be the only consideration. Hence, in Scenario 4 respondents choose between a career path that promises an "easier" life with fewer sacrifices versus one that promises posthumous impact and fame, and in Scenarios 2 and 11 they choose between a more convenient or "fun" option versus an option that might be considered "the right thing to do."

Third, once we found divergences between predicted SWB and choice, in our later surveys (the Cornell studies) we wanted to assess the magnitude of these divergences in scenarios that are representative of important decisions faced by our respondent population. For this purpose we asked a sample of students to list the three top decisions they made in the last day, month, two years, and in their whole lives.¹¹ Naturally, decisions that were frequently mentioned

¹¹The sample included 102 University of Chicago students; results were subsequently supported by surveying another 171 Cornell students. See BHKR for details and classification of responses.

by respondents revolved around studying, working, socializing and sleeping. Hence, in the resulting Scenarios 7-10, individuals have to choose between socializing and fun versus sleep and schoolwork; traveling home for Thanksgiving versus saving the airfare money; attending a more fun and social college versus a highly selective one; and following one's passion versus pursuing a more practical career path. To these scenarios we added Scenario 6, which involves a time-versus-money tradeoff tailored for a student population.

Fourth, as an informal check on our methods, we wanted to have one falsification-test scenario where we expected a respondent's choice and SWB ratings to coincide. For this purpose, we added Scenario 5, in which respondents face a choice between two food items (apple versus orange) that are offered for free and for immediate consumption. Since we carefully attempted to avoid any non-SWB differences between the options, we hypothesized that in this scenario, predicted SWB would most strongly predict choice. This scenario has the additional attraction of being similar to prevalent decisions in almost everyone's life, which is our third consideration above.

2.2.3 Main questions

Choice question. In all studies, for each scenario, the choice question is worded as in our example above. In our analysis, we convert the horizontal six-point response scale into an intensity-of-choice variable, ranging from 1 to 6, or into a binary choice variable. CNSS responses are elicited as binary choices.¹²

SWB question. While the choice question is always kept the same, we vary the

¹²CNSS responses are elicited as binary because in telephone interviews the binary format is both briefer for interviewers to convey and easier for respondents to understand.

SWB question in order to examine how choice relates to several different SWB measures. In our Denver within-subject study we ask three versions of the SWB question, modeled after what we view as three “families” of SWB questions that are commonly used in the literature (see examples in the epigraph):

- (i) life satisfaction: “Between these two options, which do you think would make you more satisfied with life, all things considered?”;
- (ii) happiness with life as a whole: “Between these two options, taking all things together, which do you think would give you a happier life as a whole?”; and
- (iii) felt happiness: “Between these two options, during a typical week, which do you think would make you feel happier?”

As in the example above, there are six possible answers, which we convert into either a six-point variable or a binary variable.

In the CNSS study, where design constraints limited us to one version of the SWB question, we ask only version (ii). As with the choice question, response is binary.

As described shortly, in our Cornell studies we ask respondents about twelve different aspects of life, of which (one’s own) happiness is only one. In those studies we use versions of (ii) and (iii) that are modified to remain meaningful, with fixed wording, across aspects. The modified (ii) and (iii) result in these two new versions:

- (iv) own happiness with life as a whole: “Between these two options, taking all things together, which option do you think would make your life as a

whole better in terms of [your own happiness]”; and

- (v) immediately-felt own happiness: “Between these two options, in the few minutes immediately after making the choice, which option do you think would make you feel better in terms of [your own happiness].”¹³

The modified response scale now includes a middle “no difference” response, and has seven possible answers (*Option 1 definitely better; Option 1 probably better; Option 1 possibly better; no difference; Option 2 possibly better, etc.*). We allow respondents to indicate “no difference” because we anticipated that in some of the scenarios, it would make little sense to force respondents to predict that all aspects would differ across the two options (e.g., “sense of purpose” in Scenario 5, “apple vs. orange”).

On the spectrum between more cognitive, evaluative SWB measures and more affective, hedonic ones (e.g., Diener et al., 2009), we view version (i) as the most evaluative, versions (iii) and (v) as the most affective, and versions (ii) and (iv) as intermediate.

Other questions. For completeness, let us briefly mention, first, that in all questionnaires of the Denver and Cornell within-subject studies, the choice question is followed by what we refer to as a meta-choice question: “If you were limited to these two options, which **would you want yourself to choose?**” Also, recall that the SWB question in all Cornell studies is modified to elicit rankings of the two scenario options in terms of eleven additional aspects of life as well as “own happiness.” For example, in versions (iv) and (v) of the SWB question, [your own happiness] may be followed by [your family’s happiness], [your health],

¹³Since our between-subject tests have less statistical power than our within-subject tests, we ask only version (i) in our Denver between-subjects surveys and only version (iv) in our Cornell between-subjects surveys.

[your romantic life], etc.¹⁴ We discuss these additional questions and the data they yield in later sections.

2.3 Do people respond to the choice and SWB questions in the same way?

In this section we look at respondents' binary ranking of Option 1 versus Option 2 in terms of hypothetical choice compared with their binary ranking in terms of predicted SWB.

2.3.1 Within-subject results

Table 2.2 reports the distribution of binary responses to our within-subject surveys' choice and SWB questions by study and scenario, along with p-value statistics from equality-of-proportions tests. The table pools responses across SWB question variants (see 2.2.3 and table 2.1 above); we discuss results by specific SWB measure below.¹⁵

¹⁴In some questionnaire versions, we separate "own happiness" from the other eleven aspects, and ask respondents first only about own happiness in each scenario, and then, representing the scenarios, we ask about the other aspects. In these versions, we refer to the question on own happiness as an "isolated" measure of SWB (see table 2.1). In other versions, where the twelve aspects appear together, we refer to the own happiness question as a "first/last in series" measure. When own happiness is "first in series," the twelve aspects appear together in the order they are listed as regressors in table 2.3 below. When own happiness is "last in series," the twelve aspects appear together in reverse order.

¹⁵Non-response in our surveys was generally low. In the Cornell studies, virtually all questions had a non-response rate below 2 percent (one Cornell respondent was excluded due to obvious confusion with instructions). In the CNSS, fewer than 5 percent of respondents answered "Do not know" or refused to answer in any of the questions. Due to the less-structured recruiting method used in our Denver doctor's office studies, some questions from those studies had non-response rates as high as 20 percent. However, the majority of this non-response is driven by respondents being called in for their appointments, alleviating concerns of selection

Table 2.2: Choice and SWB responses across studies and scenarios (within-subject data)

Choice Scenario <i>For exact phrasing see Appendix</i>	<u>Denver</u>						<u>CNSS</u>
	1	3	4	11	12	13	1
	Sleep vs Income	Abs. Inc. vs Rel. Inc.	Legacy vs Income	Concert vs Duty	Low Rent vs Short Commute	Friends vs Income	Sleep vs Income
Higher SWB: Option 1 Chosen: Option 1	58%	48%	24%	16%	52%	50%	74%
Higher SWB: Option 2 Chosen: Option 2	29%	42%	60%	65%	32%	34%	18%
Higher SWB: Option 2 Chosen: Option 1	1%	6%	2%	12%	11%	2%	1%
Higher SWB: Option 1 Chosen: Option 2	12%	4%	14%	7%	5%	14%	7%
p-value from Liddell Exact Test	0.000	0.350	0.000	0.024	0.002	0.000	0.000
	n = 425	n = 420	n = 422	n = 422	n = 425	n = 422	n = 972

Choice Scenario <i>For exact phrasing see Appendix</i>	<u>Cornell</u>									
	1	2	3	4	5	6	7	8	9	10
	Sleep vs Income	Concert vs Birthday	Abs. Inc. vs Rel. Inc.	Legacy vs Income	Apple vs Orange	Money vs Time	Socialize vs Sleep	Family vs Money	Education vs Social life	Interest vs Career
	Version 2									
Higher SWB: Option 1 Chosen: Option 1	29%	29%	41%	44%	45%	44%	62%	68%	53%	27%
Higher SWB: Option 2 Chosen: Option 2	46%	49%	43%	31%	50%	37%	15%	15%	22%	35%
Higher SWB: Option 2 Chosen: Option 1	1%	7%	14%	8%	2%	14%	17%	5%	22%	3%
Higher SWB: Option 1 Chosen: Option 2	23%	15%	2%	17%	3%	5%	6%	12%	3%	35%
Indifference for SWB	8%	14%	13%	10%	37%	22%	10%	5%	6%	6%
p-value of Liddell Exact Test	0.000	0.002	0.000	0.001	0.424	0.000	0.000	0.001	0.000	0.000
	n = 397	n = 368	n = 375	n = 387	n = 270	n = 333	n = 385	n = 409	n = 402	n = 402

Notes: Response distribution by study and scenario. For the complete text of each scenario, see the Appendix. If a scenario's phrasing changed meaningfully between surveys, the version of the scenario is indicated in the first row of the study block. The Liddell Exact Test is a paired equality-of-proportions test of the null hypothesis that mean response to choice question = mean response to SWB question. In the Cornell data, where respondents could indicate SWB indifference, responses indicating indifference were dropped before conducting the test.

The left-most column in the top section of the table reports Scenario 1 figures from the Denver within-subject questionnaires (our “sleep vs. income” scenario from the example in section 2.2). The column’s top four cells report a vertically-stacked 2×2 contingency matrix, consisting of the joint binary distribution of subjects who favor an option in the choice question and those who favor it in the SWB question. Looking at these four cells, we point out two facts that illustrate this section’s two main findings. First, the top two cells reveal that the SWB response is highly predictive of the choice response: between the two cells, 87 percent of respondents rank Option 1 versus Option 2 in the choice question the same as they do in the SWB question. Second, the next two cells reveal systematic differences across the two questions among the remaining 13 percent of respondents: while 12 percent rank Option 1 (sleep) above Option 2 (income) in the SWB question and reverse this ranking in the choice question, only 1 percent do the opposite. This asymmetry suggests that on average, respondents react to the two questions systematically differently. The fifth cell reports the p-value from a Liddell exact test, a nonparametric, equality-of-proportions test for paired data (Douglas K. Liddell, 1983). The null hypothesis—namely, that the proportion of respondents who rank Option 2 above Option 1 is the same across the choice and the SWB questions—is easily rejected.

Examining the top five rows in table 2.2 for the rest of the Denver columns verifies that the two main findings above are not unique to Scenario 1: in the remaining five scenarios, 81 to 90 percent of respondents rank the two options identically across the choice and SWB questions; yet in four out of five cases, choice-SWB reversals among the remaining 10 to 19 percent of respon-

bias. Comparing the completed responses of subjects who did not finish the survey to the responses of those who finished the entire survey, we find no evidence of a difference in average responses.

dents are asymmetric, and the equality-of-proportions null hypothesis across the two questions is easily rejected. In these cases, respondents rank income above legacy, concert above duty, low rent above short commute, and income above friends in higher proportions in the choice question than in the SWB question. There appears to be a systematic tendency among respondents to favor money in the choice question more than in the SWB question, a point we return to below. (The results for the absolute vs. relative income scenario are discussed below.)

Similarly, the CNSS column suggests that, qualitatively, Scenario 1's findings carry over from our Denver study—a pencil-and-paper survey with six-point response scales administered to a convenience sample—to the CNSS study—a telephone survey with binary response scales administered to a nationally representative sample. While the proportion of participants with no choice-SWB reversals increases to 92 percent, almost all of the rest—7 out of the remaining 8 percent—favor Option 1 (sleep) in the SWB question and Option 2 (income) in the choice question. The direction of this asymmetry is hence the same as in the Denver sample, and equality of proportions is again easily rejected.

Last among our within-subject data, results from the Cornell surveys are reported at the bottom section of table 2.2. The structure of this portion of the table is similar to the corresponding Denver and CNSS portions, with the following three differences that result from the fact that the Cornell questionnaires allow for an additional “no difference” response in the SWB question: (a) an additional row below the top four rows reports the proportion of respondents who choose the “no difference” response; (b) the top four rows report vertically-stacked contingency matrices as before, only here they exclude these “no differ-

ence” responses (their sum is normalized to 100 percent); and (c) the “no difference” responses are excluded from the Liddell tests.¹⁶

Starting again with Scenario 1 in the left-most column, choice-SWB reversals (in the third and fourth rows, 24 percent together) are still a minority, although they are almost twice to three times more common in the Cornell sample than in the Denver and CNSS samples. Nonetheless, consistent with the Denver and CNSS data, in virtually all of these reversals—23 of the 24 percent—Option 1 (sleep) is ranked above Option 2 (income) in the SWB question and below it in the choice question. Equality of proportions is, again, strongly rejected for this scenario.¹⁷

Moving to the rest of the Cornell columns reveals a similar story. Equality of proportions is strongly rejected for all the remaining nine scenarios (2-10) as well, with the exception of Scenario 5. Recall that we constructed Scenario 5 (“apple vs. orange”) as a falsification test, where—barring problems with our methods—choice and SWB should largely coincide. The results support this prediction. Indeed, only 5 percent of responses exhibit reversals in this scenario, by far the lowest fraction among the ten scenarios. Furthermore, we find no evidence that these reversals are in one systematic direction.¹⁸ As to the two other

¹⁶The distribution of choice-responses among individuals indicating “no difference” for SWB mirrors the distribution of choice-responses among the rest of the respondents reasonably closely (BHKR table A5), and, hence, the choice proportions in table 2.2 are virtually unaffected by excluding these individuals. Moreover, under the null hypothesis that choice is determined solely by predicted SWB, the distribution of choice-responses should be closer to 50-50 for individuals indicating SWB “no difference.” Hence, the responses of these respondents actually provide additional suggestive evidence against the null hypothesis.

¹⁷Comparing each of the top four cells in the scenario 1 column across the three within-subject samples reveals that the reported proportions differ dramatically between the samples. Given the very different populations and, in the CNSS study, the very different survey methods, this finding in itself is not surprising. (For example, we speculate that since a telephone survey is harder to understand, more respondents answered the two questions in the same way, taking the “artificial consistency” mental shortcut discussed in 2.3.2 below.)

¹⁸At the same time, a sizeable 37 percent of respondents indicate “no difference” in the SWB

scenarios that are used in both the Denver and Cornell studies—Scenarios 3 and 4—choice-SWB reversals maintain their direction: in both studies, (absolute) income is ranked above relative income (Scenario 3) and above legacy (Scenario 4) in the choice questions more often than in the SWB questions. While equality of proportions is rejected in the Cornell data but not in the Denver data in Scenario 3, it is rejected in both studies in Scenario 4.

Finally, in Scenarios 6 and 8, which are used only in the Cornell studies and include a “money” option, we once again find that respondents favor money in the choice question more than in the SWB question. That this tendency holds in all seven scenarios that trade off more money/income for something else—be it more sleep, higher relative income, a legacy, a shorter commute, being around friends, having more time, or visiting family—suggests that predicted SWB understates the weight of money and income in hypothetical choice.¹⁹ Of course, predicted SWB is not the same as experienced SWB, and hypothetical choice is not the same as incentivized choice. Nevertheless, unless the difference between those gaps is sufficiently negatively correlated with the systematic gap we find between hypothetical choice and predicted SWB, our results suggest that survey measures of experienced SWB do not fully capture the weight of money and income in choice.

Our two main findings—that the ranking of the two options is identical across the choice and SWB questions for most respondents and in most scenar-

question in scenario 5—by far the highest. This may suggest that scenario 5 is “cleaner” than we intended it to be: not only non-SWB aspects of life, but even own happiness is deemed by many respondents irrelevant in what they may perceive as a context of *de gustibus non est disputandum*.

¹⁹Reassuringly, this tendency in our data is consistent both with the data of Tversky and Griffin (1991) and Hsee et al. (2003), who use a scenario similar to our Scenario 3 (absolute income vs. relative income), and with their psychological theories (e.g., “lay rationalism”) mentioned in footnote 7.

ios, but that respondents react to the two questions systematically differently—hold not only in the pooled data, but also for each SWB question variant (i)-(v) separately. We show this in BHKR table A4, which reports versions of table 2.2 by SWB measure. Interestingly, we find some differences across the measures in the prevalence of choice-SWB reversals. In the Denver sample, the life satisfaction question variation (i) comes closest to matching choice, with only 11 percent reversals, averaged across all scenarios. In comparison, happiness with life as a whole (ii) and felt happiness (iii) yield more reversals—17 percent each. In the Cornell sample, own happiness with life as a whole (iv) and immediately felt own happiness (v) both yield 22 percent reversals. We return to the comparison between different SWB measures in section 2.6.1 below.

2.3.2 Between-subjects results

Our within-subject analysis above is based on both choice and SWB responses elicited from each individual. However, empirical work that uses SWB data relies on surveys that measure SWB alone, not together with choice. Thus, two potential biases could compromise the relevance of our findings to existing SWB survey data and their applications. On the one hand, asking a respondent both questions might generate an “artificial consistency” between the two responses. For example, respondents might think they ought to give consistent answers, or might give consistent answers as an effort-saving mental shortcut. On the other hand, an “artificial inconsistency” bias is also possible if respondents infer from being asked more than one question that they ought to give different answers, or if the presence of the other question focuses respondents’ attention on the contrast between the wordings.

To assess these concerns, we compare the above results from the Denver and Cornell within-subject studies with their counterpart between-subjects studies, in which respondents are asked only the choice or only the SWB question. Three of the six Denver scenarios analyzed above, and all ten of the Cornell scenarios, are repeated with identical wording in their between-subjects counterparts (see table 2.1). Across these thirteen comparable scenarios and including only the within-subject respondents who faced the SWB measure used in the between studies (i.e., variant (i) in Denver and (iv) in Cornell), the median within-versus-between absolute difference in the proportion of respondents favoring each option is 5 percentage points in the choice question (a statistically significant difference in two scenarios) and is 8 percentage points in the SWB question (statistically significant in four scenarios).²⁰ Overall, then, the within and between response distributions sometimes differ. Moreover, the direction of the differences in the choice compared to the SWB data suggests that on average, artificial inconsistency might indeed explain some of the choice-SWB reversals in the within data: in the within data, the average choice-SWB difference in proportions is 10.8 percentage points; in the between data, it is 7.4 percentage points—about two-thirds of the within difference.

While choice-SWB reversals are on average of smaller magnitudes in the between data, they remain sufficiently large to yield statistical results comparable to those in the within data. In the between data, we can reject the null hypothesis of no difference between choice and SWB proportions in four scenarios, which

²⁰Using Fisher tests and a 5 percent significance level, we reject the null hypothesis that equal proportions choose Option 2 in the within and between data for the Denver sleep vs. income scenario (1) and the Cornell interest vs. career scenario (10). We reject the null hypothesis that equal proportions anticipate higher SWB under Option 2 in the within and between data for the Denver friends vs. income scenario (13) and the Cornell money vs. time, education vs. social life, and interest vs. career scenarios (6, 9, and 10). We report the full details of the between-subjects data analysis, including all the relevant distributions and statistical tests mentioned in this subsection, in BHKR (section II.B, table 2, and table A4).

is fewer than in the within data discussed in section 2.3.1. However, one important reason is that, mechanically, the unpaired test on the between data has much less statistical power than the paired test on the within data: even with an equal number of respondents, each responds to only one question instead of two, and we cannot partial out correlated individual effects on choice and SWB in analyzing the between data. To compare the within and between data controlling for power differences, we “unpaired” our within data, matched sample sizes as closely as possible, and simulated unpaired equality-of-proportion tests that treat these data as if they were between data. We find that we can reject the no-difference null in four scenarios, exactly the same as what we find using the between data.

Our overall interpretation is that while there are differences across the between- and the within-subject studies—in particular, choice-SWB reversals are on average less pronounced in the between-subjects studies—either set of studies supports our two main findings.

2.3.3 Measurement error

Our analysis above suggests that in many scenarios, individuals do not respond to the choice and SWB questions as if they were responding to the same question. However, in a given scenario, such rejection of the null hypothesis could be explained by differences in measurement error across the two questions—for example, because it is easier to introspect about choice than about SWB, or vice versa. An individual whose “true” ranking of the options is identical across the questions is more likely to mistakenly rank the “wrong” option higher in a

question with greater measurement error, leading to ranking proportions closer to 50-50 for that question.

Looking across table 2.2's columns reveals that cross-question differences in the measurement error for choice and SWB in the same direction in all scenarios in a study cannot explain our data. For example, in the Denver data, choice proportions are closer to 50-50 in Scenarios 1, 11, and 13, but SWB proportions are closer to 50-50 in Scenarios 4 and 12.

To summarize, the two main findings in this section are (a) that most respondents in most scenarios do not exhibit choice- versus SWB-ranking reversals, and (b) that when they do, their pattern of reversals is systematic. Overall, the two findings hold up well—although with differences in relative strength—across scenarios, populations, and designs. Furthermore, these findings cannot be explained by a measurement error structure that is stable across scenarios.

2.4 Do other factors help predict choice, and by how much?

In this section we ask: Can we identify other factors that help explain hypothetical choices, controlling for predicted own SWB? We also analyze to what extent respondents' choices in our data can be explained by their predicted SWB and other aspects of life together, compared with their predicted SWB alone.

We address these questions using data from the Cornell sample, where we ask respondents to rank the options on a set of eleven additional aspects of life, in addition to ranking them on choice and own SWB (see section 2.2.3). Specifically, in addition to being asked about "your own happiness," respondents are

also asked about: your family’s happiness, your health, your romantic life, your social life, your control over your life, your life’s level of spirituality, your life’s level of fun, your social status, your life’s non-boringness, your physical comfort, and your sense of purpose. While still a limited list, it is intended to capture “functionings” proposed by economists and philosophers (Amartya K. Sen, 1985; Martha Nussbaum, 2000); non-hedonic and eudaimonic components of well-being proposed by psychologists (e.g., Matthew P. White and Paul Dolan, 2009) that are not fully captured by measures of SWB (Carol D. Ryff, 1989); as well as other factors that we thought might matter for choice besides own happiness.

The design of our Cornell between-subjects surveys allows us to also elicit within-subject data from our 201 participants. This is done by presenting subjects with the between-subjects part of the survey, followed by an additional, within-subject part.²¹ When discussing the between-subjects results in section 2.3.2 we used only data from the first, between-subjects part. In contrast, in this section we pool data from both parts, treating them as within-subject data. Further pooling these data with the original Cornell within-subject data (432 respondents) yields an augmented sample of 633 Cornell within-subject respondents, which we analyze here. As we report in section 2.5, our main results hold in the constituent subsamples.

²¹To be specific, we present the entire sequence of ten scenarios three times. First, each scenario is presented and is followed by only a choice question (for half the respondents) or only a SWB question (for the other half). Second, after respondents finish answering that question for each of the ten scenarios, the ten scenarios are presented again, each followed by only the question (SWB or choice) respondents had not seen yet. Finally, the ten scenarios are presented for a third time, with each scenario followed by the eleven additional questions about other aspects of life. Respondents are specifically instructed to answer the surveys in exactly the order questions are presented, and the experimenters verify that they do (in the rare cases where a respondent was observed to flip through the pages, she/he was promptly reminded of this instruction). With this design, excluding data collected after the first round of scenario-presentation yields between-subjects data.

2.4.1 Response distributions

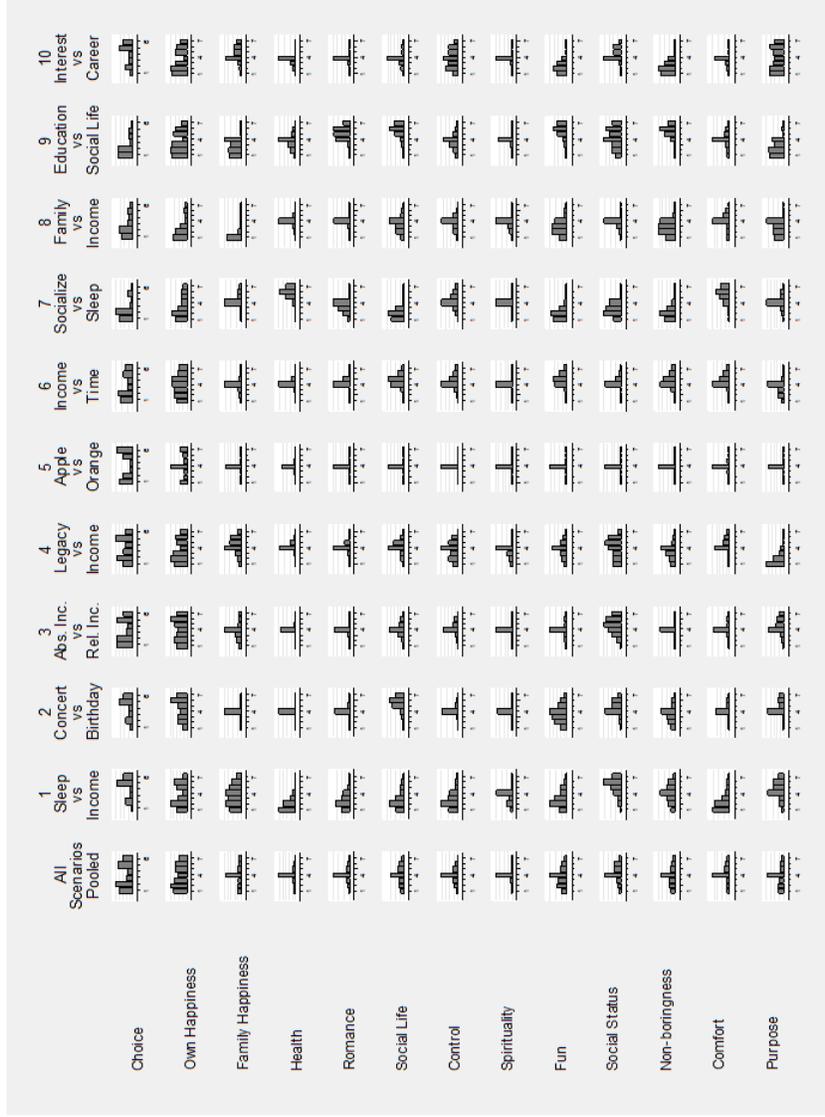
Figure 1 displays, by scenario, the histograms of raw, multi-point responses to the choice, (own) SWB, and eleven other aspect questions. Note first that the choice responses—and also the SWB responses, although to a lesser extent—tend to be bimodal with most of the mass on “definitely” or “probably,” suggesting that the choice-SWB reversals discussed in section 2.3 are not the result of widespread near-indifferences. Second, notice that we were rather successful in constructing Scenario 5 (apple vs. orange): almost everyone indicates “no difference” in the bottom eleven cells in this column. While 37 percent also indicate “no difference” on SWB, the low count of reversals in Scenario 5 suggests that for the other respondents, variation in choice is strongly related to variation in SWB. Finally, note that in many other scenarios, there is substantial variation in the eleven other aspect rankings, and that the histogram of choice responses sometimes looks rather different from the histogram of SWB responses.

