Identification of structural models in the presence of measurement error due to rounding in survey responses

Stefan Hoderlein[*]
Boston College

Bettina Siflinger[†]
University of Mannheim

Joachim Winter[‡]
University of Munich

January 19, 2015

Abstract

Distortions in the elicitation of economic variables arise frequently. A common problem in household surveys is that reported values exhibit a significant degree of rounding. We interpret rounding as a filter that allows limited information about the relationship of interest to pass. We argue that rounding is an active decision of the survey respondent, and propose a general structural model that helps to explain some of the typical distortions that arise out of this active decision. Specifically, we assume that there is insufficient ability of individuals to acquire, process and recall information, and that rational individuals aim at using the scarce resources they devote to a survey in an optimal fashion. This model implies selection and places some structure on the selection equation. We use the formal model to correct for some of the distorting effects of rounding on the relationship of interest, using all the data available. Finally, we show how the concepts developed in this paper can be applied in consumer demand analysis by exploiting a controlled survey experiment, and obtain plausible results.

Keywords: heaping; nonparametric; survey design; bounded rationality; identification.

[*] Department of Economics, Boston College, Chestnut Hill, MA 02467, USA;
e-mail: stefan_hoderlein@yahoo.com.

[†] Department of Economics, University of Mannheim, D-68161 Mannheim, Germany;
e-mail: bettina.siflinger@uni-mannheim.de.

[‡] Department of Economics, University of Munich, D-80539 Munich, Germany;
e-mail: winter@lmu.de.

1 Introduction

Motivation: The distribution of responses to quantitative survey questions often exhibits a structure that reflects specific features of both the objects to be elicited, as well as the way the question is posed. Figure 1, explained in detail below, shows a typical example of a phenomenon known as heaping or rounding.[1] When respondents in the "Health and Retirement Study" are asked about their weekly "food outside home" expenditures, the distribution of responses shows a striking degree of heaping at focal values (the figure shows the relative frequencies of the reported values), in particular multiples of 50 dollars.

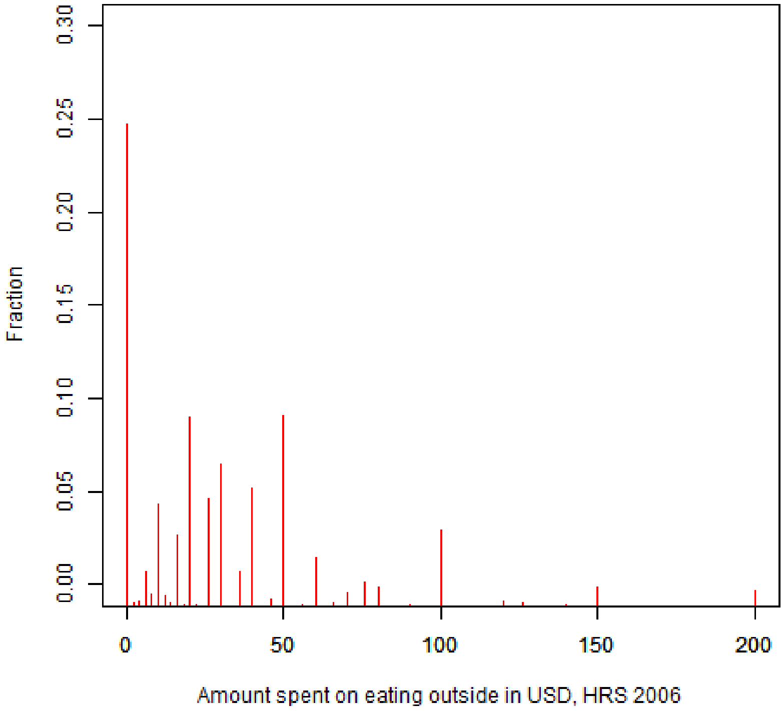

Figure 1: Spikeplot of responses to question on weekly food consumption expenditure on eating outside home (HRS 2006).

An obvious explanation for this phenomenon is rounding – in the presence of uncertainty,

[1] In this paper, we use the term rounding since it appears to be more commonly used in economics. In the statistics literature, the following definitions are used (see Heitjan and Rubin (1991)). (i) *Coarsening of data*: Only a subset of the complete-data sample space in which the true, unobservable data lie is observed. This includes as special cases rounding, heaping, censoring, missing data, etc. (ii) *Rounding*: Data values are observed or reported only to the nearest integer. (iii) *Heaping*: A dataset is said to be heaped if it includes items reported with various levels of coarseness.

individuals pick the nearest significant integer. While similar examples are ubiquitous in applied economics, this phenomenon has received, with a few exceptions, surprisingly little attention. Instead, in applications basically two strategies are being pursued. The first is to simply ignore the problem altogether, and use all the data available. The second is to discard the subsample which rounds (the "rounders", henceforth), and work with the remaining subsample (the "nonrounders").

This paper argues that, at least in some applications, both strategies may be problematic: The first strategy creates biases associated with the coarsening of information, the second, as we shall argue, potential problems of selection, and is wasteful in terms of observations. To deal with these biases, we argue that we have to first understand the mechanism which causes individuals to round, and hence parts of the observations to be rounded. More specifically, we provide a model in which optimizing individuals make an active choice about whether to round or not. It is based on a cost-benefit analysis, not unlike a Roy type treatment effect model. This model motivates us to employ an IV strategy to deal with the selection aspect of rounding, where the instruments are exogenous factors that impact the costs associated with actively remembering the exact number. We also model the limited dependent variable character of the outcome equation, so that in sum our model bears some resemblance with a structural treatment effect model with a limited dependent variable, where rounding can be thought of as a treatment. However, what distinguishes our approach from standard treatment effects models is that we are not primarily interested in the effect of the treatment (rounding), but rather want only to correct the biases rounding induces on the structural relationship of interest.

The illustration of these issues through an application from consumer demand, using in fact the same data that were used to generate Figure 1, is an important part of this paper. Throughout much of the demand literature, the structural relationship of interest is the relation between (food outside of home) expenditures and income; however, these expenditures are rounded for nearly half of the sample. Since the sample at hand consists largely of retired individuals, and the population which rounds may be less fit mentally and physically, and hence also the less active population with fewer expenditures outside home, we may expect a selection bias. Also, excluding the rounders from this application results in a large loss of information.

The results we obtain from the application confirm this intuition. After correcting for selection and the limited dependent variable character generated by application, we obtain quite similar results for the structural relationship between food outside demand and income, for both the rounder and nonrounder subpopulations. As such, this application illustrates that our approach produces sensible results in an important application.

Related literature: While, at least in our opinion, rounding did not receive the attention that it deserves in the literature, we are by no means the first to point out the implications of rounding. Indeed, Heitjan and Rubin (1991) already note that rounding implies that the measured variable is coarsened, and that hence information is lost which in turn affects statistical and econometric analysis. Although, as already mentioned, this problem is omnipresent in survey measurements of continuous variables, it is often ignored in applied work. In some situations, this might be justified if the degree of coarsening is inconsequential (Wright and Bray, 2003). Generally, however, rounding cannot be ignored. For instance, Battistin, Miniaci, and Weber (2003) and Pudney (2007) document striking amounts of rounding in self-reported consumption measures. Questions on subjective probabilities are another example of severe coarsening of data that cannot be ignored in statistical analysis (Manski, 2004; Manski and Molinari, 2010).

Heitjan and Rubin (1991) present a general model for coarsened data, including rounded, heaped, censored, and missing data. They define a “coarsened at random” condition under which the coarsening mechanism can be ignored. Heitjan (1994) defines a “coarsened completely at random” condition. In essence, these conditions ensure that the likelihood can be constructed conditionally on the coarsening and that there is no need for an explicit model of the process by which coarsening occurs. In a comparison with treatment effects, these approaches to rounding correspond to assuming treatment be randomly assigned (exogenous). However, in most experimental or survey applications, these ignorability conditions are shown not to hold

(Wright and Bray, 2003).

