Structural Behavioral Economics*

Stefano DellaVigna
UC Berkeley and NBER

June 2018

Abstract

What is the role of structural estimation in behavioral economics? I discuss advantages, and limitations, of the work in Structural Behavioral Economics. I also cover common modeling choices and how to get started. Among the advantages, I argue that structural estimation builds on, and expands, a classical behavioral tool, simple calibrations, and that it benefits from the presence of a few parsimonious behavioral models which can be taken to the data. Estimation is also well suited for experimental work, common in behavioral economics, as it can lead to improvements in the experimental design. In addition, at a time where policy implications of behavioral work are increasingly discussed, it is important to ground these policy implications in (estimated) models. Structural work, however, has important limitations, which are relevant to its behavioral applications. Estimation takes much longer and the extra degree of complexity can make it difficult to know which of a series of assumptions is driving the results. For related reasons, it is also easy to over-reach with the welfare implications. Taking this into account, I provide a partial how-to guide to structural behavioral economics, covering: (i) the choice of estimation method; (ii) the modeling of heterogeneity; (iii) identification and sensitivity. Finally, I discuss common issues for the estimation of leading behavioral models. I illustrate this discussion with selected coverage of existing work in the literature.

*Forthcoming in the 1st Handbook of Behavioral Economics, Vol.1, edited by Douglas Bernheim, Stefano DellaVigna, and David Laibson, *Elsevier*. I thank Hunt Allcott, Charles Bellemare, Daniel Benjamin, Douglas Bernheim, Colin Camerer, Vincent Crawford, Thomas Dohmen, Philipp Eisenhauer, Keith Ericson, Lorenz Goette, Johannes Hermle, Lukas Kiessling, Nicola Lacetera, David Laibson, John List, Edward O'Donoghue, Gautam Rao, Alex Rees-Jones, John Rust, Jesse Shapiro, Charles Sprenger, Dmitry Taubinsky, Bertil Tungodden, Hans-Martin von Gaudecker, George Wu, and the audience of presentations at the 2016 Behavioral Summer Camp, at the SITE 2016 conference, and at the University of Bonn for their comments and suggestions. I thank Bryan Chu, Avner Shlain, Alex Steiny, and Vasco Villas-Boas for outstanding research assistance.

1 Introduction

Behavioral economics, with its lessons regarding non-standard preferences, beliefs, and decision-making, has important applications in most fields of economics. This Handbook is a clear illustration of this broad reach, with chapters on a variety of fields, including finance, public economics, and industrial organization.

The applications in these fields employ a range of data sources—observational studies, survey collection, laboratory experiments, and field experiments, among others. The applications come with a variety of estimation methods, including simple treatment-control comparisons in experiments, correlations, instrumental variables, but also structural estimation.

In this chapter I ask: Is there an important role for structural estimation in behavioral economics, or for short *Structural Behavioral Economics*? For our purposes, I define structural as the "*estimation of a model on data that recovers estimates (and confidence intervals) for some key behavioral parameters*".[1] Further, are there special lessons for structural estimation in behavioral economics beyond the well-known advantages, such as the ability to do welfare and policy evaluations, but also the well-known pitfalls, such as the complexity of the analysis?

I argue that the answer is: Yes, and Yes. In Section 2, I discuss six advantages of structural estimation, several of which have roots in key features of behavioral research. One of the most important ones is that estimation builds on a long-standing tradition in behavioral economics of calibration of models, often through simple back-of-the-envelope calculations. Taking seriously the economic magnitude of the calibrated parameters has been the foundation for important behavioral insights, such as Rabin's calibration theorem for risk (Rabin, 2000). As I argue, estimation takes the calibration one step further, including cases in which a simple calibration is not possible.

Second, in behavioral economics there has always been a healthy exchange of ideas between theorists and applied researchers. Partly because of the focus on calibrating models, behavioral theorists have paid attention, arguably more than in some other fields, to empirical evidence. Conversely, empirical researchers, given the importance of testing the null hypothesis of the standard model, have typically followed closely the development of behavioral theory, or at least applied theory. Structural estimation builds on, and reinforces, this closeness, as it forces empirical researchers to take seriously models which they bring to the data.

Third, structural estimation also benefits from the fact that behavioral economics has a small number of widely-used parsimonious models, such as beta-delta preferences (Laibson, 1997; O'Donoghue and Rabin, 1999a) and reference dependence (Kahneman and Tversky, 1979). The presence of commonly-used models makes it more useful to test for the stability of estimates across settings.

Fourth, a key advantage of structural estimation is that one can estimate the out-of-sample performance of a model, a stronger test than in-sample fit. Indeed, we point to at at least one case in which a behavioral model and a standard model have similar in-sample fit, but the out-of-sample predictions clearly tell apart the models. Still, out-of-sample predictions appear under-utilized in behavioral economics.

A fifth advantage relates to a key feature of the behavioral field: the importance of experimental

[1] For definitions of structural estimation see Reiss and Wolak (2007), Wolpin (2013), and Rust (2014).

evidence, both from the laboratory, source of much of the initial behavioral evidence, and from the field. In experiments, there are extra advantages to estimation: paying close attention to the models at the design stage can lead to different designs that allow for a clearer test of models. In observational studies, in contrast, the design is limited by the data and the setting (though models can of course motivate the search for the right observational design). This particular advantage of models, interestingly, so far has played a larger role in laboratory experiments than in field experiments (Card, DellaVigna, and Malmendier, 2011). There is an opportunity for more work of this type.

A sixth motivation is shared by all applications of structural estimation: welfare and policy analysis. The timing for that in behavioral economics is just right. While behavioral economics has mostly shied away from policy implications until the last decade, the recent emphasis on cautious paternalism (Camerer et al., 2003), nudges (Thaler and Sunstein, 2008), and behavioral welfare economics (Bernheim and Rangel, 2009) substantially increased the policy reach of behavioral economics. Yet, many policy applications of behavioral findings do not have a fully worked out welfare or policy evaluation. Structural estimation has a role to play to ensure, for example, that we "*nudge for good*".

Having said this, should all of behavioral economics be structural? Absolutely not. To start with, many studies do not lend themselves well to structural estimation, for example because they explore a channel for which we do not have yet a well-understood model, e.g., framing effects, or the interest is on a reduced-form finding. In addition, even in cases in which there is an obvious model-data link, an alternative strategy is to derive comparative statics from the model (e.g., Andreoni and Bernheim, 2009), including in some cases even an axiomatic characterization (e.g., Halevy, 2015), to derive empirical testable predictions. This strategy allows for clear model testing, without the extra assumptions and time involved in structural estimation.

For the studies where structural estimation makes sense, in Section 3 we outline common limitations of structural estimation. These limitations are shared with applications of structural estimation in other fields, but I emphasize examples, and specific issues, within behavioral economics.

First, and perhaps most obviously, structural estimation typically takes much more time, given the number of necessary steps: the reduced-form results, spelling out the full model, the estimation strategy, and getting to reliable estimates. The estimation itself can be a very time-consuming step, and indeed much of the training for work in the structural area revolves around computational shortcuts and techniques to ensure that the results are robust. An implication is that structural analysis, being more complex, also increases the chance that programming errors may drive the results, or that the estimates may not be stable. These important time and complexity costs must be weighed against the benefits above.

A possible saving grace from this cost is well-known in the literature: sufficient statistics (Chetty, 2009). In some cases, a parameter, or a combination of parameters, can be estimated using a key statistic, or a combination of statistics, that is sufficient for estimation (hence the name). Sufficient statistics, thus, allow for structural estimation (of some parameters) using reduced-form findings, without the need for a full estimation. This is a desirable route when possible, and has been used for example to obtain estimates of some limited attention models. Sufficient statistics sometimes allow

for estimation of a structural parameter even for published papers that report only reduced-form results, as in the case of persuasion rates.

I stress an additional route to simplify estimation. When the study has an experimental nature, the researcher can alter the design to make estimation simpler. That is not an option available for observational studies, where the data is what it is, and at best the researcher can seek the best natural experiment, or data details, for model identification. But in experimental studies, the research can add treatments, or alter design features, precisely to make the structural estimation easier and more transparent. This often takes the form of some 'pricing-out' treatments, and can tip the design to within-subject designs (though not necessarily). I stress that this is another reason to work on the estimation already at the design stage, before running the experiment.

A second issue with structural estimation is that the estimates, and ensuing welfare and policy implications, are only as good as the joint set of assumptions going into the model. The estimates may be sensitive to changing some of the auxiliary assumptions, and it is often difficult to thoroughly test the robustness of the estimates. I discuss some common-sense approaches.

Relatedly, it is easy, after all the work of estimation, to take the welfare and policy implications too much at face value. At a minimum, one ought to know that the confidence intervals for the welfare implications do not allow for the model being wrong. A robust discussion of alternative assumptions, and how those could affect the implications, is important.

In the final section of the paper, I discuss some nuts and bolts of structural behavioral economics, aimed in particular at behavioral researchers interested at taking a step in that direction. I outline first the choice of the estimation method. A transparent choice is a minimum distance estimator: one identifies some moments in the data and then finds the set of model parameters that minimizes the distance between the empirical moments and the theory-predicted moments. One of the earliest papers in the Structural Behavioral Economics literature, Laibson, Repetto, and Tobacman (2007), for example, takes this route for a consumption application. A second common method, which has the advantage of using all the variation in the data (as opposed to just selected moments), is maximum likelihood: one finds the model parameters that maximize the statistical likelihood, given the data. Maximum likelihood estimation was employed on field data in another early paper in the Structural Behavioral Economics literature, Conlin, O'Donoghue, and Vogelsang (2007), and quite a bit earlier in laboratory experiments (e.g., Harless and Camerer, 1994, El-Gamal and Grether, 1995, and Costa-Gomes, Crawford, and Broseta, 2001). I also discuss non-linear least squares as another example (e.g., DellaVigna et al., 2016).

A second important choice is the modeling of the heterogeneity. When taking a model to the data, a key step is asking how the theory will match the heterogeneous behavior in the data. A common approach is to assume random utility, as in McFadden (1999): an unobserved (to the researcher) utility shock rationalizes the heterogeneity. A second approach is to assume heterogeneity of some key structural parameter, as in the random effect or mixture models: for example, the individuals may differ in their social preference parameter (DellaVigna, List, and Malmendier, 2012) or in their cognitive type (Costa-Gomes, Crawford, and Broseta, 2001). A third approach is to assume that the individuals make implementation errors, as in the trembling hand approach (e.g., Augenblick and Rabin, forthcoming).

I also discuss the distinction between key parameters and incidental parameters; while the latter ones are not of interest per se, their estimates and specifications are critical to the estimation of the key behavioral parameters. Relatedly, I discuss strategies to explore the sensitivity of structural estimates to key assumptions and to changes in moments in the data. Further, I discuss the important role of simulate-and-estimate exercises to get the ultimate structural estimation well set up.

Finally, I discuss the estimation of some common models in behavioral economics. Perhaps the most commonly used is the beta-delta model of time preferences. I summarize four key features of this model to keep in mind, especially for structural estimation: (i) timing of payoffs; (ii) money versus consumption; (iii) time period duration; and (iv) sophistication versus naiveté. On the latter point in particular, I discuss how the naive version is typically much easier to estimate while, it seems, providing an acceptable approximation in several settings. Next, I discuss the estimation of reference-dependent models à la prospect theory (Kahneman and Tversky, 1979), including a brief discussion of backward-looking versus forward-looking reference points. I also briefly discuss the estimation of social preference models.[2]

Throughout this chapter, I discuss a number of behavioral papers with structural estimation. In Table 1, I summarize key features for these papers, grouping them by behavioral features as in the DellaVigna (2009) review. For each paper, the table indicates the type of data used (observational, experimental, etc), the behavior examined (e.g, life-cycle consumption), the parameters of interest (e.g., beta and delta), some of the incidental parameters (e.g., the curvature of the cost of effort function), the estimation method (e.g., minimum distance), as well as the source of heterogeneity (e.g., random utility). Thus, the table provides a partial overview along a few dimensions of some exemplary papers in Structural Behavioral Economics.

I stress that this review, while citing some of the cases of structural estimation in laboratory experiments, is mostly focused on behavioral evidence from the field. This review is also not focused on structural estimation itself, but on its application to, and role within, behavioral economics. For coverage on structural estimation, see for example Reiss and Wolak (2007), Wolpin (2013), or Rust (2014).

2 Advantages of Structural Behavioral Economics

2.1 Calibrated Magnitudes

In a paper in the 2001 issue of the *Quarterly Journal of Economics*, Madrian and Shea provide striking evidence of inertia due to switching costs even in a setting with high stakes for getting the right choice. Changing the default for retirement savings participation in 401(k) plans from opt-in to opt-out (across two contiguous cohorts of employees) increases the participation into a 401(k) plan from about 45 percent to about 90 percent. Thus, nearly half of the workers are swayed by having to do an administrative task that is likely to take just a couple of hours. Soon thereafter, Choi et al. (2004) show that this finding holds, with almost identical magnitudes, in a number of

[2]See also the chapters in this Handbook on time preferences (Ericson and Laibson, 2018), reference dependence (O'Donoghue and Sprenger, 2018), and social preferences (Andreoni and Bernheim, 2018).

different firms.

The key issue here is that the standard model makes the right *qualitative* prediction: switching to opt-out should increase the participation rate. But are the *quantitative* magnitudes right? Building on O'Donoghue and Rabin (1999b), we walk through a simple, but fairly detailed, calibration to show that the answer is no. The benefits from switching are too large, over a lifetime of savings, for a one-time cost in the order of a hundred dollars or less to matter.

Assume that investment for retirement in an opt-in regime requires an immediate effort cost k and it involves setting aside s in retirement savings in each period, which we think of as a day. The money will earn a potential match μ from the company and will grow with interest rate r (which we assume fixed and known) all the way until retirement, which is in T periods, at which point (for simplicity) the money is distributed as a lump-sum amount. Assume also that investment is a once-and-for-all decision, so once the person starts saving, she saves s each period until retirement. To invest money s, the person must reduce consumption by s. For simplicity, we assume away taxation. The net utility of investing right away, at $t = 0$, is (up to a first-order approximation)

$$U_0 = -k + \beta\delta^T \sum_{t=0}^{T-1} (1+r)^{T-t} s(1+\mu) u'(c_T) - s u'(c_0) - \beta \sum_{t=1}^{T-1} \delta^t s u'(c_t). \tag{1}$$

The first term captures the immediate effort cost of filling in forms, followed by the benefit of retirement: the accumulated savings, inclusive of match, put aside from period 0 all the way until period $T-1$. The savings translate into increased consumption at retirement, with a benefit which we linearized locally with marginal utility $u'(c_T)$, and discounted back with $\beta\delta^T$. The next two terms capture the fact that the agent needs to cut back on consumption in order to save, starting from the current period 0 and continuing in the next periods. To simplify the calibration, we assume a constant marginal utility of consumption renormalized to 1, and also assume that the long-term discounting equals the interest rate, that is, $\delta = (1+r)^{-1}$. Under these assumptions[3], the utility of investing today is

$$U_0 = -k + s(\beta(1+\mu) - 1) + \beta s\mu \frac{\delta - \delta^T}{1-\delta}. \tag{2}$$

Under the same assumptions, waiting and investing at a future time period τ yields discounted utility, as of today, of

$$U_\tau = -\beta\delta^\tau k + \beta s\mu \frac{\delta^\tau - \delta^T}{1-\delta}. \tag{3}$$

The key difference between immediate investment and delayed investment is that in this latter case the cost of effort and consumption loss both occur only in the future.

For the exponential discounter ($\beta = 1$), the problem reduces to switching today or never, as the total expected benefit (if positive) decreases with delay, as the agent is forgoing the match rate. Given that the value of never switching is normalized to 0, she switches if $-k + s\mu\left(1-\delta^T\right)/(1-\delta) \geq 0$.

[3]This calibration thus abstracts from three factors which are likely to yield higher return to investment compared to this simple calibration: (i) the marginal tax rate is plausibly lower at retirement, when 401(k) savings are taxed, than at present, given the lower income post-retirement; (ii) if individuals are under-saving, the marginal utility from consumption may be higher after retirement than at present, and (iii) the historically high equity premium implies that plausibly $1 + r > 1/\delta$.

To calibrate this, assume a daily discount factor $\delta = .9998$, corresponding to an annual discount factor $\delta^{365} = 0.93$. The match in the Madrian and Shea (2001) case, and in many companies, is 50 percent ($\mu = 0.5$), which applies up to 6 percent contribution. We assume that individuals, if they invest, save up to the 6 percent match, which is the modal contribution in the opt-in cohort. Given an average salary annual of \$38,000, that translates into a daily saving s of about \$6 for the average worker. A 30-year old exponential worker, for whom $T = 35 * 365$, will thus invest if $k \leq \$13,835$. Even a person just one year before retirement ($T = 365$) will need to have a switching cost above \$1,000 to justify not signing up. The value of the accrued match is simply too high to forgo. Thus, an exponential individual would invest, whether under opt-in—which we modeled as $k > 0$—or under opt-out—which we can model as $k < 0$. This model thus predicts that default effects would not impact the retirement decision, in clear contradiction to the Madrian and Shea (2001) evidence.

But can a behavioral model do any better? Self-control problems would seem like a natural explanation. O'Donoghue and Rabin (1999b) show that a person with self-control problems of the beta-delta type (Laibson, 1997), so long as the person is *sophisticated* about the self-control problem, would also not delay much. This person hates to do the task at present, given the self-control problem, but is also aware that all future selves will hate it too, and thus will also be tempted not to do the task. In that battle of the selves, she knows that at some point relatively soon she has to do it, or else it is better to do it now, despite the self-control problem. While there are multiple pure-strategy Markov equilibria, O'Donoghue and Rabin (1999b) show that one can derive a bound on the maximal delay across all the equilibria. More formally, the maximal delay for a sophisticated present-bias agent ($\beta = \hat{\beta} < 1$) is given by the number of days τ that make the agent indifferent between doing the task today and doing the task in days τ. We can solve for that equating expressions (2) and (3), yielding

$$k = \frac{s\left[\beta\left(1+\mu\right) - 1 + \beta\mu\delta\frac{1-\delta^{\tau-1}}{1-\delta}\right]}{1-\beta\delta^{\tau}}. \tag{4}$$

Using a Taylor expansion for $\delta \to 1$, $(1-\delta^{\tau}) \approx (1-\delta)\tau$, then $k \approx s\left[(\beta/(1-\beta))\mu\tau - 1\right]$. Assuming a $\beta = 0.8$ and the same parameters as above, to justify a delay of 30 days requires a switching cost of $\approx$ \$350, above what is plausible. Thus, it turns out, a model of self-control problems, if individuals have rational expectations, is not really better than the standard model at explaining the magnitudes of the status-quo finding for retirement savings.

O'Donoghue and Rabin (1999b) are not finished, though. They show that things change completely if the person with self-control problems is also *naive* about the future self-control problems. That is, she expects that future selves will not have self-control problems, or at a minimum will have much less severe self-control problems than the current self has. With fully naive present-bias ($\beta < \hat{\beta} = 1$), she expects to behave like an exponential discounter tomorrow and as in the calibration above, expects to switch tomorrow.[4] Thus, using expression (4) for $\tau = 1$, the naif will decide to wait until tomorrow to switch whenever

[4]This assumes that the future self, with exponential preferences (that is, $\beta = 1$), would decide to invest. As we showed above, this is the case for reasonable switching costs, provided the agent is not just about to retire.

$$k > \frac{s\left[\beta\left(1+\mu\right)-1\right]}{1-\beta\delta},$$

or approximately when $k > \$6$ for $\beta = 0.8$. For plausible switching costs, then, a naive person engages in procrastination because she believes that future selves will do the task, and thus happily defers the task into the next period. The next period self, instead, passes it on to the next self, and so on. This model can explain why individuals do not invest under opt-in even for fairly small k, but instead invest under opt-out (since $k < 0$), thus reproducing the large observed default effects. Importantly, the idea that we may be at least somewhat naive about our self-control problems is in line with evidence on overoptimism in other areas, so the naive version of the beta-delta model is not out of line of our understanding of human psychology and behavior. Thus, a calibration provides support for naive present-bias as a leading explanation for (large) default effects in retirement savings, as well as in other applications.

This is hardly the only setting where calibrations played a key role in behavioral economics. In a very different setting, risk and insurance, behavioral economists had become intrigued by why people would buy (costly) extended warranties for small appliances, even when the risk was simply to have to buy a new, say, shaver at a moderate cost. Should people really pay to insure against this small-stake risk, when the insurance was clearly not fairly priced?

The Rabin (2000) calibration theorem provides an answer: expected-utility maximizers should not exhibit high degrees of aversion to small risk, for any plausible level of the risk aversion parameters. If individuals are averse to a sequence of small risks, they will become incredibly averse to even moderately-sized risk, as the calibration in the theorem clarifies. Said otherwise, an expected-utility maximizer should be approximately risk neutral with respect to small risks. To understand small-stake risk aversion, one needs to look elsewhere, for example at prospect theory, where the kink in utility at the reference point generates first-order risk aversion (e.g., O'Donoghue and Sprenger, 2018).

Calibrations, thus, played a key role in establishing two important results in behavioral economics, inertia and small-stake risk aversion. More generally, behavioral economists have focused on quantitative tests of standard and behavioral models in other settings as well.

If calibrations are so useful, why would one need structural estimation? Calibrations are, of course, related to structural estimation. A key difference is that calibrations can provide an order of magnitude for a parameter, but do not pin down a point estimate, nor provide confidence intervals. Confidence intervals are important because they indicate how confident we are in our inference, given a set of assumptions. Furthermore, back-of-the-envelope calibrations like the two featured above can be derived only in relatively simple models.

Calibration of More Realistic Models. A first point is that it is hard to provide back-of-the-envelope calibration for more realistic (and complex) models, and the back-of-the-envelope calibrations for the simpler cases can sometimes be misleading. Returning to the default effects example, the O'Donoghue and Rabin (1999b) calibrations are based on a deterministic switching cost k which remains constant over time. A more realistic assumption is that the switching cost k is stochastic and varies day-to-day, to reflect the fact that we are busier on certain days than on

others. In this simple dynamic programming problem, the solution consists of a threshold $\overline{k}$ such that the individual will do the task if she draws a $k \leq \overline{k}$, and wait otherwise (e.g., DellaVigna and Malmendier, 2006; Carroll et al., 2009).

Does this more realistic structure change the original back-of-the-envelope calibration in O'Donoghue and Rabin (1999b)? Several results are unchanged. The expected time delay for realistic values of the parameters is going to be short under the standard exponential discounting, and the same still holds under the sophisticated beta-delta model. The result for the naive beta-delta case, though, changes in an important way.

The naive agent decides whether to switch in period t, at the realized k, or receive the expected continuation payoff from waiting until the next period $t+1$:

$$-k + s\left(\beta\left(1+\mu\right)-1\right) + \beta s\mu \frac{\delta - \delta^{T-t}}{1-\delta} \geq \beta\delta V_{t+1}^{e}. \tag{5}$$

Since the (fully) naive individual believes that she will have exponential preferences from the next period on, she believes that the relevant continuation payoff is the same as for the exponential worker, thus the use of V_{t+1}^{e} in expression (5). Compare this to the dynamic programming problem for the exponential worker, which is $-k + s\mu + s\mu\left(\delta - \delta^{T-t}\right)/\left(1-\delta\right) \geq \delta V_{t+1}^{e}$. It is thus easy to show that the critical threshold $\overline{k}_t$ which makes the person indifferent between paying the cost now and waiting satisfies the following:

$$\overline{k}_t^{n} = \beta\left[s\mu\frac{\delta-\delta^{T-t}}{1-\delta} + s\mu - \delta V_{t+1}^{e}\right] - s\left(1-\beta\right) = \beta\overline{k}_t^{e} - s\left(1-\beta\right)$$

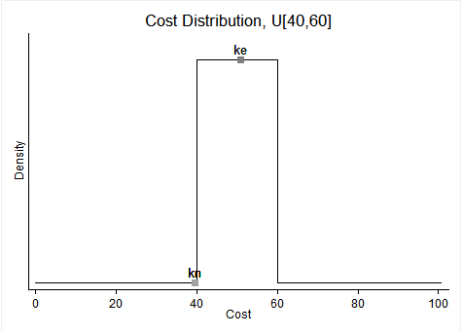
where $\overline{k}_t^{n}$ is the threshold for a naive agents and $\overline{k}_t^{e}$ is the threshold for an agent with the same parameters but $\beta = 1$. The threshold for the naive individual, $\overline{k}_t^{n}$, can be derived from the threshold for the exponential individual, $\overline{k}_t^{e}$, simply by multiplying it by the present-bias parameter β, and then applying the correction for the added disutility of money from the immediate consumption cut,$-s\left(1-\beta\right)$. Thus, having solved the problem for the exponential case, solving the naive case is very straightforward, a point we return to in Section 4.

Using this result, we can revisit a result from O'Donoghue and Rabin (2001): a naive procrastinator is likely to delay forever for realistic parameter values under opt-in. This result now depends critically on the distribution of the cost k. Figure 1a displays a case in which the cost distribution is relatively narrow, $k \sim U[40, 60]$. In this case, assuming as above $\delta = .9998$, $s = \$6$ and $\mu = 0.5$, the threshold for the exponential (which is stable far enough from retirement) is $\overline{k}^{e} = 50.94$, and the agent saves for retirement with probability $p = P\left(k \leq \overline{k}^{e}\right) = 0.547$ in each day. The expected delay for this exponential agent is short: the probability that the agent has not yet saved for retirement after 30 days is small, since $(1-p)^{30}$ is practically zero (see Figure 1b). Thus, as discussed above, an exponential agent with calibrated parameter values does not wait long.

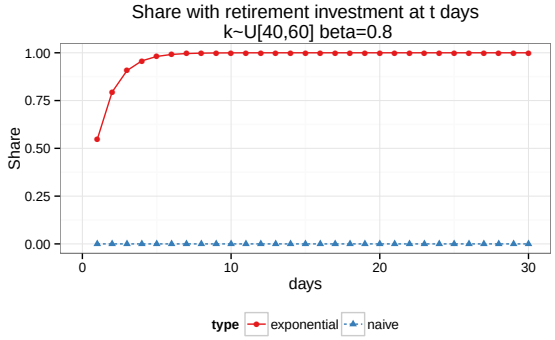
Consider now a present-bias employee with $\beta = 0.8$, a moderate degree of present bias compared to most studies (see Section 2.2). This individual will have threshold $\overline{k}^{n} = \beta\overline{k}_t^{e} - s\left(1-\beta\right) = 39.55$. Since $\overline{k}^{n} < 40$, the lower bound of the cost realization, the naive agent will never invest. Thus, this case replicates the O'Donoghue and Rabin (2001) calibration that even a moderate degree of present

Figure 1: Example cost distributions and threshold costs for naive present-biased and exponential discounters.

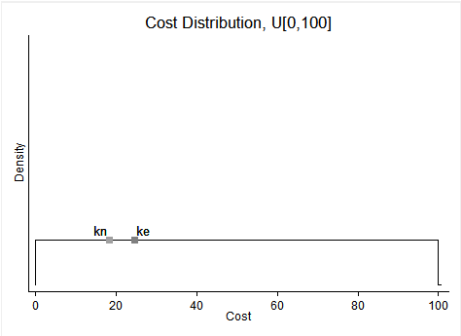
(a) Narrow cost distribution, optimal threholds.



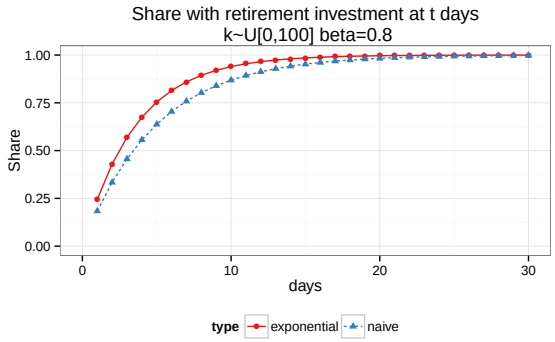
(b) Share saving for retirement over time under opt-in, narrow cost distribution.



(c) Wide cost distribution, optimal threholds.



(d) Share saving for retirement over time under opt-in, wide cost distribution.



bias generates very large delay.

Consider however now the case in Figure 1c, with a distribution of costs with the same mean, but a larger variance: $k \sim U[0, 100]$. For the parameters above, the threshold for the exponential worker (far enough from retirement) is $\overline{k}^e = 24.48$. The agent saves for retirement with probability $p = P\left(k \leq \overline{k}^e\right) = 0.245$ each day, and thus ends up saving for retirement quickly as well. The key difference is that now a present-bias employee with $\beta = 0.8$ has a threshold $\overline{k}^n = \beta \overline{k}_t^e - s\,(1-\beta) = 18.38$ and thus waits just about as long as the exponential agent (see Figure 1d). Even for a fairly low value of $\beta = 0.5$, the agent will still invest on any given day with probability $p = P\left(k \leq \overline{k}^n\right) = 0.092$ and thus the probability of delay past 30 days is very small ($(1-p)^{30} \simeq 0.05$). In fact, in this model, only for a very low $\beta \leq \frac{s}{\overline{k}^e + s} = 0.197$ will the naive individual procrastinate for sure.

Thus, the O'Donoghue and Rabin (2001) calibration for the naive procrastination case depends on the exact distribution of the cost function, something that we would not be able to easily calibrate. This is an example in which enriching a model changes the results compared to the ones obtained through a back-of-the-envelope calibration. The richer model could be estimated, provided there is a way to identify the cost distribution.[5] Thus, a first benefit of estimation is that it can handle more realistic models of decision-making compared to simple calibrations, and the more realistic models can make a difference.

Providing Point Estimates. A second benefit, as obvious as it is, is that structural estimation provides a confidence interval for the estimates, and thus can address the question of how sure we are about a particular parameter estimate. Returning to the case of inertia but in the context of employee choice among health insurance, Handel (2013) exploits a natural experiment to estimate switching costs. A firm restructured its health insurance options in year t, and it required the employees to make an active choice among the new options. The next year, the firm made some further adjustments to the insurance plans, but it did not require active choice: while employees could switch, the ones who did nothing stayed for year $t+1$ in the same plan they had in year t. The interesting part is that, as it turns out, the change in plans between year t and $t+1$ was substantial enough that, for a sub-sample of employees, the plan that they chose in year t, let's call it plan P_0, became dominated, in year $t+1$, by another plan, let's call it P_1. (The two plans only differ in the financial terms, as they grant access to the same doctors.) That is, no matter what the ex-post usage of the insurance, the person would have lost money by staying with plan P_0: the premium is over \$3,000 higher for plan P_0 and the difference in deductibles and co-pay is such that costs are always at least \$1,000 higher under plan P_0 than under plan P_1 (in year $t+1$). And yet, 80 percent of individuals who chose plan P_0 at t stick with it in year $t+1$! These two facts—the fact that one plan dominates the other by at least \$1,000 and that 80 percent of employees stick with it—implies with a simple calibration that switching costs in a standard model should be in the ballpark of \$1,000, or higher.

Yet, the case above applies to only a subset of employees. Does the finding apply broadly among the employees of this firm? Handel (2013) models the choice of health insurance in years t, $t+1$ using the predicted health needs given the health care utilization in the previous year, and assuming

[5] I am not aware of estimates of the discounting parameters for the retirement-savings decision, including the default effects.

that individuals make a rational choice given their forecasts of their future health costs and their risk aversion. If the employees switch plan from year to year, they have to pay a switching cost k (unless the firm requires active choice, as they did in year t). The maximum-likelihood estimate is $\hat{k} = \$1,729$ (s.e. 28). A series of robustness checks all provide estimates in the \$2,000 range. The structural estimates, thus, confirm the order of magnitude of the simple calibration above, but also indicate that the large magnitude is very precisely estimated. As the author remarks, it is implausible to interpret this estimate of a switching cost as a reflection of the time costs required to switch. Some behavioral explanation, like procrastination, is likely responsible.

Interestingly, this paper contributes to the Structural Behavioral Economics literature even though it does *not* spell out a behavioral model, just a standard model with switching costs. The magnitude of the estimated switching costs clearly opens the door to behavioral models, which follow-on papers can consider.

Checking for Reasonable Values. There is a third way that structural estimation complements the use of calibrations in behavioral economics: checking whether the values of the estimated parameters make sense with broadly accepted values for that parameter in the literature. An example is the paper on job search and reference dependence by myself, Attila Lindner, Balazs Reizer, and Johannes Schmieder (DellaVigna et al., 2017). The paper analyzes how the exit rate from unemployment in Hungary changes after a reform that introduces an extra step in the unemployment benefit level. The paper tests for a job search model in which the utility of consumption is reference-dependent with respect to recent income. The disutility of unemployment is particularly high when individuals have just lost a job, because they compare the unemployment benefits to the (much higher) pre-unemployment earnings on the job. As time passes, the reference point shifts to incorporate the lower income from unemployment benefits, and the disutility from unemployment is partially mitigated, leading to lower search effort. We show that this model makes unique predictions, which we test comparing the exit from unemployment under a reform of the benefit system in Hungary.

While the paper is mostly focused on the impact of reference dependence and loss aversion, it also identifies time preferences for the unemployed workers. The time preferences matter mostly because they affect the precautionary savings, and thus the degree to which consumption declines when benefits drop off. The more impatient individuals are, the more they go hand-to-mouth, and thus the more their search intensity responds to benefits decreases, just as we see in the data.