2.4.2 Explaining the variation in choice

Table 2.3 presents a variety of specifications in which we regress choice on SWB and other aspects of life, aggregating data across the ten scenarios (we discuss regressions by scenario in section 2.6.2 below).²² We want to estimate the re-

²²In the OLS and ordered probit regressions, the dependent variable is 6-point choice. In the probit regressions the dependent variable is binary choice. All regressions use 7-point ratings of aspects. Based on 633 Cornell respondents. Each observation is a respondent’s choice and aspect ratings for one scenario; there are 10 observations per respondent corresponding to the 10 scenarios in the questionnaires. Probit and ordered probit regressions include (unreported) scenario fixed effects. OLS regressions’ variables are demeaned at the scenario level, generating coefficients equivalent to those generated by including scenario fixed effects. Measurement error corrections are done using the Simulation-Extrapolation method described in section 2.4.4, under the assumption of additive measurement error. Observations with missing data in any variable are excluded from all regressions.

Figure 2.1: Raw response distributions (choice and aspects of life)



Notes: Based on 633 Cornell respondents. The histograms show the distribution of 6-point responses to the choice question (top row) and 7-point responses to the aspect questions (bottom twelve rows). The left-most column aggregates data across choice scenarios; each of the other columns corresponds to a specific scenario.

lationship from the within-scenario-rather than the between-scenario-variation in responses. For this purpose, in the probit and ordered probit specifications, we include scenario fixed effects. In the OLS specifications, we demean all variables at the scenario level. Doing so yields coefficients identical to those in a fixed-effects OLS specification, but has the advantage that the R^2 's reflect only the within-scenario explanatory power of the regressors.

The first column of table 2.3 reports an OLS regression of six-point choice on seven-point SWB. The R^2 shows that 0.38 of the variation in choice is explained by own happiness alone. In comparison, a regression of the same choice measure on our eleven other aspects (each as a seven-point variable) yields an R^2 of 0.21 (second column of table 2.3). Hence, we find that own SWB predicts choice substantially better than all of the other aspects combined. In the third column we regress choice on both own SWB and the eleven other aspects. The R^2 of 0.41 is substantially higher than that in the second column but is only slightly higher than that in the first column.²³ The pattern in these three columns is similar when we relax the linear functional form, replacing each regressor with a set of six dummy variables (not reported). In summary, when we pool data across scenarios we find that adding eleven additional aspects to the regression of choice on own SWB increases explanatory power, but the increase is rather modest. (The increase is substantial, however, in some of the individual scenarios, as we report in section 2.6.2.)

²³Bootstrapped standard errors yield the following 95-percent confidence intervals around the three respective R^2 's: [0.36, 0.40], [0.19, 0.23], and [0.39, 0.43].

Table 2.3: Regressions of choice on aspects of life

			OLS		Ordered Probit		Probit	
	None	None	None	SIMEX	None	None	SIMEX	
ME correction	None	None	None	SIMEX	None	None	SIMEX	
Own happiness	0.54*** (0.009)		0.46*** (0.010)	0.59*** (0.014)	0.37*** (0.009)	0.37*** (0.012)	0.48*** (0.019)	
Family happiness		0.15*** (0.017)	0.08*** (0.015)	0.11*** (0.026)	0.06*** (0.012)	0.09*** (0.017)	0.13*** (0.032)	
Health		0.07** (0.021)	0.00 (0.019)	0.00 (0.031)	0.01 (0.016)	0.01 (0.022)	0.02 (0.042)	
Life's level of romance		-0.00 (0.024)	-0.01 (0.021)	0.01 (0.033)	-0.00 (0.018)	-0.00 (0.025)	0.04 (0.045)	
Social life		-0.01 (0.020)	-0.03 (0.018)	-0.05 (0.028)	-0.02 (0.015)	-0.02 (0.021)	-0.04 (0.036)	
Control over your life		0.17*** (0.017)	0.08*** (0.015)	0.11*** (0.025)	0.06*** (0.012)	0.09*** (0.017)	0.13*** (0.028)	
Life's level of spirituality		-0.08** (0.024)	-0.02 (0.021)	-0.04 (0.036)	-0.02 (0.018)	-0.04 (0.025)	-0.05 (0.047)	
Life's level of fun		0.13*** (0.021)	0.05* (0.018)	0.03 (0.031)	0.04* (0.015)	0.04* (0.021)	0.03 (0.036)	
Social status		0.07*** (0.016)	0.06*** (0.014)	0.07** (0.023)	0.05*** (0.012)	0.07*** (0.016)	0.10*** (0.027)	
Life's non-boringness		0.07*** (0.020)	-0.01 (0.017)	-0.01 (0.030)	0.00 (0.014)	0.00 (0.020)	0.01 (0.037)	
Physical comfort		0.09*** (0.017)	0.04** (0.014)	0.03 (0.023)	0.04** (0.012)	0.05** (0.017)	0.04 (0.030)	
Sense of purpose		0.21*** (0.015)	0.12*** (0.013)	0.13*** (0.022)	0.10*** (0.011)	0.12*** (0.015)	0.14*** (0.025)	
Observations	6217	6217	6217	6217	6217	6217	6217	
(pseudo) R2	0.38	0.21	0.41		0.19	0.35		

Notes: Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001. See footnote 22 for more details.

2.4.3 Comparing the coefficients

In order to compare and interpret the coefficients in table 2.3, we assume that hypothetical choices in our data can be represented as maximizing a utility function $U(H(X), X)$, where H is own SWB and X is a vector of other factors that might affect choice both directly and indirectly through H .²⁴ If people choose what they think would maximize their SWB alone (as opposed to trading off their SWB for other factors), then the (vector) partial derivative $\frac{\partial U}{\partial X}$ will be identically zero. To a first-order approximation, this would require that all eleven coefficients other than that on own happiness in table 2.3's third column be zero—a hypothesis we can easily reject (F-test $p < 0.0001$). This result is robust to treating the choice measure as ordinal or as binary (table 2.3's fifth and sixth columns); to relaxing the linearity of our SWB measure by replacing it with a set of six dummy variables; and to combinations of these specifications. Furthermore, with the exception of Scenario 8 (where F-test $p = 0.086$), the result holds in each individual scenario.²⁵ All this suggests that not all the marginal utilities $\frac{\partial U}{\partial X}$ are zero, even if the first-order approximation is imperfect.

Moving from testing the null hypothesis to interpreting the magnitudes of coefficients requires additional assumptions—both standard econometric assumptions and psychological ones. Econometrically, for example, if X includes aspects we did not measure, the coefficients might be biased due to omitted

²⁴For a more thorough treatment of our empirical framework within this simple model, see BHKR.

²⁵See tables A7-A10 in BHKR for these and other specifications. Table A10 shows that this result holds by scenario even when the regressions include only aspects for which more than a trivial fraction of respondents (e.g. 15 percent) indicate answers other than “no difference.” In other words, it holds even when we include only the most reliably-estimated coefficients. Interestingly, table A10 shows that the only large and robust non-SWB coefficient in the “apple vs. orange” scenario is that on “physical comfort”; this seems consistent with the *de gustibus* interpretation of this scenario.

variables. Psychologically, the coefficients are comparable only if respondents respond to the seven-point scales similarly across the twelve aspects.

Comparing the coefficients in the third column of table 2.3, the coefficient on own happiness is by far the largest. A one-point increase in our seven-point measure of predicted SWB is associated with a highly significant 0.46-point increase in our six-point choice measure. After own happiness, the largest coefficients are on sense of purpose (0.12), control over one's life (0.08), family happiness (0.08), and social status (0.06). The relative sizes of the coefficients are similar in alternative specifications (e.g., the ordered probit column), but remember that the data are pooled across surveys that use two opposite orders in which aspects are presented, and order matters for the coefficient estimates (see section 2.5). While the rejection of $\frac{\partial U}{\partial X} = 0$ suggests that own SWB is not the only argument in the "hypothetical-choice utility function," a comparison of the coefficients suggests that the marginal utility of own happiness is several times larger than the marginal utilities of even the most significant among the other aspects we measure.²⁶

2.4.4 Measurement error

Measurement error in our measures of own happiness and the other aspects will bias the coefficient estimates and potentially also invalidate our test of the null hypothesis $\frac{\partial U}{\partial X} = 0$. In order to address these concerns, we collected repeated observations on a sub-sample (of 230) of our Cornell respondents. This enables

²⁶However, we believe that the most plausible bias from unmeasured factors exaggerates the coefficient on own happiness. In particular, an unmeasured factor whose effect on H has the same sign as its direct effect (i.e., not through H) on U will bias upward the coefficient on own happiness.

us to estimate measurement-error-corrected regressions. In particular, we use Simulation-Extrapolation (SIMEX) (J. R. Cook and Leonard A. Stefanski, 1994), a semi-parametric method that assumes homoskedastic, additive measurement error but does not make assumptions about the distribution of the regressors.²⁷ As shown in table 2.3, relative to the OLS results, the SIMEX coefficient on own happiness increases, and remains by far the most predictive regressor. However, the other aspects with largest coefficients and statistical significance in the OLS regressions remain statistically significant and also increase, suggesting that our main results in this section are not due to measurement error.

2.5 Robustness

To examine the robustness of our results from sections 2.3 and 2.4, we conduct a long list of additional analyses. Full details, including all tables and statistics, are reported in BHKR. In this section we briefly summarize our findings. Unless stated otherwise, they are based on our within-subject data from either the Denver or Cornell samples.

Are results driven by only a few individuals? We find that most respondents (both in Denver and Cornell) exhibit at least one reversal and that very few exhibit reversals in half or more of the scenarios. Moreover, to explore whether

²⁷Intuitively, the SIMEX method proceeds in two steps. First, it simulates datasets with additional measurement error and uses them to estimate the function describing how the regression coefficients change with the amount of measurement error. Then the algorithm extrapolates in order to estimate what the coefficients would be if there were no measurement error in the original data. We choose this method over several more common measurement error correction methods (such as IV or regression disattenuation) for several reasons. Primarily, the other methods are much less efficient in this setting. Moreover, the SIMEX method is flexible in its treatment of the measurement error structure, it accommodates misclassified categorical data, and it easily accommodates non-linear models such as probit or ordered probit regressions. For additional discussion of SIMEX see BHKR, and for IV results see table A12 there.

some of the respondents who do not exhibit a choice-SWB reversal in a given scenario would have done so if that scenario's tradeoff between SWB and other factors had assigned a different "price" to SWB, some Denver respondents face three versions of Scenario 4 (legacy vs. income), with three different income levels in the income option (see details in the Appendix). Ninety-one percent of these respondents monotonically rank the income option higher in both choice and SWB as the amount of income increases. Of those, 22 percent exhibit a choice-SWB reversal for at least one income level, compared to an average of 12 percent reversals at a given income level. This suggests that the fraction of reversals we observe in other scenarios is a lower bound on the fraction who would exhibit a reversal in those scenarios with some "price of SWB."

Scenario-order effects and participant fatigue. We investigate the effects of scenario order on responses with our Denver sample, where respondents face the six scenarios in one of two opposite orders (see table 2.1). Scenario-order effects could arise, for example, due to increasing fatigue or boredom among respondents. While we indeed find evidence of scenario-order effects on response patterns, they do not systematically affect the degree of choice-SWB concordance we find.

Respondents' explanations for their choice-SWB reversals. After our Cornell respondents finish responding to all the decision scenarios, we directly ask all of them additional questions, including: whether any choice-SWB reversals they might have made were a mistake (only 7 percent respond "Yes"); whether they think they would regret any choice-SWB reversal they might have made (23 percent respond "Yes"); and whether they were trying to make their choice and SWB responses consistent (20 percent respond that they were). Our results from

section 2.4 remain largely the same when the analysis excludes groups of respondents based on their responses to these questions. We also ask respondents to explain their reasoning for any choice-SWB reversals, and we view the resulting qualitative data as roughly consistent with our main results.²⁸

Self-control. To assess whether choice-SWB reversals merely reflect a self-control problem (as in David Laibson, 1997), in addition to asking participants what they would choose, we also ask some of them what they would want themselves to choose (the meta-choice question mentioned in Section 2.2.1). Aggregating across all surveys that include the meta-choice question (see table 2.1), we find reversals between choice and meta-choice in 28 percent of the cases. While self-control problems may be relevant in these cases, our main conclusions from section 2.4 are robust to either excluding these observations or to replacing choice with meta-choice as the dependent variable.

Context of choice, SWB, and other-aspect questions. Respondents' interpretations of the questions or their understanding of the meaning of the related concepts may be context-dependent.²⁹ As mentioned in sections 2.2 (see table 2.1) and 2.3, different versions of our surveys vary in whether the choice and SWB questions are asked close together or far apart, and in the order the questions are asked; they also vary in the distance between own happiness and the other eleven aspects, and in the order of the aspects. Repeating our analysis in section 2.4 by questionnaire organization indicates that order and context effects do indeed matter. For example, aspects listed earlier have larger coefficients,

²⁸For example, many respondents mention tradeoffs between their own happiness and the happiness of family and friends, or mention tradeoffs between short-lived happiness and goals like long-term career success.

²⁹Notice the important difference between this possibility and the possibility of cross-respondent differences in the interpretations or understanding of the *scenarios*. The latter possibility is a lesser concern as long as a respondent's interpretation or understanding of a scenario remains the same across the choice and SWB questions.

and own happiness as part of a twelve-aspects list has a smaller coefficient than as a stand-alone question. Yet, in all designs, aspects other than own happiness are statistically significant, and the coefficient on own happiness has the highest point estimate among the aspects.

2.6 Heterogeneity in choice-SWB concordance

We have thus far focused on characterizing the average concordance between our choice and SWB measures. However, the averages mask substantial heterogeneity: across our questionnaires (see table 2.1) and scenarios, choice-SWB coincidence ranges from well below 50 percent to above 95 percent. To provide information that may be useful for researchers and policy makers, we conduct our main analysis separately across SWB measures, scenarios, and respondent characteristics. This section briefly summarizes a more thorough treatment in BHKR.

2.6.1 Comparing SWB measures

We compare how well our different SWB question variants predict choice by comparing R^2 's from univariate OLS regressions of our multiple-point choice variable on each of our multiple-point SWB measures. As in section 2.4, we demean our variables at the scenario level. In the Denver sample, the life satisfaction question variant (i) is the best predictor of the choice question, with $R^2 = 0.65$. Happiness with life as a whole (ii) and felt happiness (iii) come second and third, respectively, with $R^2 = 0.59$ and 0.55 . The felt happiness R^2 is statis-

tically significantly lower than the life satisfaction R^2 ($p = 0.02$ calculated using bootstrapped standard errors), and the R^2 for happiness with life as a whole is not statistically distinguishable from the other two. In the Cornell sample, own happiness with life as a whole (iv) and immediately felt own happiness (v) have $R^2 = 0.39$ and 0.37 , not statistically distinguishable from each other.

These R^2 's and our findings in 2.3.1 paint a consistent picture. While in the Denver data the life-satisfaction-type SWB question is more predictive of choice than the happiness-type SWB questions, in both Denver and Cornell the felt happiness and the happiness with life as a whole questions predict choice similarly. On the evaluative-versus-affective spectrum of SWB measures (see 2.2.3 above), these results lend partial support to the notion that more evaluative measures may generate rankings more similar to hypothetical choice.³⁰

2.6.2 Comparing scenarios

For applied work, it is useful to know in which situations the assumption that people's choices maximize their predicted SWB is a better or worse approximation. Table 2.4 shows the benchmark OLS specification from table 2.3, conducted separately for each of the ten scenarios in the Cornell data. The "Incremental R^2 " row reports the difference between the R^2 's from the reported multivariate regressions and R^2 's from univariate regressions of choice on only own happiness (which are not reported).

³⁰One possible hypothesis as to why some SWB measures are better predictors of choice is that they induce participants to more accurately report the present value of instantaneous SWB flows over time. Our data do not allow us to directly test this hypothesis because we have no direct evidence on how respondents aggregate SWB over time. However, our finding that variant (v)—about happiness "in the few minutes immediately after making the choice"—is as predictive of choice as variant (iv)—about happiness in "life as a whole"—seems inconsistent

Table 2.4: OLS regressions of choice on all aspects of life, by scenario

<u>Choice Scenario</u>	All scenarios pooled	1 Sleep vs Income	2 Concert vs Birthday	3 Abs. Inc. vs Rel. Inc.	4 Legacy vs Income	5 Apple vs Orange
Own happiness	0.46*** (0.010)	0.38*** (0.031)	0.44*** (0.031)	0.52*** (0.032)	0.44*** (0.031)	0.73*** (0.036)
Family happiness	0.08*** (0.015)	0.07* (0.032)	0.01 (0.071)	0.16*** (0.046)	0.05 (0.041)	0.16 (0.159)
Health	0.00 (0.019)	-0.05 (0.055)	-0.07 (0.076)	-0.11 (0.077)	-0.04 (0.058)	0.05 (0.065)
Life's level of romance	-0.01 (0.021)	0.08 (0.059)	-0.02 (0.064)	0.07 (0.078)	-0.00 (0.066)	-0.67** (0.228)
Social life	-0.03 (0.018)	-0.02 (0.055)	0.02 (0.043)	-0.01 (0.056)	0.00 (0.058)	0.02 (0.225)
Control over your life	0.08*** (0.015)	0.02 (0.042)	0.05 (0.053)	0.04 (0.056)	0.08* (0.039)	-0.00 (0.093)
Life's level of spirituality	-0.02 (0.021)	-0.04 (0.049)	-0.00 (0.061)	-0.16 (0.090)	0.13* (0.055)	0.31 (0.221)
Life's level of fun	0.05* (0.018)	0.06 (0.042)	0.15** (0.051)	0.04 (0.066)	0.05 (0.047)	-0.08 (0.127)
Social status	0.06*** (0.014)	-0.00 (0.036)	0.04 (0.045)	0.05 (0.040)	0.04 (0.036)	-0.27 (0.227)
Life's non-boringness	-0.01 (0.017)	0.05 (0.037)	-0.03 (0.054)	0.22** (0.078)	-0.01 (0.047)	0.09 (0.121)
Physical comfort	0.04** (0.014)	0.09* (0.036)	0.00 (0.060)	-0.05 (0.054)	0.00 (0.042)	0.21** (0.066)
Sense of purpose	0.12*** (0.013)	0.17*** (0.038)	0.12** (0.047)	0.12** (0.044)	0.12** (0.041)	0.29* (0.119)
Observations	6217	615	621	620	624	624
R ²	0.41	0.46	0.43	0.53	0.41	0.58
Incremental R ²	0.03	0.06	0.03	0.04	0.04	0.02

<u>Choice Scenario</u>	6	7	8	9	10
	Money	Socialize	Family	Education	Interest
<i>For exact phrasing, see Appendix</i>	vs Time	vs Sleep	vs Money	vs Social life	vs Career
Own happiness	0.53*** (0.036)	0.31*** (0.032)	0.53*** (0.033)	0.35*** (0.029)	0.27*** (0.030)
Family happiness	0.15* (0.059)	-0.09 (0.053)	0.05 (0.050)	0.14*** (0.037)	0.21*** (0.041)
Health	0.06 (0.075)	0.18*** (0.054)	0.05 (0.057)	-0.03 (0.044)	-0.06 (0.063)
Life's level of romance	-0.10 (0.086)	0.02 (0.054)	-0.03 (0.068)	0.01 (0.053)	0.01 (0.072)
Social life	0.04 (0.071)	-0.00 (0.065)	-0.05 (0.053)	-0.04 (0.053)	0.01 (0.054)
Control over your life	0.07 (0.052)	0.15*** (0.043)	0.05 (0.049)	0.06 (0.038)	0.07* (0.035)
Life's level of spirituality	-0.15 (0.091)	-0.01 (0.076)	-0.15* (0.062)	-0.00 (0.054)	-0.01 (0.068)
Life's level of fun	0.13 (0.068)	-0.03 (0.073)	0.03 (0.059)	0.06 (0.057)	-0.00 (0.057)
Social status	-0.01 (0.061)	0.06 (0.059)	0.11 (0.060)	0.06* (0.029)	0.16*** (0.043)
Life's non-boringness	-0.03 (0.060)	0.18** (0.062)	-0.05 (0.061)	-0.02 (0.055)	0.05 (0.055)
Physical comfort	-0.00 (0.049)	0.05 (0.048)	-0.10* (0.041)	0.06 (0.040)	-0.02 (0.049)
Sense of purpose	0.05 (0.050)	0.04 (0.044)	0.09* (0.046)	0.17*** (0.037)	0.17*** (0.029)
Observations	619	622	625	626	621
R^2	0.42	0.32	0.38	0.43	0.37
Incremental R^2	0.02	0.07	0.02	0.08	0.13

Notes: Standard errors in parentheses. OLS regressions of 6-point choice on 7-point aspects of life. Based on 633 Cornell respondents. The left-most column aggregates data across choice scenarios; each of the other columns corresponds to a specific scenario. Each observation is a respondent's choice and aspect ratings for one scenario; there are 10 observations per respondent corresponding to the 10 scenarios in the questionnaires. All variables are demeaned at the scenario level, generating coefficients equivalent to those generated by including scenario fixed effects. "Incremental R^2 " is the difference in R^2 between the reported multivariate regression and a univariate regression of choice on own happiness; it represents the increased percentage of variation in choice that can be explained by including the additional aspects. Observations with missing data in any variable are excluded from the regression. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As discussed above, Scenario 5 (apple vs. orange)—which was designed to minimize choice-SWB reversals—has little variance in aspects other than own SWB and the fewest reversals (see figure 2.1 and table 2.2). As expected, the R^2 in a univariate regression of choice on SWB is the highest (at 0.56) in Scenario 5, and the incremental R^2 from adding all other aspects is the lowest (at 0.02). If this type of minor decision—which possibly comprises most decisions in life—generally features low variance in aspects other than own SWB, then the assumption that people’s choices maximize their predicted SWB might be a good approximation in such settings.

Interestingly, at the other extreme, the four scenarios we designed to be representative of typical important decisions (see section 2.2.2) facing our college Cornell sample—Scenarios 7-10 (socialize vs. sleep, family vs. money, education vs. social life, and interest vs. career)—are among the scenarios with the lowest univariate R^2 and, correspondingly, the highest incremental R^2 . Indeed, in Scenarios 7 and 10, where univariate R^2 is the lowest—at 0.25 and 0.24, respectively—incremental R^2 is 0.07 and 0.13. Here, the additional eleven aspects increase predictive power (as measured by R^2) substantially, by 28 and 54 percent. This in turn suggests that one should be especially cautious in assuming that people’s choices maximize their predicted SWB in empirical applications that focus on important life decisions.

The rest of the scenarios lie somewhere in between. They include the scenarios that were designed to explore common themes from the happiness literature and, surprisingly, those designed as situations where we most expected to find tensions between SWB and other factors.

with this view.

2.6.3 Comparing respondents

Across the Denver, CNSS, and Cornell samples, we find relatively little evidence for differences in the frequency of choice-SWB reversals across demographics that include gender, age, race, education, and income, with the exception that in the Cornell sample, black respondents are more likely than others to exhibit reversals. In the Cornell within-subject sample, we measured the “Big 5” personality traits using Oliver P. John and Sanjay Srivastava’s (1999) BFI scale. We find that a one standard deviation increase in conscientiousness is associated with a 2 percent lower likelihood of a reversal, while a one standard deviation increase in neuroticism (i.e., moody, tense) is associated with a 2 percent higher likelihood of a reversal.

2.7 Discussion

Throughout this paper, we have remained agnostic as to which survey question, if any, best elicits a person’s preferences.³¹ However, if one assumes that hypothetical choices reveal preferences, our results could help in reconciling two opposing theoretical views regarding the relationship between SWB data and preferences. The first, reflected at least implicitly in much of the economics of happiness literature, is that SWB data represent idealized revealed-preference utility in the sense of what individuals would choose if their predictions of the

³¹Note that while economists often assume that incentivized choice reveals preferences, which in turn defines economic welfare, psychologists instead often equate experienced SWB with welfare (see, e.g., Kahneman, Wakker, and Sarin, 1997). Taking this latter perspective, Hsee (1999) and Hsee et al. (2003) interpret reversals between hypothetical choice and predicted SWB as evidence of mistakes in choice behavior. For careful discussions of the appropriate notion of welfare, see, e.g., Tversky and Griffin (1991), Loewenstein and Ubel (2008), and Fleurbaey (2009).

SWB consequences of their choices were not biased. The second view, explicitly laid out in, e.g., Kimball and Robert Willis (2006) and Becker and Rayo (2008), is that even well-informed agents will be willing to trade off SWB for other things they care about, making SWB and preferences distinct. Our results suggest that people do not seek to maximize SWB exclusively, at least as it is currently measured, but that SWB is a uniquely important argument of the utility function.