There are only a few structural models of rounding in survey responses we are aware of, all of which are parametric. Pudney (2007) develops a two-stage response model in which respondents first choose a response mode (unrounded, rounded, or other heuristics) and then, if they are in the rounding mode, use interval reporting with heterogeneous degrees of coarsening. Kleinjans and van Soest (2014) propose a structural model of the response process in subjective probability questions. Their model allows for rounding (with 50% focal point responses being included separately) and item nonresponse. Ruud, Schunk, and Winter (2014) present a model of rounding in which the degree of coarsening depends on the respondent's uncertainty about the underlying quantity, a notion they support using data from a laboratory experiment where respondent uncertainty could be controlled.

The econometrics literature on measurement error has long stressed the fact that the intuitive attenuation result holds on in simple parametric models and for classical measurement error while simple solutions such as instrumental variables fail in nonlinear models; see Chesher (1991) for a concise statement of the identification problem and Wansbeek and Meijer (2000) for a textbook discussion. More recent research, reviewed by Schennach (2013), explored identification in nonlinear and nonparametric models. Important findings concern identification in the presence of non-classical covariate measurement error using instrumental variables (Hu and Schennach, 2008) and identification in the presence of classical covariate measurement error that do not require any outside information (Schennach and Hu, 2013). Hoderlein and Winter (2010) study the consequences of nonclassical measurement error in the dependent variable when errors are due to imperfect recall; the present paper is related in that it puts structure on the measurement errors that is motivated by the survey response process.

The econometric approach to rounding we propose in this paper is also related to the recent literature in the identification and estimation of treatment effects, e.g., Imbens and Angrist (1994) and Heckman and Vytlacil (2005). Since we employ a binary instrument, and consequently develop our theory for a binary IV, our approach is closer to the former than the latter. It is also related to Melly and Huber (2011) who consider a structural quantile model under a selection mechanism.

Structure of the paper: In the following section, we introduce and analyze the structural model of rounding behavior. Based on the insights we obtain, we discuss identification in the third section. The fourth section is concerned with the application: we introduce the demand setup analyzed in this paper, discuss the data, and implement our approach. Finally, an outlook concludes.

2 Structural implications from a formal model of rounding

The purpose of this paper is to model the impact of rounding due to imperfect recollection of a random variable Y. We propose that rational individuals try to balance the costs and benefits of memorizing to obtain an optimal "amount" of memory, and round if that optimal amount is below a certain threshold.[2] Formally, assume that there is a collection of infinitely many random variables (ξ_s), $\xi_s \in \mathbb{R}$, $s \in \mathbb{R}_{[0,1]}$, each of which can be thought of as giving one standardized "unit of information". Let $\mathcal{F}_m \equiv \sigma\{\xi_s | 0 \leq s \leq m\}$ denote the σ-algebra spanned by $(\xi_s)_{0\leq s\leq m}$, and note that by construction $\mathcal{F}_{m+\delta} \supseteq \mathcal{F}_m \supseteq \mathcal{F}_{m-\delta}, ..., \delta > 0$, i.e., $\{\mathcal{F}_m, 0 \leq m \leq 1\}$ is a filtration. Let $\mathbb{E}[\cdot|\mathcal{F}_m]$ denote the conditional expectation given the σ-algebra $\mathcal{F}_m$. To determine the optimal amount of information, the individual chooses now a finite number m^*, where m^* is defined as

$$m^* = \arg\max_{m\in[0,1]} \mathbb{E}\left\{\mathcal{P}_1\left[L_0 - L\left(Y, \tilde{Y}_m\right)\right] - c\left(m, \mathcal{P}_2\right)|\mathcal{F}_0\right\}, \tag{2.1}$$

where $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2')' \in \mathbb{R}\times\mathrm{P}_2$ is a (possibly infinite dimensional) individual specific parameter which may vary across the population (think of prices), such that $\sigma(\mathcal{P}) \subset \mathcal{F}_0$. Moreover, L is a standard loss function defined on $\mathbb{R}\times\mathbb{R}$ and $\tilde{Y}_m$ denotes the individual's forecast of Y for a fixed sigma algebra $\mathcal{F}_m$. Hence $\tilde{Y}_m = g\left((\xi_s)_{0\leq s\leq m}\right)$, where g is a functional mapping the

[2] The model of memorizing and survey response is related to the notion of rational inattention (Sims, 2003).

process $(\xi_s)_{0\leq s\leq m}$ into $\mathbb{R}$, and $L_0 = \mathbb{E}\left[L\left(Y, \tilde{Y}_0\right)|\mathcal{F}_0\right]$. The leading example is when L is the squared error loss so that $L\left(z, z_m\right) = \left(z - z_m\right)^2$, and $\tilde{Y}_m = \mathbb{E}\left[Y|\mathcal{F}_m\right]$ (i.e., g is the conditional expectation operator). Finally, c is a nonrandom cost function giving the minimal costs of building up memory m for every p_2.

Because of the law of iterated expectations we may rewrite the optimization problem (2.1):

$$m^* = \arg\max_{m\in[0,1]} \left\{\mathcal{P}_1 l_0(m) - c\left(m, \mathcal{P}_2\right)\right\},$$

where

$$l_0(m) = \mathbb{E}\left\{L_0 - \mathbb{E}\left[L\left(Y, \tilde{Y}_m\right)|\mathcal{F}_m\right]|\mathcal{F}_0\right\}.$$

Note that $l_0(\cdot)$ is a monotonically increasing function of m because $\mathbb{E}\left[L\left(Y, \tilde{Y}_m\right)|\mathcal{F}_0\right] \geq \mathbb{E}\left[L\left(Y, \tilde{Y}_{m+\delta}\right)|\mathcal{F}_0\right]$ for every $\delta > 0$ as the set of potential optimizers is increasing. Moreover, we maintain the assumption that $\mathcal{P}_1$ and $\mathcal{P}_2$ have no elements in common which allows us to separate both parts of the optimization problem. This is plausible since the rewards individuals obtain should not enter the cost function. Hence, for fixed m individuals first minimize the expected loss $\mathbb{E}\left[L\left(Y, \tilde{Y}_m\right)|\mathcal{F}_m\right]$ by choosing $\tilde{Y}_m$ for every fixed m, and then pick the m that minimizes the whole expression.

We now discuss the building blocks of the individual's optimization problem.

- $\Pi(m) = \mathcal{P}_1 l_0(m)$ can be interpreted as the profit associated with choosing m. For all commonly used loss functions, Π is a concave function of m. A further implication is that $\mathbb{E}\left[L\left(Y, \tilde{Y}_m\right)|\mathcal{F}_m\right] \leq \mathbb{E}\left[L\left(Y, g\left((\xi_s)_{0\leq s\leq m}\right)\right)|\mathcal{F}_m\right]$, for all other functionals g, and all m. Hence, $\tilde{Y}_m$ is the optimal predictor for fixed m, and (under some differentiability and interiority conditions) the following well known (and principally testable) first order condition holds:

$$\mathbb{E}\left[\partial_{y_m} L\left(Y, \tilde{Y}_m\right)|\mathcal{F}_m\right] = 0. \tag{2.2}$$

- The cost function $c\left(m, p_2\right)$ can be seen as the optimizer of the cost minimization problem of building up memory m, i.e., it solves the problem:

$$\min_{\zeta \geq 0} \; b'\zeta \quad \text{s.t.} \;\; \rho(\zeta, \lambda) \geq m,$$

where $\rho : \mathbb{R}^l \times \mathcal{L} \to (0,1)$ denotes the memory production function that maps the l-vector of input factors $\zeta \in \mathbb{R}^l$ and parameters $\lambda \in \mathcal{L}$ into the unit interval, and $b \in \mathbb{R}^l_+$ denotes the prices associated with these inputs. Note that $p_2 = (b', \lambda')'$, and that because of standard producer theory the factor demands $\zeta = \varphi(m, b, \lambda)$ obtain some structure, e.g. that the matrix of price derivatives for fixed λ and m is negative semidefinite or that demands be zero homogenous in b.[3]

- The parameter $\mathcal{P}$: we have chosen the letter $\mathcal{P}$ to denote parameters to emphasize the economic association between the parameters and prices. An example for $\mathcal{P}_2$ is the price or opportunity cost for the time needed to answer the survey, an example for $\mathcal{P}_1$ is the price (or the reward) an individual obtains from answering correctly. We think primarily of money, as proposed in Philipson (2001) or McFadden (2012).