We estimate two alternative models of discounting: exponential discounting with discount factor δ, and present-bias with present-bias parameter β, fixing δ at a reasonable level (to keep the number of estimated parameters the same). The two sets of estimates actually lead to a similar quality of fit: in this setting, it is not obvious how to statistically separate out present-bias from high exponential discounting. However, there is a large difference in the *plausibility* of the estimated parameters. In the exponential-discounting specification, a 15-day discount factor is estimated to be $\hat{\delta} = 0.89$, equivalent to an annual discount factor of 0.05, an extremely high degree of discounting that is at odds with discounting estimates in the rest of the literature. In the present-bias specification, instead, the estimated present-bias parameter, $\hat{\beta} = 0.58$, is well within the range of other estimates in the literature (more on this in Section 2.2). Consistency with other estimates in the literature

is an important plus, as we discuss further in Section 2.3 on parameter stability. Thus, while the two sets of estimates fit the data about equally well, one makes sense with values estimated in other settings, while the other is at odds with other estimates; in this case, this plausibility test tips the scale towards the present-bias model.

Martinez, Meier, and Sprenger (2017) on tax filing and procrastination has a similar finding. They estimate the time preferences and costs of filing that match the timing of the tax filing for a population of low-income individuals. The key observation motivating their analysis is the fact that so many people file near the deadline, despite the substantial refunds on average, especially for this low-income population. The maximum-likelihood estimates for a daily discount factor that best fits the data is $\hat{\delta} = 0.53$: fitting the delays in the data requires an extreme amount of discounting. In contrast, holding constant a plausible exponential daily discounting factor at $\delta = 0.9999$, the delay can be explained in a naive beta-delta model with a plausible present-bias parameter $\hat{\beta} = 0.92$. We return below to this paper for its out-of-sample predictions.

Thus, structural estimation builds on, and complements, calibrations in behavioral economics by (i) making it possible to consider more complex and realistic models; (ii) providing exact magnitudes for the parameters, including confidence intervals; and (iii) using previous calibrated or estimated values for the parameters to see if they make sense.

2.2 Better Understanding Models and Assumptions

As we discussed, calibrated magnitudes play an important role in behavioral economics. In fact, the emphasis on calibration is just an example of the active interaction within behavioral economics between theory and empirics. Behavioral theorists, and applied theorists in particular, have paid close attention to the applications of their models, sometimes participating in empirical studies themselves. For example, consider some of the most influential behavioral models: reference dependence (Kahneman and Tversky, 1979; Köszegi and Rabin, 2006 and 2009), present-bias models of time preferences (Laibson, 1997; O'Donoghue and Rabin, 1999a), models of social preferences (Fehr and Schmidt, 1999; Benabou and Tiróle, 2006), and k-levels of thinking (Camerer, Ho, and Chong, 2004; Costa-Gomes, Crawford, and Broseta, 2001), among others. All of the authors of these papers have also contributed to empirical studies, or have closely advised studies by others.

The converse has generally been true as well. Empirical behavioral researchers have paid close attention to the development of models, and have often tested predictions and comparative statics of the models in their empirical work. This has been particularly true in laboratory experiments, where the testing of models has played a critical role. For example, the majority of laboratory experiments published in top journals in the years 2005-10 had a model guiding the analysis, and over 30 percent of these experiments had in fact structural estimates of the parameters (Card, DellaVigna, and Malmendier, 2011). This prevalence is lower among the behavioral papers with field evidence, but nonetheless the models play an important role there too. This interaction must surely have to do with the essence of behavioral work: investigating behavior that deviates from the standard model. To define what is behavioral, it is important to be precise about the null hypothesis of the standard model, and the behavioral alternative models. One needs models for that!

This model-empirics dialogue of course happens elsewhere too, but arguably it is a stronger tradition in behavioral economics than in some other fields. In development economics, for example, a common concern is about a disconnect between the field experiments and the development models.

Strengthening the Evidence-Theory Dialogue. From this perspective, structural estimation has an important role to play in fostering further back-and-forth between the evidence and the theory. By definition, one cannot do any structural estimation without first specifying a model, since otherwise there is nothing to estimate. The link goes beyond this obvious point. For estimation, one needs to work out a number of modeling details which otherwise one may not pay close attention to, such as the behavior of the model near the boundaries (e.g., the discount factor δ close to 1 in life-cycle models), or conditions one may want to impose on the parameters (e.g., $\beta \leq 1$ in present-bias models). In fact, in preparation for estimation, one typically runs a large number of simulations of the model to understand its behavior for different parameter values; this typically teaches the researcher a lot about the model, including its range of predictions and its quantitative properties (e.g., calibrated values). We return to this in Section 4.

An important aspect is parameter identification. For structural estimation, it is critical to know which parameters are identified in a particular model. Can one, for example, estimate β separately from δ in a present-bias model of job search? Can one separate risk aversion from loss aversion λ in a reference-dependent model of insurance deductible choice? To address these questions, a researcher will want to provide a proof of identification, or lack thereof, of a parameter. To prepare the stage for a general identification result, simulate-and-estimate exercises are a perfect way to investigate these questions: one simulates a data set from a model with given parameters, and then estimates the model on this data set to see if one can recover the original parameters, and with what precision. If two parameters are not separately identified, one recovers only a combination of the two, something that one can learn through the simulation exercise.[6] The results from the simulation exercise will often show the way to the formal model identification results.

On occasion, the deep understanding of the model that comes from the study of identification can lead to insights that are of theoretical interest on their own. For example, in preparing the estimation of decisions over home and car insurance, Barseghyan et al. (2013) noticed that one could not separately identify loss aversion and probability weighting in the presence of reference dependence with Köszegi and Rabin (2009) reference points. This non-identification result is important on its own right, and it was developed as part of the estimation strategy. In fact, a key result in the paper is that loss aversion and probability weighting result in a probability distortion $\Omega(p)$ in the insurance choice decision, with $\Omega(p) = \pi(p)[1 + \lambda(1 - \pi(p))]$, where λ is the loss aversion coefficient and $\pi(p)$ is the probability weighting function. This result makes clear that one cannot separately identify the two components of prospect theory unless one is willing to make some strong assumptions about the shape of the probability weighting function. The authors advocate against doing so, and estimate $\Omega(p)$ non-parametrically. The identification occurs by comparing the insurance choice for risks occurring with different probabilities p, such as home insurance accidents versus auto

[6] One would also learn that two, or more, parameters are not separately identified using estimation on the actual data. It is best to assess identification on a known data generating process, in particular as that allows to study identification with asymptotic sample size too. I stress also that identification of a set of parameters is a separate question from how sensitive are different parameters to different moments in the data; we return to this in section 4.4.

collision accidents. Notice that the authors can, based on the data, rule out the case in which there is no probability weighting, that is $\pi(p) = p$, since in this case the expression simplifies to $\Omega(p) = p[1 + \lambda(1 - p)]$.[7]

An advantage, thus, of structural estimation is that it forces empirical researchers to come to terms with the models more closely than they would otherwise when, say, testing a qualitative prediction. There is a converse point too, in the impact on modeling. When taking a model to the data, one often has to simplify the original model. For example, several estimates and application of reference-dependent models do away with probability weighting and curvature of the value function for simplicity. In another example, in order to estimate a social-image model of voter turnout, in a paper with John List, Ulrike Malmendier, and Gautam Rao (DellaVigna et al., 2017) described more in detail below, we simplify the social image part of the model into a single parameter. The empirical test will thus not be able to distinguish between, for example, social signaling à la Benabou and Tiróle (2006), social pressure models, or other determinants of demand for social image. This simplification is the necessary cost that allows us to provide the first estimate of a model of voter turnout.

From the standpoint of the theorist, these simplifications are limitations of the test of the model, since the model is stripped down. And yet, there is important feedback to the theorist in these simplifications, since they stress which components of the model are feasible to bring to the data, and which (at least in a particular setting) are not. It also puts a premium on models with enough simplicity to be taken to the data. If the theorist takes this feedback into account for future models, the theory-empirics dialogue is strengthened, as a result of the attempts to estimate a model.

I should stress that most of these advantage in terms of model-evidence dialogue apply also to empirical evidence that explicitly tests the predictions or comparative statics of a model, even without structural estimation, a point I discuss further in Section 3.1.

Better Empirical Test. An example of the importance of a close tie between model and evidence applies to some of the earliest evidence on reference dependence, both on housing and on labor supply.

Genesove and Mayer (2001) focuses on the decision to sell a house, and provides an intuitive test of reference dependence with respect to the initial purchase price of the house. The authors compare house-owners who are currently at a loss relative to the initial purchase price, to home-owners who instead are at a gain. These differences arise because of the timing of the initial house purchase: taking a modern example, comparing two similar houses on the market in 2013, the house purchased in 2010 is likely to be on the gain side, while the house purchased in 2007 is likely to be on the loss side, given the housing market crash in 2008, followed by a housing recovery. Compared to the house-owners who are at a gain, the ones at a loss are likely to ask for a higher listing price, and thus end up waiting longer, given that they have a higher marginal utility of income (since they are on the loss side of the prospect theory function). The authors provide robust evidence of this phenomenon, with a large effect on listing price and on waiting time, and a much smaller effect on final sale price (as market forces partly discipline this bias). Importantly, in estimating this effect, the authors pay close attention to the biasing effect of unobservables that are correlated with the

[7] Confirmed in personal communication with Ted O'Donoghue.

initial purchase price of the house.

This evidence is indeed qualitatively consistent with reference dependence. The empirical specification tested in the paper, though, is not itself derived from the reference-dependent model which they cite as motivation. Indeed, a model of reference-dependence with loss aversion with respect to previous house purchase price does not yield the parametric specification in the paper (see DellaVigna, 2009). In addition, this reference-dependent model makes an additional prediction which the original paper does not test altogether: there should be bunching in the listing price at the previous purchase price, since at this point there is a jump in the marginal utility of money, given the loss aversion. To be clear: the evidence in Genesove and Mayer (2001) is consistent qualitatively with reference dependence, and is still valid and informative that way. But the empirical test would have been more efficient, and parametrized appropriately, if it had been derived from the model.

A related case is the well-known study of labor supply of cab drivers in Camerer et al. (1997), which I discuss in detail in DellaVigna (2009) and in DellaVigna and Mas (2018). In short, the original study focuses on the negative correlation between daily wages and hours worked for cab drivers, supporting a form of income targeting. The basic specification is an OLS regression of log hours worked on log wages earned at the daily level, with the wage variable instrumented with the earnings of other cab drivers the same day. In turn, this is what one would expect if cab drivers have a daily income target for their earnings, and are loss averse with respect to this target. However, writing down the stopping model with reference-dependent preferences clarifies that the initial test is not correctly specified, a point that Farber (2008) makes. Farber (2008) derives a structural model of the optimal stopping decision of cab drivers, who have to decide after having worked h hours and earned income I, whether to continue working or not. This paper allows for a stochastic reference point in income, but does not motivate the foundations for such a reference point.

The follow-up literature builds on this model, with a tighter link between the reference-dependent model and the empirical test. In particular, Crawford and Meng (2011) models reference dependence with respect to both income and hours worked, with the reference point r defined to be expectations-based along the lines of Köszegi and Rabin (2006). Expectations-based reference points have the potential to reconcile the conflicting findings: anticipated changes in the daily earnings do not lead to income targeting because they are already incorporated in the reference point, and thus gain-loss utility drops out of the utility function. Conversely, unanticipated income shocks can have non-neoclassical implications, leading to lower work effort (income targeting) with higher earnings because the reference point has not adapted.

Most recently, Thakral and Tô (2017) builds on these earlier insights and provides a new set of findings that, together with a much larger data set, has the potential to settle the earlier debate in the literature. Namely, they provide evidence that the stopping decision of cab drivers responds to income earned during the day in a manner consistent with adaptive reference points: higher income earned earlier during the day does not affect much the probability of stopping, as it is already incorporated into the reference point of the cab drivers. Higher income earned, instead, in the most recent 1-3 hours leads to a higher probability of stopping: these cab drivers are more likely to be on the gain side, as they find themselves to have earned more than expected, and the reference point has not (fully) adjusted yet. Given that they are on the gain side, the marginal return to effort is valued

less than when they are on the loss side. Notice that the standard model would make, if anything, the opposite prediction: higher recent income makes it more likely that the upcoming hours or two of work may also lead to higher than usual income, and thus should lower the probability of stopping. The Thakral and Tô (2017) results can be interpreted as evidence of adaptive reference points or of forward-looking reference points as in Köszegi and Rabin (2006), along the lines proposed by Crawford and Meng (2011).

In both of these cases, writing down the model—whether for structural estimation or just to derive the correct test—avoids the risk of estimating a specification that appears qualitatively in line with the model predictions, but may not be an actual implication of the model.

Clarifying Needed Assumptions, Real Effort Experiments. We discussed above how a model in some cases implies a different empirical specification than implemented in a reduced-form test. In other cases, a model clarifies which assumptions justify a given reduced-form specification.

An example occurs in the case of real effort tasks, in which experimental subjects decide how much effort e to put in a unit of time. The subject may be, for example, moving sliders (Gill and Prowse, 2012), solving mazes (Gneezy, Niederle, and Rustichini, 2003), or doing addition tasks (Niederle and Vesterlund, 2007). Several papers estimate variants of the OLS specification:

$$e_i = a + \beta T_i + \gamma X_i + \epsilon_i, \tag{6}$$

relating the effort e_i by subject i to treatment T_i and a set of controls X_i. The treatment T_i for example could be variation in the gender composition of the group (Gneezy, Niederle, and Rustichini, 2003) or the provision of rank information (Gill et al., forthcoming). That begs the question: under what conditions is specification (6) implied by a utility-maximization problem of the subjects?

DellaVigna et al. (2016), building on work by Bellemare and Shearer (2011) among others, show that one can derive specification (6) from a utility maximization problem. Assume that individuals maximize the utility

$$max_{e_i} s\left(T_i\right) e_i - \frac{\exp\left(\gamma e_i\right)}{\gamma} \eta_i, \tag{7}$$

The first term in expression (7) indicates the return to effort, which for simplicity we assume is linear in effort, with a marginal motivation term s which depends on the treatment T_i (hence the notation). The second part denotes the cost of effort function, which takes the exponential form $\exp\left(\gamma e_i\right)/\gamma$. The cost of effort is convex, guaranteeing an interior solution.

The final part, η_i, introduces the heterogeneity as a multiplicative factor on the cost of effort function. What kind of characteristics might we want for the distribution of η? First, we may want the cost of effort to depend on observables, X_i. Notice also that we also want to restrict η to be positive, as a negative draw implies a negative cost of effort. A log-normal distribution for η, with a mean that is a function of the observables, satisfies both of these properties. Specifically, assume that $ln(\eta_i) \sim N(\gamma k\left(X_i\right), \gamma^2\sigma^2)$. Under this assumption, (7) can be rewritten

$$max_{e_i} s\left(T_i\right) e_i - \frac{\exp\left(\gamma e_i\right)}{\gamma} \exp\left(\gamma k\left(X_i\right) - \gamma\epsilon_i\right), \tag{8}$$

with $\epsilon_i \sim N(0, \sigma^2)$.

Taking the first order conditions, shifting the terms and taking logs, one obtains

$$e_i = \frac{1}{\gamma} log\left[s\left(T_i\right)\right] - k\left(X_i\right) + \epsilon_i. \tag{9}$$

Expression (9) micro-founds the OLS estimation (6). The lognormal distribution in the cost level, once one takes the first-order condition, becomes the additive error term in (6). One implication of (9) is that changes in motivation s, such as changes in the piece rate, or competitiveness effects due to the gender mix, impact effort in log form.

Is this a desirable set of assumptions? That will depend on the researcher's assessment, and the setting. For example, one may find the assumption of a power cost of effort function $\left(e_i^{1+\gamma}\right)/\left(1+\gamma\right)$ more appealing than the assumption of an exponential cost of effort function in (8). A power cost of effort has the property that the elasticity of effort to motivation is constant and equal to $1/\gamma$. Under a power cost of effort, the implied specification is then

$$log(e_i) = \frac{1}{\gamma} log\left[s\left(T_i\right)\right] - k\left(X_i\right) + \epsilon_i, \tag{10}$$

This implies an OLS specification like (6), but with log effort as dependent variable, another expression also sometimes used in the literature. Thus, writing the model explicitly clarifies the needed assumptions for a particular reduced-form specification.

Clarifying Needed Assumptions, Gift Exchange Field Experiments. Continuing in a related line, consider the pioneering gift exchange field experiments à la Gneezy and List (2006). In these experiments, the subjects are hired for a one-day task such as coding library books. To shut down repeated-game incentives which confound the estimation of social preferences, the subjects are told (truthfully) that this is a one-time task and that they will not be hired again for this same task. Furthermore, they pay is a flat hourly rate. Therefore, their performance cannot affect their pay, nor their re-hiring.

Employees in different treatments are then exposed to different employer actions, such as surprise pay raises relative to the announced wage rate (Gneezy and List, 2006), pay cuts (Kube, Maréchal, and Puppe, 2013), and in-kind gifts (Kube, Maréchal, and Puppe, 2012). These papers find that positive monetary gifts can have a positive, if short-lived, effects on effort (Gneezy and List, 2006), though some follow-up papers do not find any effect. This positive effect on effort is larger for in-kind gifts of similar value (Kube, Maréchal, and Puppe, 2012). Also, negative "gifts" (a surprise wage cut) lower effort more than positive gifts increase effort (Kube, Maréchal, and Puppe, 2013). The differences in productivity across the treatments provide evidence about gift exchange and reciprocity, given that the effects cannot operate through reputational effects, given the one-time set-up. As such, they corroborate the evidence from a large body of evidence on gift exchange in laboratory experiments (e.g., Fehr, Kirchsteiger, and Riedl, 1998) and they provide some support for the gift-exchange conjecture of Akerlof (1982) that employee reciprocity towards the employer may provide a rationale for efficiency wages.

Can this descriptive evidence be tied back more precisely to the social preference models, in

particular to estimate reciprocity parameters? Assume that the workers put weight α on the return to the employer, as in a pure altruism model. The overall return of effort to the employer in this setting can be written as $p_E e$, that is, it is linear in the units of effort exerted. When the employee receives a gift, the altruism parameter is shifted to become $\alpha + \alpha_{Gift}$, with the parameter α_{Gift} capturing a class of reciprocity models. In a broad class of reciprocity models, generous actions (or intentions) by a first player increase the altruism weight the other player puts on this first player. Assuming a convex cost of effort, the observed effort in the control treatment and gift treatment then are

$$e^*_{Contr} = C^{'-1}(\alpha p_E) \text{ and } e^*_{Gift} = C^{'-1}((\alpha + \alpha_{Gift})p_E). \tag{11}$$

As the expressions in (11) clarify, two crucial pieces of information are missing to identify the social preference parameters α and α_{Gift}. First, we do not know what workers assume is the value of effort to the charity p_E, since they are not informed of this in the typical experiment. Second, the econometrician does not know the cost of effort function $C(e)$. Hence, it is impossible to identify the social preference parameters.

It is helpful to consider the special case with pure altruism and a power cost function: $c(e) = ke^{1+\gamma}/(1+\gamma)$. This function is characterized by a constant elasticity $1/\gamma$ with respect to the return to effort. The two solutions then reduce to:

$$e^*_{Contr} = \left(\frac{\alpha p_E}{k}\right)^{1/\gamma} \text{ and } e^*_{Gift} = \left(\frac{(\alpha + \alpha_{Gift})\, p_E}{k}\right)^{1/\gamma}$$

and thus

$$log\left(e^*_{Gift}\right) - log\left(e^*_{Contr}\right) = \frac{1}{\gamma}\left[log\left(\alpha + \alpha_{Gift}\right) - log\left(\alpha\right)\right].$$

In this particular case, thus, one need not observe the value to the employer p_E to back out the increase in altruism in log points, $log\left(\alpha + \alpha_{Gift}\right) - log\left(\alpha\right)$. Still, it is critical to know the elasticity of effort $1/\gamma$ to compute this effect. Intuitively, a 30 percent increase in effort as observed for example in Kube, Maréchal, and Puppe (2012) for an in-kind gift corresponds to just a 10 percent increase in altruism due to the gift if the task is very elastic with $\gamma = 1/3$, but it corresponds to a huge increase in altruism if the task is inelastic, say $\gamma = 3$. To identify this elasticity, it is important to vary as part of the experiment the private return to the workers to "price out" the cost of effort, as we do in DellaVigna et al. (2016), a point we return to in Section 2.5.

Further, it is not clear that the cost function will have the special feature of constant elasticity. A plausible alternative is that the elasticity decreases as effort increases, as is the case for the exponential cost function introduced above, $C(e) = k \cdot exp(\gamma e)/\gamma$. In this case, the solutions are

$$e^*_{Contr} = \frac{1}{\gamma} log\left(\frac{\alpha p_E}{k}\right) \text{ and } e^*_{Gift} = \frac{1}{\gamma} log\left(\frac{(\alpha + \alpha_{Gift})\, p_E}{k}\right).$$

We can transform the solution and divide through to obtain

$$exp[\gamma(e^*_{Gift} - e^*_{Contr})] = \frac{\alpha + \alpha_{Gift}}{\alpha}. \tag{12}$$

Expression (12) highlights another implication. Consider an experiment with a positive gift treatment, which increases output by x units, and a negative gift treatment, which decreases output by x units. Would these equal-sized impacts of the gifts on effort imply that positive reciprocity has the same magnitude as negative reciprocity? Expression (12) shows that this is not the case. Because of the steep curvature of the exponential function, the x unit increase for the positive gift would require a larger proportional change in altruism (positive reciprocity) compared to the corresponding change in altruism (negative reciprocity) for the negative gift. Intuitively, it is harder to increase effort at the margin than to reduce it. Thus, the finding in Kube, Maréchal, and Puppe (2013) that the response to a negative gift is larger than the response to a positive gift does not immediately translate into the fact that altruism α increases more for positive gifts than it decreases for negative gifts. The structural model helps to clarify this point.

2.3 Stability of Parameters

The presence of a model guiding the empirical evidence provides also a test for whether the model is well-specified. In a well-specified model, certain "deep" parameters should have comparable values across different settings. Of course any model is an approximation to a degree and never perfectly specified, but the comparison across studies and settings checks whether the model is reasonably stable. In fact, the emphasis economics places on stable preferences and models, enabling quantitative predictions within a setting and across contexts, is a key distinguishing feature of economics relative to the approach taken in psychology where the predictions are seen as largely qualitative. Models with parameter estimates are ideally suited for the purpose of quantitative predictions. This emphasis on magnitudes ties together the discussion of calibration in Section 2.1, the discussion here of parameter stability, and the discussion of out-of-sample predictions in Section 2.4.

The comparison of parameter estimates across contexts makes sense to the extent that there are parsimonious models with a small number of key parameters, which are themselves meant to capture a deep underlying behavioral feature, such as risk aversion, impatience, cognitive ability, and the like. From this perspective, behavioral economics has several leading models, each characterized by a small number of parameters that mark the deviation from the standard model: loss aversion λ and probability weighting function $\pi(p)$ for reference dependence (Kahneman and Tversky, 1979; Köszegi and Rabin, 2006 and 2009), present bias β and naiveté $\hat{\beta}$ for present-bias models of time preferences (Laibson, 1997; O'Donoghue and Rabin, 1999a), the altruism and inequity weights for inequity aversion (Fehr and Schmidt, 1999), and the share of different types in k-levels of thinking model (Camerer, Ho, and Chong, 2004; Costa-Gomes, Crawford, and Broseta, 2001). I defer to the individual chapters for a more comprehensive discussion of estimates for these leading models, but I want to highlight three cases, one regarding reference dependence, one on present bias, and one about social preferences. Before I discuss these cases on preferences, I discuss an example of apparent instability of the impact of media violence on arousal.

Stability of estimates, Media Violence and Aggression. Does media violence trigger aggression? An extensive series of laboratory experiments, surveyed for example by Anderson and

Bushman (2001), examines this link. The experimental literature exposes subjects in the laboratory (typically children or college students) to short, violent video clips. These experiments find a sharp increase in aggressive behavior immediately after the media exposure, compared to a control group exposed to non-violent clips. The laboratory findings thus document a short-run increase in aggressiveness as a result of exposure to violent media content.

In Dahl and DellaVigna (2009), we instead use field data and a natural experiment to provide evidence on this same topic. We exploit the natural experiment induced by time-series variation in the violence of movies shown in the theater. As in the psychology experiments, we estimate the short-run effect of exposure to violence, but unlike in the experiments, the outcome variable is violent crime rather than aggressiveness. We generate a daily measure of national-level box office audience for strongly violent movies (e.g., "Hannibal") and mildly violent movies (e.g., "Spider-Man"). Since blockbuster movies differ significantly in violence rating, and movie sales are concentrated in the initial weekends after release, there is substantial variation in exposure to movie violence over time. The audience for strongly violent and mildly violent movies, respectively, is as high as 12 million and 25 million people on some weekends, and is close to zero on others. We use crime data from the National Incident Based Reporting System (NIBRS) and measure violent crime on a given day as the sum of reported assaults (simple or aggravated) and intimidation.

We find that, on days with a high audience for violent movies, violent crime is *lower*, even after controlling flexibly for seasonality and for the endogeneity of movie attendance. Focusing on the nighttime hours following the movie showing (12AM-6AM), we find that for each million people watching a strongly or mildly violent movie, respectively, violent crime decreases by 1.9 and 2.1 percent. This is an economically important effect: the total net effect of violent movies is to decrease assaults by roughly 1,000 occurrences per weekend.

A possible interpretation of these contrasting findings, and indeed our own initial interpretation, is that the impact of media violence on aggressiveness is not stable across the two settings: it is consistently positive in laboratory settings, but negative when one measures it in the field with impact on violent crime. It would seem to be an example of lack of stability of an important social science finding which does not extrapolate beyond the laboratory.

This interpretation is tempting, but it is wrong. The comparison made above does not account for the fact that the two estimates capture different underlying behavioral parameters. Assume, as we do in the paper, that exposure to violent media a^v can be (causally) associated with violent behavior V : more precisely, $ln(V) = \alpha^v a^v$ with α^v denoting the impact of exposure to violent media a^v on aggression V. The laboratory evidence indicates $\alpha^v > 0$. Consider now the impact of alternative uses of time. These alternative activities a^s are also associated with crime, at a different rate: $ln(V) = \sigma a^s$; for a social activity such as playing cards, we would expect $\sigma < 0$, but for an activity such as drinking in bars we would expect $\sigma > 0$, and in fact plausibly $\sigma > \alpha^v$. That is, some of the alternative activities are plausibly worse than violent movies at making people aggressive.

As it turns out, when we identify the impact of violent movies using the variation in the number of people watching violent Hollywood movies, we identify $ln(V) = (\alpha^v - \sigma)a^v$, not $ln(V) = \alpha^v a$ as the laboratory evidence does. Thus, our estimates of a negative effect of movies on violence indicate $\alpha^v < \sigma$, not $\alpha^v < 0$. That is because, in order to go the a violent movie, individuals substitute away

from some alternative activity and the impact of crime is the net of the two alternatives. Further, the relevant alternative activity is one that applies to the kind of demographic group that watches violent movies, which is predominantly young males. For these groups with a taste for violence, the alternative activities can easily be more conducive to violence. Indeed, the evidence in our paper suggests that individuals are more likely to be sober at night, and thus less violent, if they went to a violent movie than under their alternative use of time, such as going drinking. Indeed, under some assumptions about the selection of movie-goers, we can separately identify α^v and σ; that leads us to estimate $\alpha^v > 0$, just as in the laboratory evidence.

This case stresses the importance of interpreting the evidence in light of models, since it clarifies that a superficial lab-field comparison was not comparing apples to apples. When we compare the same effect across lab and field, encouragingly we find stability of at least the sign of the effect of media violence on aggression.

Stability of estimates, Reference Dependence. An important component of prospect theory (Kahneman and Tversky, 1979) is the probability weighting function. Not surprisingly, thus, there are a number of estimates in the literature. I summarize in Panel A in Table 2 key papers estimating the probability weighting function, building on a meta-analysis in DellaVigna and Pope (2018).[8] The large majority of studies regards lottery choices in the laboratory (e.g., Camerer and Ho, 1994), but the evidence also includes field evidence, such as insurance choice (Barseghyan et al., 2013). The table shows that, in almost each paper, events of small probability, such an event with probability 0.01, are overweighted by the probability weighting function. Averaging across studies, the average probability weight for an event of probability of 0.01 is 0.06: small probabilities are overweighted by a factor of 6, a substantial degree of overweighting, along the lines of the original prospect theory (Kahneman and Tversky, 1979). This is thus, it appears, a case in which there is substantial commonality in this finding.

Panel B of Table 2 reports the result of another set of studies which are not aimed at estimating the probability weighting function, but which have clear implications for it. These studies compare the completion of a task, such as returning a survey via mail, in a certain-incentive treatment where respondents receive X to complete the task, and in a probabilistic-incentive treatment where the reward is $Y = X/p$ with probability p, with $p < 0.1$. If individuals are risk neutral, the two treatments are equivalent, but with overweighting of small probabilities (and with moderate risk aversion), the response should be larger in the probabilistic-incentive group. A study in this group is DellaVigna and Pope (2018), which examines the completion of a real-effort task on MTurk, comparing a piece rate of 1 cent to a probabilistic piece rate of \$1, obtained with probability $p = 0.01$. Halpern et al. (2011) compare the rate at which a survey is returned in a group with fixed incentive of \$10, versus with a probabilistic reward of \$2,500 with probability 0.004. Interestingly, in 4 out of 5 such studies the probabilistic-incentive treatment yields *lower*, not higher, completion, contrary to the prediction for the case of overweighting of small probabilities. Indeed, DellaVigna and Pope (2018) estimate a probability weight $\hat{\pi}(0.01) = 0.003$.

What explains this striking discrepancy, with clear commonality of results within a type of

[8] These studies are also listed in Online Appendix Table 3 in DellaVigna and Pope (2018), the references for the papers not referred to in the text are in the online appendix.

setting, but different results across settings? An ex-post conjecture is that in a field context the probabilistic rewards may not be fully trusted, with individuals expecting a lower probability than the stated one. That pushes us back to the question of what exactly determines the probability weighting: is it some form of salience of probabilities, which may be clearer in some settings than in others? A model that speaks to this, as well as further evidence to accommodate these discrepancies, would be useful.

Stability of estimates, Present Bias. Within the realm of time preferences, up to 5 years ago there was a similar discrepancy in estimates with regards to the present-bias parameter β. The earliest estimates for β are from observational field data and typically gravitate in the range 0.5-0.9. For example, using annual consumption data and matching three key sets of moments about consumption and savings, Laibson et al. (2017) estimates $\beta = 0.51$ at the annual level; the key insight here for identification is that present-bias can simultaneously explain high rates of credit card borrowing (impatience) and relatively high accumulation of illiquid assets pre-retirement (demand for illiquidity); the exponential model can fit one of these two moments, but at the cost of not fitting the other. Paserman (2008) estimates the job search model of DellaVigna and Paserman (2005) on data for unemployed workers and estimates $\beta = 0.40$ and $\beta = 0.9$, depending on the population. A number of other papers on field evidence, while not providing structural estimates, are consistent with present bias, including DellaVigna and Malmendier (2006).

At the same time, a growing number of studies estimating time preferences in the laboratory yielded either very limited evidence of present bias, or no evidence whatsoever. For example, in an influential study Andreoni and Sprenger (2012) used a design with convex budget sets and delivery of monetary payments either later the same day, or in later weeks, to estimate time preferences. The authors pay particular care to ensure the credibility of payments, since lower-credibility of later payments can induce spurious evidence of discounting. This paper delivers a precisely-estimated finding of no present bias: $\beta = 1.00\,(s.e.\,0.002)$.

This sets up an inconsistency similar to the one observed above: why is there such a large discrepancy between the laboratory experiments and the field evidence? This would appear to cast doubt on the field evidence of present bias, especially since laboratory experiments are able to control several of the confounds.

Within a short time span of 2 to 3 years, a new set of papers has largely sorted out this discrepancy, it appears. These papers take as a starting point a well-known point, that discounting should apply to the utility of consumption, not to the utility of money. An important confound of the laboratory experiments, then, is that the experimenter does not control when consumption takes place in response to the monetary payments. In fact, to the extent that most consumption does not take place immediately with the receipt of a payment, the utility associated with immediate monetary payoffs is also in the future, and thus these trade-offs do not identify the present bias. In fact, under this interpretation we would expect to find no present bias, even for immediate payments, given that (almost) all consumption takes place in the future anyway. That would explain the estimate in Andreoni and Sprenger (2012).[9] Before we turn to additional evidence, it is important to note

[9] An alternative interpretation is that the delay of a few hours in the delivery of even "immediate" payments in Andreoni and Sprenger (2012) is responsible for this result. Balakrishnan, Haushofer, and Jakiela, 2017 provide some

that this confound does not apply to the field evidence cited above, since in these applications there is a non-monetary, effort component. In the job search case of Paserman (2008), for example, the identification is based off of the fact that search effort costs are immediate, but the benefits (of receiving a job offer) delayed. The trade-off between immediate (non-monetary) effort and delayed benefits also appears for health club attendance in DellaVigna and Malmendier (2006) and for the models of default effects in Madrian and Shea (2001) (where the immediate cost is the time effort of enrolling). Thus, these examples are robust to this criticism.

This criticism suggests that a laboratory experiment with real effort, instead of monetary payments, could resolve this tension. Augenblick, Niederle, and Sprenger (2015) designs a real-effort task where subjects at time t make binding choices between how much of the effortful task to do at $t+s$ versus at $t+s+T$. The subjects make the choice for both $s>0$ (choice between two period both in the future) and $s=0$ (choice between present effort and future). From the real effort choice, Augenblick, Niederle, and Sprenger (2015) estimates $\beta=0.9$ at the daily level, replicating the field evidence of present bias. At the same time, over the same population, the time preferences elicited from monetary trade-offs provide no evidence of present bias, that is they find$\beta=1$, replicating the Andreoni and Sprenger (2012) result. The finding of present bias with a real-effort task has been replicated by Augenblick and Rabin (forthcoming), which finds similar point estimates for β with a similar design. Carvalho, Meier, and Wang (2016) also finds similar patterns (without structural estimates) using a different real-effort choice: whether to complete a shorter survey sooner, or a longer survey later.