Since hypothetical choices maximize predicted SWB (especially “life satisfaction”) for most of our respondents in most of our scenarios, our results might be interpreted as comforting for applied researchers who use SWB as a proxy for utility. We caution that the amount of choice-SWB concordance we find overstates the justification to treat SWB as a proxy for utility; applications always require additional assumptions that we do not test. For example, typical assumptions are that SWB measures are comparable and can be aggregated across individuals.³²

When comparing scenarios, our results suggest that, first, researchers should be especially cautious when interpreting SWB data as indicating what well-informed individuals would choose in settings that are perceived by those individuals to involve personally-important decisions. Second, in settings where one alternative involves higher income or more money, our survey respondents are systematically more likely to choose the money alternative than they are likely to predict it will yield higher SWB. Unless this systematic gap is sufficiently negatively correlated with the difference between the predicted-

³²Our results may also overstate the extent to which standard SWB questions provide a good measure of preferences because standard questions are asked absolutely (“How satisfied are you with your life?”), while our SWB questions are asked relatively (“Between these two options, which do you think would make you more satisfied with life?”). Different individuals may apply different scales to a greater extent for an absolute measure, making it more difficult to translate an absolute SWB measure into a meaningful utility number than might be suggested by our results.

experienced SWB gap and the hypothetical-incentivized choice gap, this finding in turn suggests that the increasingly-common practice of estimating implicit willingness to pay for non-market goods by comparing the coefficient on income with that on another variable in multivariate SWB regressions may bias these estimates upwards relative to incentivized-choice-based estimates.

Our scenario-based methodology could be usefully applied in several new directions. First, the method of assessing choice-SWB correspondence could be used to assess new SWB measures that might predict hypothetical choice better than existing SWB measures. Our findings suggest that responses to existing measures do not fully capture the weight that factors such as sense of purpose have in explaining choice. Additionally, existing SWB measures may primarily reflect current feelings, rather than also reflecting anticipated future SWB flows. In BHKR we describe pilot data we collected on two novel measures aimed at addressing these issues, neither of which appears to predict choice any better than existing measures. Nonetheless, developing new measures seems an especially promising area for further research.³³

Second, our method could be used to provide more tailored guidance for applied work by asking about scenarios that are intended to address specific issues of interest. To illustrate this point, we pilot four such scenarios at the end of our Cornell repeat-survey. For example, to reconcile the intuition that Americans today are better off than in the past with the finding that average SWB has remained flat in the U.S. over the past decades (Richard A. Easterlin, 1974, 1995; see Betsey Stevenson and Justin Wolfers, 2008, for a recent assess-

³³Since different SWB questions seem to capture distinct dimensions of well-being that correlate differently with income and other variables (e.g., Kahneman and Deaton, 2010), future research could also explore whether a combination of SWB questions predicts choice better than any individual SWB question alone (including ladder- or mountain-type SWB questions, which we do not study in this paper).

ment), we ask respondents to rank being born in 1950 versus being born in 1990 in both choice and SWB questions. Although our respondents overwhelmingly favor being born in 1990 in both questions, more choose 1990 despite believing that they would be happier in 1950 than the reverse. This result indeed suggests that some people prefer being born later even if it does not make them happier. For another example, to reconcile the intuition that expanding political and economic freedoms for women have made women better off with the finding that average SWB among women has declined in the U.S. since the 1970s, both absolutely and relative to men (Stevenson and Wolfers, 2009), we ask respondents to rank living in a world with or without these expanded freedoms for women. Again, significantly more respondents choose a world with these expanded freedoms for women in spite of believing that a world without them would make them happier than the reverse. For further examples and full details, see BHKR.

Finally, some researchers have attempted to identify the key non-SWB aspects of life that are associated with greater welfare (e.g., Sen, 1985). Others have called for an SWB-based “national well-being index” to provide a measure of welfare that captures factors not represented in economic indicators such as GDP (e.g., Diener et al., 2009). To the best of our knowledge, our paper is the first attempt to empirically estimate weights on SWB and other factors in a way that might be useful for combining them into an overall index for predicting what individuals themselves would choose. If hypothetical choices are assumed to reveal preferences that are relevant for evaluating welfare, then our method could be applied more systematically for the purpose of developing a well-being index.

2.8 Works cited

- Becker, Gary S., and Luis Rayo.** 2008. "Comment on 'Economic Growth and Subjective well-being: Reassessing the Easterlin Paradox' by Betsey Stevenson and Justin Wolfers." *Brookings Papers on Economic Activity*, Spring: 88-95.
- Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones.** 2010. "Do People Seek to Maximize Happiness? Evidence from New Surveys." NBER Working Paper 16489.
- Cook, J. R., and Leonard A. Stefanski.** 1994. "A Simulation Extrapolation Method for Parametric Measurement Error Models." *Journal of the American Statistical Association*, 89: 1314-1328.
- Deaton, Angus, Jane Fortson, and Robert Tortora.** 2010. "Life (Evaluation), HIV/AIDS, and Death in Africa." In *International Differences in Well-Being*, ed. Ed Diener, John Helliwell, and Daniel Kahneman, 105-136. Oxford: Oxford University Press.
- Diener, Ed, and Christie Scollon.** 2003. "Subjective Well-Being is Desirable, but not the Summum Bonum." University of Minnesota, Workshop on Well-Being. Minneapolis, MN.
- Diener, Ed, Richard Lucas, Ulrich Schimmack, and John Helliwell.** 2009. *Well-Being for Public Policy*. New York: Oxford University Press.
- Di Tella, Rafael, Robert J. MacCulloch, and Andrew J. Oswald.** 2003. "The Macroeconomics of Happiness." *Review of Economics and Statistics*, 85(4): 809-827.

- Easterlin, Richard A.** 1974. "Does Economic Growth Improve the Human Lot? Some Empirical Evidence." In *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, ed. Paul A. David and Melvin W. Reder. New York: Academic Press, Inc.
- Easterlin, Richard A.** 1995. "Will Raising the Incomes of All Increase the Happiness of All?" *Journal of Economic Behavior and Organization*, 27: 35-47.
- Finkelstein, Amy, Erzo F.P. Luttmer, and Matthew J. Notowigdo.** 2008. "What Good is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption." NBER Working Paper 14089.
- Fleurbaey, Marc.** 2009. "Beyond the GDP: The Quest for a Measure of Social Welfare." *Journal of Economic Literature*, 47: 1029-1075.
- Frey, Bruno S., Simon Luechinger, and Alois Stutzer.** 2009. "The Life Satisfaction Approach to Valuing Public Goods: The Case of Terrorism." *Public Choice*, 138: 317-345.
- Gilbert, Daniel T.** 2006. *Stumbling on Happiness*. New York: Knopf.
- Heffetz, Ori, and Robert H. Frank.** 2011. "Preferences for Status: Evidence and Economic Implications." In *Handbook of Social Economics*, ed. Jess Benhabib, Alberto Bisin, and Matthew Jackson, Vol. 1A, 69-91. The Netherlands: North-Holland.
- Hsee, Christopher K.** 1999. "Value-Seeking and Prediction-Decision Inconsistency: Why Don't People Take What They Predict They'll Like the Most?" *Psychonomic Bulletin and Review*, 6(4): 555-561.

- Hsee, Christopher K., Jiao Zhang, Fang Yu, and Yiheng Xi.** 2003. "Lay Rationalism And Inconsistency Between Predicted Experience and Decision." *Journal of Behavioral Decision Making*, 16: 257-272.
- Hsee, Christopher K., Reid Hastie, and Jingqui Chen.** 2008. "Hedonomics: Bridging Decision Research with Happiness Research." *Perspectives on Psychological Science*, 3(3): 224-243.
- John, Oliver P., and Sanjay Srivastava.** 1999. "The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives." In *Handbook of Personality: Theory and Research* (2nd ed), ed. Lawrence A. Pervin and Oliver P. John, 102-138. New York: Guilford.
- Kahneman, Daniel, and Angus S. Deaton.** 2010. "High Income Improves Evaluation of Life but not Emotional Well-Being." *Proceedings of the National Academy of Sciences*, 107(38): 16489-16493.
- Kahneman, Daniel, Alan B. Krueger, David A. Schkade, Nobert Schwarz, and Arthur Stone.** 2004. "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method." *Science*, 306: 1776-1780.
- Kahneman, Daniel, Peter P. Wakker, and Rakesh Sarin.** 1997. "Back to Bentham? Explorations of Experienced Utility." *Quarterly Journal of Economics*, 112(2): 375-405.
- Kelly, William E.** 2004. "Sleep-Length and Life Satisfaction in a College Student Sample." *College Student Journal*, 38(3): 428-430.
- Kimball, Miles, and Robert Willis.** 2006. "Happiness and Utility." <http://www.personal.umich.edu/~mkimball/pdf/uhap.pdf>.

- Laibson, David.** 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112(2): 443-477.
- Liddell, Douglas K.** 1983. "Simplified Exact Analysis of Case-Referent Studies: Matched Pairs; Dichotomous Exposure." *Journal of Epidemiology and Community Health*, 37(1): 82-84.
- Loewenstein, George, Ted O'Donoghue, and Matthew Rabin.** 2003. "Projection Bias In Predicting Future Utility." *The Quarterly Journal of Economics*, 118(4): 1209-1248.
- Loewenstein, George, and Peter A. Ubel.** 2008. "Hedonic Adaptation and the Role of Decision and Experience Utility in Public Policy." *Journal of Public Economics*, 92(8-9): 1795-1810.
- Luechinger, Simon, and Paul A. Raschky.** 2009. "Valuing Flood Disasters Using the Life Satisfaction Approach," *Journal of Public Economics*, 93(3-4): 620-633.
- Luttmer, Erzo F.P.** 2005. "Neighbors as Negatives: Relative Earnings and Well-Being." *Quarterly Journal of Economics*, 120(3): 963-1002.
- Nozick, Robert.** 1974. *Anarchy, State, and Utopia*. New York: Basic Books.
- Nussbaum, Martha.** 2000. *Women and Human Development: The Capabilities Approach*. Cambridge: Cambridge University Press.
- Perez-Truglia, Ricardo.** 2010. "A Samuelsonian Validation Test for Happiness Data," <http://ssrn.com/abstract=1658747>.
- Ryff, Carol D.** 1989. "Happiness is Everything, or Is It? Explorations on the

Meaning of Psychological Well-Being." *Journal of Personality and Social Psychology*, 57, 1069-1081.

Sen, Amartya. 1985. *Commodities and Capabilities*. Oxford: Oxford University Press.

Stevenson, Betsey, and Justin Wolfers. 2008. "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox." *Brookings Papers on Economic Activity*, Spring: 1-87.

Stevenson, Betsey, and Justin Wolfers. 2009. "The Paradox of Declining Female Happiness." *American Economic Journal: Economic Policy*, 1(2): 190-225.

Stutzer, Alois and Bruno S. Frey. 2008. "Stress that Doesn't Pay: The Commuting Paradox," *Scandinavian Journal of Economics*, 110(2): 339-366.

Tversky, Amos, and Dale Griffin. 1991. "Endowments and Contrast in Judgments of Well-Being." In *Strategy and Choice*, ed. Richard J. Zeckhauser. Cambridge, MA: MIT Press. Reprinted in *Choices, Values, and Frames*, ed. Kahneman, Daniel, and Amos Tversky. Cambridge, UK: Cambridge University Press.

White, Matthew P., and Paul Dolan. 2009. "Accounting for the Richness of Daily Activities." *Psychological Science*, 20(8): 1000-1008.

2.9 Acknowledgements

A previous version of this paper circulated under the title "Do People Seek to Maximize Happiness? Evidence from New Surveys." We are extremely grate-

ful to Dr. Robert Rees-Jones and his office staff for generously allowing us to survey their patients and to Cornell's Survey Research Institute for allowing us to put questions in the 2009 Cornell National Social Survey. We thank Gregory Besharov, John Ham, Benjamin Ho, Erzo F. P. Luttmer, Michael McBride, Ted O'Donoghue, Matthew Rabin, Antonio Rangel, and Robert J. Willis for especially valuable early comments and suggestions, as well as four anonymous referees for suggestions that substantially improved the paper. We are grateful to participants at the CSIP Workshop on Happiness and the Economy, the NBER Summer Institute, the Stanford Institute for Theoretical Economics (SITE), the Lausanne Workshop on Redistribution and Well-Being, the Cornell Behavioral/Experimental Lab Meetings, and seminar audiences at Cornell, Deakin, Syracuse, Wharton, Florida State, Bristol, Warwick, Dartmouth, Berkeley, Princeton, Penn, RAND, and East Anglia for helpful comments. We thank Eric Bastine, Colin Chan, J.R. Cho, Kristen Cooper, Isabel Fay, John Farragut, Geoffrey Fisher, Sean Garborg, Arjun Gokhale, Jesse Gould, Kailash Gupta, Han Jiang, Justin Kang, June Kim, Nathan McMahon, Elliot Mandell, Cameron McConkey, Greg Muenzen, Desmond Ong, Mihir Patel, John Schemitsch, Brian Scott, Abhishek Shah, James Sherman, Dennis Shiraev, Elizabeth Traux, Charles Whittaker, Brehnen Wong, Meng Xue, and Muxin Yu for their research assistance. We thank the National Institute on Aging (grant P01-AG026571/01) for financial support.

This work was previously published as:

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones.

"What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *American Economic Review*, 2012, 102(5): 2083-2110.

CHAPTER 3

CAN MARGINAL RATES OF SUBSTITUTION BE INFERRED FROM HAPPINESS DATA? EVIDENCE FROM RESIDENCY CHOICES

Daniel J. Benjamin, Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones

Abstract: We survey 561 students from U.S. medical schools shortly after they submit their choice rankings over residencies to the National Resident Matching Program. We elicit (a) these choice rankings, (b) anticipated subjective well-being (SWB) rankings, and (c) expected features of the residencies (such as prestige). We find substantial differences between choice and anticipated-SWB rankings in the implied tradeoffs between the residency features. In our data, evaluative SWB measures (life satisfaction and Cantril's ladder) imply tradeoffs closer to choice than does affective happiness, and as close as do multi-question SWB indices. We discuss implications for using SWB data in applied work.

3.1 Introduction

The marginal rate of substitution (MRS) is the magnitude that characterizes preferences: as (minus) the slope of an individual's indifference curve, it quantifies the tradeoffs that individuals are willing to make. Traditionally, MRSs are estimated from choice data. Economists must resort to alternatives, however, in settings where the relevant choices are not observed (as is often the case when externalities, non-market goods, and certain government policies are involved) or where individuals' choices are likely to reflect mistakes. An increasingly-used alternative source of data is survey responses to subjective well-being (SWB) questions—most commonly, questions about respondents' happiness, life satisfaction, or life's ranking on a ladder. In a typical application, a SWB measure is regressed on respondents' quantities of a bundle of goods, and the ratio of the coefficients on two goods yields an estimate of the goods' rate of tradeoff that would leave SWB unchanged.¹ Under the assumption that the SWB measure proxies for utility—i.e., that the SWB measure is what individuals seek to maximize—the estimated tradeoff can be interpreted as the MRS between the two goods.

The purpose of this paper is to explore empirically the extent to which tradeoffs estimated from SWB data generate MRS estimates that reliably reflect in-

¹For example, in the domain of government policy, Di Tella, MacCulloch and Oswald (2001) focus on a life satisfaction question to estimate the tradeoff between inflation and unemployment. In the domain of externalities, a large literature on social comparisons uses a variety of SWB measures to estimate the MRS between own and others' income (for a recent review, see Clark, Frijters, and Shields, 2008). In the domain of non-market goods, Deaton, Fortson, and Tortora (2010) use a variety of SWB measures, including the Cantril self-anchoring scale, to study the implied value of life in sub-Saharan Africa by comparing the coefficient on losing a family member with the coefficient on income. SWB data have been similarly used to price, for example, noise (van Praag and Baarsma, 2005), informal care (van den Berg and Ferrer-i-Carbonell, 2007), the risk of floods (Luechinger and Raschky, 2009), air quality (Levinson, 2012), benefits of the Moving to Opportunity project (Ludwig et al., 2012), and the loss of family members (Oswald and Powdthavee, 2008).

dividuals' preferences.² To that end, we elicit: (a) choice rankings over a set of options, in a setting where choice arguably reveals preferences; (b) the anticipated SWB consequences of the different choice options; and (c) the expected quantities of the goods that comprise the relevant consumption bundle under each choice option. We estimate the tradeoffs between the goods implied by SWB and those implied by preferences, and we explore the relationship between them.³

While the literature estimates the tradeoffs implied by experienced SWB, it is crucial for our purposes to compare choice tradeoffs with anticipated SWB tradeoffs in order to hold constant the conditions (including information and beliefs) under which choice is made. That way, we can attribute differences between choice and anticipated SWB tradeoffs to SWB not fully capturing the importance of certain goods in preferences. In contrast, divergences between choice and experienced SWB tradeoffs could result, for example, from mispredictions at the time of choice (e.g., Loewenstein, O'Donoghue, and Rabin, 2003; Gilbert, 2006).⁴

²The literature reflects a wide range of views regarding the relationship between SWB and preferences. On one extreme, Di Tella, MacCulloch and Oswald (2001, p. 338) explicitly identify SWB measures with utility: "The estimation describes preferences themselves." Nearer the other extreme, Deaton, Fortson, and Tortora (2010) discuss "well-being" rather than preferences, and explicitly consider the possibility that "the methods based on self-reported well-being do not tell us what we want to know" (p. 128). Moreover, they repeatedly point out that their ladder question implies dramatically different tradeoffs compared with their affective questions and hence warn against using one SWB measure, or even a combination, as an exclusive guide. Committing to neither extreme, Frey and Stutzer (2002, p. 426) write in their JEL review: "Happiness is not identical to the traditional concept of utility in economics. It is, however, closely related... [it] is a valuable complementary approach... SWB can be considered a useful approximation to utility..."

³The finding that different SWB measures imply different tradeoffs (Deaton, Fortson, and Tortora, 2010; Kahneman and Deaton, 2010) already rules out the possibility that all SWB measures simultaneously reflect preference tradeoffs accurately. However, it does not answer the questions we study, namely to what extent certain measures and combinations of measures reflect preference tradeoffs.

⁴It is logically possible that, despite the differences we find between anticipated-SWB tradeoffs and choice tradeoffs, experienced-SWB tradeoffs would nonetheless coincide with choice tradeoffs. This possibility would require that while individuals deliberately deviate, at the mo-

In section 3.2 we describe the setting we study: graduating U.S. medical students' preference rankings over residency programs. These preference rankings submitted by students to the National Resident Matching Program (NRMP), combined with the preference rankings over students submitted by the residency programs, determine which students are matched to which programs. This setting has a number of attractive features for our purposes: the matching mechanism is designed to be incentive-compatible; the choice is a deliberated, well-informed, and important career decision; the choice set is well-defined and straightforward to elicit; and due to a submission deadline, there is an identifiable moment in time when the decision is irreversibly made. We conduct a survey among a sample of 561 students from 23 U.S. medical schools shortly after they submit their residency preferences to the NRMP, so that our survey is conducted under the same conditions as the actual choice.

Section 3.3 describes our sample and survey design. We ask about each student's four most-preferred residency programs. In addition to eliciting each student's preference ranking over the four residencies as submitted to the NRMP, we also elicit her anticipated SWB rankings over the residencies, both during the residency and for the rest of her life. We focus on three commonly-used SWB measures: happiness, life satisfaction, and a Cantril-ladder measure.⁵ We

ment of making the choice, from choosing what they believe would maximize their SWB, their (realized) experienced SWB systematically differs from their anticipated SWB in a way that happens to exactly cancel out those deviations. We assume away this unlikely possibility.

⁵Examples of each of these three measures include: the National Survey of Families and Households question "Taking things all together, how would you say things are these days?" whose seven-point response scale ranges from "very unhappy" to "very happy" (e.g., Luttmer, 2005); the Euro-barometer survey question "On the whole, are you very satisfied, fairly satisfied, not very satisfied or not at all satisfied with the life you lead?" (used by, e.g., Di Tella, MacCulloch and Oswald, 2001); and the Gallup World Poll question "Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" (e.g., Deaton, Fortson, and Tortora, 2010).

also ask each student to rate each of the four residencies on each of nine features that we expected—based on our past research (in settings other than residency choice) as well as on conversations with medical school officials and with past and present students—to be the most important determinants of program choice.⁶ These include the desirability of residency location, residency prestige-and-status, expected stress level, future career prospects, and future employment opportunities.

Section 3.4 reports our analyses and results. We model residencies as bundles of attributes, and we use the choice- and SWB-rankings as alternative dependent variables in regressions where the independent variables are students' beliefs about these attributes. In our main analysis we compare the coefficients and coefficient ratios across regressions.

While we find that the coefficients of the attributes do not reverse sign, and are reasonably highly correlated, across the choice and SWB regressions, we find large and significant differences in the implied tradeoffs. For example, relative to the choice-based estimates, all anticipated-SWB estimates underweight residency prestige-and-status and residency desirability for the respondent's significant other, while overweighting social life and life seeming worthwhile during the residency. We also find that our evaluative SWB measures—life satisfaction and ladder—generally yield results closer to the choice-based estimates than our more affective happiness measure. Our results are robust to plausible forms of measurement error and biases in survey response and hold across empirical

⁶Indeed, as we report when analyzing the data, the residency attribute ratings that we elicited explain much of the within-respondent variation in residency choice rankings. In contrast, in our attempts to forecast residency choices in our data with objective, external measures such as characteristics of the city of residency and information from the Best Hospitals U.S. News Rankings, we find these measures to explain virtually none of the variation in choice (for one specification, see Web Appendix Table A11).

specifications and across subsets of our respondents.

We also explore whether multi-question SWB indices more accurately reflect revealed-preference tradeoffs. We consider three such indices: the first, a “3-SWB-measure” index, is a weighted sum of our three SWB questions; the second, a “4-period-happiness” index, consists of happiness predictions for four time intervals that together cover the rest of a respondent’s life; the third index combines the other two. While such indices have been much less commonly used to estimate tradeoffs, we are motivated by the ideas, respectively, that well-being is multidimensional (e.g., Stiglitz, Sen, and Fitoussi, 2009) and that well-being consists of instantaneous affect integrated over time (Kahneman, Wakker, and Sarin, 1997). We estimate the optimal weights of the indices as best linear predictors of choice in our data; our indices are hence constructed to perform better than those likely to be used in realistic applications, where choice data are not available. We find that while some tradeoffs based on these indices are closer to our choice-based MRSs than the tradeoffs based on the indices’ underlying questions, overall the indices do not reflect the MRSs more reliably than the single evaluative-SWB questions.

In section 3.5, we explore an alternative use of SWB data: assessing which of two concrete choice options is preferred. We find that despite the differences in implied tradeoffs between choice and SWB in our data, the two often coincide in pairwise comparisons. Because our survey elicits anticipated SWB after choice, concordance rates may be overstated in our data, while the tradeoff differences we find may be understated. We present a simple model that illustrates the relationship between pairwise predictions and tradeoffs, and we discuss the conditions under which SWB data may correctly predict choice even when the

implied tradeoffs differ.

We conclude in section 3.6. Additional results are available in the NBER working paper (Benjamin, Heffetz, Kimball, and Rees-Jones, 2013) and its accompanying web appendix, which will be referenced below.

Our work builds upon and differs from past attempts to study the relationship between choice and SWB measures in several important ways. First, while almost all existing work (Tversky and Griffin, 1991; Hsee, 1999; Hsee, Zhang, Yu, and Xi, 2003; Benjamin, Heffetz, Kimball, and Rees-Jones, 2012) compares anticipated-SWB rankings with choices that are either hypothetical or involve very small stakes, we present evidence on real, deliberated choices in a high-stakes field environment. Consequently, while in these prior studies it is unclear whether preferences are better reflected by choice or by anticipated SWB, in our setting there is a strong case for viewing choice as revealing preferences. Second, while these studies document cases where choices between pairs of options do not maximize anticipated SWB, we are the first to focus on the implications for estimating MRSs. Third, our evidence is from a setting where ordinal preferences over a well-defined and observable choice set are directly elicited. While preferences can sometimes be inferred indirectly—for example, as in Dolan and Metcalfe (2008), who, for pricing the welfare effects of an urban regeneration project, compare estimates based on contingent-valuation and hedonic-pricing methods with those based on SWB—such indirect approaches necessarily hinge on many maintained assumptions. Moreover, our paper is the first to study a field setting that allows the direct elicitation not only of preference orderings but also of anticipated-SWB rankings of the options in the choice set—an ideal setting for studying choice-SWB alignment. Fourth, while prior work considers

only single SWB questions, we also consider indices that include multiple SWB measures and multi-period affective happiness. Finally, drawing on theoretical considerations as well as on empirical results from this and previous papers, we offer guidance to applied researchers on appropriate uses of SWB data. While our findings suggest that such data are inadequate for precise inference regarding MRSs, their use in binary welfare comparisons may in some settings be on comparatively safer ground—although still subject to additional assumptions and caveats not studied in this paper (see, e.g., Adler, 2012).

3.2 Choice setting: the National Resident Matching Program (NRMP)

3.2.1 Background

After graduating from a U.S. medical school, most students enroll in a residency program. The residency is a three- to seven-year period of training in a specialty such as anesthesiology, emergency medicine, family medicine, general surgery, internal medicine, pediatrics, or psychiatry. Students apply to programs at the beginning of their fourth (and final) year. In late fall programs invite selected students to visit and be interviewed. Students subsequently submit to the NRMP their preferences over the programs where they interviewed, while programs submit their preferences over students. The NRMP determines the final allocation of students to residencies. In 2012, students were allowed to submit their preference ordering through the NRMP website between January 15 and February 22, and the resulting match was announced on March 16;

among students graduating from non-homeopathic U.S. medical schools, 16,875 submitted their preference, and 15,712 (93%) ended up getting matched (NRMP, 2012).