Our notion of bounded rationality is a formal one, and we believe that individuals still try to behave optimally, given their constraints. This assumption may indeed be criticized as requiring individuals to act overly rational – they have to solve a potentially complicated optimization problem. However, as most economic theory this should be seen as approximation of reality, where individuals choose the effort to "backcast" according to some intention.

The advantages of setting up a formal model instead of a specifying a response heuristic are twofold: First, if individuals act (at least approximately) as our model assumes them to do, then we may obtain testable implications and structural predictions. Testable implications are in particular the rational demand structure on the factors needed to build up memory, as

[3] These two elements formalize the notion of optimization, and make the "economic" association clear. There are also some parallels with existing concepts in statistical decision theory: for fixed m and p, Π_0 is formally similar to the Bayes Risk. However, note that in the Bayesian framework it is a (random) parameter that is of interest whereas in our case it is precisely the random vector Z. Moreover, the dependence on m and the focus on heterogeneity via the parameter vector P is novel.

well as the optimality condition for the optimal backcast. Structural predictions means that we may provide a welfare or money measure of the incentive we would have to provide to improve the individual's response behavior (see also McFadden, 2012). Ultimately, this may allow to assess the total costs of improving the quality of the data, and help set up a decision problem for researchers or survey field agencies that administer household surveys. Since our focus in this paper is on determining the consequences of insufficient information acquisition leading to rounding, we will leave such an analysis for future research.

More important for our analysis is hence the second advantage of this structural modelling approach: It provides economic guidance about variables that should enter the "choice of rounding" equation. As such, it provides an economic rationale for our exclusion restriction. We will use in particular the insight that these excluded exogenous variables are cost-factors in the build-up of memory. In our application, we will follow this guidance and identify such cost factors in a demand dataset.

3 Identification

This section is concerned with modelling the impact of rounding econometrically. The second section already introduced a formal model that argues that some variables impact the choice of effort, i.e., of memory, which governs the question of whether somebody rounds, while not impacting the choice decision. In econometrics terms, an exclusion restriction in the first stage (FS) selection equation is plausible. This section shows that such a restriction can be used profitably to obtain an unbiased estimate of the effect of interest.

3.1 Model and baseline assumptions

Throughout this paper, we postulate that there is a structural model out there, i.e., relationship between variables Y and X, which we want to uncover. Following the recent approach in the nonparametric identification literature, we emphasize the generality of this relationship, as well as the complexity of unobserved heterogeneity, by assuming that

$$Y = \phi(X, A),$$

for a general smooth function ϕ of a (for simplicity scalar) variable of interest X, and a high dimensional vector of unobserved heterogeneity A. The parameter of interest in this framework is

$$\mathbb{E}\left[\partial_x \phi(x, A)\right],$$

for certain values of $x \in \mathcal{X}$, i.e., the average causal marginal effect for individuals with $X = x$ in a heterogeneous population. This parameter is called local average response in Chamberlain (1984), and is related to the LASD of Hoderlein and Mammen (2007). It reduces to standard quantities in textbook models: If the model is linear, i.e., $Y = \beta_0 + X\beta_1 + A$, then it equals β_1. In case of a random coefficient model as in Hoderlein, Klemelä and Mammen (2011), i.e., $Y = X\beta_1(A)$, it becomes $\mathbb{E}\left[\beta_1(A)\right]$. However, at this point we do not want to restrict the structural function ϕ to be of any of these forms, and hence we formulate our model on this general level.

To define the entire framework formally, we make use of the following set of assumptions. For ease of notation we suppress the dependence on S :

Assumption 1. *Let $(\Omega, \mathcal{F}, P)$ be a complete probability space on which are defined the random vectors $(A, V) : \Omega \to \mathcal{A} \times \mathcal{V}$, $\mathcal{A} \subseteq \mathbb{R}^\infty, \mathcal{V} \subseteq \mathbb{R}$ and $(Y^*, X, Z) : \Omega \to \mathcal{Y}^* {\times} \mathcal{X} \times \mathcal{Z}$, $\mathcal{Y}^* \subseteq \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}, \mathcal{Z} = \{0, 1\}$, such that (i) $\mathbb{E}(Y^*) < \infty$; (ii)*

$$\begin{aligned} Y^* &= \phi(X, A) \\ D &= \mathbb{I}\left\{P(Z) < V\right\} \\ Y &= Y^* D + g(Y^*)(1 - D), \end{aligned}$$

where $\phi : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}^*$, $g : \mathcal{Y}^* {\times} \{0, 1\} \to \mathcal{Y}$ and $P : \mathcal{Z} \to \mathcal{X}$ are bounded Borel measurable function; and (iii) realizations of (Y, X, Z) are observable, whereas those of (A, V) are not.

Assumption 2. *(A, V) are independent of Z, X.*

Assumption 3. *V is absolutely continuous with respect to Lebesgue measure, s.th. $V|Z \backsim \mathcal{U}[0,1]$.*

Assumption 4. *ϕ is differentiable in x, with continuous and bounded first derivative. $g(Y^*) \neq Y^*$. Moreover, $\partial_x \phi$ is square integrable and uniformly bounded.*

Discussion of assumptions: The first of these assumptions defines the econometric structure of our model. Specifically, individuals provide the correct answer Y^* if memory is above a specific threshold; else, they provide a distorted answer, $g(Y^*)$. For instance, suppose the individuals choose values r_l (say, 50) if $Y^* \leq 75$ and r_u (say, 100) if $Y^* > 75$. Then, $g(Y^*) = 50\mathbb{I}\{Y^* \leq 75\} + 100\mathbb{I}\{Y^* > 75\}$, providing a strong version of rounding.

The second assumption specifies the dependence structure in our model. In particular, neither X nor Z are correlated with the error term; however, the fact that the population who rounds is not a random selection of the entire population causes a distortion. The third assumption in connection with the first specifies the unobservable in the selection equation to enter additively separable and be uniformly distributed. Heckman and Vytlacil (2005) contain a lucid discussion of this issue in the context of treatment effect models; in the case of binary treatment this specification turns out to be equivalent to instrument monotonicity in the LATE framework of Imbens and Angrist (1995). This assumptions implies $P(Z) = \mathbb{P}(D = 1|Z)$, and is common in the literature. In abuse of notation, we will write P instead of $P(Z)$ henceforth. Finally, the fourth assumption specifies the functions, in particular differentiability of ϕ and nontriviality of g.

3.2 What if we ignore rounding?

The first step is to ask what the mean regression identifies, if we simply ignored rounding. The following argument is instructive to understand the various effects of rounding. For simplicity, we focus on the case where $g(Y^*) = r_u\mathbb{I}\{Y^* > c\} + r_l\mathbb{I}\{Y^* \leq c\}$, i.e., rounding individuals choose r_u if they are above c, and r_l if they are below. This can be written as

$$g(Y^*) = r_l + (r_u - r_l)\mathbb{I}\left\{Y^* > c\right\},$$

and for simplicity, we choose $c = (r_u + r_l)/2$. Observe that r_u, r_l are known constants, and hence so is c. In our application r_l is 50, r_h 100, and $c = 75$.