Thus, this new set of evidence with a creative new design fully weaves in, at least for now, the different strands of the literature on the estimation of present bias. This reconciliation is especially fulfilling because it involved addressing a long-standing concern for experimental elicitation of time preference, known at least since Mulligan (1996).

Stability of estimates, Inequity Aversion. A third example of the stability of preferences is with regard to the inequity aversion model of social preferences. Fehr and Schmidt (1999) proposes a simple 2-parameter model of social preferences that allows for different social preferences whether individuals are ahead, or behind. In the notation proposed by Charness and Rabin (2002) (who have a similar model), the utility of player s (self) in a game with player o (other) is

$$U_i(x)=\begin{cases}(1-\rho)x_s+\rho x_o & \text{if } x_s>x_o\\ (1-\sigma)x_s+\sigma x_o & \text{if } x_s<x_o.\end{cases}$$

A first key assumption is that $\rho>\sigma$, that is, a player cares more about the opponent if she is ahead. Second, Fehr and Schmidt (1999) posits that players may become spiteful when behind, and have $\sigma<0$. Fehr and Schmidt (1999) calibrates the share of types in the population using observed play in a variety of games, including dictator, ultimatum, and gift exchange experiments and conclude that about 30 percent of types are selfish ($\rho=\sigma=0$), that other types are altruistic when ahead, but spiteful when behind ($\rho>0>\sigma$); furthermore, of the altruistic types, a majority is very altruistic

evidence that immediate delivery of monetary payoffs on a platform that facilitates spending leads to evidence of present bias with monetary payments.

when ahead ($\rho > 0.5$), explaining for example sharing half of the pie in the dictator (or ultimatum) game. Charness and Rabin (2002) estimates the share of types with different social preferences based on a set of discrete dictator-type experiments. A number of other papers estimate these preferences, as well as a number of other social preference models, on evidence from laboratory experiments. The more recent literature emphasizes a number of models, importantly models of reciprocity, signaling, and social norms, in which how much a player cares about another player depends on the action (or the intention) of the other player. I focus the argument here on this simple, influential early model.

Can one take the qualitative evidence on inequity aversion and altruism from the laboratory for field applications? That turns out to be difficult. Consider a charitable giving interaction: a person with high income M_s encounters a person of lower income M_o, who asks for money. If the high-income person is of the high-altruism type, $\rho > 0.5$, she should give a very large donation up to the point where the two incomes post-donation are equated (at which point she stops giving, provided $\sigma < 0.5$). This prediction clearly does not match the typical response to pan-handling requests. A further difficulty with social preference models such as inequity aversion, when applied to a field setting, is deciding to which individuals the social preferences even extend.

Overall, most models that do a good job of explaining altruistic behavior in the controlled laboratory environment, where everything is framed in terms of payoffs in the experiment and there is a small number of players, do not do well in field settings, where the definition of the relevant payoffs is murky, and the interaction is with a variety of others. To capture other-regarding behavior in the field, models of signaling (e.g., Benabou and Tiróle, 2006) and social pressure (e.g., DellaVigna, List, and Malmendier, 2012) have the potential to be more helpful.

2.4 Out-of-Sample Predictions

The test of stability of key parameter estimates across settings is an important test of the ability of a model to travel well across settings, such as across different behaviors, different experimental set-up or between the laboratory and the field. Yet, it is a very ambitious test, as much can differ across examples, with different populations, different procedures, etc. A less trying, and still very important test, is the stability within one study, either comparing estimated parameters for different behaviors of a given sample, or comparing two similar samples in a population. A very useful way to do that is with out-of-sample predictions, taking the estimates from one setting, and predicting behavior in another setting. Indeed, the ability to do out-of-sample predictions is a key advantage for structural papers, as parameter estimates are required for the prediction.

One of the earliest such examples was the work of McFadden et al. (1977). In a project to predict the ridership of the upcoming BART subway system in the Bay Area, Dan McFadden and coauthors developed the discrete-choice logit model. The study aimed at predicting, based on survey evidence and a variety of demographics, how many people would use the BART system as a function of the positioning of the stations. After the BART was built, the authors then validated their own predictions with actual ridership data, in a true out-of-sample prediction. The ridership predictions turned out to have been generally quite accurate, especially when using the models with richer demographic variables.

Other papers make predictions about the impact of a policy change, and relate it to the actual outcomes. A different form of out-of-sample prediction in the structural literature is Todd and Wolpin (2006). In this case, the authors make a prediction within a field experiment, using data in the control group to make predictions about behavior in a treatment group. Namely, Todd and Wolpin (2006) estimates a model of household behavior based on a pre-treatment sample in Mexico, and then compares it out of sample to the experimental impact of the PROGRESA intervention.

Before I get to out-of-sample prediction in the behavioral literature, I should mention a reduced-form predecessor, which is to correlate behavior in two settings within a population, with two behaviors that are thought to be related to a structural parameter. Fehr and Goette (2007) provides perhaps the first test for a sample of bike messengers that are part of a behavioral labor supply study. Namely, one (random) subgroup of bike messengers receives a higher commission rate for one month, while the remaining subgroup receives the higher commission rate in another month. Fehr and Goette (2007) then tests whether, in the month with higher commission, bike messengers actually end their shifts earlier, as one would predict if they set as reference point a fix monthly income, and are loss averse relative to this reference point. They find evidence in support of this prediction. Then, they correlate this income targeting behavior to a set of laboratory choices by the same bike messengers: the display of small-scale risk aversion, which should also depend on the degree of loss aversion λ. Fehr and Goette (2007) finds a (marginally significant) correlation between the two behaviors, consistent with the role of (heterogeneity in) loss aversion λ.

I discuss now four examples of out-of-sample predictions in the structural behavioral literature, covering work on (i) behavioral firms, (ii) reference dependence and job search, (iii) tax filing, and (iv) social image and voting.

Out-of-Sample Predictions, Behavioral Firms. An example of a structural model which is used for out-sample predictions is in the realm of *behavioral firms*: Cho and Rust (2010) estimates a model of the optimal pricing by car rental companies. Car rental companies manage the lot of used cars with two key decisions. First, the companies typically keep cars for only 2-3 years, and resell them at around 50,000-80,000 miles. Second, the company rents out all cars of a given make at a particular location at the same price, irrespective of the odometer measure. Cho and Rust (2010) estimates the depreciation of the car in terms of resale value, the cost of maintenance (which turns out to be flat in the mileage within the relevant range), and consumer preferences in terms of mileage on the car. The resulting structural estimates can be used to derive an optimal policy for resales, assuming a threshold policy for how long the car rental company keeps the cars. The estimates imply that the car rental company would do better by keeping the cars for about twice as long, until about 150,000 miles, and offering consumers a discount for rentals on older cars. The model predictions suggest that the company could increase profits very significantly, by a margin between 30% and over 100%, depending on the type of car.

Armed with this out-of-sample prediction, the authors convince the company to run a small-scale field experiment, with 6 control location and 4 treatment locations over 5 months. In the treatment locations, the company rents out also older cars, at a discount (at least for some consumers). The result of the small-scale experiment is consistent with the predictions of the structural model: the total revenue is higher in the treatment locations, along the lines predicted by the model. Despite

this, interestingly, Cho and Rust (2010) reports that the company decided to stick to its older pricing strategy, which is the traditional one in the industry, in a striking example of a firm that does not appear to be maximizing profits (*behavioral firms*).

Out-of-Sample Predictions, Reference Dependence and Job Search. DellaVigna et al. (2017) provides another example of out-of-sample predictions. Using a reform in the unemployment insurance system in Hungary, we show qualitative evidence suggestive of reference-dependent workers with adaptive reference points. That is, workers act as if they search harder when benefits have just decreased and losses loom larger, but then become (in part) habituated to such changes. Building on this reduced-form evidence, we estimate a structural model of job search and optimal consumption using the time path of the exit from unemployment. We show that a reference-dependent model provides a better fit to the data compared to a number of versions of the standard model, despite the fact that the reference-dependent model has fewer parameters (since it does not assume any heterogeneity).

Still, there is one version of the standard model, with an unusual form of heterogeneity, which does as well, or even better than the reference-dependent model. This version requires a share of the population to have an extremely high elasticity of search effort to the benefits (an elasticity around 50), which is not particularly plausible by the calibration argument outlined in Section 2.1. But is there a data-based way to compare the two models, beyond this plausibility argument? An out-of-sample test provides such a comparison. We observe an earlier, smaller benefit reform for a very similar sample, as well as the response to this same reform for workers with lower pre-unemployment earnings. We take the best estimate for the reference-dependent model and for the high-heterogeneity model in the main sample, and then make out-of-sample predictions for these two other settings, with no degrees of freedom.

The reference-dependent model, while not providing a perfect fit, does quite well across these settings, and better than the benchmark standard model. More importantly, the high-heterogeneity model has a very poor out-of-sample fit, predicting extreme responses to small benefit changes (given the high elasticity), counterfactually. An out-of-sample prediction, thus, provides a good way to compare the plausibility of models, and in this case it reinforces the message from a calibration test: a model that has implausible calibrated values (e.g., an elasticity of 50) appears to be indeed an implausible explanation for the results, making poor predictions out of sample.

Out-of-Sample Predictions, Tax Filing. As we discussed above, Martinez, Meier, and Sprenger (2017) provides evidence on the role of impatience in (late) tax filing, by estimating both an exponential discounting model, and a naive present bias model. The latter model provides much more reasonable parameter estimates, and also fits the data better. But of course it is likely possible to find alternative assumptions under which a version of the standard model would fit the data better. A strong test is to formulate out of sample predictions for the two model estimates. In particular, Martinez, Meier, and Sprenger (2017) estimate the model on the 2005-07 filing seasons, and then predict out-of-sample behavior in the 2008 season, when tax rebates altered the magnitude of the returns to filing earlier. The present-bias estimates do a better job of predicting out of sample the response to the stimulus rebate.

As a further check, along the lines of Fehr and Goette (2007), Martinez, Meier, and Sprenger

(2017) also relates the observed tax filing behavior to a measure of present bias based on the response to hypothetical choices between money at different times in the future. The individuals that are classified as present-biased based on the hypothetical survey elicitation measures exhibit more of the delay in tax filing, as predicted.

Out-of-Sample Predictions, Social Image and Voting. A different application of out-of-sample predictions takes place in other work with John List, Ulrike Malmendier, and Gautam Rao (DellaVigna et al., 2017). We designed a field experiment to test a social-image model of voting. The question of why people vote is a classical and ongoing topic for the social sciences, since pivotal voting does not provide a good explanation for large-scale elections, and various version of norms-based voting are hard to test empirically. We consider a specific social-image motivation for voting. It is common for neighbors, friends, and family to ask whether we voted. If individuals care about what others think of them, they may derive pride from telling others that they voted or feel shame from admitting that they did not vote. In addition, they may incur disutility from lying about their voting behavior. Such individuals are motivated to vote (in part) because they anticipate that others will ask if they did. If they vote, they can advertise their 'good behavior' when asked. If they do not vote, they face the choice of being truthful but incurring shame, or saying that they voted but incurring the lying cost. This trade-off is reflected in the established fact that 25 to 50 percent of non-voters lie when asked about their past turnout (Silver and Abramson, 1986).

To test this model of voting 'to tell others', we design a field experiment visiting households door to door in 2011, with the knowledge of whether a given households had voted in the congressional election of 2010. As in DellaVigna, List, and Malmendier (2012), we post flyers announcing our visit a day in advance, and we randomize the information on the flyer. In one group, the flyer simply informs households that the next day we will visit their home to ask them to complete a survey. In a second group, the flyer specifies that the survey will be about *"your voter participation in the 2010 congressional election."* We then attribute differences in the share of households opening the door and completing the survey between the two groups as reflecting the anticipated value of being asked about voting. We find that non-voters sort out significantly when informed that the survey will be about turnout and, more surprisingly, even voters sort out to some extent (though less so) when informed of the turnout questions. This evidence suggest a potentially important role for social image in the avoidance decision, if not a role for pride. But these results are not enough to identify the key parameters in the model. To estimate the value of voting 'to tell others,' we need additional counterfactual social-image values, such as the shame that voters would feel were they to say they did not vote.

The second key manipulation then takes place for the households which agreed to the ten-minute survey. For half of the respondents, we simply ask them whether they voted in the 2010 election (and then proceed with their questions). For the other (randomized) half, we inform them that the ten-minute survey will be eight minutes shorter if they state that they did not vote in the 2010 congressional election. For voters, this treatment amounts to an incentive to lie and permits us to quantify the disutility of voters were they to say (untruthfully) that they did not vote. For the 50 percent of non-voters who lie without such incentives, this treatment provides an incentive to tell the truth. The results reveal that non-voters are significantly more sensitive to these incentives than

voters. When incentivized, the share of non-voters who lie decreases significantly, by 12 percentage points, while the share of voters who lie increases only insignificantly, by 2 percentage points. The results indicate a strong preference of voters for saying that they voted.

We combine the moments from these treatments to estimate the parameters of our model. We estimate that individuals assign on average a \$7 value to being seen as a voter rather than a non-voter, when asked once by a surveyor. This social-image value does not come from pride in voting, as we saw above. Rather, they assign a quite negative value to admitting to not voting, with an estimated disutility of \$7 to lying about voting to a surveyor. The combination of social-image utility and sizable lying costs implies that the anticipation of being asked provides a reason to vote. Given that people report being asked on average 5 times whether they voted for the 2010 congressional election, we estimate a value of voting 'to tell others' of \$18 for voters and \$13 for non-voters.

The main field experiment was designed to measure the value of voting without affecting voting itself, a crucial difference from the get-out-the-vote literature (e.g., Green and Gerber, 2008). Instead, we rely on sorting, survey completion, and survey responses. But the model also suggests an obvious intervention to increase turnout: experimentally increasing how often people expect to be asked about whether they voted. In November of 2010 and of 2012, a few days before the election, we post a flyer on the doorknob of treatment households informing them that '*researchers will contact you within three weeks of the election [...] to conduct a survey on your voter participation.*' A control group receives a flyer with a mere reminder of the upcoming election. We find a small increase in turnout in response to the flyer. In 2010, the turnout of the treatment group is 1.3 percentage points higher than the control group (with a one-sided p-value of 0.06). In the higher-turnout presidential election of 2012, the turnout difference is just 0.1 percentage points (not significant). These results are consistent with the results of Rogers, Ternovsky and Yoely (2016), which also informs a treatment group that they may be called after the election about their voting behavior, and find a positive impact on turnout (of 0.2 percentage points).

We get now to the out-of-sample question. Are the results from the get-out-the-vote manipulation consistent with the results from our first set of survey-completion experiments? In the stage of revising the paper for publication, for example, a few referees wondered: the second set of results seems too small given the large estimated value of voting to tell others in the first part of the paper. While we did not have an immediate answer ready, we were able to work on the model estimation so that we would have an out-of-sample prediction of just how large we would have expected the get-out-the-vote effect to be. To do that, of course, we needed some information on the distribution of the other reasons to vote, since we need to know how many people are on the margin of the turnout decision who could be affected by the expected extra person asking about turnout.

The model estimates in the published paper provide an estimate of the distribution of these other reasons to vote, based on the first set of results. It turns out that, based on these model estimates, we predict that an announced visit to ask about voting should increase turnout by 0.3 percentage points, well within the point estimates of the estimated get-out-the-vote effect. Thus, the get-out-the-vote results are quantitatively consistent with the model, if imprecise. Of course, in retrospect this implies that the GOTV intervention was under-powered, but we could not tell ex ante. In any case, the out-of-sample prediction in this case allows us to conclude that, at least, the

two separate experimental results are in fact consistent.

2.5 Improving the Experimental Design

Structural estimation is most often used in economics for observational data, as opposed to experimental data. Structural estimation for example is common for consumption-savings papers (e.g., Gourinchas and Parker, 2002) and is the most common methodology for the estimation of pricing and competition in industrial organization (Reiss and Wolak, 2007). In comparison, field experiments with structural estimation are uncommon (Card, DellaVigna, and Malmendier, 2011).

This gap in the literature—few field experiments with structural estimation—is ironic, as experiments are ideally suited for structural estimation, as also Wolpin (2013) and Rust (2014) point out. (As we saw above, structural estimation is already quite common in laboratory experiments). First of all, in field experiments there is no trade-off between structural estimation and cleanly identified reduced-form findings. In field experiments with structural estimation, one starts from the treatment effect findings, to then move on to the estimation. Most importantly, setting up a model for structural estimation can lead to improvements in design, in a way that one cannot do with observational data, where the existing data constrains the researcher. To illustrate this point, I discuss five examples: (i) time preferences, (ii) limited attention and taxation, (iii) limited attention and energy, (iv) social preference in charitable giving, and (v) gift exchange experiments.

Time Preferences. A first case where estimation informs the design are experiments to identify time preferences. Consider a simple model of intertemporal preferences, with an exponential per-period discount factor δ. In the typical design for laboratory experiments on time preferences (e.g., Andersen et al., 2008 and Andreoni and Sprenger, 2012), a subject at time t allocates a budget M between consumption at times $t+s$ and at time $t+s'$. Suppose that there is per-period interest rate r between period s and s'. The subject aims to maximize $u\left(c_{t+s}\right)+\delta^{s'-s}u\left(\left(1+r\right)^{s'-s}\left(M-c_{t+s}\right)\right)$. Assuming an interior solution, the first-order condition which determines the allocation is

$$u'\left(c_{t+s}\right)-\delta^{s'-s}\left(1+r\right)^{s'-s}u'\left(\left(1+r\right)^{s'-s}\left(M-c_{t+s}\right)\right)=0. \tag{13}$$

Condition (13) makes clear that, provided one knows the utility function $u\left(c\right)$, for a given interest rate r the observed choice of c_{t+s} provides identification for the discount factor δ. Further, observing the choice of c_{t+s} for a variety of interest rates r provides useful identifying variation for the curvature of the utility function $u\left(c\right)$ (usually assumed to belong to a parametric class, such as CRRA or CARA utility). Essentially, varying the interest rate, which is the price of waiting, "prices out" the curvature of the utility function, and thus allows for the identification of the discounting parameter. Indeed, the elicitation of time preferences in the laboratory consists of a within-person design, with a series of consumption choices between different time periods, for varying interest rates r. As we stress later, pricing-out treatments and a within-person design are the two most common tools for structural identification of behavioral parameters in experiments.

Limited Attention and Taxation. In a pioneering paper, Chetty, Looney, and Kroft (2009) provides evidence of limited attention to state taxes using a quasi-experiment with a grocery store

chain. For a limited duration of 3 weeks, in some of the grocery stores, and just for some products, the price tags feature not just the usual pre-tax price, but also the price inclusive of the state tax. Using the diff-in-diff-in-diff structure of the experiment, the authors can estimate the change in sales of the items with the special price label, compared to the control items, compared to the control stores, compared to the control weeks. It turns out that sales decrease by 2.20 units out of an average of 25, an 8.8 percent decrease.

We can interpret this in light of a simple structural model as outlined by Chetty, Looney, and Kroft (2009) and DellaVigna (2009). Assume that the demand D is a function of the perceived price, which is $(1+t)\,p$ when the tax is made salient, but only $(1+(1-\theta)\,t)\,p$ normally, where θ indicates the degree of inattention. We can then write the percentage change in sales due to the experiment $\Delta logD$, using a log approximation, as

$$logD[(1+t)\,p] - logD[(1+(1-\theta)t)\,p] = -\theta tp * D'[(1+(1-\theta)t)\,p]/D[(1+(1-\theta)t)\,p] = -\theta t * \eta_{D,p},$$

where $\eta_{D,p}$ is the price demand elasticity, which the authors estimate to be -1.59. Thus, given that the state tax is .07375, the estimated degree of inattention is $\theta = -(-.088)/(-1.59*0.07375) \approx 0.75$. This is a case in which an experiment is designed with an eye to estimating a structural parameter using a sufficient statistic. The presence of sales allows for an estimate of the price elasticity, which in turn "prices out" the response to the tax disclosure, allowing a way to back out the degree of inattention from the reduction in sales.

Still, despite the clever design, the estimates of the degree of limited attention are quite imprecise: the week-by-week variability in the number of units sold of a product limits the precision of the estimates. Taubinsky and Rees-Jones (forthcoming) revisits the Chetty, Looney, and Kroft (2009) design with an eye to achieving a more precise prediction for the structural estimate of inattention. In turn, this higher degree of precision is used not only to estimate more precisely the average degree of inattention to taxes, but also its heterogeneity across consumers, and how it responds to the potential cost of inattention. With the structural estimation of the key behavioral parameter (inattention) in mind, Taubinsky and Rees-Jones (forthcoming) designs the experiment as a within-subject choice, as compared to the between-subjects design of Chetty, Looney, and Kroft (2009). Participants in an online platform are asked to indicate at what price they would purchase a series of items. Importantly, each person makes a choice first in a condition in which state taxes (or, depending on the treatment, 3 times the state taxes) are charged, and then in a condition in which no taxes are charged. Decisions are incentive compatible: one of their willingness to pay elicitations is used in a Becker-DeGroot-Marshak mechanism. Participants keep any unused portion of their budget and are shipped any good that they purchase.

With this design, the authors are able to observe the willingness to pay with, and without taxes, for a number of products, allowing them to obtain more precise information on attention, including information on heterogeneity. The findings confirm the Chetty, Looney, and Kroft (2009) results of substantial inattention, with much higher precision. Interestingly, the degree of inattention declines substantially in the condition where subjects are charged 3 times the tax: thus, there is a response to the cost of inattention. Further, there is vast heterogeneity in inattention: some individuals appear

to fully neglect taxes, others pay partial, or full, attention, but others yet put too much attention, acting as if taxes were significantly larger. I return in Section 2.6 to the important implications that these results have for the welfare effects of inattention.

Limited Attention and Energy. Energy savings are important, especially given the concerns about global warming. CFL light bulbs consume only a fraction of the energy that incandescent light bulbs use, saving about $40 over the life of a single light bulb compared to an incandescent. Thus their adoption is a win-win, lowering the electricity bill for the consumers, and reducing energy use. Yet, their take-up of CFL (or LED) light bulbs is still very partial. One explanation is that consumers are inattentive to the savings; another explanation is that consumers are aware of the savings, but prefer the incandescent light bulbs, perhaps because of the light quality.

Allcott and Taubinsky (2015) designs an experiment to measure the impact of inattention and test for the two explanations. Their within-subject design has participants in an online sample indicate their willingness to pay for a package of incandescent light bulbs, versus for a CFL light. This is done though a multiple price list. Using the same limited attention model as above, and for simplicity assuming energy savings of exactly $40, the WTP for a CFL versus an incandescent in this case is $WTP_{Pre} = v + (1 - \theta)\, 40$, where v is the non-monetary quality valuation for the light of CFL versus an incandescent (and is thus 0 if the consumer finds them indifferent), and θ indicates the degree of inattention to the energy savings.

Having elicited this first willingness to pay, subjects in the treatment group are exposed to information about the savings for the CFL, while subjects in a control group are exposed to alternative information. After the information, a second price list elicits the willingness to pay again. Under the assumption that in this case subjects in the treatment group are fully attentive, the new WTP is $WTP_{Post} = v + 40$. The change between the two elicited measures of willingness to pay, compared across the treatment and control group, provides a measure of how much inattention to the energy savings: $WTP_{Post} - WTP_{Pre} = \theta 40$, allowing for the estimation of the limited attention. In this design, thus, the estimation of limited attention is very straightforward. Furthermore, as the authors stress, this within-subject design allows for the estimation of the limited attention parameter θ as the individual level. In particular, as Allcott and Taubinsky (2015) shows, it is important to estimate the limited attention for different levels of the demand function in order to understand the impacts of a subsidy on CFLs. Intuitively, if consumers at the margin of indifference have inattention $\theta = 0.2$, this justifies a $0.2 * 40 = \$8$ subsidy to compensate for the limited attention. But this subsidy would be a poor welfare choice if instead the consumers on the margin of purchase are fully attentive, and consumers that are inattentive would not purchase the CFL in any case because they have a very negative v (i.e., they do not like the CFL light). In this case, the within-person WTP design is tailor-made to allow for a transparent elicitation of the distribution of limited attention at the individual level.

The authors, while finding evidence of consumer inattention, estimate it to be small: the information disclosure increases the valuation by only of about $2, suggesting a $\hat{\theta} = 0.05$; given the distribution of the θ parameter, the findings justify a subsidy of about $3, but not a CFL ban, as implemented by some countries.

Charitable Giving. As a fourth example, I discuss the role that estimation played in a field

experiment on charitable giving I ran with Ulrike Malmendier and John List (DellaVigna, List, and Malmendier, 2012).[10] The idea of the paper is to discriminate between two sets of reasons for giving to a charity, when asked. One reason is that the act of giving is associated with a utility increase, whether due to altruism, warm glow, or prestige. Alternatively, individuals may actually dislike giving money to a charity but feel worse saying no to the solicitor, and thus give due to social pressure. These two motivations for giving have very different welfare implications for the giver: giving is welfare-increasing for the donor in the first case, but welfare-diminishing for the donor in the second case.

To distinguish these two motivations for giving, for the experimental design we settled on a door-to-door campaign where we would randomize the extent to which people are informed about the up-coming fund-raising campaign. In the treatment group, but not in the control group, we posted a flyer on the door-knob of the household, informing of the upcoming fund-raiser. Households then vote with their feet, so to say, by deciding whether to open the door and then whether to give, outcomes that we could measure.

The initial plan was to run just this design. We did decide, though, to write down a model with altruism and social pressure to clarify what assumptions we were implicitly making. The dividends from writing the model were substantial. The model suggested to us new predictions—comparing the impact on larger donations which are more likely due to altruism versus on smaller donations—and also new treatments—the addition of a flyer treatment with a "Do not disturb" box to identify sorting out more easily.

Most importantly, though, we realized that we came up short in terms of identifying the altruism and social pressure parameters in the model. In the model, we assumed a cost function of shifting the probability of being at home (in response to the flyer). Without information on this function, we could not use the observed sorting to identify the parameters. Suppose for example that the flyer reduces the probability of home presence by 4 percentage points: does that translate into a significant social pressure cost, or a tiny one? That depends on how elastic people are in altering their presence at home, and thus the rate of answering the door, and we had no way in the initial design to know that parameter.

We thus decided, still in the design stage, to run a parallel field experiment specifically designed for the purpose of identifying that elasticity. We posted flyers announcing that we would conduct an X-minute survey for a payment of \$Y, and we varied across treatments the time duration X (5 or 10 minutes) and the payment Y (\$0, \$5, or \$10). The responsiveness in the presence at home with respect to the duration and the payment provided the identification to the cost of avoidance, hence allowing us to back out all other parameters.

What did we find? The reduced-form results point to the importance of social pressure for solicited donations, with the most important piece of evidence being the fact that the flyer with opt-out option lowers donations significantly, and especially small donations. As discussed above, this is a key prediction of the social pressure framework which we had not honed in until we wrote the model. As such, writing the model provided us with a tighter reduced-form test.

What do the survey treatments and the ensuing parameter estimation add to these results? They

[10] Some of this description appears also in Card, DellaVigna, and Malmendier (2011).

allow us to estimate the social pressure cost of being asked, which is sizable. Interestingly, this social pressure cost is higher, at \$4, for the local, more liked charity, compared to the out-of-state charity (\$1). As we discuss shortly, these magnitudes have implications for the welfare effects of charitable fund-raising.

Overall, getting ready for model estimation still at the design stage led us to alter the initial design and add new treatments. These additional treatments did not have to do with charitable giving—our focus—but they pinned down the cost of sorting, which itself allowed us to estimate the social preference parameters. As we discuss later, it is typical that a good part of the experimental design is geared to identifying one, or more, nuisance parameters, since without those one cannot identify the parameter of interest. Further, these additional treatments often take the form of 'pricing-out' treatments, using monetary variation to pin down a parameter in monetary terms.

From this experiment, we learned that it really helps to invest significantly in setting up the model and estimation in advance of the final design, to make more informed design choices. Notice that this approach differs from the usual power analysis, which is aimed at ensuring adequate statistical power to identify a treatment effect of, say, x percent. With model-based simulations, one can ask whether all of the key parameters are, at least in principle, identified, and with how much precision. I will return to this point in Section 4.

Gift exchange. We discussed in Section 2.2 how it is not possible to identify the relevant social preference parameters from the gift exchange experiments in the field à la Gneezy and List (2006). In DellaVigna et al. (2017), we present a design that aims to preserve, as much as possible, the features of these gift-exchange-in-the-field experiments, while at the same time allowing for parameter estimate. Recall from expression (11) that two elements are missing to identify the social preference parameters: we do not know the return to the employer p_E, and we do not know the curvature of the cost of effort function $C(e)$. The first element is, per se, easy to address. We turn to a task where we can (truthfully) inform the subjects of the average return to the employer of their effort: the subjects fold and prepare envelopes for charitable fund-raising campaigns, and we inform them that the average money raised with an envelope in similar campaigns is 30 cents. Furthermore, we also inform them that there is a donor match for some groups of the envelopes, raising the return to 60 cents. Thus, we can both state, and vary experimentally, the return to the employer (a charity in this case).

To estimate the curvature of the cost of effort function, we resort to the same "pricing out" trick as in the other experiments detailed above: we vary the piece rate to the workers p_W per envelope: sometimes it is zero (with only a flat payment), in other cases it is 10 cents per envelope (with a reduced flat payment), while in other cases it is 20 cents per envelope (with no flat payment). We find that subjects respond to the variation in the piece rate, but not very much: overall, we estimate an elasticity of 0.1. A small elasticity makes sense, since it is hard in a fixed amount of time (20 minutes of envelope folding) to fold many more envelopes, even if one wants to. Armed with an estimate of the cost of effort function and a measure of the return to the firm, we can estimate different models of social preferences of the employee towards the firm.

We estimate that the workers do have social preferences towards the employer, but that they do not put much direct weight on the exact return to the firm: their effort is about the same when they

are working with envelopes with average return to the employer (30 cents) or with twice the return to the employer (60 cents). The lack of response to the quantitative return to the firm contrasts with a clear effect on effort of the piece rate for the workers. We also show that, at least in our settings, the different gift treatments do not have significant impacts on effort, suggesting that, while baseline social preference (pre-gift) are sizable, they are not as easily affected by gifts from the employer.

2.6 Welfare and Policy

Behavioral economists in the first 25 years of history of the discipline stayed largely away from policy recommendation and controversial welfare statements. The emphasis was instead on deriving solid facts, and behavioral models to understand them.

In the mid 2000s, this initial trend started to change. Some leading behavioral economists started making a case for cautious paternalism (Camerer et al., 2003): policies that would help individuals with bias, without (significantly) hurting individuals with more standard preferences. Examples are opt-out defaults for retirement savings, or prominent displays of healthy dishes in a cafeteria. This helps with procrastination and self-control, while requiring just a minimal effort to over-ride in case one so desires. Independently, Bernheim and Rangel (2009) puts forward a framework for behavioral welfare economics, articulating a set of principles to make welfare judgments, and identifying inconsistencies, as Bernheim and Taubinsky (2018) discusses in their chapter.

Policy applications of behavioral economics then really gained traction with the book on nudges (Thaler and Sunstein, 2008), which articulated and brought to mainstream policy the logic of cautious paternalism: small interventions, like the two highlighted above, that steer people in the right direction, without imposing large costs to individuals who wish otherwise. The key difference from the previous work on behavioral policy is that behavioral policy units started to be set up, first in the UK government and then worldwide, to apply the concept of nudging to a variety of policy issues, such as increasing the take up of EITC benefits among the eligible, getting unemployed workers to show up at job fairs, and reducing the share of late tax payments (e.g., Halpern, 2015). Thus, behavioral ideas have been applied on a scale previously unheard of, by governments on the right and on the left, at the national level, and at the city and state level.

As the policy application of behavioral ideas goes mainstream in the policy world, in the academic literature there still are remarkably few cases of welfare evaluations of policy applications of behavioral findings. Yet, welfare evaluations are critical for policy choices, just as they matter for policy within the standard economic model, and in fact even more, given the conceptual difficulties of handling, for example, multiple selves in self-control models, or individuals with wrong beliefs. For example, how are we to know if we are "nudging for good", as Nobel Laureate Richard Thaler likes to put it, without a welfare evaluation? In turn, welfare evaluations require an explicit model and set of assumptions, as Bernheim and Rangel (2009) stresses. The model in particular needs to clarify which is the welfare-relevant state, or utility function, which is used to conduct the welfare evaluation. For example, in the present-bias models the assumption is typically that the welfare-relevant state is the ex-ante state, that is, the long-run self; one could, however, take a different welfare perspective, such as taking the short-run self's utility, or using a Pareto criterion on the

utility of the different selves. Once the welfare perspective is spelled out in a model, structural estimates of the parameters allow for a quantitative welfare evaluation.[11]

I review here some contributions in seven areas in which there are relevant insights for policy from a welfare evaluation based on structural estimation, including revisiting some papers discussed above: (i) health insurance, (ii) limited attention and taxation, (iii) retirement savings, (iv) firm pricing, (v) charitable fund-raising, (vi) get-out-the-vote experiments and (vii) energy policies.

Health Insurance. We discussed earlier how Handel (2013) provides evidence of substantial inertia in the choice of health insurance plans, and estimates switching costs (taken at face value) of about $2,000. With these estimates at hand, the author then asks the welfare question: what would happen if individuals were less inertial, that is, they had a lower switching cost k? Importantly, this is not just a welfare hypothetical, but rather one with policy implications. Redesigns of the health insurance choice protocol can plausibly lead to lower switching costs, and this has been discussed as an active goal of nudge policy. Within the model of Handel (2013), the welfare consequences of an individual consumer being less inertial are clear: that consumer would be better off, with the welfare benefits in the same order of magnitude (if smaller) than the switching cost reduction. But policies that lower switching costs by, say, changing the platform design would lower the costs for all the employees; what would the welfare effects then be?