The matching algorithm, described in detail in Roth and Peranson (1999), was designed to incentivize truthful preference reporting from students and to generate stable matches (in which no student and program prefer to be matched to one another over their current matches). It is based on the student-proposing deferred acceptance algorithm of Gale and Shapley (1962), which is guaranteed to produce a stable match, and where truthful reporting is a weakly dominant strategy for students. The original, simple algorithm, however, could not accommodate certain requirements of the medical matching market (such as the requirement for couples to match to residencies in the same city). The modifications to the algorithm complicate the strategic incentives and allow the possibility that no stable match exists, but simulations in Roth and Peranson (1999) suggest that effectively all students remain incentivized to truthfully reveal their preferences.

3.2.2 Key features for our study

For our purposes, medical residency choices are an especially useful context for the following reasons:

Choice versus preferences: The NRMP setup may be as close as one can get to a setting where choices reveal preferences.⁷ Residency choice is arguably one

⁷Strictly speaking, what we refer to as our choice data are survey respondents' reports on choices; we do not directly observe the actual preference ranking submitted by students to the NRMP. However, these reports seem very reliable. Among the 131 respondents who completed

of the most important career-related decisions a medical student makes, with short- and long-term consequences for career path, geographic location, friendships, and family. Because of its importance, students deliberate over their decision for months and have a great deal of information and advising available to assist them in becoming well informed. Their submitted ranking is not visible to peers or residency programs, and hence, relative to many other decisions, the scope for strategic or signaling concerns is reduced. Finally and crucially, students are incentivized by the matching mechanism to report their true preference ranking.

Identifiable moment of choice: Unlike many other important life decisions, the NRMP submission has an identifiable moment when the decision is irreversibly committed. By surveying students shortly after they submit their preference ranking to the NRMP (and before they learn the match outcome), we elicit their SWB predictions under essentially the same information set and beliefs as at the moment of making the choice.

Identifiable choice set and ranking: Unlike other decisions where observable choice data consist of only the one chosen option, often with the econometrician being uncertain as to the exact choice set from which this option was chosen, here choice data consist of a ranking over a set of residencies. Therefore, we can elicit anticipated SWB and residency features over that same set of options. Also, observing a choice ranking over multiple options confers more statistical power than observing only which option was chosen from a set.

Intertemporal tradeoff: A residency is expected to be a period of hard work, both our original and repeat surveys (see section 3.3 below), only 2 (1.5%) reported conflicting choice data. (Of the remaining 129 respondents, 5 had cross-survey differences in missing choice data but no conflicts; 2 seemed to have made easily-correctible data-entry mistakes in either survey; and 122 reported the exact same choices across the two surveys.)

long hours, and intensive training, the benefits of which will be realized once the student becomes a practicing doctor. The investment aspect of the decision allows us to distinguish instantaneous utility from lifetime utility (the present discounted value of instantaneous utility); expected lifetime utility is what determines choice. Hence we can explore whether our affective SWB question—anticipated happiness during the residency—is better thought of as related to expected instantaneous utility or to expected lifetime utility. That distinction, which we consider and discuss in section 3.4.3, is crucial for evaluating potential applications of SWB data. Currently, most papers using a happiness question rely on implicitly assuming that it proxies for expected lifetime utility.

Heterogeneity in attribute evaluations: Residency choice offers rich variation in individuals' evaluations of programs' attributes: students' assessments of fit, locational preferences, and social considerations are all reasonably idiosyncratic. This heterogeneity, together with differences in choice sets (i.e., the sets of programs where different students had interviewed), is the source of variation identifying our regression coefficients.

One limitation of residency choice for our purposes is that it is not well suited for studying tradeoffs with money—the typical numeraire used in the literature. Our original intention was to use expected income for each residency to price the other residency attributes. However, in the process of designing the survey we learned—by being explicitly told by representatives of medical schools and by medical students we consulted—that expected income is largely unrelated to this decision. The primary determinant of expected income for medical students is their choice of *specialty*, a decision typically made years before choosing a residency. Indeed, most NRMP participants apply to residencies for a single

specialty and hence should not expect their future income to vary meaningfully across their top choices. While pricing residency attributes in dollars would have been convenient, it is by no means crucial for our purposes; we instead focus on comparing MRSs and tradeoffs between the attributes directly. We elicited expected income in our survey anyway but do not use it in this paper.⁸

3.3 Sample and survey design

3.3.1 Sample

From September 2011 to January 2012, we contacted virtually all 122 U.S. medical schools with full accreditation from the Liaison Committee on Medical Education by sending an email to a school representative (typically an Associate Dean of Student Affairs) and asking for permission to survey graduating medical students. We followed up with phone calls, further emails, and/or face-to-face meetings at the Association of American Medical Colleges Annual Meeting. As a result, 23 schools (19% of our initial list) agreed to participate in our study.⁹ These 23 represent a wide range of class sizes (from 60 to 299 students in 2011) and locations, and they graduated a total of 3,224 students in 2011. The survey appendix of Benjamin, Heffetz, Kimball, and Rees-Jones (2013) reproduces the

⁸Indeed, responses to our expected-income questions are of limited usefulness. Only 40% of respondents expect any income variation across the residencies in our two expected-income questions-compared with a range of 79-96% of respondents expecting variation in the nine expected-attribute questions. Moreover, looking at the correlations between responses to a given question by a given respondent across our two survey waves, responses to the expected-income questions are among the noisiest, having within-subject correlations of 0.00 and 0.24-compared with correlations in the range 0.24-0.81 in the nine expected-attribute questions.

⁹A common reason schools gave us for not participating was that their students are already asked to participate in “too many” surveys.

initial email sent to schools, lists the participating schools, their class sizes, and the numbers of their students starting vs. completing our survey.

Between February 22 at 9pm EST (the deadline for submitting residency preferences) and March 3, students in participating schools received an email from their school's dean, student council representative, or registrar, inviting them to respond to our web survey by clicking on a link. The email is reproduced in the survey appendix. It explained, among other things, that "The results of this study will provide better information on how medical students select residency programs, and can assist in the advising and preparation of future generations of students"; that the survey is estimated to take 15 minutes to complete; and that we offer participants at least a 1/50 chance to win an iPod.¹⁰ Reminder emails were sent near the March 3 deadline. When the survey closed, at 11:59pm EST that day, we had received 579 complete responses (approximately 18% of the roughly 3,224 students contacted).¹¹ Our analysis is based on the 561 who entered name and specialty information for at least two programs (540 of whom entered information for all four programs). Due to selection, our sample—while drawn from a diverse set of schools—is unlikely to be representative of U.S. medical students. Nonetheless, if MRSs could in general be inferred from SWB data, then we would expect the same to hold in our sample.

428 of our respondents agreed, when asked at the end of the survey, to be re-contacted. They received, on a randomly-drawn date between March 7 and

¹⁰At the end of the survey, participants were thanked for their participation; were reminded that they have a 1/50 chance to win an iPod; and were asked to encourage their classmates to also participate. As an incentive for the latter, they were informed that we would increase the individual chance to win an iPod to 3/50 in schools with response rate of 70% or higher (which no school reached).

¹¹In addition to the 579 complete responses, our survey had another 680 visits that did not result in a complete response. Of these, 284 (42%) exited before proceeding beyond the first page.

9, another email inviting them to participate in a repeat survey, with a March 11 deadline. The repeat survey consisted of the same questions as the original survey, with a few new questions added at the end. Comparing responses across these two waves allows us to assess the reliability of our measures, as we do below. 133 respondents completed the repeat survey, and 131 of them (23% of our main sample) provided information for at least two programs. The median time between completion of the original and the repeat surveys was 13 days.

3.3.2 Survey design

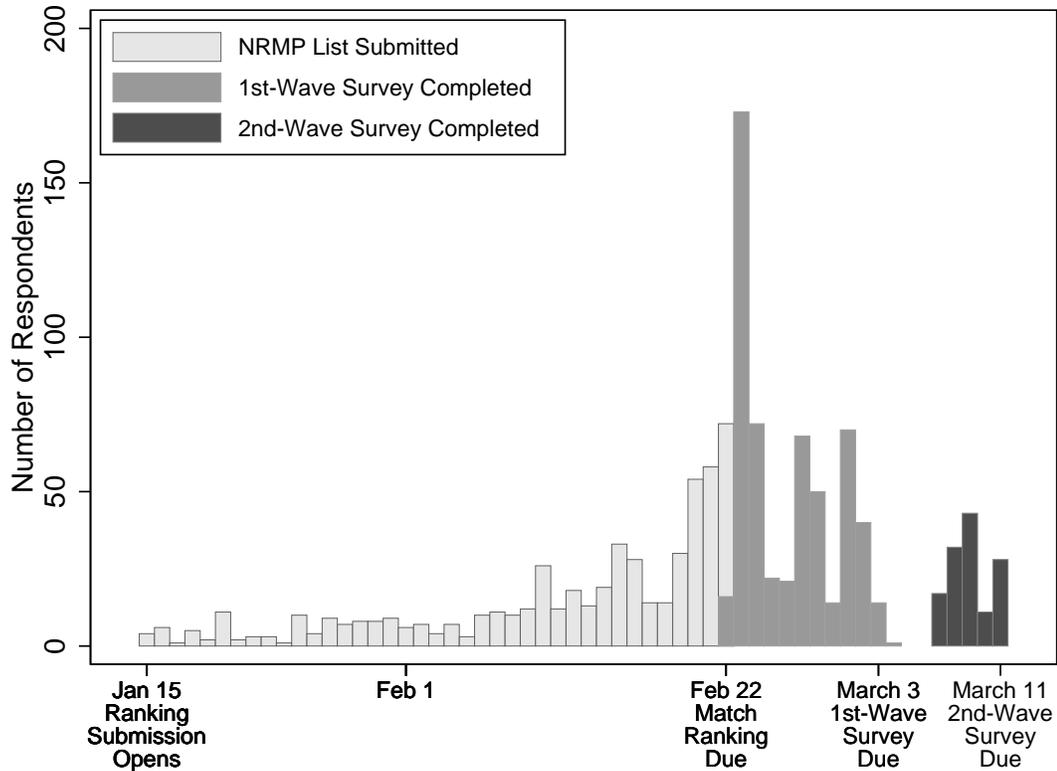
Our survey appendix provides screenshots of our survey. Here we briefly summarize important survey details. Following an introductory screen, respondents are asked: “Please enter the top four programs from the preference ordering you submitted to the NRMP.”¹² Respondents separately enter program (e.g., “Massachusetts General Hospital”) and specialty (e.g., “Anesthesiology”).

Respondents are then asked: “On what date did you submit your rank order list to the NRMP?” Figure 3.1 reports the distributions of submission dates (lighter bars) and survey response dates (gray bars) among our 561 main-sample respondents. The median number of days between choice submission and response to our survey is 11. The figure also shows the subsequent distribution of response dates for the 131 main-sample respondents who participated in our repeat survey (darker bars).

On the next screen, respondents are asked about their relationship status and

¹²While “the top four” is not the entire preference ordering, it is likely to be the relevant portion of the list for our respondents. In 2012, 83.6% of NRMP participants graduating from U.S. medical schools were matched to one of their top four choices. (First choice: 54.1%; second: 14.9%; third: 9.1%; fourth: 5.5%; NRMP, 2012).

Figure 3.1: Survey response timeline



Notes: Frequency distribution of survey responses by date. Each bar corresponds to one day. NRMP submission and 1st-wave data are for the 561 respondents in our main sample (with the exception that five respondents did not report their date of NRMP submission, and two reported invalid dates). 2nd-wave data are for the 131 respondents in the main sample who completed the repeat survey. The 1st-wave responses entered on February 22nd occurred after 9pm EST, the deadline for NRMP submission. On that date, where bars overlap, they are not stacked, and the longer bar continues behind the shorter bar.

whether they are registered with the NRMP for a “dual match.”¹³ Their answer to the relationship question determines whether the question “On a scale from 1 to 100, how desirable is this residency for your spouse or significant other?” will be included as a residency attribute on a later screen.¹⁴

Next, the following instructions appear on the screen:

For the following section, you will be asked to individually consider the top four programs you ranked. For each of these possibilities, you will be asked to report your predictions on how attending that residency program will affect a variety of aspects of your life. Please answer as carefully and truthfully as possible.

For some questions you will be asked to rate aspects on a 1-100 scale. Let 100 represent the absolute best possible outcome, 1 represent the absolute worst possible outcome, and 50 represent the midpoint.

The ranked residencies are then looped through in random order, and two screens appear for each residency. The first screen elicits respondents’ rating of the residency, using the 1-100 scale, on the main three anticipated-SWB questions and on the nine residency attributes. The second screen includes questions about expected income that we do not use in this paper.

Table 3.1 reproduces the three anticipated-SWB questions and the nine attribute questions as they appear on the first screen below the instruction: “Thinking about how your life would be if you matriculate into the residency

¹³The dual match is an option for couples trying to match to residencies simultaneously. The two submit a single list ranking pairs of programs. While 64% of our respondents indicate that they are either married or in a long-term relationship, only 7% are dual-match participants. As discussed in section 3.4.2, our main results are robust to excluding them.

¹⁴For respondents not in a relationship, this “desirable for significant other” residency-attribute variable is set to a constant in all regressions below. Doing so is appropriate since all our analysis is within-respondent, and since there is no cross-residency difference in the level of this variable for single respondents.

program in <specialty> at <program>, please answer the questions below.” The SWB and attribute questions are purposefully designed to resemble each other as much as possible in terms of language and structure, and they appear on the screen mixed together as twelve questions in random order. As a practical matter of survey design, this symmetric treatment allows us (in section 3.5 below) to compare the twelve questions on how useful each one is as a single predictor of choice, without confounds due to question language or order. Moreover, on a conceptual level it could be argued that the classification of questions as “SWB” versus “attribute” is in some cases arbitrary and has little basis in theory (a point that we return to in Section 3.6). Nonetheless, when planning our empirical strategy and prior to data collection, we set apart the three SWB questions to be compared with choice as dependent variables in regressions on the attributes (see Section 3.4 below), because in the happiness literature these are the questions that are routinely used as utility proxies.

Mixed together and arranged here roughly by the time interval they refer to, the twelve SWB and attribute questions include: three affective measures that refer to a typical day during the residency (in Table 3.1 these are labeled happiness, anxiety, and stress during residency); three evaluative/eudaimonic measures that refer more generally to the time during the residency (life satisfaction, social life, and worthwhile life during residency); one measure that refers implicitly to the time during the residency (desirability of location); one measure that refers implicitly to the time after the residency (future career prospects); one measure that simply refers to one’s “life” (ladder); and three measures that come with no specification of period (residency prestige and status, control over life, and—for respondents in a relationship—desirable for significant other).

Table 3.1: Main SWB and residency attribute survey questions

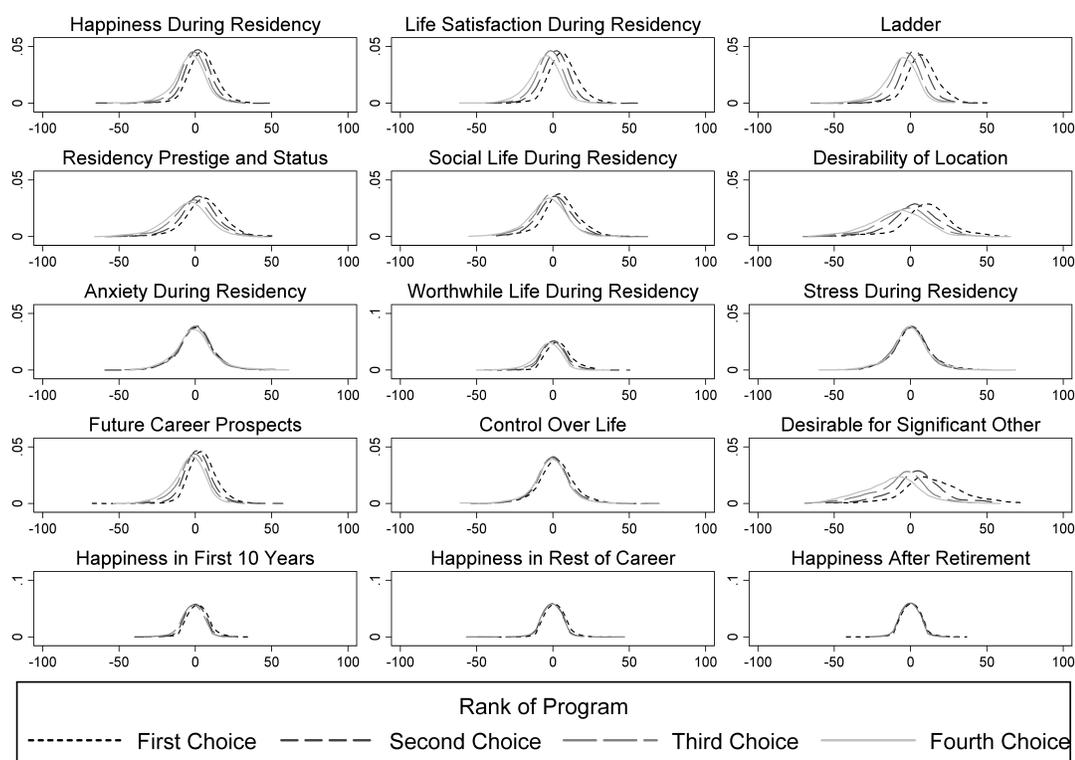
Variable label	Question prompt
	(beginning "On a scale from 1 to 100, ...")
Happiness during residency	...how happy do you think you would feel on a typical day during this residency?
Life satisfaction during residency	...how satisfied do you think you would be with your life as a whole while attending this residency?
Ladder	...where 1 is "worst possible life for you" and 100 is "best possible life for you" where do you think the residency would put you?
Residency prestige and status	...how would you rate the prestige and status associated with this residency?
Social life during residency	...what would you expect the quality of your social life to be during this residency?
Desirability of location	...taking into account city quality and access to family and friends, how desirable do you find the location of this residency?
Anxiety during residency	...how anxious do you think you would feel on a typical day during this residency?
Worthwhile life during residency	...to what extent do you think your life would seem worthwhile during this residency?
Stress during residency	...how stressed do you think you would feel on a typical day during this residency?
Future career prospects	...how would you rate your future career prospects and future employment opportunities if you get matched with this residency?
Control over life	...how do you expect this residency to affect your control over your life?
Desirable for significant other	...how desirable is this residency for your spouse or significant other?

Next, the top three residencies (rather than four, to keep the survey from becoming too long and repetitive) are cycled through again, in a new random order. For each residency we elicit anticipated happiness at different future time intervals (we provide more details when analyzing the resulting data, in section 3.4.3 below).

The survey concludes with a sequence of screens that include four questions regarding the relationship between a respondent's submitted NRMP ranking and her or his "true" preferences; a question regarding experiences with residency-program representatives' attempts at manipulating the match; and questions about gender, age, college GPA, MCAT score, and Medical Licensing Examination scores (for summary statistics, see Web Appendix Table A1). We explore these data in section 3.4.2 below. On the last screen, respondents are thanked for their participation and asked for permission to be contacted for the follow-up survey.

As a brief overview of our data, Figure 3.2 presents kernel density estimates of the distribution of our primary variables by residency rank (for means and standard deviations, see Web Appendix Table A2; for a version of Figure 3.2 demeaned at the respondent level, see Web Appendix Figure A1). As is visually clear, all have substantial variation across respondents, and many have clear differences in distribution across program ranks. For example, focusing attention on the three primary SWB measures (top row), it is clear that higher-ranked programs have higher mean anticipated SWB. Web Appendix Table A3 presents the test-retest correlations of these variables, as calculated with the repeat survey. We view the relatively high correlations of responses across waves as evidence that our survey measures elicit meaningful information.

Figure 3.2: Distributions of variables by program rank



Notes: Kernel density plots of residency attributes by preference order. (Epanechnikov; Bandwidth 5.) Based on the 561 respondents in the main sample.

3.4 Main analysis and results

3.4.1 Single SWB measures

As a first step in constructing choice-based and SWB-based tradeoff estimates, we estimate the associations of residency attributes with the choice-based and SWB-based residency rankings. The first four columns of Table 3.2 report four separate regressions of, respectively, choice, anticipated happiness, anticipated life satisfaction, and anticipated ladder questions on the nine residency at-

tributes. Each column estimates a rank-ordered logit model (Beggs, Cardell, and Hausman, 1981), which generalizes the standard binary-choice logit model to more than two ranked options. To avoid confusion, we emphasize that rank-ordered logit is different from ordered logit, an econometric technique commonly used in the happiness literature. When using rank-ordered logit, we assume that each individual i 's ordinal ranking of residencies, denoted by their rank $r \in \{1, 2, 3, 4\}$, is rationalized by a random latent index, $U_{ir} = \beta_X X_{ir} + \epsilon_{ir}$. The parameters of the latent index, β_X , are estimated by maximizing the sum of the individual-level log-likelihoods that $U_{i1} > U_{i2} > U_{i3} > U_{i4}$, the condition necessary for generating the observed ordering of residencies. The unobserved error term is assumed to follow a type I extreme value distribution, yielding a closed-form solution to the maximum-likelihood problem. We construct the regressors by dividing the attribute variables by 100 (so the regressors range from 0.01 to 1). The coefficients can be interpreted analogously to standard logit coefficients: for any pair of residencies A and B, all else equal, a one-unit increase in the difference in regressor j , $X_{(i,A,j)} - X_{(i,B,j)}$, is associated with a β_j increase in the log odds ratio of choosing A over B. We report a within-subject modification of McKelvey and Zavoina's R^2 , a statistic that measures the fraction of within-subject variation of the latent index explained by the fitted model.¹⁵

Consider Table 3.2's two leftmost columns ("Choice" and "Happiness during residency"). The first row indicates that the coefficient on residency prestige

¹⁵We modify the R^2 measure of McKelvey and Zavoina (1975) by demeaning the predicted index value \hat{U}_{ir} at the respondent level:

$$\frac{\hat{Var}(\hat{U}_{ir} - \bar{U}_i)}{\hat{Var}(\hat{U}_{ir} - \bar{U}_i) + Var(\epsilon_{ir})}$$

This ratio is the fraction of within-respondent variance in the latent index contributed by the estimated, deterministic component. The resulting measure of fit is intuitively similar to standard R^2 .

Table 3.2: Rank-ordered logit estimates: choice vs. anticipated SWB

	(1) Choice	(2) Happiness during residency	(3) Life satisfaction during residency	(4) Ladder	(5) 4-period- happiness index	(6) 3-SWB- measure index	(7) 6-SWB- question index
Residency prestige and status	2.5*** (0.3)	0.0 (0.3)	0.7* (0.3)	0.9** (0.4)	0.3 (0.4)	0.8** (0.3)	1.1** (0.4)
Social life during residency	1.6*** (0.3)	3.3*** (0.4)	2.7*** (0.4)	3.2*** (0.4)	2.6*** (0.4)	3.6*** (0.3)	3.5*** (0.5)
Desirability of location	1.7*** (0.2)	0.4* (0.2)	1.7*** (0.3)	1.9*** (0.3)	0.5* (0.3)	1.9*** (0.2)	1.6*** (0.3)
Anxiety during residency	-0.3 (0.3)	-1.3*** (0.3)	-0.5 (0.4)	-0.8** (0.3)	-1.8*** (0.4)	-0.9*** (0.3)	-1.4*** (0.4)
Worthwhile life during residency	4.4*** (0.5)	6.3*** (0.6)	7.0*** (0.6)	6.4*** (0.6)	5.9*** (0.7)	6.5*** (0.6)	6.9*** (0.8)
Stress during residency	-0.1 (0.3)	-1.0*** (0.4)	-0.7** (0.4)	-0.6* (0.3)	0.5 (0.4)	-0.7** (0.3)	0.0 (0.4)
Future career prospects	3.2*** (0.5)	0.9* (0.5)	1.8*** (0.5)	3.0*** (0.5)	1.2** (0.6)	2.6*** (0.5)	2.8*** (0.7)
Control over life	0.4 (0.3)	0.9** (0.3)	0.4 (0.3)	0.4 (0.3)	1.0** (0.4)	0.4 (0.3)	1.5*** (0.4)
Desirable for significant other	2.6*** (0.3)	0.5* (0.3)	0.7*** (0.3)	1.0*** (0.3)	0.3 (0.3)	1.2*** (0.2)	0.9*** (0.3)
# Observations	2169	2167	2169	2168	1591	2166	1590
# Students	557	557	557	557	540	557	540
McKelvey & Zavoina R^2 , within variance only	0.46	0.34	0.42	0.46	0.25	0.48	0.42
Joint significance of differences with choice coefficients		0.000	0.000	0.000	0.000	0.000	0.000

Notes: Standard errors in parentheses. Rank-ordered logit regressions of either choice (column 1) or a SWB measure (columns 2-7) on residency attributes. Only ordinal information on the dependent variables is used. Columns 2-4 use the ordinal rankings implied by the main three SWB measures. Columns 5-7 use the ordinal rankings implied by an optimal linear utility index, created by a first-stage rank-ordered logit regression of choice on the index components (reported in Table 3.4). All attribute ratings are divided by 100 before being included in the regression. Joint significance of the differences with choice coefficients (bottom row): p-value from a Wald test of the joint equality of all coefficients in the column with all coefficients in the choice column. * $p < .1$, ** $p < .05$, *** $p < .01$

and status is 2.5 in the choice regression and 0.0 in the happiness regression. This difference is highly statistically significant (Wald test p-value = 0.000). To interpret these coefficients, consider their implication for the ranking of two residency programs that are identical in all measured dimensions except for a 20-point difference in their prestige and status on the survey's 100-point scale. The choice coefficient implies that the probability of choosing the more prestigious program would be $\exp(2.5 * 20/100) / (\exp(2.5 * 20/100) + 1) = 62\%$. The happiness coefficient implies that the probability of ranking the more prestigious program higher on anticipated happiness would be 50%.¹⁶

Our estimate of the relationship between a residency's ranking and the residency's perceived prestige and status hence strongly depends on whether we use choice ranking or anticipated-happiness ranking. Examining the rest of the coefficient pairs across the choice and happiness columns reveals that, within a pair, while there are no sign reversals, there are many significant differences in coefficient magnitudes. With the exception of control over life, they are all statistically significant at the 10% level. Five of the differences are significant at the 1% level: not only residency prestige and status, but also desirability of location, future career prospects, and desirability for significant other are associated significantly more with choice than with anticipated-happiness, while the reverse is true for social life during the residency. As reported in the table's bottom row, joint equality of coefficients between the two columns is strongly rejected.