Consider now the empirical regression of Y on $X = x$. This produces

$$\begin{aligned}
& \mathbb{E}\left[Y|X=x\right] \\
= \ & \mathbb{E}\left[Y^*D|X=x\right] + r_l\mathbb{P}\left[D=0|X=x\right] + (r_u - r_l)\mathbb{E}\left[\mathbb{I}\left\{Y^* > c\right\}(1-D)|X=x\right] \\
= \ & \mathbb{E}\left[\phi(x,A)D|X=x\right] + r_l\mathbb{P}\left[D=0|X=x\right] + (r_u - r_l)\mathbb{E}\left[\mathbb{I}\left\{\phi(x,A) > c\right\}(1-D)|X=x\right],
\end{aligned}$$

This expression is intransparent, and in order to make progress, we assume in addition that the unobserved heterogeneity is additively separable, i.e., $\phi(x,a) = m(x) - a$.[4]

$$\begin{aligned}
\mathbb{E}\left[Y|X=x\right] &= m(x) + (r_l - m(x))\mathbb{P}\left[D=0|X=x\right] \\
&\quad -\mathbb{E}\left[AD|X=x\right] \\
&\quad +(r_u - r_l)\mathbb{E}\left[\mathbb{I}\left\{A \leq m(x) - c\right\}(1-D)|X=x\right]
\end{aligned}$$

Differentiating wrt x produces

$$\begin{aligned}
\partial_x\mathbb{E}\left[Y|X=x\right] &= m'(x)(1 - \mathbb{P}\left[D=0|X=x\right]) \\
&\quad +(r_l - m(x))\partial_x\mathbb{P}\left[D=0|X=x\right] \\
&\quad -\partial_x\mathbb{E}\left[AD|X=x\right] \\
&\quad +(r_u - r_l)\partial_x\mathbb{E}\left[\mathbb{I}\left\{A \leq m(x) - c\right\}(1-D)|X=x\right]
\end{aligned}$$

To make further progress, we invoke the (empirically testable) assumption that A, D indepen-

[4] Note that the under additive separability, the additive term could have been $\alpha(A)$, but then we could relabel $\alpha(A) = \tilde{A}$, implying that this formulation is without loss of generality.

dent of X, i.e., that rounding and X are not associated. Then we obtain that

$$\begin{aligned}\partial_x \mathbb{E}\left[Y|X=x\right] &= m'(x)\mathbb{P}\left[D=1\right] \\ &\quad +(r_u - r_l)\partial_x \mathbb{E}\left[\mathbb{I}\left\{A \le m(x) - c\right\}(1-D)\right]\end{aligned}$$

Adding the observation that

$$\mathbb{E}\left[\mathbb{I}\left\{A \le m(x) - c\right\}(1-D)\right] = F_{A|1-D}(m(x)-c;1)\mathbb{P}\left[D=0\right],$$

we obtain

$$\begin{aligned}\partial_x \mathbb{E}\left[Y|X=x\right] &= m'(x)\mathbb{P}\left[D=1\right] + (r_u - r_l)m'(x)f_{A|1-D}(m(x)-c;1)\mathbb{P}\left[D=0\right] \\ &= m'(x)\left\{\mathbb{P}\left[D=1\right] + (r_u - r_l)f_{A|1-D}(m(x)-c;1)\mathbb{P}\left[D=0\right]\right\}.\end{aligned}$$

This result gives a clear idea about the two effects of rounding. The first equality decomposes the effect into two separate terms. The first term is associated with the rounding population - only a proportion of $\mathbb{P}\left[D=1\right]$ displays the original effect. The second term gives the distortion of effects in the rounded sample, and is related to the difference between the two focal values, $(r_u - r_l)$, and $f_{A|1-D}(\cdot;1)$, the density of outcome unobservables given rounding, i.e., how the distribution of A changes in the rounders subpopulation. Note that the first term could be seen as a direct effect of rounding because only parts of the population round, while the second acts more like a selection effect: as x changes some of the individuals in the rounder group switch from the lower focal answer r_l to the upper r_u, these individuals have a certain value of a. As is obvious from this expression, if $(r_u - r_l)f_{A|1-D}(m(x)-c;1) > 1$, the structural marginal effect is magnified, while otherwise it is attenuated. Obviously, neither the direction nor the magnitude of bias are clear, though at least the sign does not change. However, in areas of $\mathcal{A}$ where $f_{A|1-D}$ is small, we are likely going to see attenuation. If we dispense with some of the simplifying assumptions, we obtain even more bias terms, and not even the sign needs to be preserved.

Note that just taking the non-rounders does not solve the problem in general, since

$$\mathbb{E}\left[Y|X=x,D=1\right]=\mathbb{E}\left[Y^{*}|X=x,D=1\right]=\mathbb{E}\left[\phi(x,A)|X=x,D=1\right],$$

and hence

$$\partial_{x}\mathbb{E}\left[Y|X=x,D=1\right]=\mathbb{E}\left[\partial_{x}\phi(x,A)|X=x,D=1\right]+\mathbb{E}\left[\phi(x,A)\partial_{x}S|X=x,D=1\right],$$

where S is the score $\log f_{A|X,D}(A;X,1)$ for the non-rounders. The second term is called the "heterogeneity bias" in Chamberlain (1984). If A is independent of X given D (which would be implied by joint independence of A,D from X), the second term vanishes as the score is not a function of x, and the first term becomes $\mathbb{E}\left[\partial_{x}\phi(x,A)|X=x\right]$. However, this means that there is no selection effect; the non-rounders are, in terms of their unobservables, like the population at large. Else, the first term will generally depend on D, and S will not be zero in general. The decisive issue that clarifies whether we can simply use the non-rounders is hence whether we believe there to be selection, or whether we think of the non-rounders as essentially the same individuals as the rounders, at least in terms of the unobservables that govern the outcome equation. Still, even in the case where we believe there not be a selection issue, throwing away all rounders may be very wasteful in terms of observations, and being able to say something about them may be beneficial.

3.3 How to account for selection induced by rounding

There are many applications where the researcher believes that the non-rounders are a selected sample, and we provide one such example in the application. To deal with this aspect, we make use of an IV identification strategy not unlike LATE, Imbens and Angrist (1994). Borrowing the counterfactual notation $D_1=\mathbb{I}\left\{P(1)<V\right\}$ and $D_0=\mathbb{I}\left\{P(0)<V\right\}$, and noting that the set $D_0>D_1$ defines the so-called (subpopulation of) compliers according to our treatment definition, we employ the model as defined in assumption 1. In this setup, we can think of rounding as a participation in a treatment, which individuals make in a first stage decision. The causal effect of rounding on the outcome variable is then easily seen to be a conditional (on X) LATE, i.e.,

$$\frac{\mathbb{E}\left[Y|X=x,Z=1\right]-\mathbb{E}\left[Y|X=x,Z=0\right]}{\mathbb{E}\left[D|X=x,Z=1\right]-\mathbb{E}\left[D|X=x,Z=0\right]}=\mathbb{E}\left[Y^{*}-r_{l}-(r_{u}-r_{l})\mathbb{I}\left\{Y^{*}>c\right\}|X=x,D_{0}>D_{1}\right],$$

and whether or not the conditional LATE is zero, i.e., rounding has no effect on the conditional mean on average, depends on whether or not $\phi(x,A)-r_{l}-(r_{u}-r_{l})\mathbb{I}\left\{\phi(x,A)>c\right\}$ is zero on average for the complier subpopulation.

More interesting than quantifying this effect - and at the center of this paper - is to be able to obtain an unbiased estimate of the average causal effect, $\partial_{x}\phi(x,a)$. By standard arguments from the treatment effect literature, one can show that

$$\begin{aligned}\psi_{NR}(x) &= \frac{\mathbb{E}\left[YD|X=x,Z=1\right]-\mathbb{E}\left[YD|X=x,Z=0\right]}{\mathbb{E}\left[D|X=x,Z=1\right]-\mathbb{E}\left[D|X=x,Z=0\right]}\\ &= \mathbb{E}\left[\phi(x,A)|X=x,D_{0}>D_{1}\right]\\ &= \mathbb{E}\left[\phi(x,A)|D_{0}>D_{1}\right],\end{aligned}$$

because of X indep of V,A, implying that

$$\psi_{NR}'(x)=\mathbb{E}\left[\partial_{x}\phi(x,A)|D_{0}>D_{1}\right],$$

is the average causal effect for compliers. This quantity solves the selection problem associated with rounding, and also the direct impact of rounding, but at the expense of throwing away all rounders, potentially a large fraction of the population. The obvious question is then how to use the rounders as well. To this end, we will largely use the specification $\phi(x,a)=m(x)-a$, so that

$$\psi_{NR}(x)=m(x)-\mathbb{E}\left[A|D_{0}>D_{1}\right],$$

and add the error location normalization $\mathbb{E}\left[A|D_{0}>D_{1}\right]=0$.