The welfare impact of this aggregate switching costs change is estimated to be very different: it would *reduce*, not increase, consumer welfare. To see how this is possible, consider that health insurance markets do not operate in a first-best world: adverse selection is a real issue which, as well known, can lead, at the limit, to the unraveling of health insurance markets: if only the sick employees choose the high-coverage plans, these plans get more expensive, potentially making the selection even more adverse, in a downward spiral. Pooling of consumers, in this respect, is highly advantageous. With this in the background, it is easy to see why inertia can help, in that it limits the adverse selection: once consumers choose initially, they rarely revise their subsequent choices, effectively ending in a pooling equilibrium. This result illustrates an important principle for behavioral welfare policies: in a second-best world, behavioral features can sometimes work to counteract another friction, in which case undoing the behavioral bias can be deleterious.

Health insurance choice is a setting where behavioral frictions have a first-order impact, with ensuing welfare implications. In addition to inertia, Handel and Kolstad (2015) documents that employees are remarkably poorly informed about features of different plans, and how, at least, in that setting, poor information largely explains what one could have (erroneously) estimated to be high risk aversion among employees. Bhargava, Loewenstein, and Sydnor (2017) provides evidence that, in a setting in which health insurance choice was not very transparent (a common issue), a majority of employees in fact choose a dominated plan, with large welfare costs. In the Handbook chapter, Chandra, Handel, and Schwartzstein (2018) discusses a number of other examples, and the implications.

Taxation. Another setting in which behavioral factors have a first-order impact is taxation, as the Handbook chapter by Bernheim and Taubinsky (2018) discusses in depth. As we discussed

[11] In some cases, a quantitative welfare evaluation is possible without structural estimation, such as when one elicits a measure of willingness to pay.

above, Chetty, Looney, and Kroft (2009) provides evidence of significant limited attention to state taxes using a quasi-experiment in a grocery store. What are the welfare implications of limited attention? Limited attention, in the framework of Chetty, Looney, and Kroft (2009), may in fact be beneficial, because it lowers the dead-weight loss of taxation. This is another example of two frictions counteracting each other, similar to Handel (2013).

As we saw above, Taubinsky and Rees-Jones (forthcoming) revisits this result with a field experiment on an online platform designed to provide more precise identification of the level of limited attention, its heterogeneity, and how it responds to the stake size. Using the detailed information that the within-person design provides, Taubinsky and Rees-Jones (forthcoming) provides evidence of significant heterogeneity in the limited attention parameter. This heterogeneity, it turns out, overturns the welfare result in Chetty, Looney, and Kroft (2009), which is based on a model with the same degree of inattention across consumers. If everyone has the same degree of inattention, yes, inattention can be useful in counteracting the dead-weight loss. But if individuals are as heterogeneous in their inattention as Taubinsky and Rees-Jones (forthcoming) documents, then the opposite result applies, with the heterogeneous bias lowering welfare. This occurs because, once there is heterogeneity in inattention, the goods are mis-allocated across consumers, with welfare losses that, to a first approximation, are quadratic in the heterogeneity of the bias: the individuals in the tail of heterogeneity suffer disproportionate losses due to their bias. Here, the estimated heterogeneity in attention is crucial for the result; if the heterogeneity were more limited, the Chetty, Looney, and Kroft (2009) welfare result would still apply. Thus, the structural estimate of the heterogeneity informs the welfare result.

Retirement Savings. The Madrian and Shea (2001) finding of inertia in retirement savings played an important role in behavioral economics. Even before that, the concern that Americans under-save for retirement played an important role in Laibson (1997)'s hyperbolic discounting model. It is not surprising thus that policy implications of behavioral economics for retirement savings have been an active area of research. Most of this work has been model-based, often with calibrations, but with a limited role, so far, for structural estimation.

Given the large default effects of Madrian and Shea (2001) what is the optimal default? At first, it would seem that an opt-out default, which maximizes the participation into a retirement plan, is optimal. But if opt-out is preferable, what should the default retirement plan be? Almost certainly not at the same conservative default as in Madrian and Shea (2001), with a 3 percent savings rate in a money market fund. With an opt-out scheme, individuals whose optimal savings rate differs enough from the optimal plan will need to pay a transaction cost if they switch, or stay in a sub-optimal plan. An alternative option is active choice: as Carroll et al. (2009) shows, when employees are asked to make an active choice (with no apparent default), the large majority participates in a retirement plan. The downside of active choice is that everyone needs to pay the effort cost. Carroll et al. (2009) lays out a model with stochastic switching costs akin to the one in Section 2.1 and show conditions under which active choice is better, or a default is preferable.

Bernheim, Fradkin, Popov (2015) expands the welfare lessons of Carroll et al. (2009) and structurally estimate different behavioral models on aggregate choice data in response to opt-in, versus opt-out options, as in Madrian and Shea (2001). Under a standard model, the implied switching

cost that rationalizes the data in the maximum likelihood estimates is unreasonably high, on the order of $2,000. This is along the lines of the calibrations above in Section 2.1 and in DellaVigna (2009), and has a similar magnitude as the estimated switching cost in the health setting by Handel (2013). Switching costs of this magnitude are highly improbable, as the authors point out, and point towards a behavioral explanation of the findings.

Bernheim, Fradkin, Popov (2015) considers different behavioral models, including models of time inconsistency, attention, and anchoring, and for each of the models, presents a welfare evaluation along the lines of Bernheim and Rangel (2009). While the results differ across the different behavioral models, a common conclusion is that the optimal default tends to be at the highest contribution rate matched by the employer. At these companies, the match rate is so generous (50 percent or 100 percent) that for a variety of models and distributions of parameters, it is broadly optimal to ensure that the behavioral agents do not forgo the benefit of the match. Bernheim, Fradkin, Popov (2015) also discusses the welfare criteria for the different behavioral models, a point discussed in detail in the Handbook chapter by Bernheim and Taubinsky (2018).

The choice architecture of the optimal contribution default, of course, can be more sophisticated than just the choice of a default for immediate saving at enrollment. The Save More Tomorrow plan by Thaler and Benartzi (2004) consists of a default choice that will automatically increase contributions towards retirement at the time of future pay increases. This choice responds to the preference of present-biased people to increase savings, but not at the cost of cutting consumption in the present. The take-up for this plan is high in a group of employees offered the plan, suggesting that a plan with these features may potentially offer even higher welfare. It would be interesting to see a full welfare evaluation of this Save More Tomorrow plan with estimates of the underlying behavioral model.

A related, but separate, policy question is about the optimality of existing policies for retirement, and especially social security, which is a form of forced savings, and the 10% penalty for withdrawal from 401(k) plans. In the presence of self-control problems, forced savings, or a savings device with costly withdrawal, can be optimal. Beshears et al. (2017) considers the optimal illiquidity for the case in which individuals are present-biased, but they are heterogeneous in their degree of present bias. Under a calibrated distribution of parameters, a combination of forced savings (like social security) and private savings (like IRAs) is optimal.

While Beshears et al. (2017) focuses on the overall design of the retirement system, Beshears et al. (2015) designs a portfolio choice problem to provide empirical evidence on the demand for commitment. Participants in an experiment invest money into a combination of different accounts, including accounts with illiquid features. When the liquid account and illiquid account offer the same interest rate, the most illiquid account attracts more money than other account with less commitment. The combination of this and other facts allow the authors to calibrate the percent of sophisticated, naive, and exponential agents in their sample. The demand for commitment in this more rarefied portfolio choice can at least in principle inform the optimal liquidity design referred to above.

Firm Pricing. I discussed in Section 2.4 an example of work in the area of Behavioral Firms: the study by Cho and Rust (2010) on price uniformity in the car rental market. In work with

Matthew Gentzkow (DellaVigna and Gentzkow, 2017), I explore a different form of price uniformity: pricing in retail chains—grocery chains, drug store chains, and mass merchandise chains. As is well-known, most retail chains adopt elaborate pricing schemes over time, with frequent sales occurring at staggered timing for different products. In the paper, we document a less-known, but no less pervasive, aspect of their pricing: the uniformity in pricing across stores. That is, even chains with hundreds of stores typically charge the same price, or at least very similar prices, across all of the various locations. In doing so, the chains would appear to forgo significant profits from pricing to the local elasticity. But do they? We document that we can exploit the pattern of sales over time to build a measure of price elasticity η_s for store s; this measure is highly correlated with local determinants of purchasing power like income. We then turn to testing a very simple prediction of a monopolistic pricing model with constant-elasticity demand, that the optimal price p_s should satisfy $p_s^* = c_s * \eta_s / (1 + \eta_s)$. Under the assumption that any one chain will have constant marginal cost, this very simple structural model implies that we can test it with the OLS regression

$$log\left(p_s^*\right) = \alpha + \beta log\left[\eta_s / \left(1 + \eta_s\right)\right] + \epsilon_s. \tag{14}$$

Under the structural model, we should find $\beta = 1$. Notice that this is a case in which a structural model can be tested with a very simple specification.

Instrumenting for the price elasticity η_s with per-capita income for the consumers in store s, we estimate that prices in a chain are nearly unresponsive to the local elasticity: $\hat{\beta}_{within} = 0.1$, which is much smaller than the optimal-pricing null of $\beta = 1$. This result confirms the essentially rigid prices across stores. At the same time, we find that chains do seem to act accordingly to optimal pricing as in (14) when relating the overall price level of a chain to the overall income level of the areas where it operates. In that between-chain regression for food chains over 63 chains we find $\hat{\beta}_{between} = 0.9$. We conjecture that this form of price rigidity may arise from managerial inertia: in an industry that has always priced this way (all 63 grocery chains, covering a total of nearly 10,000 stores, price in a similar way), managers may be too inertial to take the cost of experimenting. We estimate that, by forgoing pricing to market, the chains have on average 8 percent lower profits.

We then turn to a perhaps surprising policy implication of this fact: price uniformity has very regressive policy implications. Most stores that operate in lower-income areas belong to chains that operate also in middle-income areas. Given that these chains set prices to respond to the average income they operate in, they charge higher prices than they would if they were to price to market. Similarly, stores in higher-income areas often belong to chains that also operate in middle-income area, and thus charge lower prices than if they were pricing to market. Thus, price uniformity leads higher-income consumers to receive a transfer from lower-income consumers. We can use our simple structural model to estimate the extent of this transfer. We estimate that if stores were to price according to (14) with $\beta = 1$, consumers in lower-income areas would pay about 2 percent less for groceries, and consumers in higher income areas 6-8 percent more. This implication is all the more striking since fairness concerns are often voiced as justifications for price uniformity.

Charitable Fund-Raising. I return here to another paper discussed above, the field experiment on altruism and social pressure in charitable giving (DellaVigna, List, and Malmendier, 2012). As

we discussed above, our evidence suggests that social pressure is a significant determinant of giving in door-to-door fund-raising. Does this result have meaningful welfare implications? We consider in particular the question of what fund-raising does to the welfare of the households contacted about giving. This focus is different from the usual emphasis of charitable giving on the welfare of the recipients.

In a model with no social pressure, the welfare effect of a campaign can only be positive, since a donor can always costlessly say no. But in the presence of social pressure, this free-disposal condition does not hold: the benefits of a campaign for the willing donors have to be weighed against the cost non-donors pay for being asked and saying no, which we estimate to be about $4 for a local charity. The welfare impact for the non-donor is especially important, since only a small minority of the households contacted end up giving. In addition to this cost for non-donors, we estimate that as many as 50 percent of the donors would have preferred not to be asked, because social pressure induces them to give when they would not have given otherwise, or give more than they otherwise would. Taking into account these forces, our benchmark specification indicates that our door-to-door campaign induces a welfare loss of about $1 on average per household contacted (including households that were not at home and hence did not suffer a welfare loss, and not counting the benefits associated with the public good provision). An interesting and counter-intuitive result is that raising money for the local and well-liked favorite charity is associated with more negative welfare impacts than raising money for an out-of-state and lesser-known charity. More people are willing to donate to the local charity, but at the same time, the social pressure cost of saying "no" to the local charity is significantly higher, and the second force dominates.

We can also use these welfare findings to speculate about possible regulation of fund-raising campaigns. Currently, charitable campaigns are exempt from do-not-call lists. Yet our results suggest that, if calls are akin to in-person visits, the welfare impact of these calls may be negative too, suggesting that these calls may be added to the list. Our welfare results, though, suggest the potential for a less heavy-handed intervention. If charities gave a heads-up about the upcoming fund-raising, as in our "opt-out" treatment, the welfare effect for givers could only be positive (assuming they receive the notification). Perhaps more interestingly, fund-raisers could benefit from these notifications as well. While they would lose some of the donations due to social pressure, these donations are likely to be smaller, and fund-raisers could save the fund-raising time by addressing only the more-likely givers. Our estimates suggest that these benefits from the notification could outweigh the costs for the charities, at least in our context. We did not envision this result before we did the estimation and turned to the welfare impacts.

Get-Out-The-Vote. Relatedly, we can also do a welfare evaluation for the voting-to-tell others experiment (DellaVigna et al. (2017)), detailed in the previous section. In particular, we can calculate the welfare effects of a GOTV intervention based on informing potential voters that they will be asked whether they voted. This is a GOTV intervention which we designed ourselves, but, independently, a related GOTV message was used in the 2012 presidential election and is evaluated in Rogers, Ternovsky and Yoely (2016). Thus, the welfare evaluation is relevant to actual campaign material used.

First, we calculate the welfare effect of this GOTV intervention. The average value of being asked

about voting, is estimated to be -$2.8 for voters and -$5.9 for non-voters. These sizable negative welfare effects occur because on average individuals do not derive pride from voting, and they strongly dislike admitting that they did not vote. We can relate this to the predicted effectiveness of this GOTV intervention, which is estimated to increase turnout by 0.3 percentage points. Thus, to get one extra vote with this intervention, 295 people would have to be contacted. Taking these two numbers together, we estimate that this GOTV intervention would result in a dis-utility of $1,189 per additional vote yielded.

This welfare dis-utility dwarfs the cost estimates in the current literature, which typically include just the postal costs of sending the GOTV material (e.g., Rogers, Ternovsky and Yoely, 2016). As far as we know, ours is the first welfare evaluation of a get-out-the-vote intervention, an area of vast growth in political science.

Energy Policies. The estimate of the welfare implications and policy implications in the previous cases rely on the estimate of the model parameter given the observed behavior, e.g., in charitable giving, or voting. But welfare evaluations can also be set up in a more straightforward way with willingness to pay (WTP) measures. Allcott and Kessler (forthcoming) does so for the case of the Opower Energy Use reports. Following the pioneering work of Robert Cialdini (Schultz et al., 2007), utility companies have been partnering with the Opower firm to design energy use reports. The reports, which provide a comparison to the energy use of neighbors, have had remarkable success in reducing energy use in the treatment households by about 2 percent, an effect that appears to be fairly long-lasting (Allcott and Rogers, 2014). Such reports have spread quickly and have been used by a number of utilities, with millions of households receiving them.

Yet, what do we know about the welfare impact of these reports? The typical welfare impact uses the energy savings, but does not take into account costs of adjustment by consumers, or social pressure costs which are not implausible given the research discussed above (e.g., DellaVigna, List, and Malmendier, 2012). Consumers may reduce their energy use upon hearing that they do worse than neighbors, but they may resent such messages. A full welfare evaluation would include also this consumer evaluation.

The authors ofAllcott and Kessler (forthcoming) design a survey which they send to consumers who have received energy reports for a period, but for whom the energy reports might be discontinued. The survey then, effectively, asks for the WTP to continue receiving the reports. To make this plausible, the survey respondents are given a budget of $10, and are asked, for example, whether they prefer "*4 more Home Energy Reports PLUS a $5 check, OR a $10 check?*" This checks whether they value the continuing reports more than $5. They are also, among other questions, asked whether they prefer "*4 more Home Energy Reports PLUS a $10 check, OR a $5 check?*". This question measures whether the dis-utility of receiving more reports is smaller than $5. The response from a series of seven such questions provides an interval for the WTP.

The key result of the survey is that the willingness to pay is largely positive, if moderate, with a similar result for phone respondents and mail respondents. Only 15 percent of respondents report a WTP more negative than -$1, while 45 percent of respondents report a WTP above $1. Overall, the average WTP is $2.5. The authors can then take this number, integrating with the other components of the welfare evaluation (such as producer surplus) to show that altogether the Energy

Reports produce welfare effects that are positive, but smaller than in previous calculations.

An advantage of this type of welfare evaluation is that it provides a transparent welfare number with very little structure, since the WTP is a sufficient statistic under the assumption of their model. At the same time, the elicitation of WTP typically relies on survey questions, which often have a low response rate and which can also turn a natural interaction, in this case with a utility, into an artefactual setting. I should also note that the elicitation of a WTP relies on the respondents being able to correctly forecast their future behavior, in this case how they will use the Opower reports. While it is hard to know whether consumers have correct beliefs, it helps when they have had experience with the behavior in question, like in this case where they had had access to the reports for a period before the WTP.

3 Limitations of Structural Behavioral Economics

3.1 Not the Right Tool

I expanded above on a number of advantages of doing model-based structural estimation in behavioral economics. Does this mean that all of, or most of, behavioral economics should be structural? No! I discuss a number of costs of structural estimation to weight against the benefits. One should weigh the advantages against the limitations case by case. But there is an additional reason that in many cases one should not go for structural estimation: it is just not the right tool. I provide here four categories, and examples, of that.

Novel Areas and Exploratory Analysis. Estimation presupposes that there is a model to estimate. That is typically not the cases for novel areas of explorations, where the authors are breaking new ground. For example, some of the most striking results in the last decade in the area of social preferences are about moral wiggle room. Among the earliest papers, Dana, Weber, and Kuang (2007) shows that subjects choose to remain ignorant about the payoffs associated with an action in order to justify more selfish behavior. This very influential paper does not have a model, despite being published in a theory journal (*Economic Theory*). That makes perfect sense since back then there was not a good model for this set of facts. In fact, this area still remains an area under active exploration.

A different example is for results on, say, morality or culture, areas where there is a lot of interest, but typically not a model that would be useful to estimate. A further case is a paper that shows that framing effects of a particular type exist, so that two different ways to present the same identical choices lead to different results (e.g., Benartzi and Thaler, 2002). In some of these cases, a model can be useful to set up the null hypothesis that is then rejected: in the cases above, the hypothesis of standard social preferences with no moral wiggle room, or no framing effects. But it would not help at all to estimate this model, since it is just a straw man, so to say.

Descriptive Results. In some cases, the area of research is well-trodden, but the interest is on a reduced-form, descriptive finding. For example, it is important to know which factors increase the take-up of benefits among a population that is eligible. Bhargava and Manoli (2015) shows that there is a striking effect for the EITC take-up of just simplifying the form sent to recipients from

2 (crowded) pages to 1 (clearly organized) page. This is an important result, but not one that we need a model to interpret (until, at least, we have a model of ease of comprehension of language).

In other cases, we care about comparing magnitudes across studies, but the relevant magnitudes are of reduced-form effects, not of structural parameters. Consider again the finding of inertia in retirement savings of Madrian and Shea (2001). Yes, switching from opt-in to opt-out alters the choice of 45 percent of employees, which is surely impressive. But perhaps participation in retirement savings is very elastic to other factors, such as information and peer effects. That turns out not to be the case. At all. Duflo and Saez (2003) uses a clever inducement design to get employees of a university to obtain information about retirement: they mail $20 coupons to (some of) the staff in the treatment department, but not to the staff in the control department; the coupons can be redeemed at the retirement fair. The treatment is very successful in that it induces an extra 16 percentage points of employees to attend the fair. Yet this results in only a 0.9 percentage point increase in the participation in the retirement plan. Choi et al. (2006) documents another case of a company that offers a one-hour financial education class. All of the participants that are not already enrolled in a 401(k) plan state that they intend to start saving for retirement. Yet, only 14 percent do, compared to 7 percent in a group of non-attenders who also were not participating in a 401(k) plan before. The 7 percent difference between the two groups, small as it is, is an overstatement of the impact of the financial education class, given the self-selection of individuals in the retirement class. Even changes of key features of the retirement plans, such as the match rate, have much smaller effects than changes in the default investments (Choi et al., 2006). This comparison of reduced-form magnitudes makes it even clearer how strikingly large the default effects are.

Not Enough Information. In other cases, it would be interesting and valuable to estimate a set of behavioral parameters, but the data does not provide enough information to get to the estimates. Two papers on reference dependence provide ideal examples. Reference-dependent models have a small number of parameters like the loss aversion parameter λ which we are interested in comparing across settings, so it would be especially valuable to obtain model estimates. Yet, consider the case of tax filing study of Rees-Jones (2018). The paper provides evidence implying that individuals have reference-dependent preferences with zero tax due as the reference point. An individual who had $12,000 of taxes withheld will work harder to find receipts for tax deductions when the pre-deduction tax due is $12,200 (and thus the individual would owe the IRS) than when the tax due is $11,800 (and thus the individual would get a refund). Loss aversion relative to the owing-zero-taxes reference point drives the extra effort in searching for receipts. Rees-Jones (2018) makes precise predictions based on the reference-dependent model, and alternative models, and interprets the evidence in light of these models. Having set up carefully a model of reference-dependent tax filing, Rees-Jones (2018) also shows that, while it is possible to estimate how much extra tax elusion we observe due to reference dependence, it is not possible to estimate the prospect theory parameters, such as the degree of loss aversion. This is because we do not have a way to estimate the shape of the effort function to find receipts. A very similar case applies for the Allen et al. (2017) paper which documents that marathon-runners bunch at round time numbers (from the left), such as 3h59m. The observed bunching, which is a prediction of reference dependence with loss aversion, is quantitatively consistent with different values of the behavioral parameters depending on the curvature of the effort

function of running. In these cases, the model provides guidance on what can be estimated, and what not.

Observational data is a typical setting in which we may be unable to estimate the model parameters, even when the setting nicely matches the underlying behavioral model: we are limited by the existing evidence. The inability to estimate model parameters often applies also in the case of laboratory or field experiments. However, in these cases one can typically design the treatments so as to create the variation needed for estimation, as I discussed in Section 2.5. As this chapter argues, not doing so can represent a missed opportunity, though I provide a counter-argument in the next point.

Model and Axioms. Consider again the situation in which the author intends to focus on an important link between the model and the data, like in the previous case. Instead of structurally estimating the model, one can alternatively derive a series of comparative statics and predictions of the model, which are then checked with the data. These predictions of the model could be further grounded even in an axiomatic characterization. This is especially useful when the author is comparing two models which make opposite sign predictions about a particular comparative statics.

An advantage of this form of model-based empirical testing, compared to the one with structural estimation, is that to derive structural estimates one needs to make a range of additional assumptions, which jointly go into the results of the estimation.[12] The directional comparative static, or prediction, of the model can be instead derived under a broader set of conditions. I will cite here just some examples. In the realm of social preferences, in a lab setting Andreoni and Bernheim (2009) derives the implications of a model of social image and test it experimentally, while Bandiera, Barankay, and Rasul (2005) presents a simple model of altruism and its unique implication for a switch from relative pay to piece rate, as observed in the field evidence. Among the examples on reference dependence, Barberis, Huang, and Santos (2001) presents a model of implications of reference dependence for asset prices and Fehr and Goette (2007) consider the unique comparative statics of reference dependence in a model of labor supply. In the realm of time preference, DellaVigna and Malmendier (2006) tests a qualitative implication of present-biased models for gym attendance, and Kaur, Kremer, and Mullainathan (2015) presents a detailed model of the implication of self-control problems on the demand for commitment devices at work, and test the implications in a field experiment. In all of these cases, an explicit model derives unique comparative statics from the behavioral model, and compares it to a version of the standard model. In some cases, the characterization of the comparative statics of the behavioral model follows directly from an axiomatic characterization, although this is more common for laboratory experiments (e.g., Halevy, 2015) than for field evidence.

3.2 Complexity and Time Costs

For the studies where structural estimation is the right tool, there still are important limitations of structural estimation. These limitations are shared with applications of structural estimation in other fields, but I emphasize examples, and specific issues, within behavioral economics.

[12] A response to this criticism of structural estimation is the use of semi-parametric approaches, requiring fewer assumptions. In practice, this approach has not yet proven viable in a behavioral setting, given the need for much more information.

Perhaps most obviously, structural estimation typically takes much more time, given the number of necessary steps. First, one still needs to gather data, or design an experiment, and obtain the reduced-form results. Second, one needs to outline a model, as for the case of model-based empirical research. Unlike in the case in which one tests just a comparative statics or qualitative prediction, the model needs to be fully specified, which will take additional time. Third, the researcher needs to outline the estimation strategy, including tests on simulated data. Fourth, the estimation itself is typically a very time-consuming step, especially since one needs to ensure that one gets to reliable estimates that are global, rather than local, optima. Not infrequently, this last part is so computationally complex that a researcher invests a lot of time to learn computational short-cuts and techniques to speed up the estimation. While there are ways to navigate these time trade-offs efficiently (I mention some below and in the last section), there is no sugar-coating the time needed.

A secondary implication of this complexity is that, even setting time to execution aside, errors are more likely to creep up in structural estimation work than in reduced-form analysis. If, for example, one is running a regression, comparing different treatments, or estimating and instrumental variable regression, there is software ready for these purposes that is (one assumes) correct. But a researcher that sets up a consumption-savings life-cycle problem from scratch has many steps that need to be checked carefully. An implication is that structural analysis, being more complex, increases the chance that programming errors may drive the results, or that the estimates may not be stable. There are a number of strategies to guard against this, as I discuss in Section 4. Still, the risk is real, and these extra needed checks are a big reason for some of the time costs.

Sufficient Statistics. A possible saving grace from the time costs and risks of complexity is well-known in the literature: sufficient statistics (Chetty, 2009). In some cases, a parameter, or combination of parameters, can be estimated using a key statistic, or a combination of statistics, that is sufficient for estimation (hence the name). Sufficient statistics, thus, allow for structural estimation (of some parameters) using reduced-form findings, without a need to pin down all of the parameters. This is a desirable route when possible.

This approach has been used in particular to obtain estimates of simple limited attention models. We described in Section 2.5 how Chetty, Looney, and Kroft (2009) derives an estimate of the limited attention parameter taking a simple ratio of empirical moments from the data: the treatment effect of the disclosure of the taxes, and the price elasticity.

Another example within the limited attention literature is Lacetera, Pope, and Sydnor (2012) which considers how left-digit-bias affects the price at an auction site of used cars. In their model, consumers pay full attention to the left-most digit, but only partial attention to the other digits. Taking a step back on the classical application to supermarket pricing, that means that a price of \$6 is coded as \$6, but a price of \$5.99 is coded as $\$5 + (1 - \theta)\, 0.99$. To the extent that there is left-digit bias ($\theta > 0$), consumer will perceive round numbers as a significant ramp up, explaining why most prices are 99 cent price.

Lacetera, Pope, and Sydnor (2012) takes this to the context of inattention to the odometer mileage in the sale of used cars. Intuitively, a car with 20,010 miles will sell for a significantly lower price than a car with 19,990 miles. To firm ideas, assume that the perceived value of a car $\hat{V}$ is a linear function of the perceived mileage $\hat{M}$: $\hat{V} = K - \alpha \hat{M}$, and that the mileage is perceived with

the left-digit bias above. Then, the model predicts that at each 10k mileage increase, the perceived value $\hat{V}$ will jump down discretely by $-\alpha\theta 10,000$: the jump is increasing in the inattention θ and in the depreciation parameter α. For interior mileage levels, instead, the valuation of a car will decrease for each mile driven by $-\alpha(1-\theta)$, where inattention θ is attenuating the slope. Putting these two together, it is clear that one can structurally estimate the inattention θ simply by taking the ratio of the size of the jumps to the continuous slope, and applying them a simple transformation. In this case, structural estimation is no harder than an OLS regression, followed by an application of the delta method.[13]

In this case, this simple estimation procedure for inattention yields a lot of value. The inattention parameter of dealers at the auction house is sizable and precisely estimated at $\hat{\theta} = 0.31\,(s.e.\,0.01)$. But one could ask, why do dealers exhibit inattention, are they really irrational? Applying the same estimation strategy to used car sale prices, Busse et al. (2013) shows that the ultimate buyers exhibit similar inattention, implying that the dealers are just optimally responding, in an interesting example of behavioral IO. Furthermore, they can also test if there appears to be significant heterogeneity in left-digit bias among buyers by comparing purchases in high-income, versus low-income zip codes. Busse et al. (2013) finds that the inattention is a bit lower in higher-income ZIP codes, but the difference is small.

Sufficient Statistics on Published Results. Returning to the discussion of limited attention, with a simple, linear model like the one of limited attention above, one can sometimes do structural estimates of a parameter simply using published data from a paper, even if the authors did not provide structural estimates themselves. As I outline in DellaVigna (2009), an example is Hossain and Morgan (2006) which reports a field experiment to test for limited attention to shipping costs on eBay. Consider consumers with quasi-linear preferences who value an item v, minus the cost to them. In a second-price auction (eBay), they bid their value $b^* = v - (1-\vartheta)\,c$, where c is the shipping cost. If they are fully attentive, they fully subtract the shipping cost from their bid, but in presence of limited attention, they do not do so fully. Thus, the revenue to the seller is $R = b^* + c = v + \theta c$, and is increasing in the shipping cost. Hossain and Morgan sell pairs of CDs, one at very low shipping cost ($c_{Lo} = 0.01$), one at higher shipping cost ($c_{Hi} = 4.00$). By comparing the average revenue under the conditions Lo and Hi, we can obtain an estimate of the limited attention parameter, even if the paper does not do so: $\theta = (R_{Hi} - R_{Lo})/4 = 2.61/4 = 0.65$.

Another case in which one can use published information from a paper with sufficient statistics methods is to estimate the extent of confusion. While behavioral work largely investigates factors that shift people behavioral in a specific direction—for example to choose a particular health club contract, or to neglect taxes—, it is also interesting to ask how much random noise due to pure error, or confusion, there is in simple choice data. Shue and Luttmer (2009) provides evidence of random error in the choice of a political candidate among those in a ballot. In particular, they consider California voters in the 2003 recall elections and exploit the random variation in the placement of candidates on the ballot. They find that the vote share of minor candidate i, $VoteShare_i$, is

[13] Lacetera, Pope, and Sydnor (2012) actually present estimates using a non-linear least squares estimator to allow for a non-linear impact of miles driven M on the value of the car. But the linear model comes close to their estimates as depreciation is approximately linear.

significantly higher for candidates whose name on the ballot is adjacent to the name of a major candidate. In particular, denoting with $VSAdjacent_j$ the vote share of the major candidate when adjacent, they estimate

$$VoteShare_i = \hat{\beta}_0 + 0.0010\,(s.e.\,0.0002) * VSAdjacent_j + Controls. \tag{15}$$

The estimated coefficient can be interpreted as a structural estimate of the degree of confusion: about 1 in 1,000 voters intending to vote for a major party candidate mis-votes. Since there are typically about 3 minor candidates surrounding a major party candidate in a ballot, that implies that about 1 in 300 voters meaning to vote for a major candidate instead vote for a minor candidate. This structural estimate is precise enough that the authors can show that it is larger for more confusing voting methods (such as punch-cards) and for precincts with a larger share of lower-education demographics, that are more likely to make errors when faced with a large number of options. This structural estimate in this simple setting can be obtained with an OLS specification.

Interestingly, one can obtain an estimate of the degree of confusion in a very similar setting, the financial market, from published evidence in Rashes (2001). Rashes (2001) focuses on the trades of two companies, MCI and MCIC. The ticker for the MCI communication company is MCIC, while MCI is the ticker for a little-known closed-end mutual fund, Massmutual Corporate Investors. Some investors attempting to trade shares of the larger communication company confuse tickers and trade the MCI company instead, resulting in a .56 correlation between the two trading volumes. This occurs despite the difference in fundamentals: the MCIC company, for example, has only a .03 correlation in volume with the communication company AT&T. While Rashes (2001) does not run the equivalent of regression (15), we can still back out an estimate using the correlation information and the standard deviations. As DellaVigna (2009) shows, this implies a degree of confusion among investors (for this particular setting) of about 1 in 2,000 trades.

A third example is the case of persuasion rates. In our work on the impact of media bias on voting (DellaVigna and Kaplan, 2007), Ethan Kaplan and I show that in towns where Fox News is available via cable by 2000, the vote share for Bush over Gore is higher by half a percentage point. After spending much time putting together the data needed for these estimates, we wanted to at least attempt to move beyond the reduced-form result to provide a more general measure of the extent to which the media persuades an audience. We thus introduced the concept of persuasion rate with a simple model. To simplify the presentation, assume away the turnout margin: the only choice is to vote Republican or Democrat. We assume that (i) the media message reaches a share e of the population and (ii) the message converts a share f of the individuals who were going to vote Democrat. Then the share voting Republican y satisfies $y = y_0 + ef\,(1 - y_0)$, where y_0 is the vote share prior to hearing the message. If we observe the vote share and the exposure share in treatment and control areas (in this case, town where the cable company carried Fox News and towns where it did not), we can write

$$f = \frac{y_T - y_C}{e_T - e_C} \frac{1}{1 - y_0}. \tag{16}$$

The first term in the persuasion rate f is simply the reduced-form impact on the dependent

variable y, divided by the first-stage, so to say, on exposure rates. This ratio is familiar from, for example, the case of a Wald estimator. The last term resizes the effect by the persuadable population $1 - y_0$. If, for example half of exposed voters were already voting Republican, only the remaining half is persuadable. This model makes a number of simplifying assumptions, such as a monotonic and uniform persuading impact of the message on the audience. But it has the advantage that it requires very little information to be computed.