Examining the next two columns ("Life satisfaction during residency" and "Ladder") reveals that with few exceptions, these two measures' coefficients lie

¹⁶Of course, our coefficient estimates (and hence our tradeoff estimates below) may be subject to omitted-variable bias. However, if choice-based MRSs were identical to SWB tradeoffs, any resulting bias would equally affect the choice-based and SWB-based estimates. Our discussion below is therefore focused less on the point estimates themselves and more on whether they differ across choice and SWB.

between those of choice and those of happiness. These two evaluative measures seem on some attributes closer to happiness, an affective measure, and on other attributes closer to choice. For example, while on social life during the residency, the two are virtually indistinguishable from happiness, all with coefficients larger than that on choice, on desirability of location the two are indistinguishable from choice, with coefficients much larger than that on happiness. Across the rows, most of the ladder estimates appear closer to the choice estimates than the life satisfaction estimates; statistically, however, we cannot distinguish the two evaluative measures from each other. Indeed, Wald tests of the joint equality of coefficients between any pair among the four columns strongly reject the null of equality ($p = 0.000$), for all pairs except the life satisfaction and ladder pair ($p = 0.52$).

To what extent do these differences in coefficient estimates translate to differences in estimated tradeoffs? To answer this question regarding a given tradeoff—for example, between prestige and social life—one can compare, across Table 3.2's columns, the within-column ratio between the two relevant coefficients. To answer this question regarding a given attribute—for example, "How large are the cross-column differences in estimated tradeoffs between prestige-and-status and the other eight attributes?"—we could use that attribute as a numeraire and report nine tables (one per numeraire), each with relatively noisy ratio estimates. Instead, we report Table 3.3, a single table that summarizes each attribute's eight relevant within-column ratios with a single, less noisy measure that can be compared across columns. The table reports the ratio of each coefficient from Table 3.2 to the average absolute value of coefficients in its Table 3.2 column. With this normalization, each of Table 3.3's entries can be interpreted as an average weight in tradeoffs. For example, a higher normalized

coefficient on an attribute in the choice column relative to the happiness column would mean that on average, the MRS between another attribute and this one is lower in the choice column than the corresponding tradeoff estimate in the happiness column. Standard errors are calculated using the delta method.

Examining Table 3.3's first row and comparing column 1 with columns 2-4 reveals that residency prestige and status's regression coefficient in the choice column is 1.4 times the average of the nine attributes' regression coefficients; with any of the three anticipated-SWB measures, however, prestige and status's regression coefficients are below average, ranging from 0.0 to 0.4 times the average. This difference in implied tradeoffs is rather dramatic: the estimate in the choice column is more than three times larger than the largest SWB estimate.

To examine the statistical significance of this and other differences, Web Appendix Table A4 replaces each estimate in columns 2-7 of Table 3.3 with its difference from the corresponding estimate in Table 3.3's column 1 (the choice column). Table A4 also reports the p-value of each difference. Relative to the choice-based estimates, all three SWB measures underweight residency prestige-and-status and desirability for significant other, and overweight the importance of social life and life seeming worthwhile during the residency. Other attributes also show significant differences, but they appear to be less systematic. As reported in Table 3.3's bottom row, we again easily reject joint equality—in this table, of normalized coefficients—between any of the three SWB measures and choice.

Comparing across Table 3.3's SWB columns, the life satisfaction and ladder columns appear similar to each other (as in Table 3.2), with virtually all estimates in between the choice estimates and the (always equally signed) happi-

Table 3.3: Tradeoff estimates: choice vs. anticipated SWB

	(1) Choice	(2) Happiness during residency	(3) Life satisfaction during residency	(4) Ladder	(5) 4-period- happiness index	(6) 3-SWB- measure index	(7) 6-SWB- question index
Residency prestige and status	1.4*** (0.2)	0.0 (0.2)	0.4* (0.2)	0.4** (0.2)	0.2 (0.3)	0.4** (0.2)	0.5** (0.2)
Social life during residency	0.8*** (0.2)	2.0*** (0.2)	1.5*** (0.2)	1.6*** (0.2)	1.7*** (0.3)	1.7*** (0.2)	1.6*** (0.2)
Desirability of location	0.9*** (0.1)	0.3* (0.2)	1.0*** (0.1)	0.9*** (0.1)	0.3* (0.2)	0.9*** (0.1)	0.7*** (0.1)
Anxiety during residency	-0.1 (0.2)	-0.8*** (0.2)	-0.3 (0.2)	-0.4** (0.2)	-1.1*** (0.2)	-0.4*** (0.2)	-0.6*** (0.2)
Worthwhile life during residency	2.4*** (0.2)	3.9*** (0.3)	3.9*** (0.3)	3.2*** (0.3)	3.7*** (0.4)	3.1*** (0.2)	3.2*** (0.3)
Stress during residency	-0.1 (0.2)	-0.6*** (0.2)	-0.4** (0.2)	-0.3* (0.2)	0.3 (0.3)	-0.3** (0.2)	0.0 (0.2)
Future career prospects	1.7*** (0.3)	0.5* (0.3)	1.0*** (0.3)	1.5*** (0.3)	0.8** (0.4)	1.3*** (0.2)	1.3*** (0.3)
Control over life	0.2 (0.2)	0.5*** (0.2)	0.2 (0.2)	0.2 (0.2)	0.6** (0.3)	0.2 (0.1)	0.7*** (0.2)
Desirable for significant other	1.4*** (0.1)	0.3* (0.2)	0.4*** (0.1)	0.5*** (0.1)	0.2 (0.2)	0.6*** (0.1)	0.4*** (0.1)
# Observations	2169	2167	2169	2168	1591	2166	1590
# Students	557	557	557	557	540	557	540
Joint significance of differences with choice coefficients		0.000	0.000	0.000	0.000	0.000	0.000

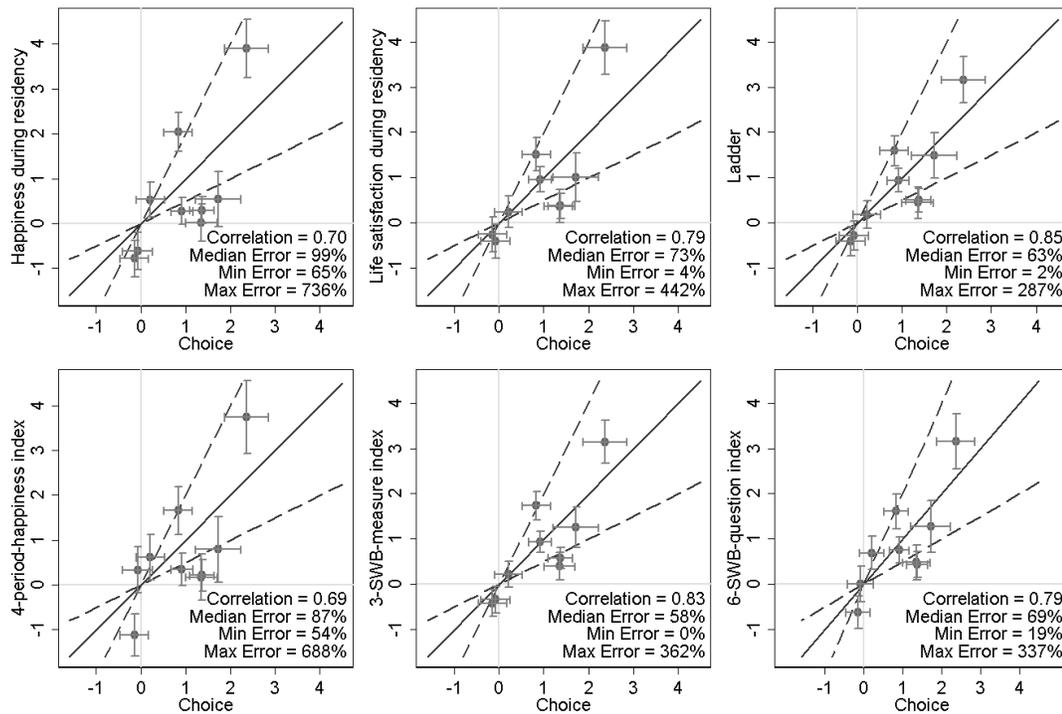
Notes: Delta-method standard errors in parentheses. Entries are coefficients from Table 3.2, normalized by taking their ratio to the average absolute value of the nine coefficients in their Table 3.2 column. Joint significance of the differences with choice entries (bottom row): p-value from a Wald test of the joint equality of all entries in the column with all entries in the choice column. * $p < .1$, ** $p < .05$, *** $p < .01$

ness estimates. Considered jointly, the coefficients in both the life satisfaction and ladder columns are again statistically different from the happiness column ($p = 0.000$) but are not distinguishable from each other ($p = 0.63$).

Since comparing the choice and SWB columns of Table 3.3 is one of the central aims of our paper, Figure 3.3 provides a visual rendering of the table. Each of the figure's six graphs is based on Table 3.3's column 1 and one other column (from among columns 2-7). Within each graph, each of the nine points represents an attribute. Each attribute's x- and y-coordinates correspond, respectively, with its choice and SWB estimates from Table 3.3, with their 95% confidence intervals represented by the horizontal and vertical capped bars. Points in the northeast or southwest quadrants hence represent cases where choice and SWB estimates have the same sign; on the solid 45-degree line, the estimates are equal. To assist in visually assessing how far a point is from the 45-degree line, the dashed lines demarcate the boundaries outside of which estimates differ by more than a factor of two.

Focusing on the top three graphs, it is visually apparent that almost all points fall in the same-sign quadrants and that, additionally, there is substantial positive correlation between the choice and SWB estimates (correlations are reported in each graph). However, there are also substantial differences between the estimates, often by a factor of two or more. To quantify these differences, we define a percentage-error measure of SWB-based estimates relative to the choice-based benchmarks: $\left| \frac{\beta^{SWB} - \beta^{Choice}}{\beta^{Choice}} \right|$, where the β s represent an attribute's estimates in Table 3.3, and the superscript SWB represents one of the SWB columns. An error of 60%, for example, corresponds to cases where the SWB estimate is either 40% or 160% of the choice estimate. Each graph reports the minimum, median, and

Figure 3.3: Tradeoff estimates: choice vs. anticipated SWB



Notes: Based on Table 3.3 estimates. Each graph presents a comparison of one SWB measure (columns 2-7 of Table 3.3) to choice (column 1 of Table 3.3). Each point represents one of the nine attributes included in the regressions, and its x- and y-coordinates correspond to the normalized choice and SWB coefficients, respectively. 95% confidence intervals are represented by the horizontal and vertical capped bars. The dashed lines demarcate the boundaries outside of which the normalized choice and SWB coefficients differ by more than a factor of two. See section 3.4.1 for discussion of the error metrics.

maximum error among the nine attributes. The median ranges from 63% for the ladder measure to 99% for the happiness-during-residency measure. While such margins of error may be tolerable for some applications—for example, applications that focus only on the sign of an effect—they are a serious limitation to the use of these measures when more precise tradeoff estimates are needed.

3.4.2 Robustness

In this subsection, we probe the robustness of our main results to several possible sources of bias.

Biases in survey response: Due to a halo effect, respondents' overall assessments of residencies might leak into their subjective assessments of either anticipated SWB or residency attributes (or both). Similarly, cognitive dissonance might lead respondents to modify their subjective assessments to rationalize the choice order they reported earlier in the survey. To the extent that the ratings of the residency attributes are affected, the coefficients in our regressions are biased upward. Such a bias, however, could not by itself explain the differences in coefficients across columns. Moreover, to the extent that the ratings of anticipated SWB measures are affected, the concordance between the SWB-based rankings and the choice ranking would increase, biasing downward any choice-SWB differences across the columns. Therefore, the differences we do observe should be viewed as a lower bound on the actual divergence between anticipated-SWB and choice rankings.

Econometric specification: The estimates in Tables 3.2 and 3.3 are based on a rank-ordered logit model, which is designed for analyses where the dependent variable is—like our choice data—a rank ordering. Using this same specification for our SWB data makes our estimates comparable across columns and allows us to avoid making assumptions regarding similar use of the SWB rating scales across respondents. In contrast, typical happiness regressions in the literature use OLS or ordered logit/probit, where dependent-variable scale use is assumed to be identical across respondents (or identical up to differences in means, in fixed-effects regressions). To examine the sensitivity of our findings

to specification, we conduct side-by-side comparisons of the SWB columns from Table 3.3 with analogous estimates using OLS with respondent fixed effects (Web Appendix Table A5) and ordered logit (Table A6). These alternative specifications yield estimates similar to the rank-ordered logit regressions and do not change our conclusions from the previous subsection. *Measurement error:* Our respondents' attribute and SWB assessments are likely subject to measurement error. To the extent that the attribute ratings are affected, the coefficients in our regressions are biased. As with the survey-response biases above, however, this bias could not explain the differences in coefficients across columns. Of greater potential concern is the possibility that anticipated SWB is affected: while classical measurement error in the dependent variable would not bias coefficient estimates in OLS, it would bias our rank-ordered logit estimates. Consequently, if anticipated SWB is a noisy measure of choice utility, then measurement error could generate differences in coefficients across the choice and SWB columns. That the coefficients from the fixed-effects OLS specification mentioned above (Web Appendix Table A5) do not meaningfully differ from those in Table 3.2 suggests, however, that such measurement error cannot drive our results.

Heterogeneity in response-scale use: Our analysis above assumes that respondents are identical in their use of the attributes' 1-100 response scales. While heterogeneity in attribute scale use could not explain the choice-SWB differences we find, we verify that our conclusions are unchanged when we re-estimate Table 3.3 after first normalizing the response scales at the respondent level (Web Appendix Table A7; each attribute is demeaned at the respondent level, and then divided by the respondent-specific standard deviation, prior to entering the regressions).

Heterogeneity in tradeoffs: Our analysis above imposes identical coefficients across respondents. Heterogeneity in coefficients could not by itself explain the choice-SWB differences we find. However, it is possible that our results are driven by a particular subpopulation, and that for many or most in the sample, the tradeoffs represented by their anticipated SWB are similar to those implied by their choices. To assess this possibility, we cut the sample along various respondent characteristics. For each sample cut, we re-estimate Table 3.3 (web appendix, pp. 19-32). Our main findings continue to hold across these sample cuts, suggesting that they are pervasive across subgroups within our sample. For example, comparing the choice column with each of the SWB columns, we reject at the 1% level the null hypothesis of jointly identical tradeoffs in each of these cross-column comparisons when cutting the sample by: gender, above and below median MCAT scores, above and below median age, whether or not the respondent agreed to be re-contacted for the follow-up survey (76% of our respondents agreed), and whether or not the respondent completed the follow-up survey (23%); and when excluding dual-match participants (7%). When cutting the sample by relationship status, we reject the null at the 5% level.

Choice versus preferences: As discussed in section 3.2.2, an important advantage of the NRMP setting is that the mechanism incentivizes students to submit their true preference ranking. However, some students may deviate from truthful reporting—for example, due to misunderstanding the mechanism. To assess this possibility, we re-estimate Table 3.3 three more times: excluding respondents who report manipulation attempts by schools (3% of our sample); excluding respondents who report that their NRMP submission did not represent their “true preference order” (17%);¹⁷ and *including* only these 17% of respon-

¹⁷Given the incentive compatibility of the mechanism, this 17% figure may seem surprisingly high. Only 5% of our sample, however, indicate that they chose their list “strategically,” and less

dents, but as dependent variable in the choice column replacing their submitted NRMP ranking with what they report as their “true preference order” (web appendix, pp. 33-35). As above, our conclusions do not change, and we continue to reject joint equality across the choice and SWB columns at the 1% level.

3.4.3 Multi-question SWB indices

Our results thus far suggest that none of our single-question anticipated-SWB measures generates tradeoff estimates that reliably reflect choice tradeoffs. However, two distinct hypotheses separately imply that combinations of questions may better capture choice utility and hence may yield more similar tradeoffs. We now explore these two hypotheses.

Happiness as instantaneous utility: When a survey respondent reports feeling happy, is her report better viewed as relating to her instantaneous utility or to her expected lifetime utility? Our evidence above suggests that happiness-during-residency tradeoffs do not reflect expected-lifetime-utility MRSs. Our findings hence pose a challenge to the interpretation of happiness regressions as estimating choice-utility MRSs (except in situations with no intertemporal considerations).

To explore an alternative hypothesis—the SWB-as-instantaneous-utility hypothesis—we examine whether anticipated happiness would better reflect choice if it integrated happiness predictions over the full expected horizon of

than 1% indicate that they felt they made a mistake. The remaining 11% indicate another reason and are free to explain in a free-response textbox. Most such explanations point to constraints based on family preferences or location, perhaps suggesting that the preferences we estimate for these respondents are best understood as those of their households, as opposed to themselves as individuals.

life, rather than over only the residency years. For that purpose, we elicit additional happiness predictions in our survey. As mentioned in section 3.3.2 above, after responding to questions about each of the top four residencies, the respondents cycle again through the top three, in a new random order. They are instructed as follows:

For the following section, you will again be asked to individually consider the top three programs you ranked. For each of these possibilities, you will be asked to report your predictions on how attending that residency program will affect your happiness during different periods of your life. Please answer as carefully and truthfully as possible.

For each residency, respondents see a screen with questions. The three primary questions read: “On a scale from 1 to 100, how happy do you think you would be on average [during the first ten years of your career]/[for the remainder of your career before retirement]/[after retirement]?” Each is followed by questions assessing the uncertainty of the forecast.

Aggregating such questions into a present-discounted-value-of-happiness index requires weighting them by appropriate discount factors (taking into account the different lengths of their respective intervals). In a field setting where choice data are not available, the researcher would have to choose weights based on her beliefs regarding the discount factor. Since we have choice data, we instead conduct a rank-ordered logit regression predicting choice with our four anticipated happiness questions, and use the estimated latent-index coefficients as our weights. This is the best linear index that could be constructed for predicting choice in our data and hence represents a best-case scenario (by this choice-prediction criterion) for a present-discounted-value-of-happiness mea-

sure that might be used in a realistic application.

The regression for constructing the index is reported in column 1 of Table 3.4. The coefficients on the happiness variables are roughly declining over time, in spite of the increase in time-interval length, consistent with steep discounting.¹⁸ However, the McKelvey and Zavoina R^2 of 0.17 indicates relatively low goodness-of-fit, suggesting that the index still omits significant amounts of choice-relevant information.

Returning to Tables 3.2 and 3.3, in column 5 we use the ordering implied by this multi-period anticipated-happiness index as the dependent variable (“4-period-happiness index”).¹⁹ In Table 3.3, on most of the attributes column 5 is slightly closer to column 1 (choice) than column 2 (happiness during residency) is, but on some of the attributes column 5 is slightly farther. Overall, the 4-period-happiness tradeoff estimates still exhibit substantial differences from the estimates in column 1 (joint significance of differences $p = 0.000$ between columns 1 and 5; $p = 0.08$ between columns 2 and 5). Moreover, columns 3 and 4—life satisfaction and ladder—seem in general closer to column 1 than column 5 is (both columns 3 and 4 are statistically different from column 5, with $p = 0.01$ or less). Indeed, while Figure 3.3 reports that the median error for the 4-period-happiness index is smaller than for happiness during residency, it is larger than

¹⁸While we do not know the exact length of three of the time intervals, we can calculate them roughly. The during-the-residency happiness measure would typically cover five years starting from the present. By definition, we know that the first-ten-years-of-career measure covers the ten years that follow. Since the average age in our sample is 27, the rest-of-career measure is expected to cover roughly another 23 years until retirement ($= 65-27-5-10$). With life expectancy roughly 80 years at that age, the after-retirement measure would cover on average another 15 years. Hence, relative to the during-the-residency measure, the first-ten-years-of-career is roughly twice as long, and the last two time windows are roughly three to five times as long.

¹⁹Since the three beyond-residency anticipated-happiness questions are elicited for only the top three residency choices, the estimates in column 5 in Tables 3.2 and 3.3 are based on a subset of the data columns 1-4 are based on. When we restrict the two tables to the 1591 observations used in column 5 (Web Appendix Tables A12 and A13), our conclusions below are unchanged.

Table 3.4: Main SWB and residency attribute survey questions

	(1)	(2)	(3)
	Choice	Choice	Choice
Happiness during residency	4.5*** (0.5)	0.6 (0.4)	0.9 (0.6)
Happiness in first 10 years	4.6*** (0.8)		3.5*** (0.9)
Happiness in rest of career	2.1** (0.9)		2.4*** (0.9)
Happiness after retirement	1.2 (0.8)		2.0** (0.9)
Life satisfaction during residency		4.4*** (0.5)	3.9*** (0.7)
Ladder		5.5*** (0.4)	5.4*** (0.6)
# Observations	1609	2192	1607
# Students	544	561	544
McKelvey & Zavoina R^2 , within variance only	0.17	0.37	0.37

Notes: Standard errors in parentheses. Rank-ordered logit regressions of choice on SWB measures. All aspect ratings are divided by 100 prior to inclusion in the regressions. Since future happiness measures are only elicited for three of the four ranked residencies, less data are available for conducting these regressions relative to those with only the primary SWB questions. However, restricting all three regressions to the same sample of 1607 observations has only minor impact on the coefficient estimates (although column 2's R^2 decreases to 0.32); see Web Appendix Table A10. * $p < .1$, ** $p < .05$, *** $p < .01$.

for life satisfaction and ladder.

In summary, we find limited support for the SWB-as-instantaneous-utility hypothesis; our four-time-period anticipated-happiness index is far from yielding reliable MRS estimates.

Multidimensional SWB: Although much of the economics literature treats different SWB questions as interchangeable, several recent papers mentioned in the introduction find that different questions have different correlates and argue that they capture distinct components of well-being. To the extent that well-being is multidimensional, a multi-question SWB index might yield tradeoff estimates that are closer to our choice-based MRS estimates than those yielded by any single measure.

To explore this possibility, we construct a “3-SWB-measure” index from our main three SWB questions, and a “6-SWB-question” index by also including the three beyond-residency happiness questions (from the 4-period-happiness index above). To maximize the predictive power of the indices for choice, we again use as weights the coefficients estimated in first-stage regressions of choice on the components of each index.

Columns 2 and 3 of Table 3.4 report our first-stage regressions. In both regressions the coefficient on happiness during the residency is indistinguishable from zero, and is substantially smaller than the corresponding coefficient in column 1 as well as smaller than the coefficients on the two evaluative measures in columns 2 and 3 (life satisfaction during the residency and ladder). In other words, once the two evaluative measures are controlled for, happiness during the residency contributes significantly less to predicting choice. The fit of the

indices in columns 2 and 3, as measured by the McKelvey and Zavoina R^2 , is substantially better than in column 1.

Returning to Tables 3.2 and 3.3, their columns 6 and 7 use, respectively, the orderings implied by each of the two SWB indices as the dependent variable in the regression. We easily reject, in both tables, joint equality of coefficients between each of the two multi-SWB regressions and: choice (see each table's bottom row), happiness ($p = 0.000$), and, less strongly, the 4-period-happiness index ($p = 0.06$ or less). Nonetheless, we cannot distinguish the two from each other or from either life satisfaction or ladder (p -values range from 0.15 to 0.97); indeed, in Figure 3.3 the four relevant graphs appear rather similar.²⁰

To summarize, we find no support for the multidimensional-SWB-as-utility hypothesis; our indices that incorporate multiple SWB measures not only fail to match the choice-based MRS estimates, but also fail to do significantly better than our single-question evaluative SWB measures. Of course, the SWB measures we include in these indices are far from exhausting every conceivable measurable dimension (and time period) of well-being, and hence we cannot rule out the possibility that an index based on a sufficiently rich set of questions might yield reliable MRS estimates—indeed, an index that captured all the aspects that our respondents consider when making decisions should, by construction, match choice quite closely. Nonetheless, since the SWB measures we use in this paper are modeled after those most common in existing social surveys and applied research, our results suggest that a straightforward exten-

²⁰It may seem surprising that, relative to single-question life satisfaction or ladder questions, the two indices do not in general yield coefficients and tradeoff estimates that are closer to those based on choice, since the indices are better predictors of choice by construction. This finding is directly related to the fact that while a measure may be highly correlated with choice, it may not necessarily yield tradeoff estimates similar to those implied by choice. See section 3.5 for discussion.

sion of current practices—using a linear combination of a few commonly-used SWB measures—would not be a substantial improvement for estimating MRSs.

3.5 From slopes to orderings: predicting choice ranking from anticipated-SWB ranking

While our finding of substantial differences between the tradeoffs implied by widely-used SWB measures and those revealed by choices calls into question the practice of using SWB data to estimate MRSs, SWB data could be used instead for assessing which among a set of options is most preferred. We begin this section by exploring the usefulness of our anticipated-SWB data in predicting pairwise choices.

Table 3.5 examines all possible within-respondent pairwise comparisons of residency programs. Each row corresponds to a single SWB or attribute question (top two panels) or a multi-question index (bottom panel). Columns 1, 2, and 3, respectively, report the percent of cases where the program that is ranked higher in choice is ranked higher, the same, or lower than the other program by the row's measure. We assess each measure's usefulness in predicting choice with two yardsticks: the "correct-prediction rate" (another way to think of column 1) and the "conditional correct-prediction rate" (column 4). The latter equals column 1 divided by the difference between 100% and column 2; it is the share of cases where choice and a row's measure yield the same ranking, conditional on the measure ranking one option above the other.