3.4 Using the rounders

By similar arguments as in the previous subsection,

$$\psi_R(x) = \frac{\mathbb{E}\left[Y(1-D)|X=x,Z=1\right] - \mathbb{E}\left[Y(1-D)|X=x,Z=0\right]}{\mathbb{E}\left[D|X=x,Z=1\right] - \mathbb{E}\left[D|X=x,Z=0\right]} = r_l + (r_u - r_l)\mathbb{E}\left[\mathbb{I}\left\{\phi(x,A) > c\right\}|D_0 > D_1\right]$$

To make further progress, we impose the additional structure, $\phi(x,a) = m(x) - a$. We then obtain that

$$\psi_R(x) = r_l + (r_u - r_l)F_{A|D_0>D_1}(m(x) - c).$$

Assuming as above that $c = (r_u + r_l)/2$, a known constant, letting $\tilde{m}(x) = m(x) - c$, this becomes a standard single index model, i.e.,

$$\frac{\psi_R(x) - r_l}{(r_u - r_l)} = F_{A|D_0>D_1}(\tilde{m}(x)).$$

Usually, identification is resolved at this stage by normalization. We have to be careful at this point, however, because the non-rounders already yield identification of the model. In particular, note that under the specification $\phi(x,a) = m(x) - a$, arguments as in the previous subsection yield that $F_{A|D_0>D_1}(a)$ is identified through

$$\lambda(x) = \frac{\mathbb{E}\left[\mathbb{I}\left\{Y - m(x) < a\right\}D|X=x,Z=1\right] - \mathbb{E}\left[\mathbb{I}\left\{Y - m(x) < a\right\}D|X=x,Z=0\right]}{\mathbb{E}\left[D|X=x,Z=1\right] - \mathbb{E}\left[D|X=x,Z=0\right]},$$

from the non-rounder sample, and thus can be treated as known. If A is continuously distributed for the compliers, we hence obtain that

$$m(x) = (r_u + r_l)/2 + F^{-1}_{A|D_0>D_1}\left(\frac{\psi_R(x) - r_l}{(r_u - r_l)}\right)$$

in the rounder subsample. This opens up the way for nonparametric estimation, and this is indeed the route that we consider in the application. For more general cases of rounding, however, we suggest a more semiparametric estimator. To understand its' structure, we have to first understand the general structure of identification.

3.5 More than two rounding values, but one degree of rounding

In this subsection, we introduce the first generalization. We allow for more than two focal values, but the degree of rounding stays the same. For instance, individuals round to the values $50, 100$ and 150, so the steps between rounded values are known. Observe that the general model structure does not change, i.e, we still have

$$\begin{aligned} Y^* &= \phi(X, A) \\ D &= \mathbb{I}\left\{P(Z) < V\right\} \\ Y &= Y^* D + g(Y^*)(1-D), \end{aligned}$$

but the function $g(y^*)$ is now given by

$$g(Y^*) = \sum_{k=0,\ldots,K} r_k \mathbb{I}\left\{Y^* \in I_k\right\},$$

where $I_0 = (-\infty, y_1)$, $I_K = [y_K, \infty)$, and for $k = 1, ..., K-1$, we have $I_k = [y_k, y_{k+1})$. Moreover, note that $y_k - y_{k-1} = c$, where c is known constant, independent of k (in our example, 50), and note that $r_k = (y_{k+1} + y_k)/2$, $r_0 = r_1 - c$, and $r_K = r_{K-1} + c$. Finally, for ease of notation, we choose the specification $\phi(x, a) = \mu(x) + a$, which is without loss of generality compared to the previous section, and just employs a different normalization.

We first note, that the analysis of nonrounders does not change, including the fact that (and the means by which) we can obtain $F_{A|D_0>D_1}$. Next, for the rounders, we obtain again by standard arguments, that

$$\begin{aligned} \psi_R(x) &= r_K + \sum_{k=1,\ldots,K-1} (r_{k+1} - r_k) F_{A|D_0>D_1}(r_k - m(x)) \\ &= r_K + c \sum_{k=1,\ldots,K-1} F_{A|D_0>D_1}(r_k - m(x)), \end{aligned}$$

implying that

$$\frac{\psi_R(x) - r_K}{c} = \sum_{k=1,\dots,K-1} F_{A|D_0>D_1}(r_k - m(x)).$$

Since all of these objects are known (r_K and c), or identified from data ($\psi_R, F_{A|D_0>D_1}$), it gives us a second, usually numerical, way to solve for $m(x)$. We can use this result to estimate m.

Finally, what if people round to different degrees, e.g., part of the population do not round at all, some round to the nearest 10, some to the nearest 50? Obviously, this complication corresponds to multiple treatment with endogeneity. We conclude from the analysis of Imbens and Angrist (1995) that there is no good solution in this setup, which is why we defer this topic until further progress has been made in the treatment effects literature.

3.6 Using the entire sample in a semiparametric estimator

While this shows identification using information from both rounded and nonrounded observations as embodied in ψ_R and ψ_{NR} in a nonparametric fashion, we now propose to use a parametric specification, i.e., $m(x) = x'\theta_1$. Note that

$$\psi_{NR}(x;\theta_1)\mathbb{P}(NR) + \psi_R(x;\theta)(1 - \mathbb{P}(NR)),$$

where $\mathbb{P}(NR)$ is the fraction of nonrounders in the population, and

$$\psi_R(x,\theta) = r_K + c \sum_{k=1,\dots,K-1} G(r_k - x'\theta_1;\theta_2),$$

and G is a parametric cdf, e.g., probit (normal cdf) with parameter θ_2. The estimators are then obtained as minimizers of the distance between nonparametric estimates of ψ_{NR} and ψ_R and this expression, i.e.

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta\in\Theta}\{\int \left(\hat{\psi}_{NR}(x) - x'\theta_1\right)^2 \pi(NR) \\ &\quad + \left(\hat{\psi}_R(x) - r_K + c \sum_{k=1,\dots,K-1} G(r_k - x'\theta_1;\theta_2)\right)^2 (1 - \pi(NR))f_X(x)dx\} \end{aligned}$$

where $\pi(NR)$ is the fraction of nonrounders in the sample.

4 Application: Correcting for rounding in consumer demand

4.1 A quick comparison with the literature

Applications involving consumer demand have a long history in economics, and date back at least to the early work of Stone (1954). Key milestones were the (parametric) flexible functional forms demand systems (e.g., the Translog, Jorgenson et al.,1980; the Almost Ideal, Deaton and Muellbauer, 1980; and the extension by Blundell, Banks and Lewbel, 1996). Most of these approaches use budget shares, i.e., they divide the expenditure for a given product by the total expenditure for all nondurable products. The same is true for nonparametric approaches that are close to our model, because they involve heterogeneity in an explicit fashion, e.g., Lewbel (2001) and Hoderlein (2011). Since the divisions by total expenditure have the tendency to obscure rounding, like in the early work of Stone (1954) and Jorgenson et al (1982), we use total expenditure. To the best of our knowledge, there are no papers in consumer demand that attempt to correct for rounding; as already mentioned, it is standard practice to form budget shares and subsequently ignore the problem. Since we focus on food consumption outside the home, however, our result can be compared with many papers. A common finding is that this aggregate good is a luxury, see, e.g., Lewbel (1999) for an overview. Hence, we expect similar quantitative results.