In DellaVigna and Kaplan (2007), we provide an estimate for the persuasion rate using evidence on audience rates from the media company Scarborough. In a review paper (DellaVigna and Gentzkow, 2010), we apply the persuasion rate to a variety of published papers not only in the area of the media and voting, but also on charitable giving, and consumer demand. The persuasion rate estimates provide at least the order of magnitude for the persuasive impact of a message. We find that persuasion rates in the order of 10 percent ($f = 0.10$) are a good approximation in a variety of studies on the media and voting, suggesting a substantial persuasive power of the media. The persuasion rate is an example of a very simple structural model with assumptions that are clearly falsified in many settings. For example, it assumes that everyone is equally persuadable, while surely there must be heterogeneity in the willingness to change behavior. And yet, its simplicity makes it portable across settings, and thus allows one to do at least a first step of comparability across settings.

Outside of these examples (and especially the limited attention one), sufficient statistics are not yet commonly used to estimate structural parameters in behavioral economics. One natural use would be using bunching estimators for models with loss aversion and a fixed, known reference point: the extent of bunching at this point should, like in public finance application, provide information on the degree of loss aversion. Yet, as I discussed in Section 2.2, several papers on reference dependence do not make use of the bunching prediction. Rees-Jones (2018) does examine bunching of tax-payers at the zero-amount-due reference point, and shows that in this case the amount of bunching, which he measures, does not directly reveal the degree of loss aversion.

Simplicity by Experimental Design. I stress an additional route to simplify estimation. When the study has an experimental nature, the researcher can alter the design to *make* the estimation simpler. That is not an option available for observational studies, where the data is what it is, and at best the researcher can seek the best natural experiment, or data details, for model identification. But in experimental studies, the research can add treatments, or alter design features, precisely to make the structural estimation easier and more transparent.

I discussed already several of these examples in Section 2.5. The laboratory experiments on time preference (e.g., Andersen et al., 2008, Andreoni and Sprenger, 2012, and Augenblick, Niederle, and Sprenger, 2015) estimate the discount factor by observing choices for money now or later at different discount rates. Taubinsky and Rees-Jones (forthcoming), in their experiment on limited attention and taxation, takes advantage of within-subject variation in the degree of taxation. DellaVigna, List, and Malmendier (2012), in order to identify altruism and social pressure in charitable giving, adds extra survey treatments to estimate the response of home presence to monetary incentives. DellaVigna et al. (2016), in a real effort experiment, varies within subject the return to the charity and the piece rate to estimate social preferences in the workplace.

In these cases, the structural estimation becomes quite straightforward. In time preferences experiments like Andersen et al., 2008 and Andreoni and Sprenger, 2012, identification often takes place with a simple tobit or non-linear least squares specification, or simple applications of maximum likelihood. In Taubinsky and Rees-Jones (forthcoming), estimation of most of the results is a variant of linear estimation. In DellaVigna et al. (2016), the estimation is with non-linear least squares. Only DellaVigna, List, and Malmendier (2012) requires more structure and minimum-distance estimation.

What is common across all these examples? In all these experiments the key is a 'pricing-out' treatments, which intuitively provides a way to norm the results in dollar values. In the time preference experiments, it is the variation in the discount rate that alters the price of the future versus the present. In Taubinsky and Rees-Jones (forthcoming), the variation in the tax rate allows one to identify whether the willingness to pay goes down one-by-one with the tax rate, or not. In DellaVigna et al. (2016), the variation in the piece rate makes it possible to price out the curvature of the cost of effort, as at the margin workers equate the marginal cost of effort to the marginal return of effort. In DellaVigna, List, and Malmendier (2012), the flyers for the survey treatments with different advertised survey payments price out the cost of sorting in and out of the home.[14]

There is also a second component to nearly all the cases above: a key within-subject treatments, whether in the interest rate (the multiple price lists), in the level of the tax, or in the piece rate for effort (in the real-effort experiment). The one exception is DellaVigna, List, and Malmendier (2012) where the variation is between people. The within-subject comparison is not a necessary condition, unlike the 'pricing out' treatments, but it helps much with statistical power in identifying the key parameters, as it allows one to compare a choice under multiple conditions, holding constant the person 'fixed effects'. The trade-off with this within-subject structure is that it is difficult to do such within-subject manipulations in field experiments, without revealing the experimental nature of the interventions.

3.3 Robustness to Assumptions and Uncertainty

Structural estimates typically take a long time and, because of the complexity, run additional risk that bugs in the code or improper convergence of the estimation may be responsible for some of the results. But even setting aside time costs and the risk for errors, another issue is that the estimates, and the ensuing welfare and policy implications, are only as good as the joint set of assumptions going into the model. The set of assumptions, as we discuss more in the next Section, includes the estimation methods, assumptions about the error term, and assumptions about functional form and the way auxiliary parameters affect the key predictions, among others. The estimates may be sensitive to changing some of the auxiliary assumptions, and it is often difficult to test thoroughly the robustness if the estimation is very time consuming.

Consider some of the examples discussed thus far. In our study of charitable giving and social pressure (DellaVigna, List, and Malmendier, 2012), would the results change if we used a different

[14]The 'pricing out' idea is also present in papers with reduced-form results. In Bertrand et al. (2010), for example, the authors estimate the impact of various psychological treatments (such as a simplified comparison) on the take up of a loan product. They then compare the impact to the impact of a reduction of 1 percent in the interest rate (one of the experimental arms). Thus, they can measure the effect of the psychological interventions in interest-rate equivalents.

cost function for sorting in and out of the home, or if we estimated the results by maximum likelihood, instead of by minimum distance? In the analysis of consumption-savings choice of Laibson et al. (2017), are the results sensitive to the underlying assumptions about the income process, or to the choice of moments? In the study of inertia in health insurance choice of Handel (2013), are the estimates of switching cost sensitive to the assumption made about predicted health shocks? In the estimate of inattention with respect to energy costs of Allcott and Taubinsky (2015), how important are the assumptions about limited attention?

Questions like the above are both legitimate and important. I discuss three main routes to address questions of this type: (i) exploring, and documenting, extensive robustness analysis; (ii) using calibrated magnitudes, and (iii) assessing the uncertainty of the estimates.

Documenting Robustness. As obvious as this step is, it is critical to have an extensive discussion of robustness within the paper, with estimates of a variety of alternative models, which these days mostly belong in online appendices. The audience at paper presentations, referees, and editors will surely contribute to a list of alternative specifications that one should pay attention to.

Consider for example the systematic set of robustness checks in the analysis of consumption-savings by Laibson et al. (2017). The authors document (i) how the results change if one only uses a subset of the moments for estimation; (ii) how the estimates for time preferences are affected by the degree of risk aversion (if it is assumed instead of estimated); (iii) robustness to a variety of assumptions about the income process, about the rate of return of the various assets, and the degree of illiquidity of the housing asset; (iv) robustness to using different weighting schemes in the minimum-distance estimator (more on this below); (v) robustness to allowing for unobserved heterogeneity in the time preference parameters. Of these checks, the risk aversion parameter is the most influential: assuming low risk aversion weakens the evidence of present bias (that is, leads to present bias coefficients closer to 1). Other factors do not affect as much the key conclusion of significant present bias.

Importantly, this set of robustness checks highlights the key critical ones that one would like to see in structural papers. In the same orders as listed above, one would like to examine the robustness to: (i) *(moments)* using a different set of moments, or subset of evidence, to probe which part of the evidence is identifying what; (ii) *(key parameters)* alternative assumptions about a key parameter, or functional form; (iii) *(incidental parameters)* alternative assumptions about incidental parameters; (iv) *(statistical estimation)* the statistical estimation method; (v) *(random effects)* fixed versus random parameters. A similar list applies to most cases of structural estimation, see for example the lists of steps to estimation in Reiss and Wolak (2007). Which robustness checks are most informative will depend on the paper.

In our analysis of altruism versus social pressure on charitable giving (DellaVigna, List, and Malmendier, 2012), we go through a similar list: (i) we use a different set of moments; (ii) we assume alternative distributions of the altruism parameters; (iii) we allow for asymmetric costs of sorting; (iv) we use different weights in the minimum distance estimation, and (v) we allow for two-type heterogeneity in social pressure. The most informative set of robustness checks in this case was regarding (i), the set of *moments* used. As discussed in Section 2.5, we introduced in the design a set of door-to-door survey treatments where we advertised with a flyer the day before the

payment for the survey, ranging (across treatments) from unpaid to $10. These treatments were designed to identify the cost of sorting in and out of the home, which in turn we needed to identify the social preference parameters. So we ran the estimation of the experimental results excluding the survey treatments, expecting that without such treatments we would not be able to identify any of the key parameters. That was, after all, why we did these survey treatments in the first place! It turns out that we were wrong, as Online Appendix Tables 3 and 4 in (DellaVigna, List, and Malmendier, 2012) show. The estimate excluding the moments from the survey treatments actually provides a comparable estimate of the cost of sorting out (if less precise) and a social pressure cost that is also similar to the benchmark one. The sorting observed in the different charity treatments, and the amount given under the different treatments are enough, parametrically, to identify the sorting parameter. But we certainly have more trust in the estimated cost of sorting that uses the survey moments, and we find it very reassuring that the estimated sorting costs using just the survey moments and using only the giving moments are quite close. This turned out, ex post, to amount to a check of the stability of the estimated parameters, a form of over-identification test.

It is especially important to report alternative assumptions which have more impact on the results, in order to explore the limits of one's own results. An example we discussed above is in the reference-dependent job search model of DellaVigna et al. (2017). In the paper, we go through a similar list of robustness checks, and a rather exhaustive list of more standard models of job search, including habit formation models and models with more extensive heterogeneity in the cost of search. Almost all the alternatives we brought to the data did not match the fit of the reference-dependent model, despite typically having as many, or more, parameters (since we estimate the reference-dependent model with no heterogeneity). One form of heterogeneity which we examined for completeness is heterogeneity in the elasticity of the cost of search. This is not a form of heterogeneity examined in previous papers, as far as we can tell, nor was it requested by reviewers or an editor; but it is a possible variation which we considered as part of a large set of robustness checks. It turns out that this particular form of heterogeneity fits the data as well, and in fact better, than the reference-dependent model. At the same time, this form of heterogeneity did not make much sense to us, with implausible elasticities needed to fit the data. We asked ourselves as coauthors: what should we do? If we had not included this specification in the paper, likely no one would have asked for it. But it would not be right, as we would be de-emphasizing the best-performing alternative model to our preferred model, even if we did not personally find it a credible alternative explanation.

As I detailed in Section 2.4, we decided to present this model in detail in the paper, and we use two sets of out-of-sample predictions to compare this model to the reference-dependent model, and alternative standard models. We strongly believe that this was the right choice, even if it took a couple extra months of work for the team of 4 authors and 2 research assistants, and it took precious space in the paper to explain. Consider the benefits: we provided a fairer portrayal of the set of estimates, and we presented a novel model with unobservable heterogeneity that the literature can further explore in the future (in addition of course to our main contribution, the reference-dependent model). As an extra benefit, we thought harder about out-of-sample predictions of the model than we would have otherwise done, in what turned out to be an important addition to the paper.

Another example of robust test of the assumption for a model arises with Allcott and Taubinsky (2015). As discussed, Allcott and Taubinsky (2015) estimates the extent of limited attention with respect to the energy savings of a CFL light bulb. The estimated level of inattention justifies a subsidy of about \$3, but not a CFL ban, as implemented by some countries. Could alternative assumptions change this result? In this case, examining the robustness of the result is especially important given the focus on the policy implications (the level of the optimal subsidy for CFL light bulbs). In presence of direct policy implications, one ought to be extra careful. Allcott and Taubinsky (2015) considers a rather standard list of alternative assumptions for the WTP elicitation, but then they do something more unusual: they call into question a key feature of their data, and its impact for the welfare conclusion. In particular, Allcott and Taubinsky (2015) finds that close to half of consumers are near indifferent between the CFL versus the incandescent light bulb, that is, their WTP for one versus the other differs by no more than \$2. In the treatment condition, this is less true, but still about 35 percent of consumers are still in that range. Ex ante, one would not expect such a large mass: a CFL bulb saves around \$40 compared to an incandescent light bulb, so to have so many people that are near indifferent between the two types of bulbs implies that a large number of consumers place a disutility of around \$40 to the quality of light of the CFL (the non-pecuniary component of valuation).

A possible interpretation of this aspect of the findings is that, at least in part, consumers may be unsure and drawn to zero, the middle of the scale, by some form of bounded rationality. Allcott and Taubinsky (2015), in a key robustness check, takes this head on, and uses a clever computation of excess mass, based on the part of the distribution that is not near zero, to generate a counterfactual. The authors show that under this alternative, the subsidy would be larger, and a ban would be justified. It is unusual, and refreshing, to see a full discussion (over a page in the text) of a scenario that implicitly calls into question some of the validity of the WTP elicitation used in the paper. Structural papers, and reduced-form papers alike, benefit from such frank discussion of the key potential limitations of a paper.[15]

Examining the role of alternative assumptions in structural estimation is especially important as sometimes seemingly innocuous modeling choices can affect the estimates of key parameters. For example, Apesteguia and Ballester (2018) demonstrates that using random utility models, such as the logit and probit, poses identification problems for the estimation of, for example, risk and time preferences. These models can violate monotonicity under the commonly assumed CARA and CRRA utility functions: the probability of choosing a risky over a safe option is initially decreasing in the risk aversion, but at some level of risk aversion the noise component will begin to dominate, driving probability of selecting either choice to 50%. Thus, there are multiple levels of the risk aversion parameter that rationalize a given observed preference for the risky lottery. The authors show that random-parameter versions of the same models are instead immune to this identification issue. They use data from Andersen et al. (2008) to demonstrate how the estimated risk and time preference parameters differ under the two approaches.

[15]Such frank discussion would surely be more common if referees take a realistic view that all papers have limitations, and that it is best to have the authors discuss these limitations, as opposed to instead taking any such limitation to recommend rejection of a paper.

Calibration. As we just discussed, it is critical to present as much robustness as possible for key model assumptions. At the same time, a complementary approach to assessing the credibility of the results is by appealing to simple calibrations, just as we discussed in Section 2.1. Let's return for example to the estimate of switching costs in health insurance by Handel (2013). Handel (2013) shows that across a variety of alternative specifications the estimated switching cost is in the order of $2,000, clearly too much to be consistent with a standard model where the cost of effort captures the value of time spent. An alternative way to build credibility for the results is the motivating fact for the paper which I discussed in Section 2.1: one health plan dominates another plan for a subset of employees, with a minimum loss from picking the dominated plan of over $1,000; and yet, a majority of people stay with the dominated plan. It is very simple in this case to infer that the switching cost on average has to be at least as high as $1,000, thus providing a reduced-form validation for the results. Having this simple reduced-form counterfactual reassures the reader that the structural estimates are summarizing the same key features of the data that are apparent in the reduced-form analysis, as opposed to being driven by some hidden auxiliary assumption.

Another good example of this comes from the time preference experiments, whether on monetary payments (e.g., Andreoni and Sprenger, 2012) or on real-effort choices (e.g., Augenblick, Niederle, and Sprenger, 2015). In both cases, the estimate of the discounting comes transparently from the comparison of intertemporal choices as the rate of interest between the earlier and the later period is varied. The time discounting parameters can be calibrated off of simple figures of the results, such as those in Andreoni and Sprenger (2012) and in Augenblick, Niederle, and Sprenger (2015).

Assessing the Uncertainty of the Estimates. Above, we emphasized that it is important to examine how the point estimates are affected by assumptions about the model. A related point is about the precision of the estimates. How sure are we about the confidence interval of the structural estimates? This is especially important in cases in which the estimates are used to formulate policy and welfare implications.

A simple, but key, point to be made here is that structural estimates do *not* incorporate uncertainty about the model being wrong. That is, the confidence models reflect exclusively the uncertainty about the parameters, or error term, explicitly modeled in the paper. And yet, arguably the larger source of uncertainty is often about whether the model being estimated is the right one to capture the setting at hand.

One can often get a sense of this outside-the-model uncertainty by comparing structural estimates of a parameter in a paper under two different alternative (non-nested) sets of assumptions. It is quite common that under either of the assumptions the parameter is precisely estimated, but the estimates under the different assumptions are really quite different. This suggests some meta-confidence interval that in some sense takes into account also the uncertainty across the different sets of assumptions. Indeed, this issue is precisely a key reason to present estimates under a broad set of alternative assumptions, including presenting the robustness checks that affect the results the most, as I discussed earlier in this section.

The reference-dependence job search paper referred above (DellaVigna et al., 2017) provides an example of the case above. In our benchmark specification of reference dependence with adaptive expectations, we assume that the reference point is the average income in the previous N periods,

and we estimate a loss aversion parameter $\lambda = 4.54\,(s.e.\,0.25)$. This parameter is quite precisely estimated and, given that it uses the Köszegi and Rabin (2006) formulation, not far from the usual consensus value of 2.25.[16] In one of the robustness checks, we estimate the same model, but assuming adaptive expectations with an AR(1) process. The fit of the model is quite similar, but the resulting estimate of the loss aversion parameter is really quite different: $\lambda = 16.9\,(s.e.\,4.08)$. A relatively minor change in the assumptions leads to a large change in a behavioral parameter. This change demonstrates that the job search setting is not the ideal setting to estimate the level of loss aversion, since unemployed individuals are always on the loss side of the utility function. This is not to say that we cannot identify evidence of reference dependence; we believe that we do. But we would overreach if we sold heavily our point estimate of the loss aversion parameter. Indeed, the standard error in that main estimate is too small, and one gets a better sense of the precision of the estimate considering a number of alternative assumptions.

To stress once again the main point I am making here, standard errors are only about the precision of the inference *under the assumption* that a particular model is correct. It is then useful to consider how the point estimates vary as one varies the identifying assumptions.

4 Nuts and Bolts of Structural Behavioral Economics

In this section, I discuss some nuts and bolts of structural behavioral economics, aimed in particular at behavioral researchers interested at taking a step in that direction. I discuss first the choice of estimation method and the modeling of heterogeneity, the two building blocks of a structural model. I then highlight the distinction between key parameters and incidental parameters, discuss the sensitivity of the parameter estimates to the empirical findings, and a few other issues arising in structural estimation.

The coverage in this section is meant to be just a first introduction to the methodological issues and should be seen as a teaser for more comprehensive treatments. Some of the relevant references to dig deeper in the structural estimation literature are Reiss and Wolak (2007), Wolpin (2013), and Rust (2014), cited earlier, as well as Judd (1998) and Adda and Cooper (2003).

4.1 Estimation Method

4.1.1 Minimum Distance.

A transparent choice is a minimum distance estimator: one identifies some moments in the data and then finds the set of model parameters that minimizes the distance between the empirical moments and the theory-predicted moments. We discuss here together the case of classical minimum distance and the case of simulated minimum distance, a distinction we return to later.

[16] The Köszegi and Rabin (2006) formulation allows for consumption utility and gain loss utility, that is, $v\,(c|r) = u\,(c) + \eta\,[u\,(c) - u\,(r)]$if $c > r$ and$v\,(c|r) = u\,(c) + \eta\lambda\,[u\,(c) - u\,(r)]$ if $c < r$. The original Kahneman and Tversky (1979) prospect theory formulation does not have consumption utility and is just $v\,(c|r) = u\,(c) - u\,(r)$if $c > r$ and$v\,(c|r) = \lambda\,[u\,(c) - u\,(r)]$ if $c < r$. A loss aversion of 2.25 as estimated in Tversky and Kahneman (1992) in the traditional prospect theory formulation translates into a loss aversion of about 3.25 in the Köszegi and Rabin (2006) formulation, assuming a consumption utility weight $\eta = 1$.

Consumption-Savings Example. To make things concrete, consider the case of one of the earliest papers in the Structural Behavioral Economics literature, Laibson, Repetto, and Tobacman (2007), now Laibson et al. (2017). This paper documents two sets of facts: (i) individuals borrow substantially on credit cards; and yet, (ii) when they come close to retirement they have substantial wealth (including housing and 401(k)s and IRAs). To be more precise, the first moment is the share of 21-30 years olds with a credit card: $\hat{m}_1 = 0.81$ ($s.e.$ 0.02). The second moment is the share of annual income borrowed on a credit card: $\hat{m}_2 = 0.20$ ($s.e.$ 0.02). The third moment is the wealth held by 51-60 year olds, measured in units of annual income: $\hat{m}_3 = 5.34$ ($s.e.$ 0.22). Thus, the large majority of households borrow on high-interest credit cards, and yet accumulate wealth by retirement. Which model parameters can explain these three moments?[17]

To answer this question, the next step is to write down a model that generates as output, given a set of parameter values θ, predictions for the three moments $m_1(\theta), m_2(\theta), m_3(\theta)$. In the case at hand, consumers solve a lifetime consumption problem, including how much to consume and save, and whether to borrow on a credit card. The solution depends on a set of parameters, which is the combination of the present-bias β, the long-term discounting δ and the risk aversion ρ; thus, $\theta = (\beta, \delta, \rho)$. For any given value of the parameters, there is a prediction for the three moments.[18] Given the computational complexity of the dynamic programming problem, the solution for $m(\theta)$ is obtained by simulating a large number of consumers with a given set of parameters θ, and averaging over the implied moments. (The result of the simulations differs for each consumer because of different draws, for example of the income process.) Thus, this is an application of simulated minimum distance in that the moments are simulated.

To take a relevant case, consider an exponential discounter with $\theta_0 = (\beta = 1, \delta = 0.89, \rho = 1)$. For this impatient, though time-consistent, individual, the implied set of moments is $m_1(\theta_0) = 0.70$, $m_2(\theta_0) = 0.20$, $m_3(\theta_0) = -0.03$. Thus, this combination of parameter fits well the first two moments, with high credit card borrowing, but predicts slightly negative wealth at retirement, in strong contrast to the substantial (and precisely estimated) accumulation of assets. Consider instead the alternative set of parameters with present bias $\theta_1 = (\beta = 0.50, \delta = 0.99, \rho = 1.25)$; in this case, the implied set of moments is $m_1(\theta_1) = 0.60$, $m_2(\theta_1) = 0.23$, $m_3(\theta_1) = 5.02$. At this set of parameters, one can approximately fit all three moments.

The estimation strategy is to simply take the three moments above, and find the model parameters that fits those best. More formally, a classical minimum distance estimator solves the following problem:

$$min_\theta \left(m(\theta) - \hat{m}\right)' W \left(m(\theta) - \hat{m}\right), \tag{17}$$

were $\hat{m} = (\hat{m}_1, \hat{m}_2, \hat{m}_3)$ is just the vector of the empirical moments chosen for estimation, $m(\theta)$ indicates the model implied moments, and W is a weighting matrix. The simplest case is one in

[17] For sake of simplicity, I am simplifying here, since in the paper the authors have as separate moments the three moments above for each of 4 age groups, thus 3*4=12 moments, not just 3 moments.

[18] In addition to these key parameters, there is an additional set of parameters $\widetilde{\theta}$, like the distribution of the income shocks, which are not estimated but calibrated based on additional data, or prior results in the literature. Having this moments calibrated, or estimated separately, simplifies the estimation process for the main set of parameters.

which the weighting matrix is the identity matrix, $W = I$, in which case (17) becomes

$$min_\theta \sum_i \left(m_i\left(\theta\right) - \hat{m}_i\right)^2 = \left(m_1\left(\theta\right) - \hat{m}_1\right)^2 + \left(m_2\left(\theta\right) - \hat{m}_2\right)^2 + \left(m_3\left(\theta\right) - \hat{m}_3\right)^2, \tag{18}$$

that is, simply the sum of the squared distance between the empirical moments and the model predictions. So for example for the set of exponential parameters θ_0, that would equal $\left(0.70 - 0.81\right)^2 + \left(0.20 - 0.20\right)^2 + \left(-0.03 - 5.34\right)^2$ =28.85. In comparison, for the alternative set of parameters with present bias θ_1, that would equal $\left(0.60 - 0.81\right)^2 + \left(0.23 - 0.20\right)^2 + \left(5.02 - 5.34\right)^2 = 0.15$. Clearly, the second set of parameters, θ_1, does much better. This example also illustrates a first motivation for a weighting matrix: the moments may not even be in the same scale, as is the case here (share of people for the first moment versus share of annual income for the other moments). There is actually an optimal set of weights, given by the inverse of the variance-covariance matrix of the moments. Since using the full matrix can lead to instability in the estimates (consider all those off-the diagonal terms), a common choice is the inverse of the diagonal of the variance-covariance matrix. For this simple case, the minimum-distance case in (18) becomes

$$min_\theta \sum_i \frac{1}{\sigma_i^2}\left(m_i\left(\theta\right) - \hat{m}_i\right)^2,$$

that is, each moment is weighted by its precision (the inverse of the empirical variance). intuition is the same as for weighted least squares.

Returning to the problem in general, the minimization takes place over the parameters θ. This often involves setting up a grid for values of the parameters, simulating the model for each of the parameters, storing the values of the objective function, and comparing them, exploring the space of parameters in directions that (locally) improve the objective function. In this case, it turns out that the solution is the set of parameters θ_1 highlighted above (with present bias).

The classical minimum-distance also provides standard errors for the parameters, given by the square root of the diagonal of the estimated variance

$$\frac{\left(\hat{G}'W\hat{G}\right)^{-1}\left(\hat{G}'W\hat{\Sigma}W\hat{G}\right)\left(\hat{G}'W\hat{G}\right)^{-1}}{N}, \tag{19}$$

where $\hat{G} \equiv N^{-1}\sum_{i=1}^{N}\nabla_\theta m_i\left(\hat{\theta}\right)$ and $\hat{\Sigma} \equiv Var[\hat{m}_i]$. Expression (19) simplifies for the case of the optimal weighting matrix, in which $W = \hat{\Sigma}^{-1}$. In this case, it simplifies to

$$\frac{\left(\hat{G}'\hat{\Sigma}^{-1}\hat{G}\right)^{-1}}{N}.$$

To gain intuition on this expression, consider a very simple case with just two moments and two parameters, with each moment depending on only one parameter. That is, assume $m_1(\theta_1, \theta_2) = f_1(\theta_1)$ and $m_2(\theta_1, \theta_2) = f_2(\theta_2)$, where the functions are differentiable. Further assume that the moments are uncorrelated, such that the variance-covariance matrix is diagonal. In that case, the

matrix G equals $G = \begin{pmatrix} \frac{df_1}{d\theta_1} & 0 \\ 0 & \frac{df_2}{d\theta_2} \end{pmatrix}$, and the standard error for the estimated $\hat{\theta}_i$ is $\sqrt{\sigma_i^2 / \left(\frac{df_i}{d\theta_i}\right)^2}$. This expression is intuitive[19]: the structural parameter $\hat{\theta}_i$ is more precisely estimated the smaller is the variance σ_i^2 in the moment that identifies it, and the larger is the responsiveness of the moment to the structural parameter. Intuitively, if the moment is very unresponsive to the value of the parameter at the best estimate, the parameter will be imprecisely estimated.

In general, of course, a moment will depend on multiple parameters, and thus the precision of the estimate of a parameter will depend on the precision of several moments.

Charitable Giving Example. Another case of application of minimum-distance is the field experiment by DellaVigna, List, and Malmendier (2012) on altruism and social pressure in charitable giving. In this setting, the set of moments $\widehat{m}$ is simply the set of experimental findings in each of the treatments. For each charity treatment T, we observe three key moments, the probability of answering the door, $P(H)_T$, the probability of giving, $P(G)_T$, and the amount given, which we code as the probability of giving in a certain dollar range, e.g., $P(0 < G < 10)_T$. The survey treatments also provide two moments each, the probability of answering the door, $P(H)_T$, and the probability of completing the survey, $P(SV)_T$. These moments, together with moments on probability of opting out $P(OO)_T$ (for the opt-out treatments) provide the inputs to the structural estimation, for a total of 70 moments, all in the form of probabilities. The minimum distance estimation thus transparently takes as inputs the experimental results in the treatment and control groups.

The model-based moments $m(\theta)$ are a function of a set of 15 parameters θ. The first six are the behavioral parameters of interest, indicating the social preference: the social pressure cost of saying no in person, S, and the mean and standard deviation, μ_a and σ_a, of the altruism weight a, separately for each of the two charities involved in the fund-raising experiment. In addition, the model includes also 9 incidental parameters. These parameters are not of interest *per se*, but the estimation of the model depends on the value of these parameters. For example, a critical component in the paper is the cost of sorting in and out of the home, since a set of outcome variables is the probability of answering the door $P(H)_T$, and how that responds to various forms of motivation. We specify the cost of sorting as $c(h) = (h - h_0)^2 / 2\eta$, where h is the (optimally chosen) probability of staying at home (that is, answering the door). The assumption is that the further the household goes from the baseline probability of home presence h_0, the more costly the adjustment is; further, η indicates the (inverse of) how costly such adjustment is. The two parameters to be identified are h_0 and η. The identification of the first parameter h_0 is straightforward, as it depends on the share answering the door in the control (no flyer) group. The identification of η, instead, depends on a variety of moments, but especially on the survey treatments. We return to the discussion of identification and sensitivity below.

For each value of the 15 parameters θ, we solve analytically or by numerical approximation the various implied probabilities of the 70 moments, $m(\theta)$, and we search for the values of the parameters that minimized the distance as in (17), just as in Laibson et al. (2017).

Pros and Cons. Minimum distance has the advantage of transparency: the estimation fits a

[19] If we can not write an analytical solution, a numerical derivation of G is not complicated for implementation as numerical differentiation.

set of moments that the author picks, and the moments are (in good practice) clearly presented for the readers to inspect. As the charity field experiment above illustrates, the choice of the set of moments is often transparent in an experiment, where it is the value of a key dependent variable in the various treatment groups.

In the era of administrative data sets which typically cannot be publicly posted, another advantage of minimum distance estimation is that it makes it possible to post the moments, even when the underlying data must remain confidential. This allows researchers to replicate the structural findings, provided one takes the moments as given. This is the case for example of the paper on unemployment benefit reform in Hungary (DellaVigna et al., 2017): the underlying data is confidential, but the moments are posted together with the estimation code.

The choice of the moments, which provides for transparency of the analysis, however also generates a drawback. Unlike in the case of maximum likelihood, minimum distance does not use all of the information in the data, by focusing on just some moments. For example, in the consumption setting one could wonder, what if one used different information in the consumption data? Or for the charity paper, what if one used a different breakdown of the probability of giving in a particular dollar range?

4.1.2 Maximum Likelihood

A second common method is maximum likelihood: one finds the model parameters that maximize the statistical likelihood, given the data.

Consider a model of behavior that predicts, for a given vector of parameters θ, a likelihood that one would observe the realization x, that is, $L(x|\theta)$. For example, take a very simple real-effort task with only one treatment, with an (observed) piece-rate incentive for effort p and (unobserved) cost of effort $e^2/2\varphi$, where φ is a productivity parameter. Straightforwardly, the first order condition leads to optimal effort $e^* = \varphi p$. Assume that φ has a log-normal distribution, $\varphi \sim \exp(X)$, where $X \sim N\left(\mu, \sigma^2\right)$. In this case, the set of unknown parameters are the mean and variance underlying the cost of effort distribution, that is, $\theta = \left(\mu, \sigma^2\right)$. Then x is simply the data, the vector of effort choices by each individual i in the experiment: $x = (e_1, e_2, ..., e_N)$.

Maximum likelihood is based on asking: given the observed data realization for x, what is the set of parameters θ that would have generated the data with the highest probability? That is, maximum likelihood solves

$$max_\theta L(x|\theta).$$

Take for example the real-effort case. Suppose that the vast majority of subjects exert little effort, despite incentives, with little variability across subjects. Then it is unlikely that the data x is generated by set of parameters with very high mean productivity μ; rather, the most likely parameters will have low μ and low σ^2, given the limited heterogeneity. Conversely, if the data reveals a wide heterogeneity, we would infer a large σ^2.

Real Effort and Time Preference Example. Consider more in detail a paper that uses maximum likelihood for the estimation of the discounting parameters using the real-effort paradigm mentioned (Augenblick and Rabin, forthcoming).

Present-biased subjects choose at time k how many tasks to complete at time t, e_t, for a piece rate w paid at time T.[20] Utility is assumed to be linear in the monetary payments with cost of effort function $C(e) = \frac{1}{\varphi \cdot \gamma}(e+10)^{\gamma}$, where the 10 is from the required 10 tasks that must be completed. This power cost of effort function has, as discussed above, a constant elasticity property. Optimal effort is then given by

$$e^* = argmax_e \delta^{T-k} \cdot (e \cdot w) - \frac{1}{\beta^{1(k=t)}} \cdot \delta^{t-k} \frac{1}{\varphi \cdot \gamma} (e+10)^{\gamma},$$

where $1(k=t)$ is an indicator that the decision for effort is occurring in the same period. In this case, costs are immediate and benefits delayed, and thus the future benefits are discounted by β; notice that to obtain the above expression we divided the expression by $\beta^{1(k=t)}$. This implies the solution (when interior)

$$e^* = \left(\frac{\delta^{T-k} \cdot \varphi \cdot w}{\frac{1}{\beta^{1(k=t)}} \cdot \delta^{t-k}} \right)^{\frac{1}{\gamma - 1}} - 10.$$

Notice that up to this point, there is no source of heterogeneity in the model. That is, for a given set of parameters $(\delta, \beta, \varphi, \gamma)$, the model predicts a deterministic level of effort e^*. To match the heterogeneous behavior in the data, the authors assume an implementation error: the observed effort levels are normally distributed around the predicted effort, with standard deviation, σ. Then the likelihood of any observed effort e_j is given by

$$L(e_j) = \phi \left(\frac{e_j^* - e_j}{\sigma} \right)$$

where ϕ is the pdf of a standardized normal distribution. This, together with handling the censoring (for the cases in which effort is at the boundaries of the choice set), closes the model. The authors then maximize the likelihood with respect to 5 parameters, $(\delta, \beta, \varphi, \gamma, \sigma)$, using the experimental induced variation in timing t of the effort choice and wage rate w.