As can be seen in the top panel of the table, the ladder question has the

Table 3.5: Predicting binary choice from anticipated-SWB and attribute questions

	Preferred program rates higher	The two programs have same rating	Preferred program rates lower	Conditional correct-prediction rate	# Pairwise program comparisons
Happiness during residency	52%	27%	21%	71%	3240
Life satisfaction during res.	59%	23%	18%	77%	3244
Ladder	65%	18%	17%	80%	3245
Residency prestige and status	56%	16%	28%	67%	3244
Social life during residency	52%	20%	28%	65%	3247
Desirability of location	61%	14%	25%	71%	3241
Anxiety during residency	38%	29%	33%	53%	3236
Worthwhile life during res.	44%	40%	16%	73%	3235
Stress during residency	40%	26%	34%	54%	3236
Future career prospects	49%	30%	21%	70%	3247
Control over life	40%	30%	30%	57%	3235
Desirable for significant other	65%	16%	19%	77%	2087
Avg. hap. in first 10 years	34%	53%	13%	72%	1603
Avg. hap. in rest of career	28%	56%	16%	64%	1603
Avg. hap. after retirement	22%	64%	14%	62%	1605
4-period-happiness index	62%	10%	28%	69%	1592
3-SWB-measure index	75%	3%	22%	77%	3233
6-SWB-question index	76%	2%	22%	78%	1588
12-question index	81%	0%	19%	81%	3179
15-question index	82%	0%	18%	82%	1566

Notes: Based on only the ordinal ranking of the variable in each row. All six binary comparisons among the top four programs are considered. Columns 1-3 sum to 100% in each row. Column 4 reports the correct prediction rate in cases where a prediction is made; that is, excluding cases of indifference (column 2). Column 5 reports sample size.

highest correct-prediction rate (65%) among our three SWB and nine residency attribute questions. It also has the highest conditional correct-prediction rate [80%]. Among the 64% of respondents in a relationship, the next-best predictor is desirability to one's partner (correct-prediction rate 65%)[conditional correct-prediction rate 77%]. In decreasing order of correct-prediction rate, the other questions are: desirability of location (61%)[71%]; life satisfaction during residency (59%)[77%]; residency prestige and status (56%)[67%]; happiness during residency (52%)[71%]; social life during residency (52%)[65%]; future career prospects (49%)[70%]; worthwhile life during residency (44%)[73%]; stress during residency (40%)[54%]; control over life (40%)[57%]; and anxiety during residency (38%)[53%]. Due to potential biases in survey response such as the halo effect and cognitive dissonance discussed above (in Section 3.4.2), we interpret these rates as upper bounds and focus not on their absolute magnitudes but rather on comparing them across questions.

Regardless of whether we assess usefulness by the conditional or unconditional correct-prediction rate, we find that the evaluative SWB questions—ladder and life satisfaction—as well as desirability to one's significant other, are among the single-question measures that match choice most closely. At the other extreme, anticipated negative feelings—anxiety and stress during the residency—do not predict choice well (with a conditional correct-prediction rate only slightly better than a 50-50 guess).

The middle panel of Table 3.5 analyzes the three beyond-residency happiness questions. While for happiness in the first ten years of one's career, the conditional correct-prediction rate is nearly the same as for happiness during the residency [72% vs. 71%], the unconditional rate is much lower (34% vs.

52%), reflecting many ties (column 2). For happiness measures further in the future, both rates are lower. Therefore, these measures are of relatively limited usefulness as single-question predictors of pairwise choices.

Finally, for comparison with these single-question measures, the bottom panel of the table examines our three multi-question indices (discussed in III.C) and two additional indices that incorporate the nine attribute questions into the multidimensional SWB indices. The weights in these two additional indices are estimated from regressions analogous to those in Table 3.4 (Web Appendix Table A8). The 4-period-happiness index's conditional correct-prediction rate is slightly below that of the happiness-during-the-residency question [69% vs. 71%], but, with far fewer ties (column 2), the index's unconditional rate is much higher (62% vs. 52%). The rest of the indices, which are based on increasing numbers of questions (3, 6, 12, and 15), have relatively high (and increasing) conditional correct-prediction rates [77%, 78%, 81%, and 82%, respectively]. As including more questions in an index yields fewer ties, the indices' unconditional rates are much higher than that of any single question (75%, 76%, 81%, and 82%, respectively).

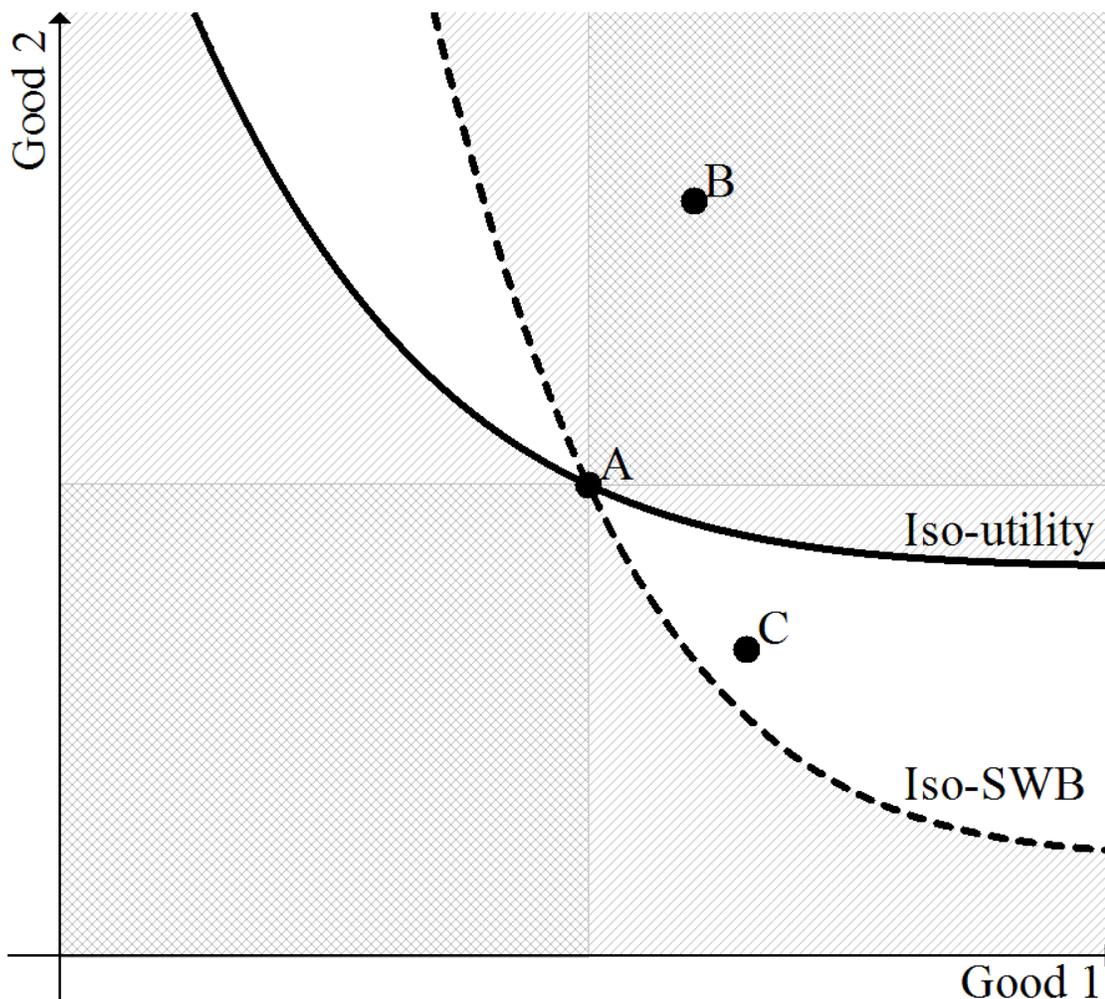
It may seem puzzling that the evaluative SWB questions and, to an even larger extent, the 3- and 6-question indices correctly predict choice at relatively high rates, in light of our finding that the tradeoffs they imply are so different from the MRSs implied by choice. Figure 3.4 presents a simple model with two attributes that illustrates the relationship between pairwise predictions and tradeoffs. We orient the attributes so that both are "goods": preferences are monotonically increasing in each. We assume, consistent with our no-sign-reversals finding in the previous section, that anticipated-SWB is also monotoni-

cally increasing in each good. The solid line represents an individual's iso-utility curve, while the dashed line represents her anticipated iso-SWB curve; we assume these curves satisfy standard regularity conditions. The respective slopes at choice option A differ: the SWB tradeoff does not coincide with the MRS. Indeed, while option A is preferred to option C, SWB is higher at C than at A. In contrast, despite the difference in slopes, option B ranks higher than option A in both choice and SWB. More generally, a SWB-based comparison of option A with any option in the unshaded areas—the “discordance region”—would favor the wrong option from a preference point of view; while a SWB-based comparison of A with an option in any of the shaded areas—the “concordance region”—would yield the right choice. Locally, the discordance region is larger the greater is the difference in slopes. Globally, it is always strictly limited to the northwest and southeast quadrants—the quadrants where an alternative to A involves sacrificing one good for the other, i.e., where neither option vector-dominates the other.

More generally, with any number of goods, the “closer” one choice option is to vector-dominating the other, the more likely it is that the alternative to A lies in the concordance region. In our data weak vector dominance (i.e., weak inequality component by component) occurs in 16% of binary comparisons—a high percentage relative to what might be expected with nine independently and symmetrically distributed attributes ($2 \times \frac{1}{2^9}$, assuming no ties). Indeed, with the exception of stress and anxiety during residency, within-respondent attribute ratings are generally moderately positively correlated across residencies (Web Appendix Table A9). These positive correlations may help explain the reasonably high rates of concordance that we find.²¹

²¹Another implication of this model is that, under reasonable assumptions regarding the distribution from which the alternative to A is drawn, when the alternative lies on a more distant

Figure 3.4: Implications of iso-utility and iso-SWB curves for ordinal prediction



Notes: This figure illustrates the implications of different tradeoffs in choice-utility and anticipated SWB for binary comparisons. The solid line represents an individual's iso-utility curve, while the dashed line represents her iso-SWB curve. Comparing option A to options in any of the shaded areas (for example, option B), the iso-utility and iso-SWB curves imply the same binary ordering. Comparing option A to options in the unshaded areas, the curves imply different orderings (option C, for example, has higher SWB but is less preferred).

The empirical settings where SWB comparisons would be most needed for drawing inferences about the preference ranking of options, however, are settings that involve sacrificing some goods for others. For example, Gruber and Mullainathan (2005) conduct SWB-based welfare comparisons among smokers who face higher versus lower cigarette taxes—a setting that involves an inherent tradeoff between health and wealth, and where SWB data could potentially be useful because, in the presence of self-control problems, choices may not reveal preferences. In such no-vector-dominance settings, the model above does not make a clear prediction on whether choice and SWB would yield the same ranking. Hence, in the absence of evidence like ours from specific settings of interest, it is hard to assess the usefulness of SWB data for welfare comparisons in those settings.

Due to the inherent difficulty of observing choice and anticipated SWB in many of the situations where SWB data might be useful, Benjamin, Heffetz, Kimball, and Rees-Jones (2012) study hypothetical choices and anticipated SWB in thirteen settings designed to have no vector dominance. They find an overall correct-prediction rate of 83%, with wide variation across choice settings, and they identify features of the settings that are associated with higher rates. Evidence from more settings is needed before we would be confident in drawing generalizations regarding the reliability of SWB data for inferring preference rankings.

(i.e., much higher or much lower) iso-utility curve, it is more likely to lie in the concordance region. In Web Appendix Tables A14-A16 we report three additional versions of Table 3.5, restricting the underlying data to three respective subsets of pairwise program comparisons: only first- versus second-, only first- versus third-, and only first- versus fourth-ranked programs. We find, as expected, that virtually all of our measures are better predictors of choice as the ranking difference increases. For example, ladder's conditional correct-prediction rate increases from 78% to 87% to 90%.

3.6 Concluding remarks

Scholars and lay people alike have long been fascinated by happiness and its correlates. By regressing SWB measures on bundles of goods and examining coefficient ratios, researchers have been able to compare in common units the associations between SWB measures and a wide variety of goods. Such comparisons have generated a large and growing number of interesting findings. To what extent do these coefficient ratios represent economists' notion of MRSs?

Our main finding is that, among the medical students in our sample, the MRSs of residency program attributes implied by their preference rankings are far from equal to the tradeoffs implied by their anticipated-SWB responses—regardless of whether we use a happiness, a life satisfaction, or a ladder measure of SWB; a simple combination of such measures; or a simple combination of anticipated happiness over the near and distant future. At the same time, we find no sign reversals between choice and our SWB measures in their association with any of the nine attributes; we find relatively high correlations across the nine attributes between their choice-regression and SWB-regression coefficients; and we find relatively high choice-SWB concordance rates in binary residency comparisons.

Of course, our sample of medical students is a convenience sample, our evidence is limited to the specific context of residency choice, and the nine residency attributes that constitute our bundle of goods are far from exhaustive. Nonetheless, we view our real-choice field evidence as an important advance over and a complement to existing evidence from prior work. When we consider them together, some common themes emerge across the findings in this

paper and those in previous work that studies hypothetical choices in a range of realistic scenarios (Benjamin, Heffetz, Kimball, and Rees-Jones, 2012; henceforth BHKR) and abstract scenarios (Benjamin, Heffetz, Kimball, and Szembrot, 2012; henceforth BHKS). We highlight four such themes. First, our main conclusion that anticipated-SWB tradeoffs differ from choice MRSs is consistent with results from the earlier papers that attributes of the options help to predict hypothetical choices, controlling for anticipated SWB. Second, as mentioned above, our finding of high concordance rates between choice and anticipated-SWB in binary comparisons is similar to BHKR's finding. Third, all three papers conclude that evaluative SWB measures are closer to choice than affective happiness measures.²² Finally, all three papers find that measures of family well-being—family happiness (in the previous work) and residency desirability to one's spouse or significant other (in this paper)—are among the strongest predictors of choice.

We believe that each of these findings has practical implications for empirical researchers who consider using SWB measures to proxy for utility. We list four such implications, in respective order paralleling the four themes above. First, SWB tradeoffs should not be interpreted as MRSs. Second, binary SWB rankings may in some settings be highly predictive of preference rankings—even when SWB tradeoffs are far from MRSs. This of course also means that high choice-SWB concordance in pairwise comparisons should not be interpreted as justifying the use of SWB data to estimate MRSs. Third, evaluative SWB measures may more reliably reflect preferences than affective happiness measures—even

²²However, BHKR examine life satisfaction and happiness with life as a whole as their evaluative measures and do not study the ladder measure, and BHKS find that, in contrast to other evaluative measures, the ladder question predicts hypothetical choice less well than many other measures they study, in regressions that control for other measures. The potential discrepancy between that finding and the finding reported here makes us reluctant to draw a strong conclusion regarding the ladder question *per se*.

when happiness is integrated over several time periods. Finally, measures of family SWB may in some settings reflect preferences at least as reliably as evaluative measures of own SWB. Such family-SWB measures are not commonly used in empirical applications but warrant exploration.

While we hope that researchers find these practical implications useful, we also caution that using SWB data in empirical work typically requires additional assumptions, often heroic—for example, about interpersonal comparability of SWB survey responses (see, e.g., Adler, 2012)—that we do not evaluate in this paper.

From a theoretical perspective, if different aspects of well-being are all viewed as inputs into preferences, then in principle the specific aspects captured by traditional SWB measures should not be treated differently from other inputs a priori. From this point of view, rather than regressing SWB on other goods, estimating preferences requires regressing choice on a bundle that includes SWB measures together with those other goods. BHKR and BHKS run such regressions with hypothetical choice. The findings from those papers and this paper suggest that while the well-being aspects captured by traditional SWB measures are among the most important inputs into preferences, they are not the only important inputs. Consequently, consistent with the view expressed by Deaton, Fortson, and Tortora (2010), it seems unlikely that one SWB question or even a combination of a small number of them would capture enough of the important inputs to be sufficient as an all-purpose utility proxy.

If tradeoffs estimated from SWB data are not MRSs, how should they be interpreted? From the above theoretical perspective, SWB tradeoffs may be interpreted as technical rates of substitution (TRSs) that characterize the production

function for SWB (as in Kimball and Willis, 2006, and Becker and Rayo, 2008). Just as it is valuable for economists and policymakers to estimate TRSs for other important utility inputs such as health, estimates of TRSs for SWB have generated and will likely continue to generate valuable insights into the production of SWB.

3.7 Works cited

- Adler, Matthew D.** 2012. "Happiness Surveys and Public Policy: What's the Use?" *Duke Law Journal*, forthcoming.
- Becker, Gary S., and Luis Rayo.** 2008. "Comment on 'Economic growth and subjective well-being: Reassessing the Easterlin Paradox' by Betsey Stevenson and Justin Wolfers." *Brookings Papers on Economic Activity*, Spring: 88-95.
- Beggs, S., S. Cardell, and J. Hausman.** 1981. "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics*, 16: 1-19.
- Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones.** 2012. "What Do You Think Would Make You Happier? What Do You Think You Would Choose?" *American Economic Review*, 102(5): 2083-2110.
- Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones.** 2013. "Can Marginal Rates Of Substitution Be Inferred From Happiness Data? Evidence from Residency Choices" NBER working paper No. 18927.
- Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot.** 2012. "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based

on Stated Preference." NBER Working Paper No. 18374.

Clark, Andrew, Paul Fritters and Michael Shields. 2008. "Relative Income, Happiness and Utility: An Explanation for the Easterlin Paradox and Other Puzzles." *Journal of Economic Literature*, 46(1): 95-144.

Deaton, Angus, Jane Fortson, and Robert Tortora. 2010. "Life (Evaluation), HIV/AIDS, and Death in Africa." In *International Differences in Well-Being*, edited by Ed Diener, John Helliwell, and Daniel Kahneman, Oxford: Oxford University Press, 105-136.

Di Tella, Rafael, Robert J. MacCulloch, and Andrew J. Oswald. 2001. "Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness." *American Economic Review*, 91(1): 335-341.

Dolan, Paul, and Robert Metcalfe. 2008. "Comparing Willingness-To-Pay and Subjective Well-Being in the Context of Non-Market Goods." CEP Discussion Paper No 890.

Frey, Bruno, and Alois Stutzer. 2002. "What Can Economists Learn from Happiness Research?" *Journal of Economic Literature*, 40(2): 402-435.

Gale, David, and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly*, 69: 9-15.

Gilbert, Daniel. 2006. *Stumbling on Happiness*. New York: Knopf.

Gruber, Jonathan, and Sendhil Mullainathan. 2005. "Do Cigarette Taxes Make Smokers Happier?" *B.E. Journal of Economic Analysis and Policy*, 5(1).

Hsee, Christopher K. 1999. "Value-Seeking and Prediction-Decision Inconsis-

tency: Why Don't People Take What They Predict They'll Like the Most?" *Psychonomic Bulletin and Review*, 6(4): 555-561.

Hsee, Christopher K., Jiao Zhang, Fang Yu, and Yiheng Xi. 2003. "Lay Rationalism and Inconsistency Between Predicted Experience and Decision." *Journal of Behavioral Decision Making*, 16: 257-272.

Kahneman, Daniel, and Angus S. Deaton. 2010. "High Income Improves Evaluation of Life but not Emotional Well-Being." *Proceedings of the National Academy of Sciences*, 107(38): 16489-16493.

Kahneman, Daniel, Peter P. Wakker, and Rakesh K. Sarin. 1997. "Back to Bentham? Explorations of Experienced Utility." *Quarterly Journal of Economics*, 112(2): 375-405.

Kimball, Miles, and Robert Willis. 2006. "Utility and Happiness." Unpublished, University of Michigan.

Levinson, Arik. 2012. "Valuing Public Goods Using Happiness Data: The Case of Air Quality." *Journal of Public Economics*, 96: 869-880.

Loewenstein, George, Ted O'Donoghue, and Matthew Rabin. 2003. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics*, 118(4): 1209-48.

Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sonbonmatsu. 2012. "Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults." *Science*, 337: 1505-1510.

- Luechinger, Simon, and Paul A. Raschky.** 2009. "Valuing Flood Disasters Using the Life Satisfaction Approach." *Journal of Public Economics*, 93: 620-633.
- Luttmer, Erzo.** 2005. "Neighbors as Negatives: Relative Earnings and Well-being." *Quarterly Journal of Economics*, 120(3): 963-1002.
- McKelvey, Richard, and William Zavoina.** 1975. "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." *Journal of Mathematical Sociology*, 4: 103-120.
- National Resident Matching Program.** 2012. "National Resident Matching Program, Results and Data: 2012 Main Residency Match?." National Resident Matching Program, Washington, DC.
- Oswald, Andrew, and Nattavudh Powdthavee.** 2008. "Death, Happiness, and the Calculation of Compensatory Damages," *Journal of Legal Studies*, 37(S2): S217-S252.
- Roth, Alvin, and Elliot Peranson.** 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review*, 89(4): 748-780.
- Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi.** 2009. Report by the Commission on the Measurement of Economic Performance and Social Progress. www.stiglitz-sen-fitoussi.fr
- Tversky, Amos, and Dale Griffin.** 1991. "Endowments and Contrast in Judgments of Well-Being." In *Strategy and Choice*, ed. Richard J. Zeckhauser. Cambridge, MA: MIT Press. Reprinted in *Choices, Values, and Frames*, ed. Kahneman, Daniel, and Amos Tversky. Cambridge, UK: Cambridge University

Press.

Van den Berg, Barnard, and Ada Ferrer-i-Carbonell. 2007. "Monetary Valuation of Informal Care: The Well-Being Valuation Method." *Health Economics*, 16: 1227-1244.

Van Praag, Bernard, and Barbara Baarsma. 2005. "Using Happiness Surveys to Value Intangibles: The Case of Airport Noise." *Economic Journal*, 115(500): 224-246.

3.8 Acknowledgements

We thank Al Roth for valuable early suggestions, and Matthew Adler, Greg Besharov, Aaron Bodoh-Creed, Angus Deaton, Jan-Emmanuel De Neve, Dan Gilbert, Sean Nicholson, Ted O'Donoghue, Andrew Oswald, and Richard Thaler for valuable comments. We are grateful to participants at the Cornell Behavioral Economics Research Group, Cornell Behavioral/Experimental Lab Meetings, UCLA/UCSB Conference on Field Experiments, Michigan Retirement Research Center Annual Meeting, Stanford Institute for Theoretical Economics, AEA Annual Meeting, and Duke Law School New Scholarship on Happiness Conference, as well as seminar audiences at Chicago Booth and Cornell for helpful comments. We thank Allison Ettinger, Matt Hoffman, and Andrew Sung for outstanding research assistance. We are grateful to NIH/NIA grants R01-AG040787 and R01-AG020717-07 to the University of Michigan and T32-AG00186 to the NBER, and to the S. C. Johnson Graduate School of Management, for financial support.

APPENDIX A

APPENDIX TO “LOSS AVERSION MOTIVATES TAX SHELTERING”

A.1 Proofs

Proposition 1. *In the final-wealth-dependent sheltering model, if $m^{FWD}(\cdot)$ is twice continuously differentiable and the PDF of b^{PM} is continuous, then the PDF of $b = b^{PM} - s$ is continuous.*

Proof. Let $s^*(b^{PM}|w)$ denote the optimal sheltering solution. Assume first that $s^*(b^{PM}|w)$ is an interior solution for all b^{PM} , in which case it is determined by the first order condition $m'(w - b^{PM} + s^*(b^{PM}|w)) = c'(s^*(b^{PM}|w))$. The implicit function theorem guarantees that $s^*(b^{PM}|w)$ is continuously differentiable. As a result, we can express final balance due as a continuously differentiable function of pre-manipulation balance due: $b(b^{PM}) = b^{PM} - s^*(b^{PM}|w)$. The convexity of $c(\cdot)$ guarantees that $b(b^{PM})$ is strictly increasing, and thus invertible. Denote the inverse function as $\psi(b)$. The CDF of b may be expressed in terms of the CDF for b^{PM} by the relationship $F_b(x) = F_b^{PM}(\psi(b))$. Differentiating yields $f_b(x) = f_b^{PM}(\psi(b))\psi'(b)$, which expresses the PDF of b as a product of continuous functions.

The above assumed the case of an interior solution. Note that in the case where $c'(0) \leq m'(w - b^{PM})$, the first-order conditions do not hold, $s^* = 0$, and the result immediately follows.

Proposition 4. *Consider the mixed-type sheltering model. If r is in the support of the balance due distribution for the loss-averse type, then there exists a threshold value c such that $E[s^{*LA}(\lambda, \eta, \theta^{LA}|b^{PM} = 0)] - E[s^{*FDW}(w, \theta^{FWD}|b^{PM} = 0)] > c > 0$ implies:*

a) $E[b^{PM}|b = r] > \lim_{b \rightarrow r^+} E[b^{PM}|b]$ and $E[b^{PM}|b = r] > \lim_{b \rightarrow r^-} E[b^{PM}|b]$; and

b) $E[s|b = r] > \lim_{b \rightarrow r^+} E[s|b]$ and $E[s|b = r] > \lim_{b \rightarrow r^-} E[s|b]$.

Proof. The following proves the claim in part b above. Part a follows immediately from part b since $b = b^{PM} - s$.

To simplify notation, let p^r denote the probability an agent is loss averse conditional on reporting zero balance due. Let p^+ denote $\lim_{b \rightarrow r^+} Pr(LA|b)$, and p^- denote $\lim_{b \rightarrow r^-} Pr(LA|b)$. Let $\bar{s}^{FWD} = E[s|type = FWD, b = r]$, $\bar{s}_{low}^{LA} = E[s|type = LA, b < r]$, $\bar{s}_{mid}^{LA} = E[s|type = LA, b = r]$, and $\bar{s}_{high}^{LA} = E[s|type = LA, b > r]$.

Considering right continuity: Note that $\lim_{b \rightarrow r^+} E[s|b] = p^+ \bar{s}_{high}^{LA} + (1 - p^+) \bar{s}^{FWD}$ and $E[s|b = r] = p^r \bar{s}_{mid}^{LA} + (1 - p^r) \bar{s}^{FWD}$.