4.2 Data and data clearance

To illustrate our method, we use data from the Health and Retirement Study (HRS) and from the Consumption Activity Mail Survey (CAMS). The HRS is a longitudinal panel study that biennially surveys a representative sample of US-Americans aged 50+. The HRS collects information on various topics of US daily life, including health and cognition, income and expenditures. In off years, the HRS collects information on household consumption and spending in a supplemental survey, the Consumption and Activities Mail Survey (CAMS). Both data sets have been used before for the analysis of self-reporting errors in surveys (see Manski and Molinari, 2010; Hoderlein and Winter, 2010). We analyze data from the 2006 wave of the HRS, and from the 2007 wave of CAMS. In 2006, the HRS interviewed 18,469 individuals from 12,288 households, and 4,572 individuals (3,392 households) were also participating in the CAMS survey 2007.

In this application, the outcome of interest is the self-reported amount that respondents spend per week on eating food outside home, which is our Y variable. The variable is elicited in the following way: "About how much do you (and other family members living there) spend eating out in a typical week, not counting meals at work or at school?". Memorizing the exact quantity in an interview situation requires a considerable amount of effort. Thus, instead of reporting exact expenditures, respondents may decide to facilitate the answering process and provide a rounded value. Depending on the true latent amount of these expenditures people may round to different focal values such as 10, 25, 50, 100, 150 and so on. This implies that respondents use different rounding strategies e.g. multiple of USD 50 for different expenditure levels.

We restrict the sample to only individuals who report food outside expenditures between USD 25 and USD 125 for several reasons. First, by only considering focal values of USD 50 and USD 100 we reduce the complexity involved by different rounding strategies, e.g. rounding to a multiple of 10. Such respondents may use strategies which is beyond the scope of our theoretical model. An extension of the model that deals with multivariate rounding strategies is a topic for future research. Second, our range of values excludes the possibility to round down to zero expenditures, preventing us from running into additional selection issues between the two populations. Third, most data lie between USD 20 and USD 100 (see figure 1). Using the entire range of data implies that we do not have enough nonrounder observations for high expenditure levels. The range of values between 25 and 125, however, provides sufficient data points to make the rounder and the nonrounder population comparable to each other.

The rounding mechanism which is assumed in our model is as follows: A respondent strictly rounds up to the value of 50 if the true value for food expenditures lies in the interval $[26, 49]$, and she rounds off for true values of $[51, 75]$. Using the identical strategy reporting food expenditures of 100 is the result of rounding up for values $[76, 99]$ or rounding off for values $[101, 125]$. For a true, latent value between $r_l = 50$ and $r_u = 100$, the decision for rounding up or off is made at the threshold value $c = \frac{r_l+r_l}{2} = 75$[5].

In terms of the causal explanatory variable in whose effect we are interested in, a subset of X in our notation, we use log total weekly expenditure, again in line with the literature (see Lewbel (1999)). We drop outliers in total weekly spending reports (upper and lower 0.2% percentile). In order to control for household characteristics, we compute an indicator from valid reports on the respondents' marital status, race, gender, age and labor force status[6]. As a third control, we include the total interview time in our analysis. Eventually, we obtain an analytic sample of 2,467 individuals. These three variables together form the set of explanatory variables, i.e., X in our notation. A would then be unobserved heterogeneity that shifts the causal relationship.

As introduced in the section before, D denotes the treatment indicator for non-rounding. It takes the value 1 if an individual reports any exact value but the focal values 50 and 100, and is zero otherwise. Rounding is not exogenously determined but the result of a decision process, leading to a selection bias. We instrument the treatment by exploiting a unique feature of our data. In 2006, the HRS randomly assigns respondents to a module on physical health measures and biomarkers, such as blood samples or high blood measures. While about 50 percent of the respondents participate in this module in 2006, the other 50 percent received the identical module in 2008. This induces extra time to the total interview time, thus increasing the respondents' costs of memorizing the exact value of weekly food expenditures. The additional costs may decrease the amount of information provided by respondents. Accordingly, respondents who entered the module on physical health measures in 2006 are assumed to hold a higher

[5] We neglect the case of USD 50 or USD 100 being exact expenditures for food outside home.

[6] Dimensionality of the single covariates is reduced by conducting a principal component analysis, providing us with a single continuous measure of household characteristics.

probability of rounding in food expenditures in 2006 than those who receive the module two years later[7]. This defines our instrument Z, and since the allocation was random, we feel that the exogeneity assumption is well justified.

4.3 Econometric specifications

To formalize this issue, we use precisely the specification outlined in assumption 1 in conjunction with the other assumptions. We then use this structure to correct for the influence of selection associated with rounding by first estimating the structural function

$$\psi_{NR}(x) = \frac{\mathbb{E}\left[YD|X=x, Z=1\right] - \mathbb{E}\left[YD|X=x, Z=0\right]}{\mathbb{E}\left[D|X=x, Z=1\right] - \mathbb{E}\left[D|X=x, Z=0\right]}$$

All conditional expectations in this expression are estimated by nonparametric local linear regressions, using a second order Gaussian Kernel[8]. More specifically, we first split the sample according to whether respondents participated in the 2006 module of health measures and biomarkers, $Z = 1$, or not, $Z = 0$, and then apply local polynomial regression, using $\widetilde{Y}_{NR} = YD$ (and analogously $\widetilde{Y}_R = Y(1-D)$ later for $\psi_R(x)$) as new dependent variables. The estimates for $\psi_{NR}(x)$ are then obtained by forming the ratio of the numerator and the denominator at a grid of x-values containing 45 values for log total spending, and at the sample mean of household characteristics and the total interview time. Clearly, the estimation of $\psi_{NR}(x)$ is based on the entire sample of HRS respondents, rounder and nonrounders. For the sake of brevity, because we have only nonzero values for nonrounders in this expression, we refer to estimates of $\psi_{NR}(x)$ as to nonrounder regressions, while we define estimates for $\psi_R(x)$ as rounder regressions. As regards the details of $\hat{\psi}_{NR}(x)$, the selected bandwidths are provided by table 1.

This provides us with an estimate for the structural demand function m, using $\hat{m} = \hat{\psi}_{NR}(x)$.

The estimation of $\hat{\psi}_R(x)$ proceeds analogously. However, for the estimation of m using the rounders equation, as outlined in section 3.4 we require first an estimate of $F_{A|D_0>D_1}(a)$. This is

[7] It is important to note that the physical health measures were taken before respondents were asked about their weekly food expenditures.

[8] Nonparametric estimation was performed using the "np" package for the statistical software R (Hayfield and Racine, 2008).

Table 1: Bandwidth selection for the nonparametric estimation of $\psi_{NR}(x)$

	log total spending	HH characteristics	total interview time
$E(YD\|X=x,Z=1)$	2.00	1.30	1300
$E(YD\|X=x,Z=0)$	2.40	2.20	2100
$E(D\|X=x,Z=1)$	1.80	0.70	1600
$E(D\|X=x,Z=0)$	1.9040	1.20	2800

straightforwardly obtained from the residuals in the nonrounders regression, i.e., using the fact that $\hat{m} = \hat{\psi}_{NR}$, we calculate $Y_i - \hat{\psi}_{NR}(X_i) = \hat{A}_i$, providing us with the sample residual distribution. Using the values of equispaced five percent percentiles, $a_k = 5, 10, 15, ..., 85, 90, 95$, as thresholds, we compute 19 dummy variables which take the value zero as long as the individual residual value $\hat{A}_i$ is absolutely larger than the percentile residual value a_k. Multiplying with the treatment dummy D allows us to estimate $\hat{\lambda}(x) = \hat{\lambda}_{a_1}(x), \hat{\lambda}_{a_2}(x), \ldots, \hat{\lambda}_{a_{19}}(x)$[9]. Using the results of the rounder regression, the focal values $r_l = 50/r_u = 100$, and the threshold value $c = \frac{r_l + r_u}{2} = 75$, we transform each value of $\hat{\psi}_R(x)$ to a probability value and map this to the corresponding percentile value of the unobserved heterogeneity distribution $F_{A|D_0>D_1}$. Finally, we compute

$$\hat{m}_R(x) = (r_u + r_l)/2 + \hat{F}^{-1}_{A|D_0>D_1}\left(\frac{\hat{\psi}_R(x) - r_l}{(r_u - r_l)}\right)$$

4.4 Empirical Results

We start by considering figure 2, which illustrates the first stage results. It shows estimates of $E(D|X = x, Z = 0)$ and $E(D|X = x, Z = 1)$, which are the building block for the difference $E(D|X = x, Z = 1) - E(D|X = x, Z = 0)$ in the denominator. The solid line corresponds to respondents who receive the module on physical health measures and biomarker in 2008. For this group of respondents the probability of reporting an exact value is always higher than for the other group. The differences between the two estimates is negative and between 5–12 percent, depending on the level of total weekly expenditures. It suggests that the extra interview time devoted to this module increases the probability of reporting rounded values which is consistent

[9] As for the estimation of the structural function we use a nonparametric local linear estimator with a second order Gaussian Kernel. The bandwidth values of the denominator are the same as in table 1. For each λ_{a_k} with $k = 1, \ldots, 19$ an arbitrary bandwidth was selected.