Pros and Cons. An advantage of maximum likelihood is that it uses all the information in the data, given the modeling assumptions. As such, it is the most efficient and it can lead to more precise estimates of parameters compared to minimum distance, in which only some moments are used for estimation. Intuitively, the MLE chooses the optimal moments for efficient estimation. Furthermore, unlike in the Minimum Distance case, there is no choice of the weighting matrix since the likelihood function already incorporates the optimal weighting.

At the same time, this advantage—that maximum likelihood uses all the information in the data and all the structure in the model—can turn into a disadvantage too, making the Maximum Likelihood results less robust. A canonical issue with maximum likelihood is that observations with a low ex-ante likelihood according to the model can be influential outliers. These "rare" observations are very informative according to the model and thus, when they occur, can lead to a large shift in the estimates. Whether this is a feature or a bug depends on how much faith one has in the functional forms that lead to the prediction of low likelihood in that region. Often, one may not

[20] The paper focuses on the estimation of the naiveté parameter $\hat{\beta}$, which I neglect here for simplicity.

want the estimates to rely on these observations in the same way that we worry about the impact of outliers in OLS regressions. I consider three examples of this for (i) real effort experiments, (ii) charitable giving, and (iii) left-digit inattention.

Consider first real-effort experiments in which subjects have to do an effortful task, such as moving sliders, and the experimental outcome is the number of units of effort e (e.g., Gill and Prowse, 2012, DellaVigna and Pope, 2018). As we discussed in Sections 2.2, a natural assumption for these real-effort tasks is to posit a power cost of effort function $c(e) = \left(e^{1+\gamma}\right)/(1+\gamma)$, since this parametric form is associated with a constant elasticity $1/\gamma$ of effort to motivation. This assumption implies equation (10), which relates $log(e)$ to motivation. Consider now the case, for example, of DellaVigna and Pope (2018). In this paper, we estimate the impact of different behavioral interventions, such as gift exchange and peer information, on effort using a typing task that yields on average effort of around 1,800 units (a-b presses per 10 minutes). In this experiment, while the vast majority of subjects exert effort between 500 and 2,500 units, a small number does just a couple presses and then stops. Under a power function, the observations with, say, 1 or 2 presses are extremely informative because the dependent variable is $log(e)$. In this case, moving from 1 to 3 presses has the same impact on the likelihood function as moving from 1,000 to 3,000 presses. Thus, a treatment which happens to have 5 or 6 more subjects who put in just a couple presses would be estimated to induce lower motivation, even if for hundreds of subjects the impact of the treatment is to, say, increase effort from 1,800 to 2,200 (a 0.5 s.d. shift). In this case, rather than taking all predictions of the model literally, which is what maximum likelihood does, it seems sensible to focus on key statistics that are less sensitive to such outliers, such as average effort in a treatment (which is our first strategy in DellaVigna and Pope (2018)), or estimate using effort binned in groups, say 0-100, 101-200, etc (which is our other strategy in DellaVigna and Pope (2018)).[21]

A second example concerns estimates of the altruism or warm glow functions from charitable giving contributions (e.g., DellaVigna, List, and Malmendier (2012)). A single, very large contribution in one of the treatment groups can shift the results on the estimated generosity. In DellaVigna, List, and Malmendier (2012), we deal with this issue by generating as moments for our minimum-distance estimator the share of donations in binned amounts.

A third example is from the work of Shlain (2018) on left-digit bias and pricing response. Assume that consumers have left-digit bias, that is, they perceive a price of \$5.00 as more than a cent larger than a price of \$4.99. Then, as Shlain (2018) shows, for sufficiently high left-digit bias, optimizing retailers will not only never set a price of \$5.00, but in fact should also never set a price of \$5.03 or \$5.12; there should be a gap in the price distribution and prices should restart at, say, \$5.30. Shlain (2018) measures directly the consumer inattention from the consumer response to different prices and then estimates what degree of consumer inattention the retailers appear to perceive, given the price endings of products in stores. The difficulty, and here I come back to the point, is that in this model, even a single price of \$5.00 is an influential outlier, since it should never occur for any positive degree of limited left-digit bias. If one were to use maximum likelihood for estimation, it would be important to explicitly model some measurement error in the data. Alternatively, a

[21] There are a number of alternative strategies including using maximum likelihood but explicitly modeling an added shock which can lead, for example, to very low effort.

minimum-distance estimation using bins of prices can more easily address the issue.

4.1.3 Non-Linear Least Squares

As Table 1 makes clear, minimum distance and maximum likelihood are the two most common options in the literature. A third case that appears in the literature is non-linear least squares.

Real Effort Example. Reconsider the case of real effort experiments discussed in Section 2.2. Assume that individuals have an exponential real effort cost function $c(e) = (exp(\gamma e)/\gamma) * \eta$ as in (7), with a log-normal distribution for the error term η. Furthermore, assume that the return to effort equals $(s+p)e$, where p is a piece rate for each unit of effort, and s is a measure of intrinsic motivation, perhaps capturing a (per-unit-of-effort) warm glow in doing a task. Then, following the derivation in (9), we can write the optimal effort as

$$e_i = \frac{1}{\gamma} log[s+p] - k + \epsilon_i. \tag{20}$$

Notice that expression (20) could be estimated by OLS if it were not for the fact that the parameter s appears inside a non-linear term together with the incentive term p, which varies by treatment. Expression (20) can be estimated by non-linear least squares, recovering the parameters γ, s, k, σ, where σ^2 is the variance of the error term ε.

Indeed, the advantage of setting up the error term as a log-normal distribution multiplying the cost of effort function is to obtain a simple estimating equation. It does not hurt that the assumption of a multiplicative, positive error term is plausible, with heterogeneity across people, and across times, in the marginal cost of effort. I also pointed out above that a parallel equation to (20) but with log effort as dependent variable obtains if the cost of effort function has a power form.

I use this non-linear least squares estimation with coauthors in work on gift exchange (DellaVigna et al., 2017) and in the study of different motivators of effort (DellaVigna and Pope, 2018). Non-linear least squares is also used in the estimate of limited attention to the odometer and left-digit bias (Lacetera, Pope, and Sydnor, 2012), as well as in the estimate of time preferences in Andreoni and Sprenger (2012).

Pros and Cons. The advantage of non-linear least squares is that it is very straightforward. However, it is an approach that applies only in some cases where the model lends itself to create predictions of that functional form.

4.2 Modeling of Heterogeneity

In the previous section, I discussed a couple leading approaches to statistical estimation. Across any of these approaches, a researcher needs to make a critical modeling decision: how to model the heterogeneity in the data. To give a stark example, return to the simple case above of a real-effort task with piece rate incentive p, with solution for optimal effort $e^* = \varphi p$. In the simplest model, all individuals have the same productivity φ, and thus all would display the same effort e^* for a given incentive p. And yet, the data always confronts us with heterogeneity in behavior. The econometrician, or modeler, thus has to decide how to model the heterogeneity.

There are three main approaches, which I briefly describe. A first common approach is to assume **random utility**: an unobserved (to the researcher) utility shock rationalizes the heterogeneity. A second approach is to assume **random parameters** or a **mixture model**, that is, heterogeneity of some key structural parameter(s). A third approach is to assume that the individuals make **implementation errors**. The approaches are not exclusive, as one could model heterogeneity along multiple dimensions.

4.2.1 Random Utility

Consider the case of an individual considering alternative choices for house deductibles as in Barseghyan et al. (2013). Assume that deductible A is associated with expected utility (or prospect-theory utility) Eu_A while deductible B is associated with expected utility (or prospect-theory utility) Eu_B. Under random utility, the assumption is that there is an extra additive component of the utility function, ϵ, that is unobservable to the econometrician, thus the name random utility. Under random utility, individuals thus choose A over B if

$$Eu_A + \epsilon_A \geq Eu_B + \epsilon_B.$$

As McFadden (1974, 1999) show, under the assumption that the error term has a type 1 extreme value distribution with scale parameter σ, the probability of choice of A can be written very conveniently in a logit form as

$$P(A) = \frac{exp(Eu_A/\sigma)}{exp(Eu_A/\sigma) + exp(Eu_B/\sigma)}.$$

The random utility model underlies static models of discrete choice, including models of insurance choices as in Barseghyan et al. (2013) and models of risk and time preferences identified using multiple price lists (e.g., Andersen et al., 2008). It also underlies models of dynamic discrete choices, such as the tax filing case (Martinez, Meier, and Sprenger, 2017). See Train (2009) for a good introduction to discrete choice models.

4.2.2 Random Parameters

In the random parameters case, the researcher assumes that a structural parameter is heterogeneous and has a distribution. An example of this approach which we already discussed is the paper on altruism and social pressure (DellaVigna, List, and Malmendier, 2012). In that paper, we assume that (for simplicity) there is a homogeneous social pressure cost S, but we allow for heterogeneity in the altruism parameter a, assuming that it has a normal distribution. (We also consider the case in which it has a negative exponential distribution for robustness). The presence of a random parameter generates heterogeneity of predictions in the data. Indeed, in this paper we assume no other source of heterogeneity, that is, there is no error term, and all the heterogeneity loads on the altruism term.

Similarly, in the work on social image in voting (DellaVigna et al., 2017), we assume heterogeneity in the social image value of telling others that one voted, s_V, and heterogeneity in the social image value of telling others that one did not vote, s_N. In another example, Laibson et al. (2017) does the

main analysis of the consumption-savings problem assuming homogeneous time preference parameters, but later allow for random parameters, assuming unobserved heterogeneity in the present-bias parameter β.

Notice that the case of random parameters amounts to unobserved heterogeneity in a parameter. One can of course also allow for observable heterogeneity, that is, for a parameter to be different as a function of some observables. For example, Barseghyan et al. (2013), in their estimate of insurance choice, does the main analysis in a random utility logit framework assuming no heterogeneity in the key parameters, but later relax it to assume both observed and unobserved heterogeneity in the key object of interest, the probability weighting function.

A related approach to assuming a distribution of parameters is the approach in *mixture models*, which posits a discrete number of types, and then classifies the population into types. For example, Costa-Gomes, Crawford, and Broseta (2001) allows for a mixture of cognitive types and classifies the players using information from observing behavior of the same experimental subjects across 18 games. Observing a given subject making a number of decisions, as is the case in Costa-Gomes, Crawford, and Broseta (2001), enables the classification into types. In another example, Fisman, Kariv, and Markovitz (2007) estimates social preferences from a sequence of 50 choices, each with different slopes in a budget line for giving. With this detailed data set, Fisman, Kariv, and Markovitz (2007) also presents individual-level estimates. This type of individual-level zeroing-in is less common in most field settings, though it may become more common in this era of Big Data.

4.2.3 Implementation Errors

Unlike in the random parameter case, in the implementation error case the assumption is that the agents implement an optimal choice with some error. Returning to the simple real-effort case above, with solution $e^* = \varphi p$, we could assume that this solution is implemented with error, leading to a specification $e^* = \varphi p + \epsilon$, where ϵ has some distribution, e.g., normal. Notice the difference from the random parameter case: there may not be any optimization problem that implies the solution $e^* = \varphi p + \epsilon$, as the term ϵ is tacked on to the optimal choice.

An example that we discussed above of implementation errors is Augenblick and Rabin (forthcoming): the assumption of a normally distributed implementation error allows the authors to set up the likelihood and is the only source of assumed heterogeneity in the data.

4.3 Key parameters and Incidental Parameters

A paper with parameter estimation represents a natural next step compared to model-based papers which derive comparative statics, or prediction of the model, and test it. A key difference is that estimation of the model requires a full specification not just of the key part of interest of the model, the key parameters, but also of the incidental parameters. We highlight this distinction for the papers in Table 1.

Consider the case of consumption-savings in Laibson et al. (2017). The paper focuses on the time preference and risk aversion parameters, which are the key parameters. And yet, it is not possible to identify those without pinning down a long list of incidental parameters, like the noise in the

income process and bequest motives, since these parameters and functions determine the optimal consumption path. In Laibson et al. (2017), the estimation process proceeds in two stages: in a first stage the incidental parameters are estimated or calibrated using additional data sources. Then, in a second stage, the key parameters are estimated. In this case, it is important in the second stage to take into account the error in estimation of the incidental parameters when deriving the confidence intervals of the parameters at hand.

In general, the two groups of parameters will be estimated jointly, as it typically will not be possible to separate out the two sets of parameters. For example, in our study of charitable giving (DellaVigna, List, and Malmendier, 2012), while we focus on the social preference parameters, we need to pin down a number of incidental parameters, like the cost of avoidance. The incidental parameters take all kinds of forms depending on the setting. In the study of risk preferences (and time preferences), an important incidental parameter is the estimated background consumption, which goes to determine the curvature of the utility function. This parameter is sometimes estimated (Andreoni and Sprenger, 2012) and other times is simply assumed, or calibrated. In all of the experiments with real effort tasks, the parameters related to the cost of effort function are critical incidental parameters (e.g., Augenblick, Niederle, and Sprenger, 2015; Augenblick and Rabin, forthcoming; DellaVigna et al., 2017; DellaVigna and Pope, 2018. In the study of insurance choice, the claim probability by observables plays a similarly important role (Barseghyan et al., 2013).

An important special case is one in which one can derive sufficient statistics for the key parameters. In this case, effectively, one can cancel out the incidental parameters by virtue of a particular setting, or by design. Return for example to the case of left-digit bias and the sale prices of used cars as a function of odometer pricing (Lacetera, Pope, and Sydnor, 2012). Remember that the idea of the paper is that the perceived value of a car $\hat{V}$ is a linear function of the perceived mileage $\hat{M}$: $\hat{V} = K - \alpha\hat{M}$, and that the mileage is perceived with the left-digit bias described previously. The model predicts that at each 10k mileage increase, the perceived value $\hat{V}$ will jump down discretely by $-\alpha\theta 10,000$: the decrease is increasing in the inattention θ, the key parameter, and in the depreciation parameter α, the incidental parameter. If that were all that we observe, we would need to identify the incidental parameter α in order to get the key parameter θ. But the same model shows that for interior mileage levels, the valuation of a car will decrease for each mile driven by $-\alpha\left(1-\theta\right)$, where inattention θ is attenuating the slope. Taking the ratio of the jump to the slope gets rid of the incidental parameter α and leaves an expression that is only a function of the parameter of interest, θ. Thus, in this case it is not necessary to identify the incidental parameter (though it turns out that it is easy to do so) in order to get to the parameter of interest.

The case of sufficient statistics like this is convenient and important, in that one does not need to worry about the exact value of the incidental parameters, and estimating them correctly, and can focus on the parameters of interest.

4.4 Identification and Sensitivity

An important question, and maybe *the* key question for structural papers, is: what identifies the estimates of the key parameters? For example, what identifies present bias in the consumption data

of Laibson et al. (2017) and in the real effort choice of Augenblick, Niederle, and Sprenger (2015)? What identifies reference dependence in job search of DellaVigna et al. (2017) and in insurance choice of Barseghyan et al. (2013)? What pins down the evidence of limited attention in energy choice of Allcott and Taubinsky (2015) and with respect to taxes of Chetty, Looney, and Kroft (2009)?

Sufficient Statistic. We discussed some of the answers above. In some of the papers, the identification relies on a sufficient statistic, typically based on simple pricing-out variation. This is the case, for example, for the identification of limited attention in all the papers mentioned above. The identification in some of these papers relies on an intervention that calls attention to the shrouded feature. Then one can compare the valuation of a good with limited attention, versus when attention is drawn to the particular feature, such as the tax (Chetty, Looney, and Kroft, 2009), or the energy savings (Allcott and Taubinsky, 2015). In other cases the identification is based on experimental variation of the shrouded attribute, like the tax, inferring back the inattention from the willingness to pay (Taubinsky and Rees-Jones, forthcoming). The identification of time preferences in experiments such as Andreoni and Sprenger (2012) and Augenblick, Niederle, and Sprenger (2015) is also essentially based on sufficient statistics, comparing the impact of delay to the impact of interest rate variation.

Calibration. A related case is one in which the identification of the structural model is not as simple as that above, but one can check the identification with the help of a calibration. This is the case for example for the estimate of switching cost in health insurance in Handel (2013): it is not obvious to show why the estimated switching cost is around $2,000, but the author can point to an example of a combination of plans in which the individuals in one plan lose at least $1,000, and more employees do not switch; this calibration indicates the right order of magnitude for the structural estimate. Another case is for the identification of deductible choice in Barseghyan et al. (2013). While the estimate of the slope of the probability weighting function relies on the full structure, the descriptive evidence from Sydnor (2010) and from Barseghyan et al. (2013) clearly suggests that individuals are willing to pay much higher premia, in order to avoid the chance of having to pay a higher deductible than predicted by the standard model. In particular, Sydnor (2010) presents a calibration that suggests that a probability weighting function that doubles or triples small probabilities would go a long way to explain the observed behavior. This simplified calibration lends further credibility to the structural estimates.

Sensitivity. In other cases, the complexity of the problem at hand makes it infeasible to develop summary statistics, or a simple calibration. For example, an optimal consumption-savings problem as in Laibson et al. (2017) and an optimal job search choice with heterogeneous types as in Paserman (2008) and DellaVigna et al. (2017) does not fit into either of the cases above. In these cases, a comprehensive set of robustness checks helps highlight what identifies the model, by highlighting the difference in the estimates as one removes one piece of the estimate. A particularly useful robustness exercise is examining what happens to the estimate if one uses only a subset of the moments, or only part of the data. For example, in Laibson et al. (2017), the estimates provide more limited evidence of present bias ($\beta = 0.88$ versus $\beta = 0.50$ in the benchmark specification) if the moments include only the borrowing on credit card, and do not include the (significant) wealth accumulation. In this

case, the credit card borrowing can also be accommodated by high exponential impatience; indeed, in this case the estimated exponential discount factor is $\delta = 0.94$ versus $\delta = 0.99$ in the benchmark case.

Andrews, Gentzkow, and Shapiro (2017) proposes a formal way to undertake the above exercise, highlighting which features of the data are driving each parameter estimate. As such, the authors provide more formal evidence of an exercise like the one detailed above. While the above discussion refers to those questions loosely as "identification", one ought to distinguish between the formal definition of *identification* (the existence of a unique value of parameters that solve the optimization problem) versus what they call *sensitivity* of the parameters. Sensitivity is how alternative model assumptions, and/or alternative moments, will change the parameter estimates. The authors propose that researchers can present a sensitivity matrix, Λ, in their papers to allow readers to conduct alternative hypotheses. The sensitivity Λ describes how the estimated parameters will change in response to local perturbations in the model. Formally,

$$\Lambda = \left(G'WG\right)^{-1} G'W$$

where G is the Jacobian of the moment conditions and W is the weighting matrix (both as in equation 19). Hence, the sensitivity matrix is built from components that are already calculated in many cases. Andrews, Gentzkow, and Shapiro (2017) shows that then Λ can be used, asymptotically, to translate the alternative assumptions about the model (or moments) into the bias in parameters

$$E\left(\tilde{\theta}(a)\right) = \Lambda E\left(\tilde{g}(a)\right)$$

where $\tilde{\theta}(a)$ is the vector of parameters under the alternative a and $\tilde{g}(a)$ is the vector of the moment statistics (e.g., in the minimum distance case, the difference between model-based moments and the moments in the data) under the alternative a.

To illustrate, we reprise the simple example from above (see 4.1.1) in which we have two moments and two parameters, with each moment depending on only one parameter. In this very simple case, the sensitivity matrix, given that we already found $G = \begin{pmatrix} \frac{df_1}{d\theta_1} & 0 \\ 0 & \frac{df_2}{d\theta_2} \end{pmatrix}$ and $W = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$, is

$$\Lambda = \begin{pmatrix} \frac{1}{\frac{df_1}{d\theta_1}} & 0 \\ 0 & \frac{1}{\frac{df_2}{d\theta_2}} \end{pmatrix}$$

In this case, a change in moment i affects only the estimate for parameter i and not for parameter j, as we would expect. Further, a change of 1 in moment i affects the estimate of parameter i by $\left(df_i/d\theta_i\right)^{-1}$: the more responsive locally the moment is to a change in the value of the parameter, the less sensitive the parameter estimate is to a change in the moment. Recall that Λ is a $k \times n$ matrix, where k is the number of parameters and n is the number of moments.

Sensitivity, Charity Paper. To illustrate this sensitivity measure, we return to the charity paper (DellaVigna, List, and Malmendier, 2012), which is one of three applications of the methodology

discussed by Andrews, Gentzkow, and Shapiro (2017). Using the sensitivity methodology, one can ask how much the estimates for the social preference parameters and for the incidental parameters would change if the moments were to change. To start from a simple case, consider the sensitivity of the incidental parameter h_0, which is the baseline probability of answering the door. We expect this parameter to be mostly identified by the share answering the door in the control group (the one that does not receive a flyer), since the probability of being at home in the control group equals $m^{Home}_{Control}(\theta) = h_0$. A slight complication is that the value of h_0 also affects the share answering the door in the flyer treatments. As the first vertical panel of Figure 2 shows, the estimated Λ indicates that the estimate for h_0 is mostly sensitive to the share answering the door in the control group, in the direction that one would expect. For example, the second row indicates that an increase of 1 percentage point in the observed probability of opening the door in the no-flyer treatment for the \$0, 5-minute survey would increase the estimated $\hat{h}_0$ by about 0.3 percentage points. Why does $\hat{h}_0$ increase by only 0.3 for each percentage point increase in the observed moment? That is because that is only one of two no-flyer treatments, and the probability of being at home h_0 parameter is also identified by other moments. This sensitivity analysis examines the shift in parameters as one varies one moment at a time.

A more interesting case is the sensitivity of the social pressure cost of saying no to a charity S to the various moments. As we discussed above, there is not one single moment of the data pinning down this parameter which will depend on the sorting in and out of the home, and the different amounts given to the charity. What does the sensitivity matrix Λ say at the model estimates? Andrews, Gentzkow, and Shapiro (2017) reports the results. The social pressure cost S is indeed sensitive to the share answering the door in the flyer treatments and to the share giving \$0-\$10 and the share giving exactly \$10. Here, Andrews, Gentzkow, and Shapiro (2017) points out the influence of the latter moment, since the model predicts bunching at such level of giving due to social pressure. But what if we think that some of the giving of \$10 is not due to social pressure but due to, say, heaping at round numbers? In this case, the estimated social pressure cost would be lower. This allows a potential reader to assess locally the impact of shift in one moment on the parameter estimates. An alternative way to examine this point, the one we had pursued in the paper, was to redefine the set of moments so as not to rely on bunching at \$10. In Online Appendix Table 3 we indeed considered the impact of grouping the smaller giving into giving of \$0-\$10, without singling out giving of exactly \$10; indeed, this leads to less precise estimates of social pressure. It is important to thoroughly examine the sensitivity of key parameters to the data, as documented in this case.

In the third vertical panel of Figure 2 we consider the sensitivity of the estimated value of time for a 1-hour survey, in dollar terms. The figure indicates that the completion rate of the unpaid, 5-minute survey is positively correlated with the value of time. Holding constant the completion rate for the 10-minute survey, a higher completion rate for the 5-minute survey indicates that respondents care more about the value of time.

Sensitivity, Consumption Paper. Laibson et al. (2017) applies the same methodology to their consumption paper. The paper has only 3 parameters to be estimated, and 12 moments (share borrowing, mean borrowing, and wealth, repeated over 4 age groups). Thus the Λ matrix,

Figure 2: Sensitivity of selected parameters from DellaVigna, List, and Malmendier (2012)

The figure shows a subset of the sensitivity matrix Λ as defined in Andrews, Gentzkow, and Shapiro (2017), applied to DellaVigna, List, and Malmendier (2012). The sensitivity of three parameters, the probability of being at home (in 2009), the cost of giving 0 dollars when asked in person (social pressure) and the value of 1 hour of survey, is displayed. Since sensitivity is not scale-invariant, some rescaling is needed: The probability of being at home is not rescaled; the value of 1 hour of survey completion, is scaled by one-hundredth to make it in percentage points; and the social pressure is scaled by one-tenth (divided by 100 to make in pp and multiplied by 10 since the estimate is in per-dollar terms for 10 dollars total).

in Appendix Table 1 in Laibson et al. (2017), has dimension 3*12. The matrix shows several intuitive relationships: a higher share of younger consumers borrowing on visa, or more borrowing by this group would lead to estimates of more present bias (i.e., lower β). Similarly, more wealth accumulation would lead (other than for the 51-60) to less present bias (higher β).

Some associations, however, are not as obvious to understand. For example, higher wealth accumulation among the 51-60 year olds would be associated with a higher discount factor δ but lower present bias β. This is a case in which, for any given moment, the two key parameters—β and δ—move the moments in the same direction, so separating apart the effect on the two of them is not obvious. Also, this sensitivity methodology is well set-up to consider shifts in one moment. But it is not obvious what it means to increase the share borrowing on a visa without also increasing the amount borrowed by that group. In cases such as this, the sensitivity analysis is less straightforward. Nonetheless, this is a useful tool that complements the other sensitivity analysis mentioned above.

4.5 Making It Work

In this subsection I discuss a few additional important points in structural estimation including some more practical issues.

Simulate-and-Estimate. In the process of successfully estimating a model, no step is more important than a robust set of simulate-and-estimate exercises. Consider a researcher who is interested in self-control and wonders if an expensive data set on credit card take-up and usage would identify the self-control parameters she is interested in. Or a second researcher who is about to launch a field experiment to test for loss aversion among teachers using claw-back incentives. Both researchers have written down behavioral models of the setting at hand, but are concerned whether the data can tell apart the model they are considering from alternative models. Is it worth acquiring the data set? And does the field experiment have the needed treatment arms?

A simulate-and-estimate exercise is typically the best tool to answer questions such as these. The researchers can take the model they have written down, including assumptions about the heterogeneity and values for the key parameters and for the auxiliary parameters. The model is then fully specified, and ready for a set of simulations. The simulations will then produce values for the set of moments for a minimum-distance estimation, or data sets to estimate with maximum likelihood. The researcher can then estimate the model on the simulated moments and data. Is it possible to recover the initial parameters? With what precision? Is a set of parameters not separately identified? A key advantage of the simulate-and-estimate exercise is that there is a correct answer, and the researcher knows it. If even under these favorable conditions it is not possible to identify the desired present bias or loss aversion parameter, the real data will not help either (or if it does, it may do so for the wrong reasons).

A second key advantage, as obvious as it is, is that simulate-and-estimate exercises can be done even before any data is available. As such, they are ideal at the stage of experimental design. Of course, having some data from a pilot run, or from previous related studies helps to assign more plausible values for key sets of parameters. For example, how much noise should one assume in the logit error term? How responsive are teachers to incentives in previous studies?

I learned the importance of such simulate-and-estimate exercises the hard way together with my team of coauthors on the "Voting to Tell Others" field experiment DellaVigna et al. (2017). This paper is based on a door-to-door survey run in the Summer and Fall of 2011, following the 2010 congressional elections. In fact, we had also run a similar experiment the year before, in the summer of 2010, asking people about turnout in the 2008 election. We had designed the experiment in light of our social-image model of voting (as in the ultimate published version of the paper); the design had flyers that alerted households about an upcoming survey about their voter participation, or just about a survey; in addition, we cross-randomized the incentives to respond to the survey. This part of the design for the 2010 experiment was just like the one that we also later did in 2011.

But when we attempted to estimate our model on the data, things did not square. In particular, a key set of parameters was very noisily estimated and appeared to be collinear. It was at that stage that, with a puzzle at hand, we did a systematic set of simulate-and-estimate exercises. These exercises revealed to us what we should have known: we were missing a key set of treatments. For voters, our experiments revealed their utility of saying that they voted, s_V, but did nothing to uncover the counterfactual disutility of saying that they did not vote, s_N, or the lying cost L for that matter, given that voters (typically) saw no need to lie about turnout, having voted. It was at that point that we came up with the idea of supplementing the previous design with a lying incentive among households agreeing to the survey. For a control group, we simply asked if they voted. For another group, we told them that we would ask whether they voted, but also informed them if they said that they did not vote, the survey would be (truthfully) 8 minutes shorter. This manipulation would allow us to uncover some of those other parameters that we needed to estimate. Had we done this exercise earlier, we would have come to the field with the full design. We thought that we had done our homework by writing down the model, but we had overly trusted our (faulty) intuition about identification of the model.

Simulate-and-estimate exercise are useful at all stages. At the early stage, they can answer questions of identification and questions of statistical power for a parameter (under certain assumptions). With the data at hand, they can be useful to examine a puzzling phenomenon in the data.

Analytical Solution versus Simulations. A second important issue is how to solve the model, whether to solve it analytically (or numerically), or whether to use simulations to derive the moments. In the minimum distance paradigm, this corresponds to the difference between classical minimum distance and simulated minimum distance (for a reference on the latter, see McFadden, 1989).

In some cases, the model is really quite straightforward, and it makes sense to derive a closed-form solution for the estimation. This is the case, for example, in most experiments on time and risk preferences (e.g., Andreoni and Sprenger, 2012; Augenblick and Rabin, forthcoming) or real-effort experiments (DellaVigna et al., 2017; DellaVigna and Pope, 2018). In these cases, the experiments were designed so that the estimation would be relatively straightforward. But in other cases, the model becomes more complicated to solve. For example, in our study of altruism and social pressure in charitable giving (DellaVigna, List, and Malmendier, 2012), in order to solve for the share answering the door and giving in the flyer treatments, we had to solve for the optimal donation and optimal probability of being at home for a particular altruism type a, and then integrate the solution

over all the altruism types. Since we assumed a normal distribution of altruism, we were able to solve this, partly analytically and partly numerically. After all, in that paper differences in altruism were the only source of heterogeneity, so deriving an optimal solution was not as complicated.

In our paper on voting-to-tell-others (DellaVigna et al., 2017), we started off in much the same way. In that case, though, the decision of whether to stay at home and answer the survey depended on the baseline willingness to do a survey s, as well as on the social-image parameters, s_V and s_N. Each of these three parameters was assumed to be heterogeneous across the population. Thus, even under the assumption that the three sources of heterogeneity were i.i.d., the moments on answering the door and completing the survey required the solution of triple integrals, since we needed to aggregate over the three sources of heterogeneity. To be precise, for the treatments in which a flyer informed the users about a generic survey, given that the cost of avoidance is assumed to be $c(h) = (h - h_0)^2/2\eta$, the solution was $h^* = max[min[h_0 + \eta max(s - c, -S), 1], 0]$. The outside max and min terms are simply taking care of the corner solutions for the probability of being at home h, so the key role here is played by the willingness to do surveys in general, s, net of the time cost c. Notice also the social pressure term $-S$, since the individual anticipates that she can always decide to pay the social pressure cost, rather than doing the survey. The only stochastic term in this expression is $s \sim N\left(\mu_s, \sigma_s^2\right)$. Thus, with some care integrating over the different regions, the integral was not too hard. In the treatments in which the flyer announces a survey about whether the individual voted, though, there is in addition the term about utility of telling others about voting. In this case, the solution for a non-voting individual is $h^* = max[min[h_0 + \eta max(s + z - c, -S), 1], 0]$, with $z = max(s_N, s_V - L)$. The term z captures the fact that the individual can tell the truth and get the social image utility of being a non-voter s_N or lie and get the social image utility of being a voter, but also incurring the disutility L from lying. Since also s_V and s_N are heterogeneous, to compute the observed moment on answering the door now requires a triple integral over s, s_V, and s_N. Even assuming independent draws, this is computationally more demanding. While we still derived this, at some point in the revision of the paper we also decided to model the initial selection into voters and non-voters. That made it even harder to derive even numerical approximations for the solutions, as the triple integral needed to take into account the initial selection.

So we pivoted to a simulation-based solution for the moments. For each combination of parameters, to compute the moments we drew a population of, say, 1,000 individuals, each with different draws or s, s_V and s_N. For each individual, given their draw, we would solve for the implied moment, and then average across the 1,000 moments to compute the average moment. Given that the moment is based on a finite sample averaging rather than an analytical solution, there is an approximation error introduced, which one can take into account in the standard errors. One would think that this approach would be computationally hard-going since the solution for any parameter requires 1,000 draws. And yet, this approach did not take much longer than our best attempts at an analytical solution, given that it did not require any triple integrals and the like. Also, this approach reduced the chance of errors in the code, since the model just involves the solution for a particular drawn type (which is pretty straightforward), followed by sample averaging across the draws. As such, the code was much more straightforward. Simulation-based estimates are used, for example, also in Laibson et al. (2017) where the complicated dynamic programming problem of optimal consumption-savings

can only be solved for a particular draw of the variables.

Starting Points. Even taking all the above into account, how can one get started in this literature? A good place to look is existing papers, as summarized in Table 1. Nearly all of the published papers in this Table have data and code posted, thanks to the good norms on data posting in our discipline.[22]

5 Working with Common Behavioral Models

5.1 Present Bias

I summarize four key features of the model of present bias (Laibson, 1997; O'Donoghue and Rabin, 1999a) to keep in mind, especially for structural estimation: (i) timing of payoffs; (ii) money versus consumption; (iii) time period duration; and (iv) sophistication versus naiveté.

Timing of Payoffs. In the present-bias model, the key distinction is between the present time period t and future time periods $t+1, t+2, ...$ Thus two questions are critical: first, what is present and what is future, and second how long does the present last. We discuss the first issue here, and return to the second issue shortly.