Define threshold $c = \frac{p^+}{p^r - p^+} (s_{high}^{LA} - s_{mid}^{LA})$. If $\bar{s}_{mid}^{LA} - \bar{s}^{FWD} > c$, then $\lim_{b \rightarrow r^+} E[s|b] - E[s|b = r] = p^+ \bar{s}_{high}^{LA} + (1 - p^+) \bar{s}^{FWD} - p^r \bar{s}_{mid}^{LA} - (1 - p^r) \bar{s}^{FWD} = p^+ \bar{s}_{high}^{LA} - p^r \bar{s}_{mid}^{LA} + (p^r - p^+) \bar{s}^{FWD} = p^+ (\bar{s}_{high}^{LA} - \bar{s}_{mid}^{LA}) + (p^r - p^+) (\bar{s}^{FWD} - \bar{s}_{mid}^{LA}) < 0$. This implies $E[s|b = r] > \lim_{b \rightarrow r^+} E[s|b]$.

Considering left continuity: Note that $\lim_{b \rightarrow r^-} E[s|b] = p^- \bar{s}_{low}^{LA} + (1 - p^-) \bar{s}^{FWD}$ and $E[s|b = r] = p^r \bar{s}_{mid}^{LA} + (1 - p^r) \bar{s}^{FWD}$.

If $\bar{s}_{mid}^{LA} - \bar{s}^{FWD} > 0$, then $\lim_{b \rightarrow r^-} E[s|b] - E[s|b = r] = p^- \bar{s}_{low}^{LA} + (1 - p^-) \bar{s}^{FWD} - p^r \bar{s}_{mid}^{LA} - (1 - p^r) \bar{s}^{FWD} = p^- \bar{s}_{low}^{LA} - p^r \bar{s}_{mid}^{LA} + (p^r - p^-) \bar{s}^{FWD} = p^- (\bar{s}_{low}^{LA} - \bar{s}_{mid}^{LA}) + (p^r - p^-) (\bar{s}^{FWD} - \bar{s}_{mid}^{LA}) < 0$.

This implies $E[s|b = r] > \lim_{b \rightarrow r^-} E[s|b]$.

A.2 Indirect tests of tax evasion

To supplement this evidence of sheltering-related behaviors reported in section 1.4, I will now explore two additional behaviors motivated by existing literature.

First, I explore the amount of charitable contributions reported by itemizing tax filers. Previous literature (e.g. Bakija and Heim, 2011) has documented a substantial elasticity of charitable giving to perceived tax incentives. Figure A.7 graphs the average amount of reported charitable contributions, conditional on any being reported, for each dollar-amount of balance due ranging from -100 to 100. Again, a marked spike in this behavior is visible at zero. Donors reporting zero balance due have an average contribution of \$1,726, in contrast to the average contribution of \$955 for other donors in this graph.

I will proceed to use these data on charitable giving to calculate a metric of tax evasion developed in Feldman and Slemrod (2007), hereafter called FS. This work builds off of two basic intuitions. First, different sources of income have different associated propensities for underreporting, largely associated with the ease of detection of evasion. For example, table 1 of FS lists the voluntary reporting percentage (the fraction of earned income that is declared to the IRS) by different income types. Over 99% of wages and salaries were voluntarily reported in tax years 1987 and 1988. This high rate is attributable to the fact that this amount of income is confirmed by your employer, and underreporting on the part of the taxpayer is easily detectable. For other types of income, particularly business or farm earnings, detection of underreporting is markedly more difficult and costly for the IRS, and thus underreporting is far more common. FS estimate these rates of underreporting by utilizing the charitable donations

claimed in itemized deductions. The basic assumption is that your charitable giving is a function of your true income. If we observe that an individual's charity depends on which of these types of income they earn, it either means that individuals who are in the position to successfully evade taxes are particularly charitable, or (more plausibly) that individuals in the position to successfully evade taxes are in fact evading taxes; they only seem to be giving a large portion of their income to charity because their reported income is much smaller than their true income.

Either the legal use of charitable giving as a tax shelter, or the inference of illegal tax evasion, is consistent with the reference-dependent model. To expand on the implied amount of tax evasion which would rationalize this behavior, I employ the quantitative methods of FS, which assume the effect is entirely driven by evasion.

Following the primary regression approach of FS, I decompose income into the portions from schedules C, D, E, and F. The remainder of income is denoted as the "visible" portion, which is assumed to be correctly reported due to the extremely high voluntary reporting rates on these components.

The primary specification is a non-linear least squares regression:

$$\ln(C + 100) = \alpha_0 + \alpha_1 \ln(V + \sum_{ih} k_{ih} R_{ih} + \sum_j b_j S_j) + \alpha_2 \ln(MTR) + \epsilon \quad (\text{A.1})$$

$$\text{where } i = \text{Schedule C, D, E, and F} \quad (\text{A.2})$$

$$j = \text{Schedule C and F} \quad (\text{A.3})$$

$$h = \text{Positive, Negative} \quad (\text{A.4})$$

C denote charitable giving reported on schedule A, S_j is a dummy variable

indicating the presence of schedule j , and MTR denotes the marginal tax rate, which captures the tax benefits gained by marginal changes to charitable giving activities.¹ True wealth is captured by the term $V + \sum_{ih} k_{ih} R_{ih} + \sum_j b_j S_j$; visible income V is assumed to be correctly reported, and the k_{ih} terms estimate the rate of underreporting for different alternative sources of income, with a coefficient of 1 implying no underreporting. Dummy variables for submitting schedule C and F are included to capture the idea that individuals submitting one of these schedules are likely to have more evasion than individuals who do not, even if their submitted schedule reports zero liability. This regression approach assumes that the elasticity of charitable giving with respect to income is invariant to the income source, and uses that structure to estimate the rates of underreporting on these different sources.

With these model estimates I can forecast true income by multiplying the income from various schedules by their estimated underreporting rate. This yields an evasion metric, calculated as the difference between their reported total income and their total income predicted by this model.

Figure A.8 graphs the mean of this evasion metric for each dollar amount of balance due from -100 to 100. Consistent with previous results, the evasion metric is sharply spiked at zero balance due. The average predicted evasion amount for observations with non-zero balance due between -100 and 100 is \$4,674 (95% CI: \$4,381 - \$4,967), as opposed to the significantly higher value of \$15,799 (95% CI: \$4,840 - \$26,759) for individuals precisely at zero. Overall, the analysis of charitable giving data suggests a significant increase in tax sheltering activity among individuals at zero balance due.

¹An error in data recording occurred for the MTR in the 1987 model year. Data from this model year is excluded from this analysis.

Another approach to detecting tax sheltering among unaudited returns was suggested by Slemrod (1985), based on the effect that slight discontinuities in the tax schedule have on sheltering behavior. In practice, most taxpayers do not determine their final tax due by a direct calculation using marginal tax rates. Instead, they use a tax table like the one produced below, determine their income bin, and input the resulting number.

1990 Tax Table—Continued

If line 37 (taxable income) is—		And you are—				If line 37 (taxable income) is—		And you are—				If line 37 (taxable income) is—		And you are—			
At least	But less than	Single	Married filing jointly *	Married filing separately	Head of a household	At least	But less than	Single	Married filing jointly *	Married filing separately	Head of a household	At least	But less than	Single	Married filing jointly *	Married filing separately	Head of a household
Your tax is—		Your tax is—				Your tax is—		Your tax is—				Your tax is—		Your tax is—			
5,000						8,000						11,000					
5,000	5,050	754	754	754	754	8,000	8,050	1,204	1,204	1,204	1,204	11,000	11,050	1,654	1,654	1,654	1,654
5,050	5,100	761	761	761	761	8,050	8,100	1,211	1,211	1,211	1,211	11,050	11,100	1,661	1,661	1,661	1,661
5,100	5,150	769	769	769	769	8,100	8,150	1,219	1,219	1,219	1,219	11,100	11,150	1,669	1,669	1,669	1,669
5,150	5,200	776	776	776	776	8,150	8,200	1,226	1,226	1,226	1,226	11,150	11,200	1,676	1,676	1,676	1,676
5,200	5,250	784	784	784	784	8,200	8,250	1,234	1,234	1,234	1,234	11,200	11,250	1,684	1,684	1,684	1,684
5,250	5,300	791	791	791	791	8,250	8,300	1,241	1,241	1,241	1,241	11,250	11,300	1,691	1,691	1,691	1,691
5,300	5,350	799	799	799	799	8,300	8,350	1,249	1,249	1,249	1,249	11,300	11,350	1,699	1,699	1,699	1,699
5,350	5,400	806	806	806	806	8,350	8,400	1,256	1,256	1,256	1,256	11,350	11,400	1,706	1,706	1,706	1,706
5,400	5,450	814	814	814	814	8,400	8,450	1,264	1,264	1,264	1,264	11,400	11,450	1,714	1,714	1,714	1,714
5,450	5,500	821	821	821	821	8,450	8,500	1,271	1,271	1,271	1,271	11,450	11,500	1,721	1,721	1,721	1,721
5,500	5,550	829	829	829	829	8,500	8,550	1,279	1,279	1,279	1,279	11,500	11,550	1,729	1,729	1,729	1,729
5,550	5,600	836	836	836	836	8,550	8,600	1,286	1,286	1,286	1,286	11,550	11,600	1,736	1,736	1,736	1,736
5,600	5,650	844	844	844	844	8,600	8,650	1,294	1,294	1,294	1,294	11,600	11,650	1,744	1,744	1,744	1,744
5,650	5,700	851	851	851	851	8,650	8,700	1,301	1,301	1,301	1,301	11,650	11,700	1,751	1,751	1,751	1,751
5,700	5,750	859	859	859	859	8,700	8,750	1,309	1,309	1,309	1,309	11,700	11,750	1,759	1,759	1,759	1,759
5,750	5,800	866	866	866	866	8,750	8,800	1,316	1,316	1,316	1,316	11,750	11,800	1,766	1,766	1,766	1,766
5,800	5,850	874	874	874	874	8,800	8,850	1,324	1,324	1,324	1,324	11,800	11,850	1,774	1,774	1,774	1,774
5,850	5,900	881	881	881	881	8,850	8,900	1,331	1,331	1,331	1,331	11,850	11,900	1,781	1,781	1,781	1,781
5,900	5,950	889	889	889	889	8,900	8,950	1,339	1,339	1,339	1,339	11,900	11,950	1,789	1,789	1,789	1,789
5,950	6,000	896	896	896	896	8,950	9,000	1,346	1,346	1,346	1,346	11,950	12,000	1,796	1,796	1,796	1,796

Across the years of my sample, for all but extremely low income filers, the width of income bins was \$50, as seen in the image above. In the absence of attempts to manipulate tax liability, we would assume an individual's position in the \$50 bracket to be roughly uniformly distributed. However, if some individuals are responding to the manipulation incentives induced by these minute discontinuities, they will manipulate their liability just enough to cross these thresholds. Thus we would see "too many" people with income close to the top

of these 50 dollar brackets, as was indeed documented in Slemrod (1985).²

Figure A.9 graphs two metrics related to Slemrod's analysis, restricted to observations that were calculated using tax tables and which fell within \$100 of zero balance due. The first panel graphs the mean position of AGI within the table bracket, and the second panel graphs the mean propensity to be in the top fifth of the table bracket. In contrast to previous metrics, there is no evidence of any discontinuity in these measures at zero. This lack of response might be expected, due to the fact that bunching at zero appears primarily driven by high-income individuals, for whom the tax tables are not used for the calculation of final tax due.

A.3 Supplemental tables and figures

²For example, 23.7% of filers with a high marginal tax rate and fungible income items appeared in the top 20% of their 50 dollar bracket.

Table A.1: Sample size across SSN codes and years

SSN Code	A	B	C, D, E	Total
1979	8852	9013	26935	44800
1980	9107	9205	27709	46021
1981	9131	9282	27825	46238
1982	9129	0	0	9129
1983	9389	9514	0	18903
1984	9636	0	0	9636
1985	9948	10013	0	19961
1986	9990	0	0	9990
1987	10362	10543	0	20905
1988	10627	10707	0	21334
1989	10952	11054	0	22006
1990	11122	11230	0	22352
Total	118245	90561	82469	291275
Unique Taxpayers	15950	15919	32158	64027

Notes: This table presents the number of responses over time by different SSN groups. Five randomly-determined four-digit SSN endings were chosen to form the sample, labeled A-E. Group A was sampled from 1979-1990. Group B was not sampled in 1982, 1984, or 1986. Groups C, D, and E were sampled only for the first three years of the data collection.

Table A.2: Structural parameter estimates

	79	80	81	82	83	84
μ : Standardized AGI	246.74*** (17.52)	332.34*** (23.36)	861.40*** (51.56)	331.65*** (98.79)	-501.26*** (77.52)	-351.48*** (100.35)
μ : Constant	-701.67*** (13.45)	-717.64*** (16.95)	-1219.97*** (48.81)	-1241.50*** (139.89)	-1274.53*** (58.27)	-1185.27*** (67.88)
$\ln(\nu)$: Standardized AGI	0.22 (0.32)	-0.29 (2.04)	0.61*** (0.18)	0.24 (0.52)	-2.02* (0.82)	0.17 (0.42)
$\ln(\nu)$: Constant	0.85 (0.51)	-0.95 (1.89)	0.96 (0.55)	2.15* (1.10)	3.75*** (0.59)	3.02*** (0.82)
ψ : Standardized AGI	208.99*** (21.92)	309.34*** (29.62)	926.75*** (55.34)	469.04*** (111.99)	-350.12*** (68.58)	-236.84** (89.25)
ψ : Constant	-433.01*** (13.66)	-422.99*** (16.40)	-827.97*** (46.56)	-816.18*** (120.24)	-875.82*** (49.18)	-832.36*** (57.04)
$\ln(\sigma_1)$: Standardized AGI	1.24*** (0.05)	1.40*** (0.04)	1.37*** (0.04)	1.54*** (0.10)	1.46*** (0.11)	1.11*** (0.13)
$\ln(\sigma_1)$: Constant	7.17*** (0.03)	7.24*** (0.03)	7.30*** (0.03)	7.46*** (0.08)	7.55*** (0.08)	7.40*** (0.09)
$\ln(\sigma_2)$: Standardized AGI	0.96*** (0.02)	1.04*** (0.02)	1.06*** (0.02)	0.92*** (0.04)	1.00*** (0.03)	0.90*** (0.04)
$\ln(\sigma_2)$: Constant	5.95*** (0.01)	6.01*** (0.02)	6.10*** (0.02)	6.14*** (0.04)	6.09*** (0.02)	6.07*** (0.04)
θ_1 : Standardized AGI	0.18* (0.07)	0.05 (0.08)	-0.26*** (0.07)	-0.23 (0.20)	0.10 (0.12)	0.51*** (0.14)
θ_1 : Constant	-0.79*** (0.05)	-0.33*** (0.05)	-0.21*** (0.05)	-0.58*** (0.12)	-1.06*** (0.08)	-0.83*** (0.11)
γ : Standardized AGI	2.74*** (0.15)	3.16*** (0.18)	3.81*** (0.19)	2.55*** (0.41)	1.11*** (0.16)	1.42*** (0.27)
γ : Constant	-0.34** (0.10)	-0.37** (0.13)	-1.97*** (0.15)	-1.84*** (0.45)	-1.34*** (0.13)	-1.09*** (0.18)
N	33679	33969	34496	7023	14168	7417

Standard errors in parentheses.

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

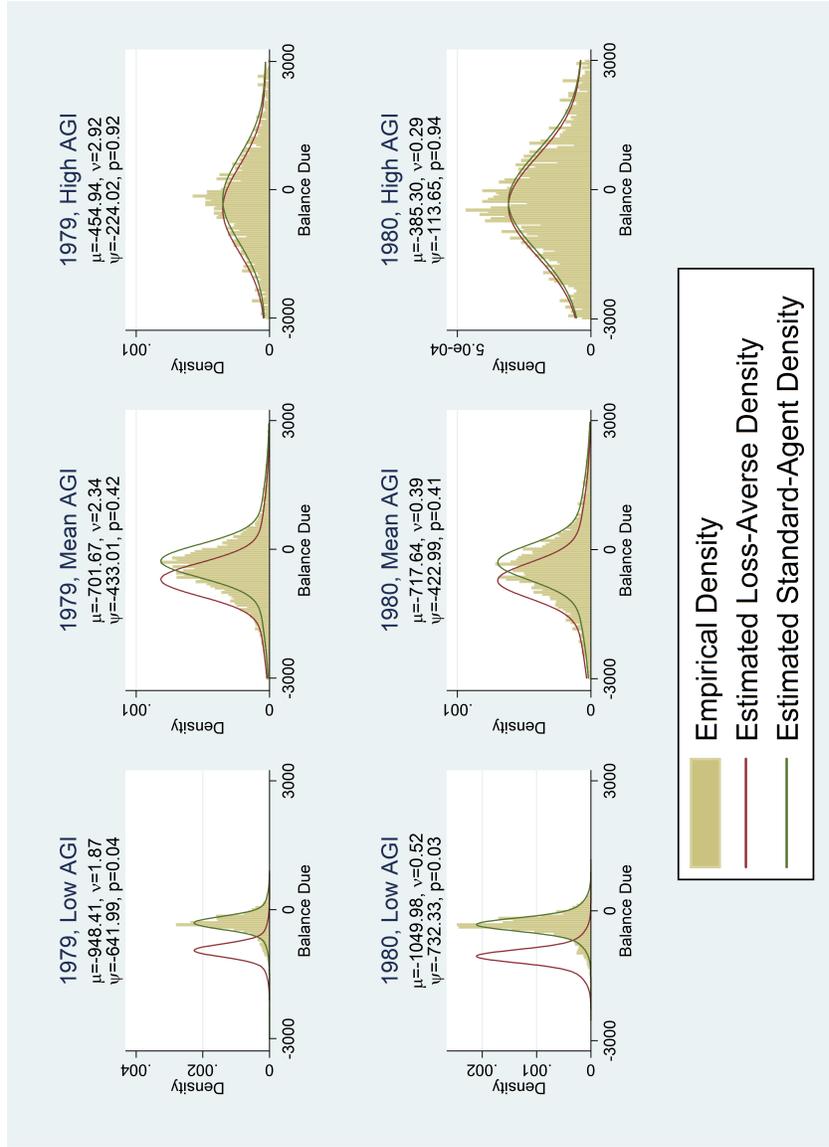
Table A.3: Structural parameter estimates (cont.)

	85	86	87	88	89	90
μ : Standardized AGI	-494.55*** (61.31)	-551.51*** (86.53)	236.83*** (44.76)	531.84*** (30.55)	606.15*** (35.88)	584.42*** (39.62)
μ : Constant	-1327.02*** (46.34)	-1378.99*** (68.36)	-1116.09*** (34.32)	-998.82*** (22.42)	-1042.87*** (25.24)	-1087.56*** (28.74)
$\ln(\nu)$: Standardized AGI	-0.49 (0.92)	1.14** (0.40)	0.52* (0.21)	0.35 (0.27)	-0.40 (0.59)	0.26 (0.25)
$\ln(\nu)$: Constant	3.37** (1.07)	-0.55 (2.22)	2.19*** (0.54)	2.37*** (0.48)	2.79*** (0.59)	2.72*** (0.45)
ψ : Standardized AGI	-335.38*** (56.16)	-393.59*** (78.35)	601.98*** (50.42)	763.67*** (39.62)	857.65*** (45.93)	823.09*** (49.63)
ψ : Constant	-920.49*** (40.81)	-962.37*** (59.46)	-520.75*** (37.95)	-536.86*** (27.70)	-539.20*** (31.93)	-590.56*** (35.45)
$\ln(\sigma_1)$: Standardized AGI	1.37*** (0.08)	1.53*** (0.13)	1.73*** (0.12)	1.77*** (0.12)	1.48*** (0.10)	1.53*** (0.11)
$\ln(\sigma_1)$: Constant	7.52*** (0.06)	7.74*** (0.10)	7.92*** (0.09)	8.03*** (0.10)	7.79*** (0.07)	7.81*** (0.09)
$\ln(\sigma_2)$: Standardized AGI	1.01*** (0.03)	1.07*** (0.04)	1.04*** (0.02)	0.91*** (0.02)	0.88*** (0.02)	0.90*** (0.02)
$\ln(\sigma_2)$: Constant	6.11*** (0.02)	6.18*** (0.03)	6.37*** (0.02)	6.36*** (0.02)	6.35*** (0.02)	6.38*** (0.02)
θ_1 : Standardized AGI	0.28** (0.11)	0.51** (0.16)	0.68*** (0.14)	0.78*** (0.14)	0.73*** (0.11)	0.57*** (0.12)
θ_1 : Constant	-0.88*** (0.07)	-0.85*** (0.09)	-0.99*** (0.07)	-0.85*** (0.07)	-0.69*** (0.07)	-0.85*** (0.07)
γ : Standardized AGI	1.05*** (0.13)	1.08*** (0.18)	1.91*** (0.21)	2.36*** (0.17)	2.06*** (0.17)	1.61*** (0.16)
γ : Constant	-1.33*** (0.10)	-1.26*** (0.14)	-1.15*** (0.16)	-0.64*** (0.13)	-0.98*** (0.13)	-1.05*** (0.13)
N	14927	7552	15542	15648	16076	15944

Standard errors in parentheses.

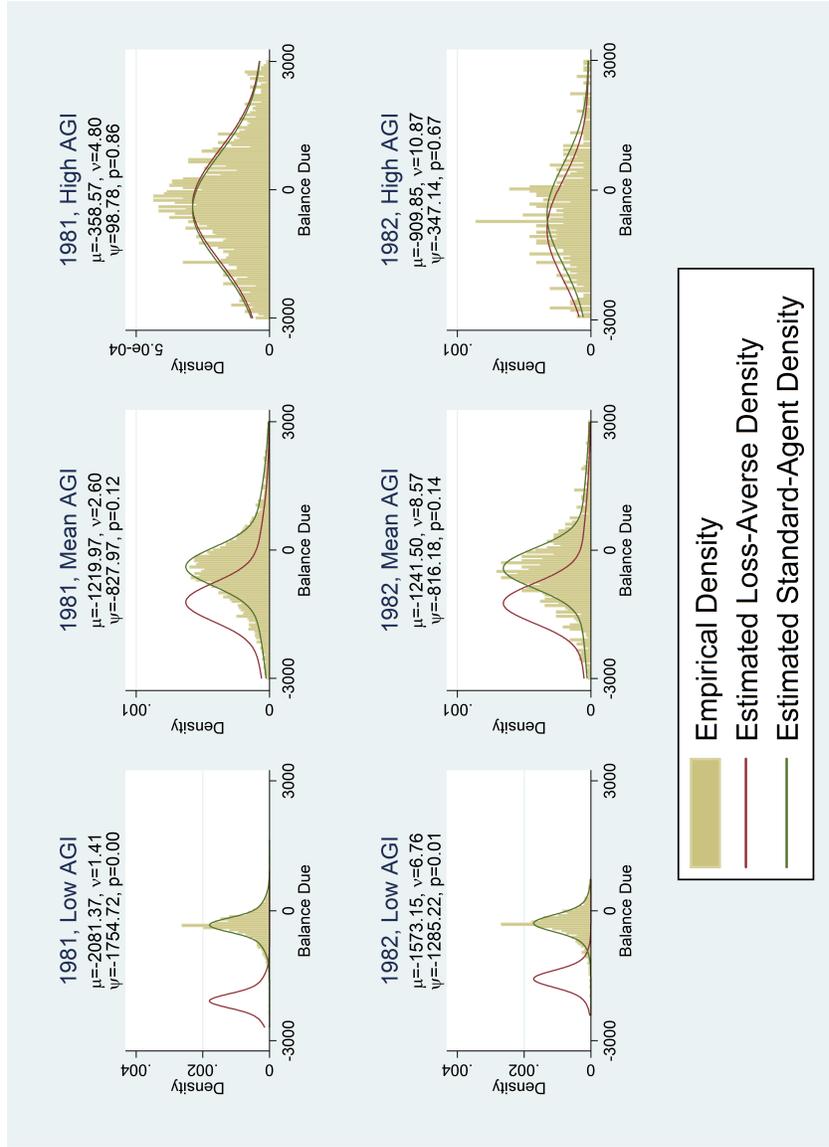
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A.1: Graphs of fit of structural estimates by year and AGI level:
1979-1980