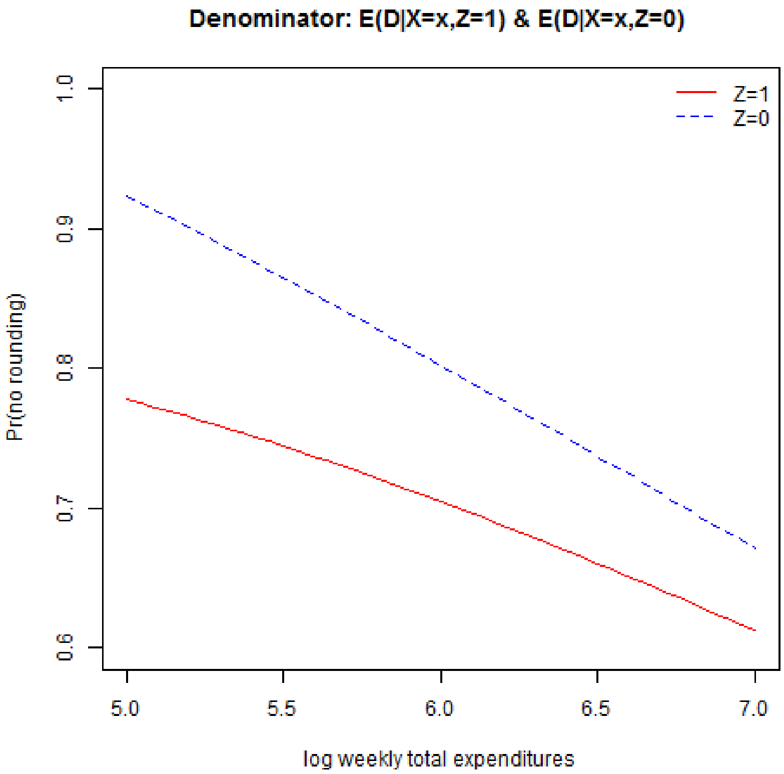

Figure 2: First stage results from nonparametric regression on the probability of not rounding

with our assumption of increasing costs of memorizing. Moreover, it also shows that higher total expenditures, which are associated with higher life cycle income, increase the probability of rounding as well, indicating higher opportunity costs (recall that we are conditioning on socio-economic characteristics such as education). The decrease in the difference between the two probabilities makes sense as well, if one assumes decreasing marginal effects of opportunity costs. Finally, note also that our instrument is quite informative, as the relative magnitude of the change is reasonably large (e.g., an decrease in the probability from 0.85 to 0.75 is substantial).

We next move to the results that show $\hat{\psi}_{NR}$ and $\hat{\psi}_R$. Both, the denominator and the numerator take negative values, so that $\hat{\psi}_{NR}$ and $\hat{\psi}_R$ are strictly positive. Note that under our assumptions, the former provides and estimate of the structural function m, while the latter does not have a structural interpretation. It is hence instructive to look at the two graphs in comparison. The left graph in figure 3 shows the results of the nonrounder regression. For

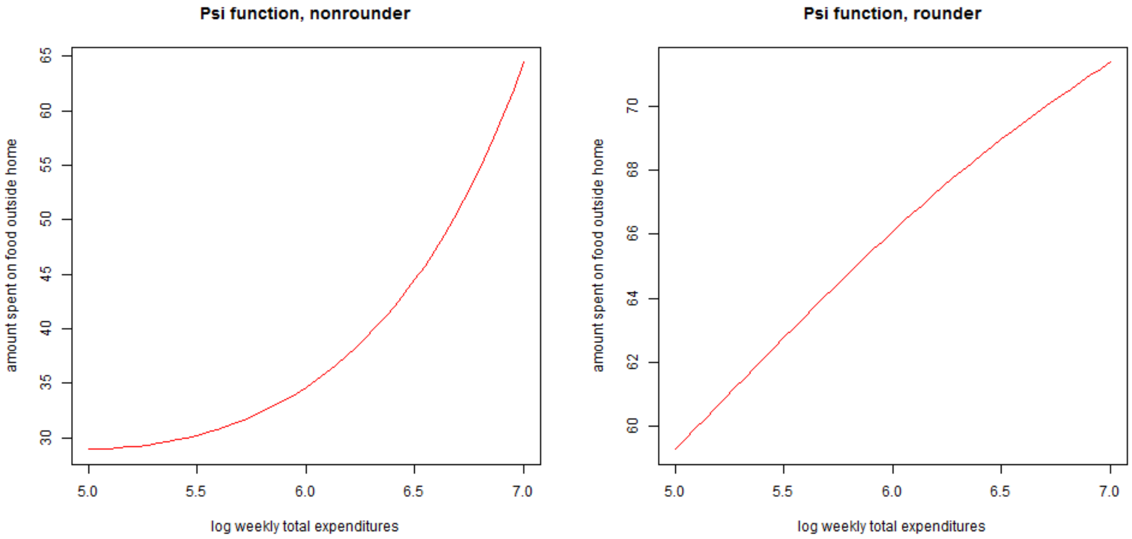

Figure 3: Results nonparametric (non)rounder regressions

low expenditures the amount spent on food outside home raises only moderately with an one percent increase in total expenditures. In contrast, the reaction of food expenditures to an one percent increase in total expenditures is very strong for high total expenditures. Altogether, the estimated function $\hat{\psi}_{NR}$ indicates that the demand for food outside home increases more than proportionally with raising total expenditures. Thus, eating outside home can be characterized as a luxury good, which makes a lot of sense and is in line with other findings in the literature, see Lewbel (1997). Note that in our framework $\hat{\psi}_{NR}(x)$ has a structural interpretation since it solves the selection problem associated with rounding and is not affected by the direct impact of rounding. However, the disadvantage is that the rounder information is not properly used[10].

From the right graph in figure 3 it becomes clear that the estimation of the relationship between weekly expenditures for eating outside home and log total weekly expenditures is different for the rounder regression. First, the values of $\hat{\psi}_R(x)$ range between roughly 55 and 70, while the range for $\hat{\psi}_{NR}(x)$ is roughly between 30 and 65. Second, the functional form is almost linear, indicating that $\hat{\psi}_R(x)$ does not identify the structural relation between food spending and total expenditures in the rounder sample. This illustrates nicely that even accounting for the selection effect of rounding, there is a pronounced difference between $\hat{\psi}_R(x)$ and the

[10] The fraction of rounder is significant. In our empirical example dropping rounders corresponds to excluding about 33 percent of the sample, raising general concerns of representativeness.

structural function which is due to the bias associated with rounding: Since the estimate is a weighted average of the values of the rounded observations, i.e., 50 and 100, it is bound to be confined to lie in this interval (as would any mean regression). However, the estimate of the structural function $\hat{m} = \hat{\psi}_{NR}$ reveals that there are far more observations that are actually rounded up to 50 (from below), then rounded off to 50 (or to 100), especially at lower incomes. As such, we have a sizable distortion in particular at the lower end, which also conflicts with the character of eating outside home as a luxury good. Indeed, if $\hat{\psi}_R(x)$ is true consumers would almost uniformly spend around 65 USD a week, even if their income triples. This illustrates nicely that the rounders results cannot be used directly, and motivates the need for a procedure like the one advocated in this paper.