For the first question, a key distinction is between goods with immediate costs and future benefits (*investment goods*) and good with immediate benefits and future costs (*leisure goods*). Individuals with self-control problems consume investment goods less than they would like from an ex ante perspective, and consume leisure good more than they would like (DellaVigna and Malmendier, 2004; DellaVigna, 2009).

In a given setting, the timing of payoffs should be clear so as to distinguish between these two cases. For example, doing bureaucratic work to save for retirement, to file taxes, or change health insurance plans implies an immediate effort cost with delayed benefits, qualifying these activities clearly as investment goods. Importantly, the immediate cost is *not* a monetary cost but a dis-utility of effort of going through the bureaucratic steps. Similarly goes for applying for jobs for unemployed workers, studying hard for an upcoming exam, working hard at the workplace or, for most people, getting oneself to exercise. In other cases, the opposite timing makes it clear that a good is a leisure good, such as for example eating a tempting dessert, smoking a cigarette, drinking excessively, or entering a store and purchasing a fancy piece of jewelry or clothing. In these cases, there is an immediate non-monetary benefit, the satisfaction of a craving.

In other cases the timing is not as clear. Consider the case of ordering a mail-order item: is the consumption of a good that will be received 2-3 days later immediate enough to qualify as present? In some cases, it is hard to guess, and sometimes our intuition as economists can be off. An interesting example is in the Augenblick, Niederle, and Sprenger (2015) paper which introduced real-effort tasks to estimate time preferences. Their main task, an annoying task of transcription, was an investment good by design, and by all accounts, people saw it as is. But the authors also designed a similar effort task that was designed to be much more fun, similar to Tetris, with an

[22] Psychology for example is only now getting to comparable availability of data and programs for papers published in top journals.

eye to having a placebo task, or even a leisure good task. It turns out that, judging from behavior, participants in the experiment saw this task just as costly in terms of effort as the other one.

The key point is that, in any model of an activity based on present-bias, the intertemporal timing of costs and benefits should be quite clear, as otherwise all results will be off.

Money versus Consumption. A related issue, that we raised above in Section 2.3, is the fact that the timing of payoffs refers to consumption, not money receipt. For example, if I receive a paycheck today and I plan to spend it next month to purchase the newest iPhone, all of the utility of consumption associated with today's paycheck is presumably in the future. If one instead models present bias over monetary payments, one would erroneously set up things otherwise. I should mention that, given how cash constrained many consumers are, there is often a fairly tight link between monetary payments and consumption (e.g., Ganong and Noel, 2017 for the case of spending out of unemployment benefits). But still, one needs to be very careful.

Overall, the large majority of the papers on present bias, and nearly all the ones based on field evidence, handle this issue appropriately, as the timing of payoffs in most setting typically relies on either a bureaucratic annoying immediate effort cost (investment goods), or on an immediate satisfaction of a craving (leisure goods). Still, to flag an example of how this could come up, consider in my own work the assumed timing of payoffs for health club attendance in DellaVigna and Malmendier (2006): individuals attending the gym on day t pay an immediate effort cost c and earn a delayed health benefit b, both non-monetary; but there is also a per-visit fee p (assumed to be paid immediately). Thus from a perspective of the time t-self, we model the payoff as $-c-p+\beta\delta b$. The assumption that the dis-utility cost of the per-visit fee is immediate is clearly open to challenge. One could as easily assume that the cost in terms of consumption occurs in the future, since that is when any displaced consumption is likely to occur. That is correct. Importantly, though, the timing of the payment of p is not what defines health-club attendance as an investment good, which depends on c and b. Working out the example with a delayed payment of p leads to similar qualitative results. The general point is that questioning the timing of consumption versus monetary payments is critical for this literature, and certainly for structural estimates of these models.

A we discussed in Section 2.3, the biggest case of this issue occurred for laboratory tests of present bias with monetary payments. Across these experiments, one had to maintain that the utility of consumption happened in the same period as the payment from the experiment. The real-effort elicitation of time preferences pioneered by Augenblick, Niederle, and Sprenger (2015) removes the need to make this strong assumption and, indeed, strongly suggests that violations of this assumption can explain previous inconsistencies between the different experimental results (see Section 2.3).

Time period duration. A reasonable objection to present-bias models is that they are more sensitive to how we model the frequency of the periods than exponential-discounting models. Some papers, for example, model the periods as days, and thus assume that, from today's perspective, tomorrow's payoff is in the future, and thus discounted with $\beta\delta$. Other papers instead model the periods as years, and thus assume that all of this year's payoffs are in the present. Who is right, and does it matter?

A first issue is to better understand the psychology of present bias, and when things are processed

as being in the present, versus in the future. Here we have very limited evidence except for Augenblick (2017), in which the author uses the same real-effort task as in Augenblick, Niederle, and Sprenger (2015), but varies systematically the time duration between the decision on effort, and when effort has to take place. This paper, thus, has the ability to estimate the exact timing of decay of discounting, at least within a particular setting. Augenblick (2017) finds that at least a third of the discounting occurs already within a few hours, and the majority of it occurs within a day or two. Thus, at a minimum, it is not wrong, when modeling decisions at the daily level, to assume that tomorrow is already into the future. Balakrishnan, Haushofer, and Jakiela (2017) provides concordant evidence from an experiment with monetary trade-offs.

A separate issue is whether modeling the decision at the daily or quarterly or annual level matters. That depends on the situation, but certainly for some models this modeling decision can have a big impact. Return to the procrastination model of O'Donoghue and Rabin (1999b) discussed in Section 2.1. Procrastination occurs because the naive individual incorrectly thinks that she will do in the next period the task that she is unwilling to do this period. What we define the next period to be is critical to the calibration. If the next period is tomorrow, or even next week, the loss due to delaying saving for retirement is small enough to justify waiting (given the wrong beliefs about doing it next time). But if we model the decision as taking place at the quarterly or annual level, even a naive agent would plausibly not procrastinate for plausible parameters, since the cost of delaying that one (long) period is too high. The structural estimates in Shui and Ausubel (2005) of switching credit card offers and in Fang and Silverman (2009) of welfare take-up, for example, are developed under the assumption that individuals take decisions once a quarter (for credit card switching) or once a year (for welfare take-up); the results likely would be quite different if the decision was modeled at the daily level.

Two final remarks on this. First, the level at which the decision is modeled has nothing to do with the frequency with which a decision is observed. Assume that a researcher studies credit card switching using quarterly data. Still, the researcher can model switching at the daily level, then create quarterly averages to compare to the data. Second, given that we do not know how frequently people are thinking of, and taking the decisions, one could model the decisions as taking place at different horizons, and let the data tell which horizon best explains the data.

Sophistication versus naiveté. Perhaps the most important decision in the estimation of models of present bias is the assumption of sophistication or naiveté. In the O'Donoghue and Rabin (2001) language, are present-bias individuals ($\beta < 1$) aware that their future selves have self-control problems too (sophisticates, $\hat{\beta} = \beta$) or do they expect the future selves not to have self-control problems (naives, $\hat{\beta} = 1$)? These are just the polar cases, with intermediate cases of partial sophistication for $\beta < \hat{\beta} < 1$. If it were possible, of course, estimating the general case of partial sophistication would be ideal. But if it is computationally infeasible, where does the evidence currently stand on the extent of naivete', and does it make a large difference computationally and economically?

We do have some evidence on the first question and, on net, we have at least as much evidence of naiveté as we have of sophistication. For one thing, we have vast evidence of inertia effects which naive procrastination provides a parsimonious explanation for (if not, certainly, the only one).

Second, we now have a large literature on demand for commitment devices, which only make sense in the presence of sophistication. With very few exceptions (e.g., Schilbach, forthcoming), there is only very moderate demand for commitment. Third, most of the field evidence on present bias makes the most sense with the naiveté interpretation (e.g., Ausubel (1999) and DellaVigna and Malmendier (2006)). Finally, and most importantly, we have direct evidence measuring the $\hat{\beta}$ parameter in a real-effort task modeled a la Augenblick, Niederle, and Sprenger (2015). In this experiment, Augenblick and Rabin (forthcoming) finds that $\hat{\beta}$ is close to 1, suggesting that full naiveté is a convenient approximation to their findings.

Given this evidence tilting the scale towards naiveté, the next issue is whether computationally it matters. The answer is clear: the assumption of naiveté often helps tremendously in simplifying the estimation. To see why, consider again the example in Section 2.1. Assume that an agent with exponential discounting is solving a dynamic programming problem involving a maximization of the type

$$max_{c_t^e} u\left(c_t^e\right) + \delta V^e\left(c_t^e\right),$$

where V^e is the value function for the exponential agent, which will be in general a function of the current choice (for example, since that can affect the assets available next period). The naive present-biased agents will solve

$$max_{c_t^n} u\left(c_t^n\right) + \beta\delta V^e\left(c_t^n\right).$$

The key point is that the naive agent believes that the value function from the exponential agent will apply to her. Thus, to solve the naive problem one can solve the exponential problem, solving for the value function V^e(which is typically the time-consuming part), and then to obtain the consumption for the naive c^n it is a quick step typically. The case for the sophisticated is different, as one must solve a different programming problem, with the war of attrition between the sophisticated selves. In some cases, such as general consumption problems, the sophisticated case has poorly-behaved solutions in ways that the naive solution does not (Laibson, 1997; Laibson et al., 2017).

This is not to say that the naive present-bias assumption is a perfect approximation and it should be applied everywhere. It certainly, for example, would not be appropriate for a paper that studies the demand for commitment. But, given where the evidence currently is on $\hat{\beta}$, at least starting off with the naive case provides a good compromise of computational complexity. Indeed, a few recent papers in the area take this route (e.g., Laibson et al., 2017; DellaVigna et al., 2017).

5.2 Reference Dependence

Perhaps the most influential model in behavioral economics, prospect theory (Kahneman and Tversky, 1979), has been applied to the estimation of risk preferences on lottery choice, and beyond that to a variety of applications to field evidence. As well-known, the key components of prospect theory are: (i) a reference point; (ii) loss version relative to the reference point; (iii) diminishing sensitivity of the value function around the reference point, and (iv) a probability weighting function that magnifies small probabilities (see O'Donoghue and Sprenger, 2018). The applications of this reference-dependent model to field evidence fall broadly into two categories, about effort targeting

a reference point, and about first-order aversion to risk.

Two Classes of Field Evidence. Underlying the first set of applications, most of which I already discussed, is an (often implicit) model of costly effort with a target (the reference point), with loss aversion with respect to the target. Considering for example the labor supply applications (Camerer et al., 1997; Farber, 2008; Crawford and Meng, 2011; Thakral and Tô, 2017), cab drivers are likely to put more effort and thus work longer if they are still trying to make their daily target, and thus are still on the loss side of the value function. The applications of costly effort up to a target (the reference point) include tax elusion to achieve zero taxes due (Rees-Jones, 2018; Engström et al., 2015), heightened job search given the loss relative to recent income (DellaVigna et al., 2017), and marathon running to achieve a round-number goal (Allen et al., 2017). A second set of applications does not involve effort, but rather price setting to achieve a target: house sales to make the previous house purchase price (Genesove and Mayer, 2001) merger prices to make the 52-week high, an industry benchmark (Baker, Pan, and Wurgler, 2012), and the willingness to pay for the endowment effect (Kahneman, Knetsch, and Thaler, 1990). In this first set of applications, risk and uncertainty is not central, unlike in the motivating examples for prospect theory.

A second set of applications is, instead, exactly focused on risk. In finance, Nicholas Barberis, in a series of influential papers, has made the case that prospect theory can make sense of a variety of anomalies, including the equity premium and the performance of IPOs (Barberis, Huang, and Santos, 2001; Barberis, 2018). For insurance choice, Sydnor (2010) and then Barseghyan et al. (2013) put forward a model to explain the preference of insurers for low-deductible plans even when the extra premium is high.

Interestingly, the two sets of applications focus on almost orthogonal parts of the reference-dependent model. The first set of applications, on effort relative to a target, typically assumes a simplified version of prospect theory with a piece-wise linear function around a reference point, with loss aversion. In these examples, diminishing sensitivity and probability weighting are typically assumed away for simplicity.

In the second set of applications, loss aversion still plays a role, especially since it generates first-order aversion to risk around the reference point. But diminishing sensitivity and especially probability weighting play a key role. Barseghyan et al. (2013) for example attributes much of the preference for low-deductible insurance to the overweighting of small probabilities. Barberis (2018) makes the case that probability weighting can explain preference for skewed investments, such as IPOs.

While both types of applications are clearly model-driven, few applications provide structural estimates of the underlying behavioral parameters; among the exceptions that we discussed are Crawford and Meng (2011), Barseghyan et al. (2013), DellaVigna et al. (2017), and Thakral and Tô (2017). As the number of such estimates increases, we will have a better sense of the stability across contexts of the loss aversion parameter, of the curvature of the value function, of the probability weighting function, and of the reference point formation, which is the focus of my next point.

Reference Point. These papers also differ in another important way: the reference point determination. Early applications of prospect theory focused on cases in which there is a salient reference point, like a current, or recent, situation, such as the purchase price for a house (Genesove

and Mayer, 2001) or a stock (Barberis, Huang, and Santos, 2001), or the initial endowed objects (Kahneman, Knetsch, and Thaler, 1990). Recent papers have often followed this lead, focusing on a benchmark number for mergers (Baker, Pan, and Wurgler, 2012), salient relevant round number, like zero taxes due (Rees-Jones, 2018; Engström et al., 2015) or round finishing minutes for marathon running (Allen et al., 2017), or recent average earnings for the unemployed (DellaVigna et al., 2017). Indeed, the presence of a single, salient number is typically crucial in these papers to establish the evidence, as in the bunching evidence for tax filing (Rees-Jones, 2018) and marathon running (Allen et al., 2017). This bunching evidence would not be possible if the reference point target were more uncertain. The plausibility of these targets comes from the fact that, for one reason or another, they are salient: it is hard to forget the purchase price of one's own home, and easier to brag about going under 4 hours in marathon running.

Yet, the downside of these cases is that the reference point is, in ways, arbitrary and context-dependent. Is it possible to have a model of reference point that is more portable across contexts? Köszegi and Rabin (2006) did just that assuming that the reference point is a forward-looking expectation of what is likely to happen in a given situation. In a forward-looking reference point, the gain-loss utility is computed relative to the (stochastic) distribution of realized outcomes according to the solution concept of personal equilibrium. I refer to the chapter on reference dependence by O'Donoghue and Sprenger for details on the concept, but I will stress the status of its application, especially as far as structural estimates.

There is some evidence along the lines of forward-looking reference points. The occurrence of domestic violence during American football matches, for example, is heightened for losses, but only when the loss is unexpected, suggesting a role for expectations (Card and Dahl, 2011). Abeler et al. (2011) presents a clever design using a real-effort task that implies some forward-looking component in the reference point. Finally, at least one paper on the endowment effect (Ericson and Fuster, 2011) provides evidence of Köszegi-Rabin reference points, in that the extent of endowment effect is affected by the expected trading probability (as predicted by the model).

And yet, the evidence for forward-looking reference points is quite weak, as even some of the evidence supporting it has not stood up to re-examination. In a follow-up paper to the Abeler et al. (2011) real effort design, Gneezy et al. (2017) shows that, while the authors can replicate the original finding, the Köszegi-Rabin pattern in the effort data does not appear, and in fact reverses sign, for other values of the parameters, and a different design. Similarly, in the case of the endowment effect, Heffetz and List (2014) and Goette, Harms, and Sprenger (forthcoming) find no evidence that the pattern of trades is affected by the expected trading probability, as predicted by Köszegi-Rabin reference points. To the opposite, the evidence points to an endowment effect predicted by the initial endowment. That is, the reference point largely appears to be the status quo, as opposed to expectations.

In my view, the pendulum has swung back and a growing number of papers takes seriously the alternative hypothesis that the reference point is mostly backward-looking, either the status quo or a recent average of recent outcomes. This is the case for example in Heffetz and List (2014) and Goette, Harms, and Sprenger (forthcoming) where the reference point appears to be the status-quo assignment. One advantage of status-quo-type reference point is that they are easier to test for in

field data. One can, for example, test for bunching at zero balance due for taxes, or at round-number finishing times for runners, as these numbers are focal points. For forward-looking reference points, instead, there is no sharp prediction for observed behavior: any bunching is smoothed out by the stochastic nature of the reference point, so one most construct a deliberate design for the purpose, typically in the laboratory.

In the future, it would be great to have more examples of evidence in which the model allows for both forward- and backward-looking reference points, and tests for both, and estimates parameters. One example I am aware of is my own work on job search. Most of the estimation is focused on backward-looking reference points given by recent earnings, which helps explain the pattern of exit from unemployment. We also, however, consider a forward-looking reference point, where the reference point is given by the distribution of earnings from the perspective of the previous period: that is, with some probability a person will have found a job, but most likely the person will still earn the UI benefits. We show that this forward-looking reference point does not help at all explain the patterns in the data. Among the laboratory evidence, the endowment effect experiments like Goette, Harms, and Sprenger (forthcoming) implicitly provide joint evidence on both status quo and forward-looking reference points. The literature could really benefit from more papers estimating the path of the reference point comparing backward-looking and forward-looking cases.

5.3 Social Preferences

As I mentioned in Section 2.3, there is a large literature on estimates of social preferences in laboratory experiments, especially for models of inequity aversion and reciprocity (e.g., Charness and Rabin, 2002, Bellemare, Kroger, and Van Soest, 2008), but also other models of altruism and estimates at the individual level (e.g., Andreoni and Miller, 2002 and Fisman, Kariv, and Markovitz, 2007).

There is not a parallel literature on field evidence with estimation of the same models. Rather, there are different strands of literature which use different classes of models, typically testing qualitative statics, as opposed to estimating the parameters. The long-standing literature on charitable giving typically is motivated by models of warm glow (Andreoni, 1989); papers in this literature typically test predictions such as the degree of crowd out, and do not estimate structural parameters. A number of papers in the behavioral labor literature (DellaVigna and Mas, 2018) provide qualitative evidence of social preferences with respect to, for example, horizontal or vertical pay equity. In a few cases, a specific model of social preferences is spelled out, for example in studies of social preferences and conflict in Kenya (Hjort, 2014) and on relative pay versus piece rate for fruit pickers (Bandiera, Barankay, and Rasul, 2005). These papers consider a simple model of pure altruism a la Becker (1974), with weight α on the relevant other person's utility. This simple model is used to derive a set of predictions on how fruit pickers respond to the ethnicity of the co-worker under a combination of compensation schemes (Hjort, 2014), and how productivity varies under piece rate versus relative pay (Bandiera, Barankay, and Rasul, 2005). In neither of these cases are the parameters estimated. The lack of structural estimates is consistent with the view that the models of social preferences used in these applications—warm glow and pure altruism—are seen more as illustrative than describing

fully the underlying social preferences.

In my work on social preferences, my coauthors and I have argued that, even when a model is simplified, providing quantitative evidence on it helps in several respects. For example, in our work on charitable giving (DellaVigna, List, and Malmendier, 2012), we take a simple model of altruism and warm glow, as well as a simple model of social pressure: individuals pay a disutility cost S if they turn down an in-person request to be helpful, whether by not giving money to charity or not doing a survey. We do not view the social pressure cost S as a deep structural parameter that will be constant across all settings, but rather as a reduced-form parameter that captures a variety of avoidance motivations, such as social signaling. Still, even this reduced form models allows us to achieve useful goals, with an order-of-magnitude assessment of the disutility of being asked, and the ability to do welfare evaluations of this situation. Similarly, in our social-image model of voting DellaVigna et al. (2017), we do not view the social image parameters s_V (the utility of saying that one voted, when asked) and s_N (the utility of saying that one did not vote, when asked) as deep, invariant parameters. Indeed, we would expect them to vary across elections and across other setting where one wants to signal good intentions (such as voting in this case). Still, even this simpler model allows us to get a fuller picture of the impact, for example, of get-out-the-vote interventions based on telling others that they will be asked.

I would like to single out a venue for work on social preferences in the field: designing simplified field situation that, while not taking place in the laboratory, allow for more control and thus more detailed tests of social preferences. The field experiments on gift exchange starting from Gneezy and List (2006) are designed with an eye to this: create a mini-job of 6 hours, varying across conditions the generosity of the employer. While papers in this literature do not provide model estimates, we show in DellaVigna et al. (2016), as discussed above, how to add treatment conditions so as to identify the social preference parameters. In particular, we can identify both the baseline social preference weight α towards the employer, and how much this changes in response to a "gift" from the employer ($\alpha + \alpha_{Gift}$). Further, we attempt to distinguish between a model of pure altruism where the employee cares about the utility of the employer with weight α, from a model in which the employee cares about the effort benefiting the employer, but not the exact return. This second model, which we call warm glow, itself captures a variety of social preference motivations, such as sense of duty and some form of signaling. While the experiment does not pin down these separate models, it already provides an indication of magnitudes, and a test of the null hypothesis of the pure altruism model.

A different example is the paper on social campaign by Dubé, Luo, and Fang (2017). The authors worked with a marketing campaign for the sale of movie tickets with a field experimental design. In particular, different individuals received a text message offering them the option to purchase a ticket for a movie at different discount levels. Crossed with the discount amount, the experiment randomized how much money the firm would donate to a charity if the ticket is purchased. A simple altruism model predicts that the share purchasing movie tickets would be increasing in both the discount (for the obvious reasons) and the donation to the charity. The authors find a more nuanced pattern that that. The share purchasing tickets increases in the charitable donation, but only when there is no discount on the price of a ticket, or a small discount. When the price of a ticket is steeply

discounted, instead, a higher charitable discount *lowers* the take up of the movie ticket, a pattern that is consistent with crowd out of motivation: a buyer would like to signal that she cares about the charity, which is a motivation to purchase the ticket. But when the ticket is steeply discounted, purchasing it does not deliver this signal.

Dubé, Luo, and Fang (2017) estimates a model of social signaling a la Benabou and Tiróle (2006). Assume that the consumer has consumption utility $V + \alpha p + \gamma a$ where V is the utility from a movie, and γ is the measure of social preferences, with $a = 0/1$ indicating if the individual purchases the ticket at price p. In addition, the person puts weight on the inferred value of altruism given her decision: $\lambda_\gamma E(\gamma|a, p, y)$. This is the key component from Benabou and Tiróle (2006): the person cares about the inference about the social preference that one can draw from the action. Notice that in this case no one was observing the action, but the signaling can also be to oneself, assuming that the person remains unsure about one's own social preferences. The final part of the utility function is the signaling utility with respect to the marginal utility of money $\lambda_\alpha E(a|a, p, y)$. This part has a less obvious interpretation, but is necessary, it turns out, to produce the crowd-out pattern above. Then the individual purchases the ticket if

$$U(1) = V + \alpha p + \gamma a + \lambda_\alpha E(\alpha|a, p, 1) + \lambda_\gamma E(\gamma|a, p, 1) \geq U(0) = \lambda_\alpha E(\alpha|a, p, 0) + \lambda_\gamma E(\gamma|a, p, 0) \quad (21)$$

Denoting with $\Delta(a, p) = \lambda_\alpha E(\alpha|a, p, 1) + \lambda_\gamma E(\gamma|a, p, 1) - \lambda_\alpha E(\alpha|a, p, 0) - \lambda_\gamma E(\gamma|a, p, 0)$ the net ego utility, this implies that the updating on the social preferences is as follows:

$$E(\gamma|a, p, 1) = E\left[\gamma|\gamma > -\frac{V + \alpha p + \Delta(a, p)}{a}\right],$$

which can be solved jointly with the other conditions. This condition, together with the other conditional expectation conditions, can be fed back into (21) and solved as a fixed point problem. Thus, the solution of the social signaling model need not be that complicated. Dubé, Luo, and Fang (2017) shows that this model matches very well the observed moments, unlike the altruism model which, as mentioned above, would predict monotonicity.

This paper illustrates the promise of models of signaling: these models can fit observed behavior that does not fit well with the traditional models of social preferences, including the moral wiggle room behavior observed in lab experiments (Dana, Weber, and Kuang, 2007). It will probably take a while, though, for the literature to converge on models of social preferences estimated in field settings.

6 Conclusion

What is *structural behavioral economics*? What role can it play? I argued in this chapter that papers with point estimates of some behavioral parameters—the definition of structural behavioral—have a natural role in the behavioral literature. Behavioral economics has benefited from a close relationship between behavioral theory and empirics, which structural estimation can build on. Behavioral economics has also made important use of calibration of magnitudes of effects, and

structural estimates take calibrations one step further. Experimental evidence has also played a key role in behavioral economics, and model-based designs can provide useful input already at the design stage of the experiment. Structural estimates of behavioral parameters allow to test for the stability of the parameters and for out-of-sample predictions. Further, model-based estimates of behavioral parameters help with welfare and policy evaluations, an area of fast growth within behavioral economics.

I also presented an important set of considerations to keep in mind. To start with, one needs to be mindful of the settings in which structural estimation can add substantial value, and the ones in which it does not. For example, many times one is interested in a reduced-form result, or there is no adequate or relevant model to test. Even when there is a model to test, a test of model-based qualitative predictions often achieves many of the same goals that structural estimation aims to achieve, without all the structure that estimation requires. For the settings where structural estimation has advantages—and as I argued above, it is many—, it is important to keep in mind the drawbacks from the additional complexity, and sensitivity of the results to the number of assumptions that go into the estimation. With this in mind, I presented a number of strategies to keep the estimation relatively simple, or at least make it simpler, and to examine the sensitivity of the results.

Hopefully, that leaves some researchers with a desire to get started in this direction. Thus, I discussed a number of the modeling choice and considerations to keep in mind for structural behavioral work. Among the most important ones are the choice of estimation method, and the modeling of the heterogeneity (i.e., the source of noise in the data). I also highlighted how it is important to think not only of the key model parameters, but also of ancillary parameters and assumptions. Finally, I discussed common issues and strategies in the estimation of some of the most common behavioral models.

Where does the structural behavioral literature currently stand? In Table 1 I summarize some key modeling choices—such as the estimation method, the source of heterogeneity, and the behavioral parameters of interest—for the structural behavioral papers that we discussed in this chapter. I emphasize that this list is not exhaustive and reflects, of course, my particular approach to the topic. Having said this, it does provide some examples and it illustrates the variety of topics, and methods chosen. I am so curious to see what a revised version of Table 1 will look like in 10 years.

References

Abeler J, Falk A, Goette L, Huffman D. Reference Points and Effort Provision. *American Economic Review.* 2011; 101(2): 470–492.

Adda, J. and Cooper, R. (2003). Dynamic Economics: Quantitative Methods and Applications. MIT Press, Cambridge, MA.

Akerlof GA. Labor Contracts as Partial Gift Exchange, *Quarterly Journal of Economics*. 1982. 97(4): 543-569.

Allcott H and Kessler J. The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons. *AEJ: Applied Economics*. Forthcoming.

Allcott H, and Rogers T. The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. 2014. *American Economic Review,* 104(10): 3003-37.

Allcott H, Taubinsky D. Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market. *American Economic Review*, 2015; 105(8): 2501-38.

Allen EJ, Dechow PM, Pope DG, Wu G. Reference-Dependent Preferences: Evidence from Marathon Runners. *Management Science.* 2017; 63(6):1657-72.

Andersen S, Harrison GW, Lau MI, Rutström EE. Eliciting risk and time preferences. *Econometrica.* 2008; 76(3): 583-618.

Anderson, CA, Bushman BJ. Effects of Violent Video Games on Aggressive Behavior, Aggressive Cognition, Aggressive Affect, Physiological Arousal, and Prosocial Behavior: A Meta-analytic Review of the Scientific Literature. *Psychological Science*, 12 (2001), 353-359.

Andreoni J. Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. *Journal of Political Economy.* 1989; 97(6), 1447-1458.

Andreoni J, Bernheim BD. Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica.* 2009; 77(5): 1607-1639.

Andreoni J, Bernheim BD. Theories of Social Preferences. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics.* Elsevier; 2018.

Andreoni J, Miller J. Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica.* 2002; 70(2), 737-753.

Andreoni J, Sprenger C. Estimating time preferences from convex budgets. *American Economic Review.* 2012; 102(7), 3333-3356.

Andrews I, Gentzkow M, Shapiro J. Measuring the Sensitivity of Parameter Estimates to Estimation Moments. *Quarterly Journal of Economics*, 2017, 132, 1553-1592.

Apesteguia J, Ballester MA. Monotone Stochastic Choice Models: The Case of Risk and Time Preferences. *Journal of Political Economy*, 2018, 126, 74-106.

Augenblick N. Short-Term Discounting in Unpleasant Tasks. 2017 Working paper.

Augenblick N, Niederle M, Sprenger C. Working over Time: Dynamic Inconsistency in Real Effort Tasks. *Quarterly Journal of Economics.* 2015; 130(3): 1067-1115.

Augenblick N, Rabin M. An Experiment on Time Preference and Misprediction in Unpleasant Tasks. *Review of Economic Studies.* Forthcoming.

Ausubel LM. Adverse Selection in the Credit Card Market. 1999 Working Paper.

Baker M, Pan X, Wurgler J. The effect of reference point prices on mergers and acquisitions. *Journal of Financial Economics,* 2012; 106(1): 49-71.

Balakrishnan U, Haushofer J, Jakiela P. How Soon Is Now? Evidence of Present Bias from Convex Time Budget Experiments. 2017, NBER Working Paper #23558.

Bandiera O, Barankay I, Rasul I. Social Preferences and the Response to Incentives: Evidence from Personnel Data. *Quarterly Journal of Economics,* 2005; 120(3): 917-962.

Barberis N. Psychology-based Models of Asset Prices and Trading Volume. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics,* Volume 1, Elsevier; 2018.

Barberis N, Huang M, Santos T. Prospect Theory and Asset Prices. *Quarterly Journal of Economics*, 2001; 116(1): 1-53.

Barseghyan L, Molinari F, O'Donoghue T, Teitelbaum JC. The Nature of Risk Preferences: Evidence from Insurance Choices. *American Economic Review.* 2013; 103(6): 2499-2529.

Becker GS. A Theory of Social Interactions, *Journal of Political Economy.* 1974. 82(6), 1063-1093.

Bellemare C, Kröger S, Van Soest A. Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities. *Econometrica.* 2008; 76(4), 815–839.

Bellemare C, Shearer BS. On the Relevance and Composition of Gifts within the Firm: Evidence from Field Experiments, *International Economic Review.* 2011; 52(3), 855-882.

Bénabou R, Tirole J. Incentives and Prosocial Behavior. *American Economic Review.* 2006; 96(5): 1652-78.

Benartzi S and Thaler R. How Much Is Investor Autonomy Worth?. *Journal of Finance*; 2002. 57(4): 1593-1616.

Bernheim BD, Fradkin A, Popov I. The Welfare Economics of Default Options in 401(k) Plans. *American Economic Review.* 2015; 105(9): 2798-2837.

Bernheim BD, Rangel A. Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *Quarterly Journal of Economics.* 2009; 124(1): 51-104.

Bernheim, BD, Taubinsky D. Behavioral Public Economics. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics,* Volume 1, Elsevier; 2018.

Bertrand M, Karlan D, Mullainathan S, Shafir E, Zinman J. What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment. *Quarterly Journal of Economics.* 2010; 125 (1): 263-306.

Beshears J, Choi JJ, Harris C, Laibson D, Madrian BC, Sakong J. Which Early Withdrawal Penalty Attracts the Most Deposits to a Commitment Savings Account? 2017. Working paper.

Beshears J, Clayton C, Choi, J, Harris C, Laibson D, Madrian BC. Optimal Illiquidity. 2017. Working paper.

Bhargava S, Loewenstein G, Sydnor J. Choose to Lose: Health Plan Choices from a Menu with Dominated Option. *Quarterly Journal of Economics.* 2017; 132(3): 1319–1372.

Bhargava S and Manoli D. Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment Bhargava. *American Economic Review.* 2015; 105(11): 3489-3529.

Busse M, Lacetera N, Pope D, Silva-Risso J, and Sydnor J. Estimating the Effect of Salience in Wholesale and Retail Car Markets. 2013. *American Economic Review Papers and Proceedings* 103(3): 570-574.

Camerer C, Babcock L, Loewenstein G, Thaler R. Labor Supply of New York City Cabdrivers: One Day at a Time. *Quarterly Journal of Economics.* 1997; 112(2): 407–441.

Camerer C, Issacharoff S, Loewenstein G, O'Donoghue T, Rabin M. Regulation for Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism". *University of Pennsylvania Law Review* 2003; 151(3): 1211-1254

Camerer C, Ho TH. Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty.* 1994; 8(2): 167-196.

Camerer, C, Ho TH, Chong JK. A Cognitive Hierarchy Model of Games. *Quarterly Journal of Economics.* 2004; 119(3): 861-898.

Card D, Dahl GB. Family violence and football: The effect of unexpected emotional cues on violent behavior. *Quarterly Journal of Economics.* 2011; 126(1): 103-143.

Card D, DellaVigna S, Malmendier U. The Role of Theory in Field Experiments. *Journal of Economic Perspectives.* 2011; 25(3): 39-62.

Carroll GD, Choi JJ, Laibson D, Madrian BC, Metrick A. Optimal Defaults and Active Decisions. *Quarterly Journal of Economics.* 2009; 124(4): 1639-1674.

Carvalho LS, Meier S, Wang SW. Poverty and economic decision-making: Evidence from changes in financial resources at payday. *American Economic Review.* 2016; 106(2): 260-284.

Chandra A, Handel B, Schwartzstein J. Behavioral Health Economics. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics.* Elsevier; 2018.

Charness G, Rabin M. Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics.* 2002; 117(3): 817-869.

Chetty, R. Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics.* 2009; 1(1):451-488.

Chetty R, Looney A, Kroft K. Salience and Taxation: Theory and Evidence. *American Economic Review.* 2009; 99(4): 1145-77.