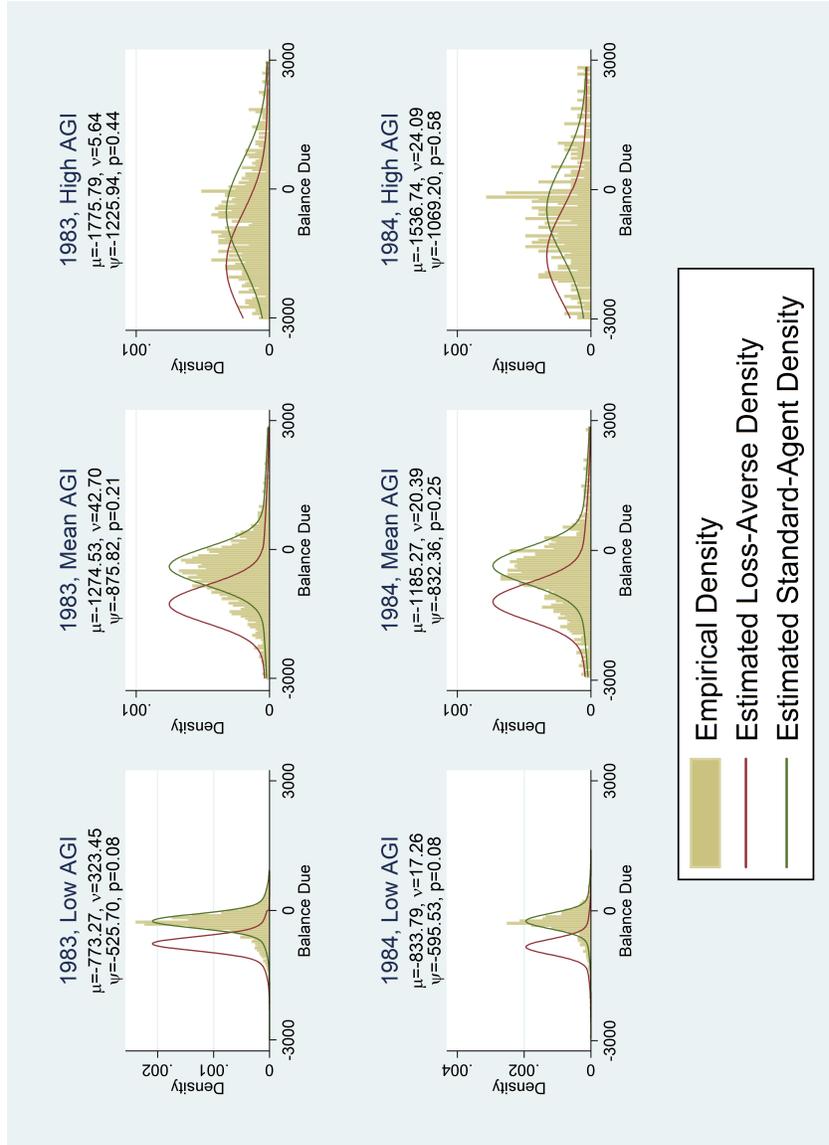
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.2: Graphs of fit of structural estimates by year and AGI level:
1981-1982



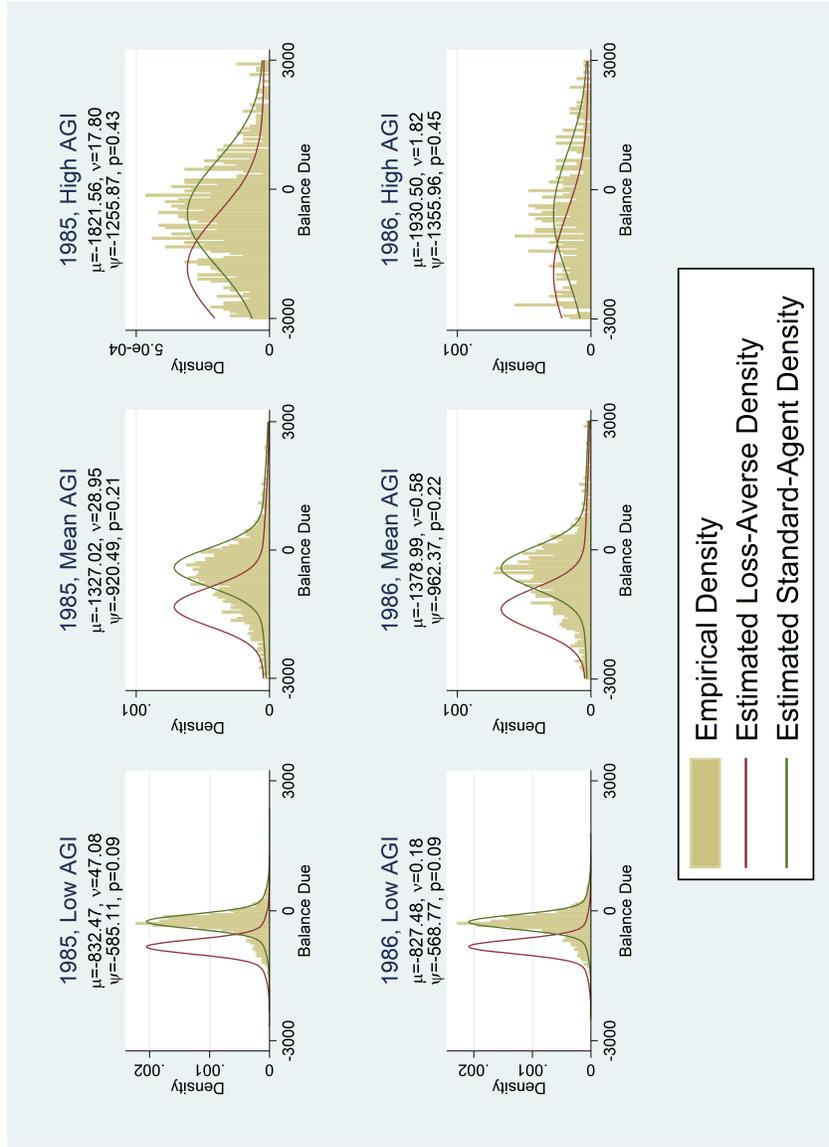
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.3: Graphs of fit of structural estimates by year and AGI level:
1983-1984



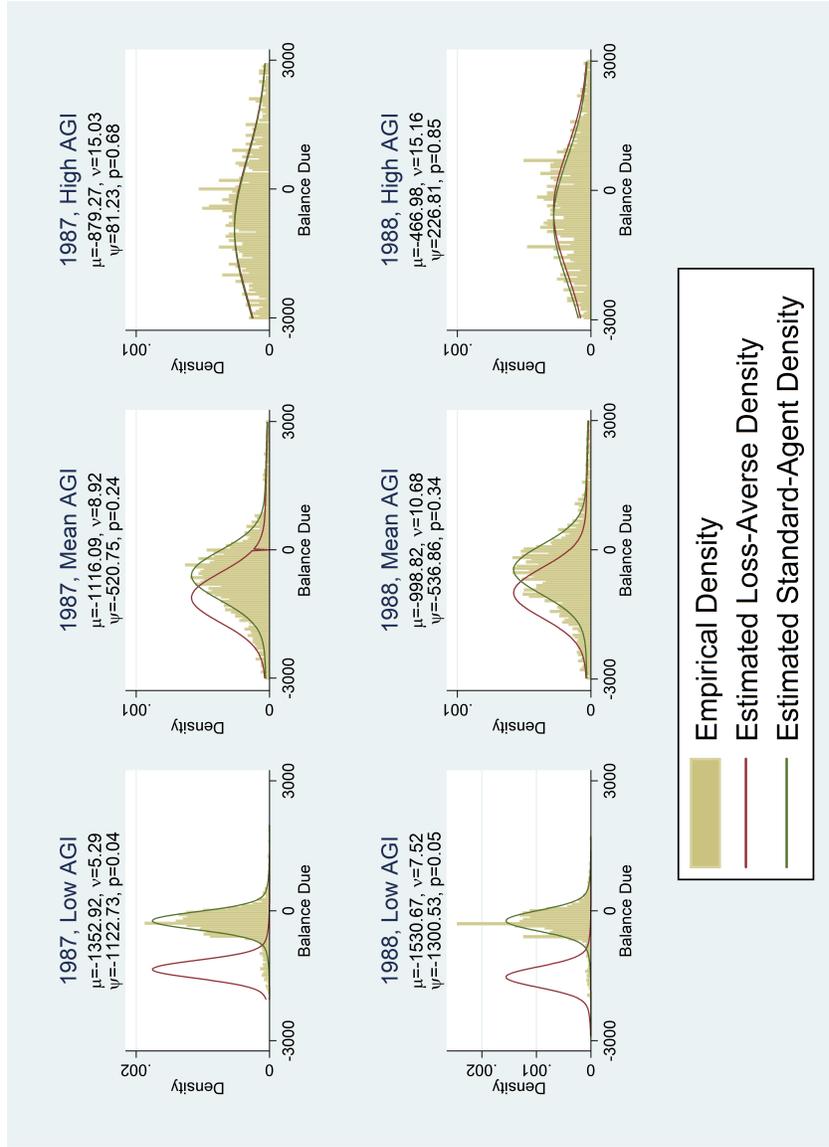
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.4: Graphs of fit of structural estimates by year and AGI level:
1985-1986



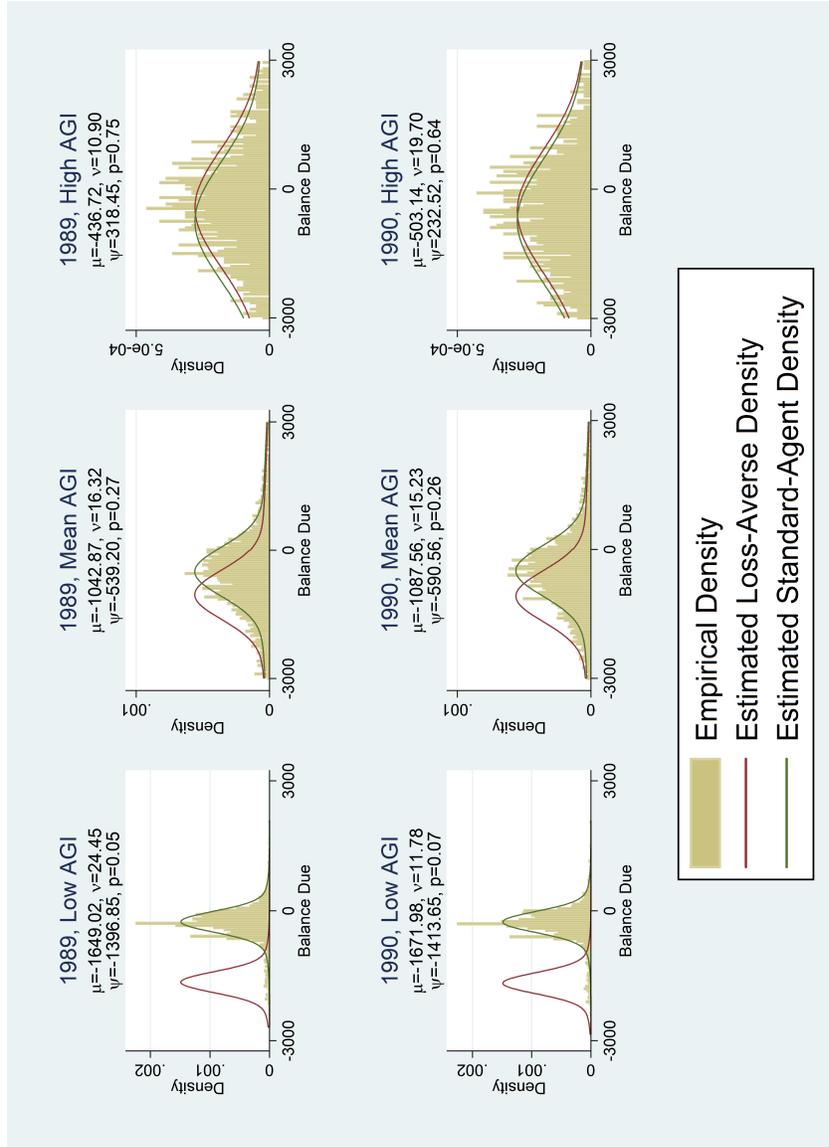
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.5: Graphs of fit of structural estimates by year and AGI level:
1987-1988



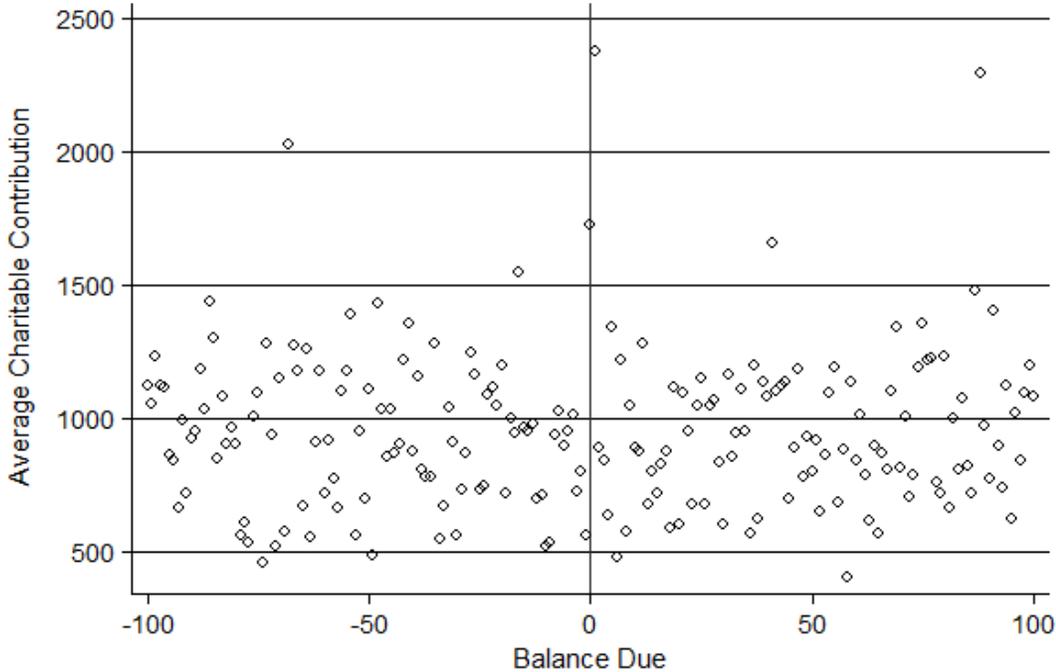
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.6: Graphs of fit of structural estimates by year and AGI level:
1989-1990



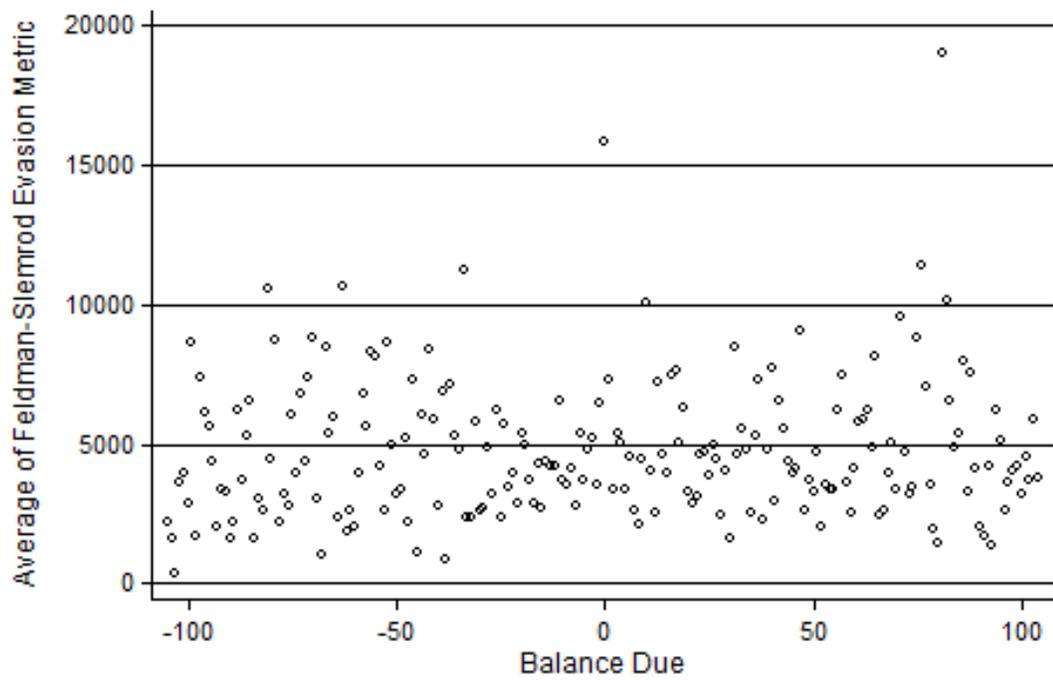
Notes: Graphs of the fit of the structural model, by year and by income group. The histograms have bins of width \$50. The low, mean, and high agi groups correspond to year-specific normalized agi levels of -1, 0, and 1, respectively. Histograms are restricted to individuals with year-specific normalized AGI within .2 of the relevant level.

Figure A.7: Charitable giving in the vicinity of zero balance due



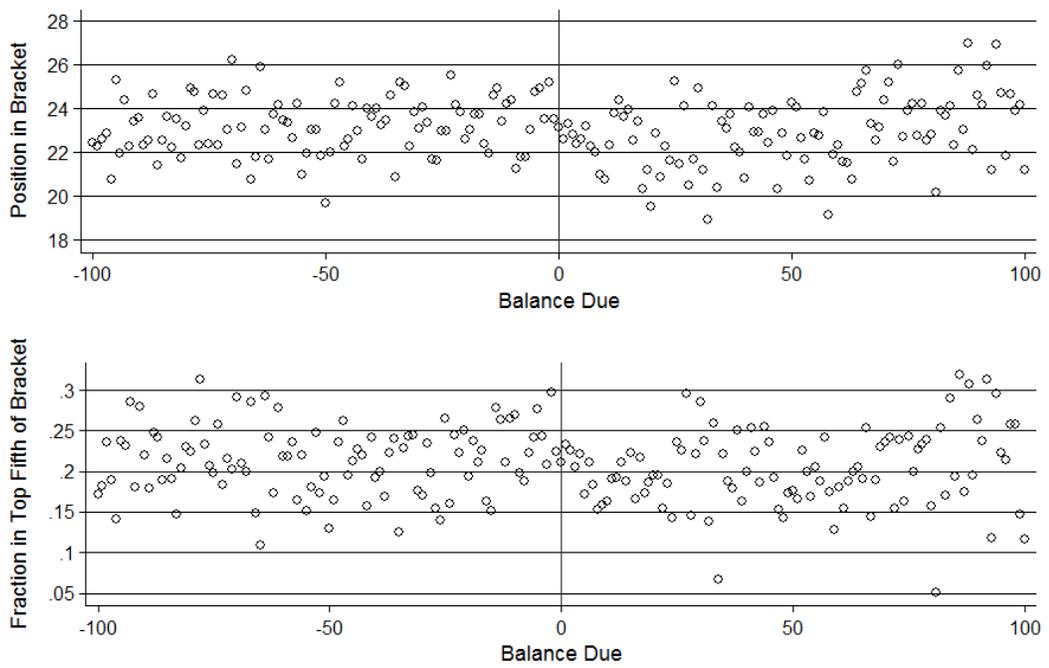
Notes: Restricted to individuals with positive charitable giving.

Figure A.8: Feldman and Slemrod (2007) evasion proxy values in the vicinity of zero balance due



Notes: Restricted to the estimation sample of the regression calculating these evasion proxies.

Figure A.9: Slemrod (1985) evasion proxy values in the vicinity of zero balance due



Notes: Restricted to individuals using tax tables with \$50 bins.

APPENDIX B

APPENDIX TO “WHAT DO YOU THINK WOULD MAKE YOU HAPPIER? WHAT DO YOU THINK YOU WOULD CHOOSE?”

Scenario 1: Sleep vs. Income

Say you have to decide between two new jobs. The jobs are exactly the same in almost every way, but have different work hours and pay different amounts.

Option 1: A job paying \$80,000 per year. The hours for this job are reasonable, and you would be able to get about 7.5 hours of sleep on the average work night.

Option 2: A job paying \$140,000 per year. However, this job requires you to go to work at unusual hours, and you would only be able to sleep around 6 hours on the average work night.

Scenario 2: Concert vs. Birthday

Suppose you promised a close friend that you would attend his or her 50th [“21st” in student samples] birthday dinner. However, at the last minute you find out that you have won front row seats to see your favorite musician, and the concert is at the same time as the dinner. This is the musician’s last night in town. You face two options:

Option 1: Skip your friend’s birthday dinner to attend the concert.

Option 2: Attend your friend’s birthday dinner and miss the concert.

Scenario 3: Absolute Income vs. Relative Income

Suppose you are considering a new job, and have offers from two companies. Even though all aspects of the two jobs are identical, employees' salaries are different across the two companies due to arbitrary timing of when salary benchmarks happened to be set. Everyone in each company knows the other employees' salaries. You must choose one of the two companies, which means you must decide between the following two options:

Option 1: Your yearly income is \$105,000, while on average others at your level earn \$120,000.

Option 2: Your yearly income is \$100,000, while on average others at your level earn \$85,000.

Scenario 4: Legacy vs. Income

(Phrasing in Denver within-subject study): Suppose you are a skilled artist, and you have to decide between two career paths for your life.

Option 1: You devote yourself to your own style of painting. This would require a number of sacrifices, such as having less time for friends and family, and making less money. For example, you expect that selling your paintings will give you an income of \$40,000 a year. If you choose this path, you don't expect that your work will be appreciated in your lifetime, but posthumously you will make an impact on the history of art, achieve fame, and be remembered in your work.

Option 2: You become a graphic designer at an advertising company. This would give you more money and more time with friends and family than Option 1. The company is offering you a salary of \$60,000 a year, which will afford

you a much more comfortable lifestyle, but you will have no impact and leave no legacy to be remembered.

(Version 2: Phrasing in Denver between-subjects study and Cornell studies): Suppose you are a skilled artist, and you have to decide between two career paths for your life. There are two styles of painting that you consider to be your own style, and you enjoy both equally. Style 1 happens to be much less popular than Style 2 today, but you know it will be an important style in the future.

Option 1: You devote yourself to Style 1. You expect that selling your paintings will give you an income of \$40,000 a year. If you choose this path, you don't expect that your work will be appreciated in your lifetime, but posthumously you will make an impact on the history of art, achieve fame, and be remembered in your work.

Option 2: You devote yourself to Style 2. You expect that selling your paintings will give you an income of \$60,000 a year, but you will have no memorable impact. [In the Denver between-subjects study, each subject saw this question three times, with different income levels in Option 2. Income levels could be \$42,000, \$60,000, \$80,000, or \$100,000.]

Scenario 5: Apple vs. Orange

Suppose you are checking out a new supermarket that just opened near where you live. As you walk by the fresh fruit display, you are offered your choice of a free snack:

Option 1: A freshly sliced apple.

Option 2: A freshly sliced orange.

Scenario 6: Money vs. Time

Suppose that due to budget cuts, the school implements a “student activities fee” of \$15 dollars a week to help pay for maintenance of facilities used for extracurricular student activities. However, the school allows you to not pay the fee if instead you put in 2 hours of service a week shelving books at the library. You face two options:

Option 1: Spend 2 hours a week shelving books.

Option 2: Pay \$15 a week.

Scenario 7: Socialize vs. Sleep

Say you are hanging out with a group of friends at your friend’s room. You are having a really good time, but it is getting to be late at night. You have to decide between two options.

Option 1: Stay up another hour. It is likely you will feel tired all day tomorrow, but this particular evening you are having an especially fun time.

Option 2: Excuse yourself from the group, and go to bed. You will be disappointed to miss the fun, but you know you will feel better the next day and be more productive at paying attention in class and doing your homework.

Scenario 8: Family vs. Money

Imagine that for the first time in three years, your parents (or if your parents are gone, your closest relatives who are older than you) have arranged for a

special family gathering that will happen the day after Thanksgiving, with everyone also invited to Thanksgiving dinner. You face two options. Would you choose to go to the family gathering the day after Thanksgiving (and maybe to Thanksgiving dinner) if getting there required a \$500 roundtrip plane ticket for plane flights that were 5 hours each way?

Option 1: Go to the thanksgiving gathering, which requires a \$500 round trip plane ticket.

Option 2: Miss the thanksgiving gathering, but save the money.

Scenario 9: Education vs. Social Life

Suppose you have decided to leave Cornell, and are transferring to a new school. You have been accepted to two schools, and are deciding where to go. The first school is extremely selective and high quality, but is in a small town out in the country with a less active social scene. The second school is in a major city with a great social scene, but is slightly less renowned. Which would you choose?

Option 1: Highly selective school, isolated socially and geographically.

Option 2: Less selective school, socially active and in a major city.

Scenario 10: Interest vs. Career

Suppose you are considering two summer internships. One is extremely interesting and involves work you are passionate about, but does not advance

your career. The other will likely be boring, but will help you get a job in the future. Which would you choose?

Option 1: Interesting internship which does not advance career.

Option 2: Boring internship which will help you get a job.

Scenario 11: Concert vs. Duty

Say you are driving by yourself to see your favorite musician in concert on their last day in town. You are five minutes away, and the concert starts in ten minutes. On the drive, you witness a truck hit a parked car, causing roughly \$500 in damages, and then drive away without leaving their information. You notice the truck's license plate, and you are the only witness. You face two options:

Option 1: Keep driving and get to the concert on time.

Option 2: Call the police, in which case you will have to wait around the parked car to give a testimony. This would take about half an hour. You would have trouble finding a seat and might miss the whole concert.

Scenario 12: Low Rent vs. Short Commute

(Phrasing in Denver within-subject study): Say you are moving to a new town. You are trying to decide between two similar apartments which you could rent. The two apartments are identical in almost everything - including floor plan, amenities, neighborhood character, schools, safety, etc. However, they have different rents and are located at different distances from your work.

Option 1: An apartment which requires a 45-minute drive to work. The rent is about 20% of your monthly income.

Option 2: A similar apartment, with only a 10-minute drive. The rent is about 40% of your monthly income.

(Version 2: Phrasing in Denver between-subject study): Say you are moving to a new town. The new town is known for its terrible traffic jams, and driving there is widely considered to be unpleasant. You are trying to decide between two similar apartments which you could rent. The two apartments are identical in almost everything - including floor plan, amenities, neighborhood character, schools, safety, etc. However, they have different rents and are located at different distances from your work.

Option 1: An apartment which requires a 45-minute drive each way to work. The commute has heavy traffic almost the whole way. The rent is about 20% of your monthly income.

Option 2: A similar apartment which requires a 10-minute drive each way to work. The commute has heavy traffic almost the whole way. The rent is about 40% of your monthly income.

Scenario 13: Friends vs Income

Say you have been reassigned at your job, and will be moved to a new location. There are two offices where you could request to work. One office is in a city where many of your friends happen to live, and pays 20% less than your current salary. The other office is in a city where you don't know anyone, and pays 10% more than your current salary. Your job will be exactly the same at

either office. You must decide between the following two options:

Option 1: Make 20% less than your current salary and move to the city with your friends.

Option 2: Make 10% more than your current salary and move to a city where you do not know anyone.