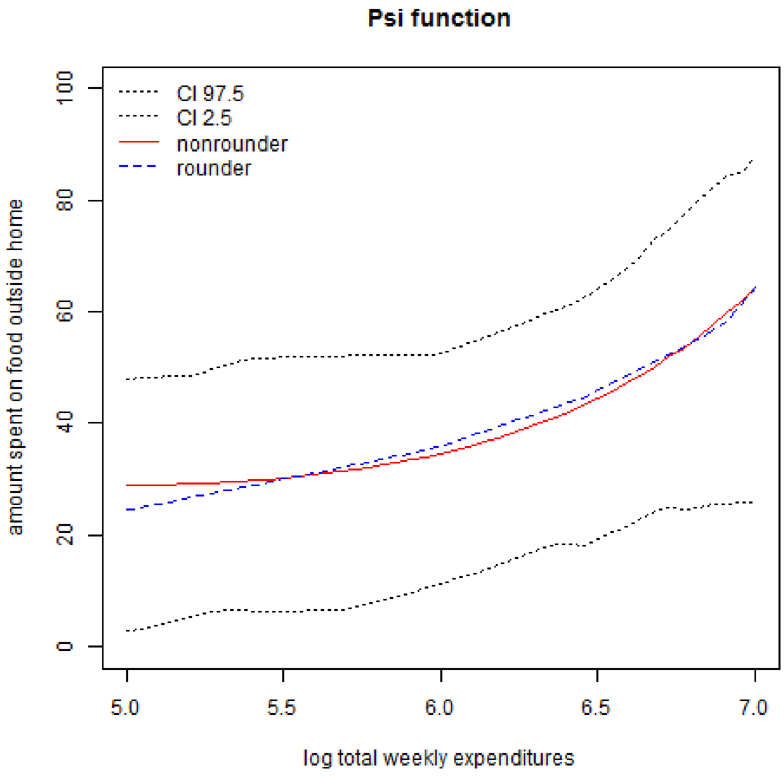

Figure 4: $\hat{m}(x)$ and $\hat{\psi}(x)$, with 95% c.i. for nonrounder regression

We now turn to the estimation of m using the nonrounders regression $\hat{\psi}_R(x)$ in the way outlined in the previous subsection. Figure 4 presents the corrected function $\hat{m}(x)$, and the results of the nonrounder regression as well as the naive bootstrap 95 percent confidence intervals for $\hat{\psi}_{NR}(x)$ [11]. The results show that the proposed correction method that employs the estimated cdf of the residuals, produces coherent results: The estimated relationship between spending for food outside home and total weekly expenditures, as expressed by $\hat{\psi}_{NR}$, is very close to the relationship from the rounder regressions, and – together with economic plausibility arguments – raises the confidence that the result really displays the structural relationship. A final estimate of m may be obtained as a pointwise weighted average of these two regressions with weights that are inversely related to the pointwise variance, but since they do not differ much, we desist from displaying this here.

5 Conclusion

This paper introduces a flexible framework that allows to consider both the selection and the information reduction aspect of rounding in survey responses. We provide a formal model that explains why individuals round. The model introduces a structural Roy-type cost-benefit analysis, which suggests the use of cost factors as instruments in a threshold crossing treatment effects approach. We develop this approach by showing identification first in the simplest and stylized case and then extend our insights to a more realistic setting. Finally, we apply our approach to a consumer demand problem.

An open issue in this framework is how to deal with different degrees of rounding, which corresponds to the multivalued treatment case. Since this case is poorly understood in the treatment effects literature, we presume that a general partial identification approach may be pursued in our setup as well.

[11] Confidence intervals of $\hat{\psi}_{NR}(x)$ are computed by bootstrap resampling (100 times). We trim the denominator in the bootstrap estimations against zero, using a cut off value of 0.05.

References

Battistin, E., R. Miniaci, and G. Weber (2003): What do we learn from recall consumption data? *Journal of Human Resources*, 38(2), 354–385.

Chesher, A. (1991): The effect of measurement error. *Biometrika*, 78(3), 451–462.

Deaton, A. and J. Muellbauer (1980): An almost ideal demand system. *American Economic Review*, 70, 312–326.

Hayfield T. and J. S. Racine (2008): Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Heckman, J. J. and E. Vytlacil (2005): Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73, 669–738.

Heitjan, D. F. (1994): Ignorability in general incomplete-data models. *Biometrika*, 81, 701–708.

Heitjan, D. F. and D. B. Rubin (1991): Ignorability and coarse data. *Annals of Statistics*, 19(4), 2244–2253.

Hoderlein, S. (2011): How many consumers are rational? *Journal of Econometrics*, 164, 294–309.

Hoderlein, S. and E. Mammen (2007): Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75, 1513–1518.

Hoderlein, S., J. Klemelä and E. Mammen (2010): Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26, 804–837.

Hoderlein, S. and J. Winter (2010): Structural measurement errors in nonseparable models. *Journal of Econometrics*, 157, 432–440.

Hu, Y. and S. M. Schennach (2008): Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1), 195–216.

Huber, M. and B. Melly (2011): Quantile regression in the presence of sample selection. Discussion Paper no. 2011-09, St.Gallen.

Hurd, M., and S. Rohwedder (2005): The Consumption and Activities Mail Survey: Description, data quality, and first results on life-cycle spending and saving. Working Paper, RAND, Santa Monica.

Imbens, G. W. and J. D. Angrist (1994): Identification and estimation of local average treatment effects. *Econometrica*, 62, 467–475.

Imbens, G. W. and J. D. Angrist (1995): Two-stage least squares estimation of average causal effect models with variable treatment intensity. *Journal of the American Statistical Association*, 90, 431–442.

Jorgenson, D., L. Lau and T. Stoker (1982): The transcendental logarithmic model of individual behavior. In: R. Basman and G. Rhodes (eds.): *Advances in Econometrics*, Volume 1. JAI Press.

Juster, F. T., and R. Suzman (1995): An overview of the Health and Retirement Study. *Journal of Human Resources*, 30, S7–S56.

Kleinjans, K. J. and A. van Soest (2014): Rounding, focal point answers and nonresponse to subjective probability questions. *Journal of Applied Econometrics*, 29, 567–585.

Lewbel, A. (1999): Consumer demand systems and household expenditure. In H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics.* Blackwell.

Lewbel, A. (2001): Demand systems with and without errors. *American Economic Review*, 611-618.

Manski, C. F. (2004): Measuring expectations. *Econometrica*, 72(5), 1329–1376.

Manski, C. F. and F. Molinari (2010): Rounding probabilistic expectations in surveys. *Journal of Business and Economics Statistics*, 28, 219–231.

McFadden, D. (2012): Economic juries and public project provision. *Journal of Econometrics*, 166, 116–126.

Philipson, T. (2001): Data markets, missing data, and incentive pay. *Econometrica*, 69(4), 1099–1111.

Pudney, S. (2007): Heaping and leaping: Survey response behavior and the dynamics of self-reported consumption expenditure. Unpublished manuscript, University of Essex.

Ruud, P., D. Schunk, and J. Winter (2014): Uncertainty and rounding in survey responses: A laboratory experiment. *Experimental Economics*, 17, 391–413.

Schennach, S. M. (2013): Measurement error in nonlinear models – a review. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics: Tenth World Congress*, Volume 3, 296–337. Cambridge University Press.

Schennach, S. M. and Y. Hu (2013): Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108, 177–186.

Sims, C. A. (2003): Implications of rational inattention. *Journal of Monetary Economics*, 50, 665–690.

Wansbeek, T. and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics.* Amsterdam: Elsevier.

Wright, D. E. and I. Bray (2003): A mixture model for rounded data. *The Statistician*, 52(1), 3–13.