Cho S, Rust J. The Flat Rental Puzzle. *Review of Economic Studies*, 2010; 77(2): 560-594.

Choi JJ, Laibson D, Madrian BC, Metrick A. For better or for worse: Default effects and 401(k) savings behavior. In: Wise D, ed. *Perspectives on the Economics of Aging.* University of Chicago Press; 2004.

Choi JJ, Laibson D, Madrian BC, Metrick A. Saving for Retirement on The Path of Least Resistance, Chapter 11 in Ed McCaffrey and Joel Slemrod (eds.), *Behavioral Public Finance: Toward a New Agenda*, New York: Russell Sage Foundation, 2006, 304-351.

Conlin M, O'Donoghue M, Vogelsang TJ. Projection Bias in Catalog Orders. *American Economic Review.* 2007; 97(4): 1217-49.

Costa-Gomes M, Crawford VP, Broseta B. Cognition and Behavior in Normal- Form Games: An Experimental Study. *Econometrica.* 2001; 69(5): 1193-235.

Crawford V, Meng J. New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational- Expectations Targets for Hours and Income. *American Economic Review.* 2011; 101(5): 1912–1932.

Dahl G, DellaVigna S. Does Movie Violence Increase Violent Crime? *Quarterly Journal of Economics*, 2009; : 677-734.

Dana J, Weber R and Kuang J. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory.* 2007; 33(1), 67-80.

DellaVigna S. Psychology and Economics: Evidence from the Field. *Journal of Economic Literature.* 2009; 47(2), 315-372.

DellaVigna S, Gentzkow M. Persuasion: Empirical Evidence. *Annual Review of Economics*, 2010, 2.

DellaVigna S, Gentzkow M. Uniform Pricing in US Retail Chains. 2017. NBER Working paper w23996.

DellaVigna S, Kaplan E. The Fox News Effect: Media Bias and Voting. *Quarterly Journal of Economics*, 2007, 122, 1187-1234.

DellaVigna S, Lindner A, Reizer B, Schmieder JF. Reference-Dependent Job Search: Evidence from Hungary. *Quarterly Journal of Economics,* 2017, 132, 1969-2018.

DellaVigna S, List JA, Malmendier U. Testing for Altruism and Social Pressure in Charitable Giving. *Quarterly Journal of Economics.* 2012; 127(1): 1-56.

DellaVigna S, List JA, Malmendier U, Rao G. Estimating Social Preferences and Gift Exchange at Work. 2016 NBER Working Paper 22043.

DellaVigna S, List JA, Malmendier U, Rao G. Voting To Tell Others. *Review of Economic Studies.* 2017; 84(1):143-181.

DellaVigna S, Malmendier U. Contract Design and Self-Control: Theory and Evidence. *Quarterly Journal of Economics.* 2004; 119(2): 353-402.

DellaVigna S, Malmendier U. Paying Not to Go to the Gym. *American Economic Review.* 2006; 96(3): 694-719.

DellaVigna S, Mas A. Behavioral Labor Economics. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics.* Elsevier; 2019.

DellaVigna S, Paserman MD. Job Search and Impatience. *Journal of Labor Economics.* 2005; 23(3): 527-588.

DellaVigna S, Pope D. What Motivates Effort? Evidence and Expert Forecasts. *Review of Economic Studies.* 2018; 85(2): 1029-1069.

Dubé JP, Luo X, Fang Z. Self-Signaling and Pro-Social Behavior: A Cause Marketing Experiment. *Marketing Science.* 2017; 36(2):161-186.

Duflo E, Saez E. The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. *Quarterly Journal of Economics*, 2003; 118(3): 815-842.

El-Gamal MA, Grether DM. Are People Bayesian? Uncovering Behavioral Strategies, *Journal of the American Statistical Association*, 1995, 90(432): 1137-1145.

Engström P, Nordblom K, Ohlsson H, Persson A. Loss Compliance and Tax Aversion. *AEJ: Economic Policy.* 2015; 7(4):132-164.

Ericson KM, Fuster A. Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments. *Quarterly Journal of Economics.* 2011; 126(4): 1879–1907.

Ericson KM, Laibson D. Intertemporal Choice. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics.* Elsevier; 2018.

Fang H, Silverman D. Time-Inconsistency and Welfare Program Participation: Evidence from the NLSY. *International Economic Review.* 2009; 50(4): 1043-1077.

Farber H. Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers. *American Economic Review*. 2008; 98(3): 1069–1082.

Fehr E, Goette L. Do Workers Work More if Wages are High? Evidence from a Randomized Field Experiment. *American Economic Review*. 2007; 97(1): 298-317.

Fehr E, Kirchsteiger G, Riedl A. Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*. 1998. 42(1): 1-34.

Fehr E, Schmidt KM. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*. 1999; 114(3): 817-868.

Fisman R, Kariv S, Markovits D. Individual Preferences for Giving. *American Economic Review*. 2007; 97(5): 1858-1876.

Ganong P, Noel P. Consumer Spending During Unemployment: Positive and Normative Implications. 2017 Working Paper.

Genesove D, Mayer C. Loss Aversion and Seller Behavior: Evidence from the Housing Market. *Quarterly Journal of Economics*. 2001; 116 (4): 1233-1260.

Gill D, Kissová Z, Lee J, Prowse V. First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*. Forthcoming.

Gill D, Prowse V. A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*. 2012; 102(1): 469–503

Gneezy U, Goette L, Sprenger C, Zimmermann F. The Limits of Expectations-Based Reference Dependence. *Journal of the European Economic Association*, 2017, 15(4), 861-876.

Gneezy U, List JA. Putting Behavioral Economics to Work: Field Evidence of Gift Exchange. *Econometrica*. 2006; 74(5): 1365-84.

Gneezy U, Niederle M, Rustichini A. Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*. 2003; 118(3): 1049-1074.

Goette L, Harms A, Sprenger C. Randomizing Endowments: an Experimental Study of Rational Expectations and Reference-dependent Preferences. *American Economics Journal: Microeconomics*, Forthcoming.

Goldfarb A, Xiao M. Who Thinks about the Competition: Managerial Ability and Strategic Entry in US Local Telephone Markets, *American Economic Review*. 2011, 101(7), 3130-3161.

Gourinchas PO, Parker JA. Consumption over the life cycle. *Econometrica*. 2002; 70(1): 47-89.

Green D, Gerber A. Get Out the Vote: How to Increase Voter Turnout, 2008. Brookings Press.

Halevy Y. Time Consistency: Stationarity and Time Invariance. *Econometrica*. 2015; 83(1): 335-352.

Halpern SD, Kohn R, Dornbrand-Lo A, Metkus T, Asch DA, Volpp KG. Lottery-Based versus Fixed Incentives to Increase Clinicians' Response to Surveys. *Health Services Research.* 2011; 46(5): 1663-1674.

Halpern D. Inside the Nudge Unit: How Small Changes Can Make a Big Difference. London, UK: WH Allen; 2015.

Handel BR. Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts. *American Economic Review.* 2013; 103(7): 2643-2682.

Handel BR, Kolstad JT. Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare. *American Economic Review.* 2015; 105(8): 2449–2500.

Harless DW and Camerer CF. The Predictive Utility of Generalized Expected Utility Theories. *Econometrica.* 1994; 62(6), 1251-1289.

Heffetz, O and List, J. Is the Endowment Effect an Expectations Effect? *Journal of the European Economic Association.* 2014. 12(5): 1396-1422

Hjort J. Ethnic divisions and production in firms. *Quarterly Journal of Economics.* 2014; 129(4): 1899–1946.

Hossain T and Morgan J. ...Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay. *Advances in Economic Analysis and Policy.* 2006. 6(2): 1429-1429.

Judd, K. L. Numerical Methods in Economics. MIT Press, 1998

Kahneman D, Knetsch JL, and Thaler RH. Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal of Political Economy.* 1990, 98(6): 1325-1348.

Kahneman D, Tversky A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica.* 1979; 47(2): 263-91.

Kaur S, Kremer M, Mullainathan S. Self-Control at Work. *Journal of Political Economy.* 2015; 123(6), 1227-1277.

Köszegi B, Rabin M. A Model of Reference-Dependent Preferences. *Quarterly Journal of Economics.* 2006; 121(4): 1133-65.

Köszegi B, Rabin M. Reference-dependent consumption plans. *American Economic Review.* 2009; 99(3): 909-936.

Kube S, Maréchal MA, Puppe C. The Currency of Reciprocity: Gift Exchange in the Workplace *American Economic Review*, 2012. 102(4), 1644-1662.

Kube S, Maréchal MA, Puppe C. Do Wage Cuts Damage Work Morale? Evidence from a Natural Field Experiment. *Journal of the European Economic Association.* 2013; 11: 853–870.

Lacetera N, Pope DG, Sydnor JR. Heuristic Thinking and Limited Attention in the Car Market. *American Economic Review*. 2012; 102(5): 2206-36.

Laibson D. Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*. 1997; 112(2): 443-477.

Laibson D, Repetto A, Tobacman J. Estimating Discount Functions with Consumption Choices over the Lifecycle. 2007 NBER Working Paper 13314.

Laibson D, Maxted P, Repetto A, Tobacman J. Estimating Discount Functions with Consumption Choices over the Lifecycle. 2017 Working Paper.

Madrian BC, Shea DF. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior. *Quarterly Journal of Economics*. 2001; 116(4):1149-1187.

Malmendier U, Nagel S. Depression babies: do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*. 2011; 126(1), 373-416.

Martinez SK, Meier S, Sprenger C. Procrastination in the Field: Evidence from Tax Filing. 2017 Working Paper.

McFadden D. Conditional Logit Analysis of Qualitative Choice Behaviour. in: P. Zarembka, Ed., *Frontiers in Econometrics*, Academic Press, New York, 1974.

Mc Fadden D, Talvitie AP, and Associates. Validation of disaggregate travel demand models: Some tests. 1977. In Urban Demand Forecasting Project, Final Report, Vol. V, Institute of Transportation Studies, University of California, Berkeley.

McFadden, D. L. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 1989, 57(5): 995–1026.

McFadden D. Computing Willingness-to-Pay in Random Utility Models. In: Melvin R, Moore C, Riezman R, eds. *Trade Theory, and Econometrics*. Routledge; 1999.

Mulligan C. A Logical Economist's Argument Against Hyperbolic Discounting. 1996 University of Chicago mimeo.

Niederle M, Vesterlund L. Do women shy away from competition? Do men compete too much?. *Quarterly Journal of Economics*. 2007; 122(3), 1067-1101.

O'Donoghue T, Rabin M. Doing It Now or Later. *American Economic Review*. 1999; 89(1):103-124.

O'Donoghue T, Rabin M. Procrastination in Preparing for Retirement. In: Aaron HJ, ed. *Behavioral Dimensions of Retirement Economics*. Brookings Institution Press and Russell Sage Foundation; 1999.

O'Donoghue T, Rabin M. Choice and Procrastination. *Quarterly Journal of Economics*. 2001; 116(1): 121-160.

O'Donoghue T, Sprenger C. Reference-Dependent Preferences. In: Bernheim BD, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics,* Volume 1, Elsevier; 2018.

Paserman MD. Job Search and Hyperbolic Discounting: Structural Estimation and Policy Evaluation. *The Economic Journal*. 2008; 118(531): 1418–1452.

Rabin M. Risk Aversion and Expected Utility Theory: A Calibration Theorem. *Econometrica*. 2000; 68(5): 1281-92.

Rashes MS. Massively Confused Investors Making Conspicuously Ignorant Choices (MCI--MCIC). *Journal of Finance*. 2001; 56(5): 1911-1927.

Rees-Jones A. Quantifying Loss-Averse Tax Manipulation. *Review of Economic Studies*. 2018; 85(2): 1251-1278.

Reiss PC, Wolak FA. Structural Econometric Modeling: Rationales and Examples from Industrial Organization, *Handbook of Econometrics*, 2007; 6(A), 4277-4415.

Rogers T, Ternovski J, and Yoeli E. Potential follow-up increases private contributions to public good. *Proceedings of the National Academy of Sciences*. 2016; 113(19), 5218-5220.

Rust J. The Limits of Inference with Theory: A Review of Wolpin (2013). *Journal of Economic Literature*. 2014; 52(3): 820-850.

Schilbach F. Alcohol and Self-Control: A Field Experiment in India. *American Economic Review*, Forthcoming.

Schultz W, Nolan J, Cialdini R, Goldstein N, Griskevicius V. The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science.* 2007; 18(5): 429–434.

Shlain A. More than a Penny's Worth: Left-Digit Bias and Firm Pricing. 2018, working paper.

Shue K, Luttmer EFP. Who Misvotes? The Effect of Differential Cognition Costs on Election Outcomes. *American Economic Journal: Economic Policy.* 2009; 1(1), 229-257.

Shui H, Ausubel L. Time Inconsistency in the Credit Card Market. 2005 Working Paper.

Silver B, Anderson B, Abramson P. Who Overreports Voting? *American Political Science Review.* 1986; 80(2), 613-24.

Sydnor J. (Over) insuring modest risks. *American Economic Journal: Applied Economics.* 2010; 2(4), 177-199.

Taubinsky D, Rees-Jones A. Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment. *Review of Economic Studies*, Forthcoming.

Thakral N, Tô L. Daily Labor Supply and Adaptive Reference Points. 2017 Working Paper.

Thaler RH, Benartzi S. Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of political Economy.* 2004. 112(S1), S164-S187.

Thaler R, Sunstein C. Nudge. New Haven, CT: Yale University Press; 2008.

Todd PE, Wolpin KI. Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility, *American Economic Review*. 2006; 96(5): 1384-1417.

Train KE. Discrete choice methods with simulation, Publisher Cambridge University Press; 2009.

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

Wolpin, K. *The Limits of Inference Without Theory*. 2013; Cambridge, Mass. and London: MIT Press.

Table 1. Representative Studies in Structural Behavioral Economics, Panel A

Paper	Type	Behavior / Choices	Parameters of Interest	Incidental Parameters	Estimation Method	Source of Heterogeneity
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Time Preferences/ Present Bias						
Laibson, Repetto, and Tobacman (2007); then Laibson, Repetto, Maxted, and Tobacman (2017)	Observational Data	Lifecycle consumption decisions	Discount factor (delta); present-bias (beta, assuming naivete'); relative risk aversion (rho)	Calibrated: pre-retirement income (mean and variance), retirement age, post-retirement income (mean and variance), household size, credit limit, interest rates	Simulated Minimum Distance	Noise in income stream
Andersen, Harrison, Lau, and Rutström (2008)	Laboratory Experiment	Lotteries over time	Mean and s.d. of risk av. (r); mean and s.d. of exp. discount (delta); hyperbolic discount (gamma)	Estimated: noise (mu, nu); types shares (pi). Assumed: cons. smoothing periods (lambda); wealth / background consumption (omega)	Maximum likelihood	Random Utility; Random Parameters (risk aversion and discounting)
Paserman (2008)	Observational Data	Unempl. and accepted wages	Discount factor (delta); present-bias factor (beta, assuming sophistication)	Value of leisure (b_0); wage distr. (mu_i and sigma_i); s.d. of meas. error in wages (sigma_w); scale of effort cost (k). Assumed: curvature of cost fn. (eta); layoff prob. (q)	Maximum likelihood	Random Parameters (2-types in scale of cost function and mean wage)
Fang and Silverman (2009)	Observational Data	Welfare takeup	Discount factor (delta), present-bias factor (beta) for both naivete' and sophistication	Stigma; home prod.; wage and skill; Calibrated: welfare benefits; fertility; Assumed: perceived present-bias (beta-tilde)	Maximum likelihood	Random Utility; Random Parameters w/ discrete types
Andreoni and Sprenger (2012)	Laboratory Experiment	Allocation of monetary payment	Discount factor (delta); present-bias (beta); CRRA risk aversion (alpha) / CARA risk aversion (rho)	Assumed / Estimated: background consumption (omega_1, omega_2)	NLS and Tobit ML on f.o.c.s and solution functions	Implementation Error (Additive noise in consumption)
Augenblick, Niederle, and Sprenger (2015)	Laboratory Experiment, real effort	Allocation of unpleasant task or money	Discount factor (delta); present-bias factor (beta)	Monetary curvature parameter (alpha); power cost of effort curvature (gamma). Assumed: Stone-Geary background parameter (omega)	Tobit regression	Implementation Error (Additive error in log allocation ratio)
Augenblick and Rabin (forthcoming)	Laboratory Experiment, real effort	Unpleasant task completion	Present-bias factor (beta), perceived present-bias factor (beta-hat), discount factor (delta), projection bias (alpha)	Cost of effort function	Maximum Likelihood	Implementation Error (Additive noise in effort)
Martinez, Meier, and Sprenger (2016)	Observ. Data + Laboratory Exp. in field	Tax filing delay	Discount factor (delta); present-bias (beta assuming naivete'); filing costs (c)	Assumed: refunds delay periods (k); noise parameters (lambda, mu)	Maximum Likelihood	Random Utility (Dynamic Logit)
Augenblick (2017)	Laboratory Experiment, real effort	Unpleasant task completion	Discounting: exponential (delta), quasi-hyperbolic (beta, delta, nu), hyperbolic (kappa), general hyperbolic (kappa, alpha)	Cost of effort function	Maximum Likelihood	Implementation Error (Additive noise in effort)

Table 1. Representative Studies in Structural Behavioral Economics, Panel B

Paper	Type	Behavior / Choices	Parameters of Interest	Incidental Parameters	Estimation Method	Source of Heterogeneity
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B. Reference Dependence						
Camerer and Ho (1994)	Laboratory Experiment	Lottery Choices	Probability Weighting Curvature (gamma); assume reference point of 0 earnings	Curvature of value function (alpha)	Maximum Likelihood	Random Utility (Logit)
Farber (2008)	Observational Data	Taxi driver length of time working	Loss aversion (delta); reference point mean and variance (theta, sigma)	Continuation function based on other factors (e.g., hours worked, day of week, time of day, weather, location, day FE)	Maximum Likelihood	Random Utility (Probit for stopping) & Random Parameter (Heterogeneity in reference point)
Crawford and Meng (2011)	Observational Data	Taxi driver length of time working	Loss aversion over hours (eta* (lambda_h-1)) & income (eta* (lambda_i-1)); assume expectation-based ref. pt. based on sample average	Cost of work effort fn. parameters (theta, rho); Predictors of stopping (driver FE, day of week, hour of day, location, weather)	Maximum Likelihood	Random Utility (Probit for stopping)
Gill and Prowse (2012)	Laboratory Experiment, real effort	Effort in slider task	Distr. of disappointment aversion (lambda_2 and sigma_lambda)	Quadratic cost of effort with individual heterogeneity and round effects (b, kappa, delta_r, Phi_mu, phi_mu, Phi_pi, phi_pi)	Simulated Minimum Distance	Random Parameter (Heterogeneity in loss aversion)
Barseghyan, Molinari, O'Donoghue, and Teitelbaum (2013)	Observational Data	Choices of deductibles on auto and home insurance	Probability weighting function (semi non-parametric), absolute risk aversion (r); assume Koszegi-Rabin reference point	Claim probabilities by observables and claim type (mu_j, result of a poisson estimation); scale of choice noise by claim type (sigma_j)	Maximum Likelihood; Sieve for probability weighting function	Random Utility (Additive noise in deductible choice)
Engström, Nordblom, Ohlsson and Persson (2015)	Observational Data	Tax non-compliance	Loss aversion (lambda); assume reference point=0 tax due	Effect of age, employment income, and gender on likelihood of taking deduction	Sufficient Statistic (Share claiming in loss domain and in gain domain)	
DellaVigna, Lindner, Reizer, and Schmieder (2017)	Observational Data; Natural Experiment	Job search and consumption smoothing	Time preferences (delta, beta); loss aversion (lambda); adjustment speed of reference-point (N); adaptive reference point	search cost function level and curvature (k, gamma); share of types	Minimum Distance	Random Parameters (3-type heterogeneity in cost of effort function)
Thakral and To (2017)	Observational Data	Taxi driver length of time working	Loss aversion over hours (eta * (lambda_h-1)) and income (eta * (lambda_i-1)); Speed of adaptation of reference point	Cost of work effort function parameters (theta, rho); Additional predictors of stopping (driver FE, day of week, hour of day, location, weather)	Maximum Likelihood	Random Utility (Probit for stopping)

Table 1. Representative Studies in Structural Behavioral Economics, Panel C

Paper	Type	Behavior / Choices	Parameters of Interest	Incidental Parameters	Estimation Method	Source of Heterogeneity
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel C. Social Preferences						
Charness and Rabin (2002)	Laboratory Experiment	Choices in distribution games	Inequity aversion (rho and sigma), and reciprocity (theta)	Noise parameter (gamma)	Maximum Likelihood	Random Utility (Logit)
Andreoni and Miller (2002)	Laboratory Experiment	Self versus other in convex budget sets	[Type level] weight on self vs other (alpha), CES subutiltiy of self vs other (rho)	pre-estimation categorization of subjects into types (selfish, Leontief, perfect substitutes) then estimating params for "weak" types	Two-limit Tobit Maximum Likelihood	Additive noise on budget shares
Fisman, Kariv, and Markovits (2007)	Laboratory Experiment	Self versus others - convex and non-convex budget lines	Individual-level weight on self vs other (alpha), and on other vs other (alpha'), CES subutility between self and others (rho), and between other and other (rho')	None	Non-linear Tobit Maximum Likelihood (2-stage estimation for 3 person)	(1) Parameters estimated at individual level (2) Budget share on self implemented with additive noise
Bellemare, Kroger, and Van Soenst (2008)	Laboratory Experiment	Dictator and Ultimatum Games	Inequity Aversion model Parameters (alpha and beta, including quadratic terms)	Distribution of noise parameters	Simulated Maximum Likelihood	Random Utility; Random Parameters (Heterogeneity in inequity aversion)
DellaVigna, List, and Malmendier (2012)	Field experiment	Charitable giving	Mean and s.d. of altruism (mu_a, sigma_a); curvature of altruism (G); social pressure cost of giving 0 to charity (S_ch)	Baseline prob. of opening door (h_0); prob. of seeing flyer (r); responsiveness of home presence (eta); mean and s.d. of utility of survey (mu_s, sigma_s); time value (v_s); social pressure cost of saying no to survey (S_s)	Minimum Distance	Random Parameters (Heterogeneity in altruism)
DellaVigna, List, Malmendier, and Rao (2017)	Field experiment	Answering survey questions on voting	Mean and standard deviation of social image utilities (mu_v, mu_n, sigma_si); lying cost (L)	Baseline prob. of opening door (h_0); prob. of seeing flyer (r); responsiveness of home presence (eta); mean and s.d. of utility of survey (mu_s, sigma_s); time value (v_s); social pressure cost of saying no to survey (S_s); mean and s.d. of residual value of voting (epsilon)	Simulated Minimum Distance	Random Parameters (Heterogeneity in willingness to do survey and social-image of saying one voted)
Dube, Luo, and Fang (2017)	Field experiment	Social campaign with movie ticket sales	Altruism (gamma); self-signaling on donations (lambda_gamma), price sensitivity (lambda_alpha), and utility from movie (lambda_V)	Utility from movie (V), price sensitivity (alpha); mixing probability (omega)	Maximum likelihood (MPEC estimator)	Random Parameters except signaling parameters (two types)
DellaVigna, List, Malmendier, and Rao (2016)	Experiment, Real Effort	On-the-job effort, preparing mailings	Altruism (alpha); warm glow (alpha); Change in soc. Pref. with gift (a_gift, alpha_gift)	Exponential and power cost function: individual fixed effects (k), curvature (s), time trend (f)	Non-linear least squares	Random Parameters (Heterogeneity in cost of effort function)

Table 1. Representative Studies in Structural Behavioral Economics, Panel D-E

Paper	Type	Behavior / Choices	Parameters of Interest	Incidental Parameters	Estimation Method	Source of Heterogeneity
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel D. Limited Attention						
Chetty, Looney, and Kroft (2009)	Field exp. + Observational Data	Grocery shopping + alcohol cons.	Tax underraction (inattention) - (theta)	Price elasticity of demand; excise tax elasticity of demand	Sufficient Statistic (DDD estimate and price elast.)	
Lacetera, Pope, Sydnor (2012)	Observational Data	Car purchases	Inattention to lower digits	Seventh degree polynominal mapping mileage to value, car specific valuations (estimated in first stage OLS)	Non-linear Least Squares	
Handel and Kolstad (2015)	Observational Data + Survey	Choice of health insurance	Constant absolute risk aversion (mean and variance), frictions (information, hassle costs, inertia)	Wealth, health type (expected health costs), individual marginal tax rate	Simulated Maximum Likelihood	Random Utility; Random Parameters (Heterogeneity in risk aversion)
Allcott and Taubinsky (2015)	Online Experiment	Choice of Light Bulbs	Inattention to energy savings of CFL bulbs	Hedonic value places on incendescent lightbulbs versus CFL bulbs	Sufficient Statistic (Change in WTP)	
Taubinsky and Rees-Jones (forthcoming)	Online Experiment	Purchasing goods (with different sales-tax)	Average tax underraction - conditional on tax size (theta(t)), and its variance	Order effects, person fixed effects	GMM	No noise, but type heterogeneity
Shlain (2018)	Observational Data	Grocery Shopping+Price setting	Consumers left-digit bias (theta), Firm perceived left-digit bias (theta-hat)	Price elasticity of demand; Promotion, seasonality, and product effects; Shape of cost distribution	Sufficient Stat.; NLS; Minimum Distance	Variation in Costs
Panel E. Behavioral Firms						
Cho and Rust (2010)	Observational Data; Experiment	Pricing of Rental Cars	Profitability of current used car lease contract	Depreciation of rental car (type-specific), distr. of miles driven under different contracts; duration of rental spells, transition prob	Maximum likelihood	Random Utility (Dynamic Logit); Random Parameters
Goldfarb and Xiao (2011)	Observational Data	Firm entry in cell phone market	Mean (tau) of Poisson distribution determining share of types in k levels model	Firm expectation about profitability; effect of competition. Noise level	Simulated Maximum Likelihood	Random Parameters (Heterogeneity across Markets)
DellaVigna and Gentzkow (2017)	Observational Data	Pricing of Retail Stores	Impact of Store-Level Income and Elasticity on Prices		OLS, Instrumental Variable	

Table 1. Representative Studies in Structural Behavioral Economics, Panel F-G

Paper	Type	Behavior / Choices	Parameters of Interest	Incidental Parameters	Estimation Method	Source of Heterogeneity
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel F. Other Categories						
Conlin, O'Donoghue, and Vogelsang (2007)	Observational Data	Purchases and returns of cold-weather items	Projection bias (alpha)	Estimated: individual tastes, cost of return, sensitivity of utility to temperature. Assumed: cost of return function, discount factor, threshold temperature when utility is derived	Maximum likelihood	Random Utility (probit); Random Parameters (heterog. in preferences on order and return date)
Handel (2013)	Observational Data; Natural Experiment	Choice of health insurance	constant absolute risk aversion (mean and variance), switching cost	Wealth, health type (expected health costs), individual marginal tax rate	Simulated Maximum likelihood	Random Utility; Random Parameters (Heterogeneity in risk aversion)
Malmendier and Nagel (2011)	Survey	Financial decisions, risk preferences, and expectations	Sensitivity to experienced returns (beta), shape of weighting function over experiences (lambda)	Effect of income, household characteristics, liquid assets, age, and year	Ordered probit maximum likelihood, non-linear least squares	
Panel G. Horse Race of Models						
Bernheim, Fradkin, and Popov (2015)	Natural Experiment	401(k) contributions	Distr. of opt-out costs (gamma), and behavioral opt-out parameters (calibrated from gamma*D(f), fn. of incidental parameters and a calibrated gamma - the "real" opt-out costs), anchoring (zeta)	Retirement saving shift parameter (alpha), firm-level mean utility weights of savings (mu_i), common sd of utility weights (sigma), fraction of employees with zero opt-out costs (lambda_1), opt-out cost distribution parameter (lambda_2)	Maximum likelihood	Random Utility (Probit); Random Parameters (Individual level preferences parameters rho and opt-out type)
DellaVigna and Pope (2018)	Online Experiment, Real Effort	Pressing a-b buttons on keyboard	Intrinsic motivation, time pref. (beta, delta), social pref. (altruism, warm-glow), gift exchange, crowd-out of low-pay, prob. weighting, reference dependence (loss aversion)	Parameters of effort cost functions (two functional forms assumed)	Minimum Distance, Non-linear least squares	No noise in method of moments (errors granted by bootstrapping underlying sample), Random Parameter (heterogeneity in cost of effort) in NLLS

Notes: This Table summarizes select papers in the *Structural Behavioral Economics* literature. We list the papers in order of publication within a topic area, followed by the working papers.

Table 2a. Evidence for Probability Weighting, Structural Estimates

Paper	Setting	Type of Probability Weighting Function	Parameter Estimate	Implied Probability Weight for 1% Prob.
(1)	(2)	(3)	(4)	(5)
Panel A. Studies Designed to Estimate Probability Weighting Function				
Tversky and Kahneman (1992)	Lottery Choice	Kahneman-Tversky	0.61	0.055
Gonzalez and Wu (1999)	Lottery Choice	Linear-in-log-odds	-	0.093 (0.003)
Camerer and Ho (1994)	Lottery Choice	Kahneman-Tversky	0.56	0.067
Wu and Gonzalez (1996)	Lottery Choice	Kahneman-Tversky	0.71	0.036 (0.002)
Harrison, List and Towe (2007)	Lottery Choice	Kahneman-Tversky	0.83	0.022
Kilka and Weber (2001)	Stock Forecasts	Linear-in-log-odds	-	0.181 (0.013)
Abdellaoui (2000)	Lottery Choice	Linear-in-log-odds	0.6	0.040 (0.001)
Tversky and Fox (1995)	NBA/NFL/Weather Forecasts	Linear-in-log-odds	-	0.031
Donkers, Melenberg and van Soest (2001)	Lottery Choice	Prelec	0.435	0.143 (0.011)
Harrison, Humphrey and Verschoor (2010)	Lottery Choice	Kahneman-Tversky	1.384	0.002 (0.000)
Bruhin, Fehr-Duda and Epper (2010)	Lottery Choice	Linear-in-log-odds	-	0.141 (0.003)
de Brauw and Eozenou (2014)	Crop Choice	Kahneman-Tversky	1.37	0.002 (0.000)
Liu (2013)	Lottery Choice	Prelec	0.69	0.057 (0.014)
Tanaka, Camerer and Nguyen (2010)	Lottery Choice	Prelec	0.74	0.045
Barseghyan, Molinari, O'Donoghue and Teitelbaum (2013)	Insurance Deductible Choice	Semi-nonparametric	-	0.07
Snowberg and Wolfers (2010)	Horse Race Data	Prelec	0.928	0.020
Aruoba and Kearny (2011)	State Lotteries	Prelec	0.89	0.020
Kliger and Levy (2009)	Financial Markets	Kahneman-Tversky	0.622	0.053 (0.001)
Average Probability Weight from Meta-Analysis				**$\pi(0.01) = 0.060$**

Notes: This Table, adapted from a meta-analysis in DellaVigna and Pope (2018), lists papers providing an estimate of the probability weighting function with the setting and type of probability weighting function used (Columns 2 and 3), and the estimated parameter for the probability weighting function, when available (Column 4). Column 5 reports the implied probability weight for a 1 % probability, given the estimated weighting function in the study. The standard errors, when available, are computed with the delta method. At the bottom of the table we report the parameter for the meta-analysis, equal-weighting across the studies.

s

Table 2b. Evidence for Overwighting of Small Probabilities, Studies with Probabilistic Incentives

Paper	Subjects	Effort Task	Sample Size	Treatments (Certain Reward vs. Probabilistic Reward with low p)	Effort with Certain Reward, Mean (S.D.)	Effort with Probabilistic Reward, Mean (S.D.)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B. Field Studies Comparing Certain Reward to Probabilistic Reward						
DellaVigna and Pope (2018)	Mturk	Button Presses	555 (P), 558 (C)	1% chance of winning US$1 (P) vs. fixed payment of US$0.01 (F) per 100 presses	2029 (27.47)	1896 (28.44)
Halpern et al. (2011)	Resident Physicians in a US Database	Survey Response	358 (P), 400 (C)	0.4% chance of winning US$2500 (P) vs. fixed payment of US$10 (F) for response	0.558 (0.497)	0.511 (0.500)
Thirumurthy et al. (2016)	Men aged 21 to 39 years old in Kenya	Uptake of Circumcision	302 (P), 308 (C)	Mixed lottery with expected retail value of US$12.50 (P) vs. food voucher worth US$12.50 (F)	0.084 (0.278)	0.033 (0.179)
Diamond and Loewy (1991)	Undergraduates in State University	Recycling	78 (P), 113 (C)	5% chance of winning $5 and 1% chance of winning $25 (P) vs. $0.50 voucher for campus store (F)	0.212 (0.409)	0.308 (0.462)
Dolan and Rudisill (2014)	16 to 24 year olds in England	Return Test Kit via Mail	247 (P), 549 (C)	10% chance of a 50 GBP Tesco voucher (P) vs. 5 GBP Tesco voucher (F)	0.732 (0.443)	0.706 (0.455)

Notes: This Table, adapted from a meta-analysis in DellaVigna and Pope (2018), lists papers examining the impact on effort (broadly defined) of deterministic incentives, or of probabilistic incentives. The expected value of incentives in the two treatments is the same (or very similar), and in the probabilistic treatments the probability of the incentive is quite low, so probability weighting would predict that the probabilisity is overweighted. We report the subject pool (Column 2), the effort task and sample size (Columns 3 and 4), the treatments (Column 5), and the effort in the treatment with certain reward (Column 6) versus in the treatment with probabilistic reward (Column 7).