NONPARAMETRIC SURVEY RESPONSE ERRORS

Rosa L. Matzkin*

Department of Economics

Northwestern University

and

Division of the Humanities and Social Sciences

California Institute of Technology

June 2007

---

* The support of the National Institute of Aging through grant F014613//AG012846-12S1 and of NSF through grants BCS-0433990 and SES-0551272 is gratefully acknowledged. This paper has greatly benefitted from the input of Daniel McFadden and Joaquim Winter. I am also grateful for the comments and suggestions of Whitney Newey, Stefan Hoderlein, two anonymous referees, the research assistance of Gabriel Katz, and the comments of participants at the joint UC Berkeley/RAND Workshop on Response Errors in Surveys of the Elderly and Internet Interviewing (January 2004), the UC Berkeley Conference in Honor of Daniel McFadden (May 2005), and the University College London Conference on Microeconometrics: Measurement Matters (June 2007).

Abstract

We present nonparametric methods to identify and estimate the biases associated with response errors. When applied to survey data, these methods can be used to analyze how observable and unobservable characteristics of the respondent, and characteristics of the design of the survey, affect errors in the responses. This provides a method to correct the biases that those errors generate, by using the estimated response errors to "undo" those biases. The results are useful also to design better surveys, since they point at characteristics of the design and of subpopulations of respondents that can provide identification of response errors. Several models are considered.

1. Introduction

Surveys have been extensively used in economics, marketing, sociology, and political science, among other fields. They provide a quick and relatively inexpensive method for gathering data on individuals. Some of this data might be very difficult to get through any other way. In surveys, a representative sample of individuals is asked, either verbally or in written form, to respond to a series of questions. These may include questions about factual aspects of the respondent's life, such as age, gender, and marital status; hypothetical questions, such as what the respondent would do in a future situation; or opinions such as approval or disapproval of some government action.

As with any other method, surveys have their own drawbacks. Survey responses are

typically plagued by response errors. (Battistin (2003), Bound and Krueger (1991), McFadden, Schwarz and Winter (2004), Philipson (1997, 2001), Poterba and Summers (1986), Schwarz, Hippler, Deutsch, and Strack (1985), Tourangeau, Rips and Rasinski (2000), and the references in Bound, Brown, and Mathiowetz (2001) are some of the works that provide strong evidence for the existence of response errors.) Survey errors may be affected by the length of the recall period (Neter and Waksberg (1964), Cannell et al. (1965), Gems et al. (1982)), the salience or importance of the behavior to be retrieved (Waksberg and Valliant (1978), Chase and Harada (1984)), the social desirability or acceptability of their correct answers (Loftus (1975)), aspects of the design of the survey, such as wether it is face-to-face or telephonic or characteristics of the interviewer (Groves (1989)). Unless biases in the responses are not dealt with, the conclusions obtained from survey data might be far from the valid ones. (See McFadden (2007) for a discussion of statistical issues about surveys.)

In this paper, we develop several nonparametric methods to deal with the identification and estimation of survey response errors. The methods will allow one to identify the form of and to measure the noise generated from different sources. Hence, these methods can then be used to (i) predict biases, due to response errors, in new surveys for either the same or a new population of respondents, (ii) "undo" the biases due to response errors, by using the measured errors, and (iii) design surveys that either minimize survey response errors or that allow one to identify and estimate the response errors.

One of the methodologies that is commonly used as a first step when analyzing response errors uses descriptive tools to analyze the relationship between responses and a few observable variables. This analysis may look at how the average response to some question varies as the value of some observable characteristics, either of the respondents or of the

survey, change. Many results in experimental psychology, some of which are described in Tourangeau, Rips, and Rasinski (2000), are presented in this way. Schwarz et al. (1991) is one such example. These authors provided evidence that the category range offered in the answer to a survey question affects reported behavior. In their study, they asked respondents for their daily use of television. Two different scales for the answers were used, one ranging from "up to half an hour" to "more than two and a half hours," and the other ranging from "up to two and a half hours" to "more than four and a half hours". The analysis of the responses concluded that respondents faced with the former range reported less use of television than those faced with the latter. Similarly, in economics, response errors often are analyzed as a function of observable variables. For example, Bollinger (1998), following Bound and Krueger (1991), estimated a nonparametric regression of the error in reported income as a function of true income, using data from the Social Security Administration to match data from the Current Population Survey.

The above type of studies are extremely useful to provide evidence about the existence of a particular response error, uncover relationships among a few observable variables, and to make simple response error predictions. However, it can seldom be used to predict quantitatively what will happen in a new situation, such as when data from a different survey is used. A more structural model is typically needed if one wants to measure the relationships, incorporate unobservable variables, model the interaction among different errors, and analyze and measure the effect of different characteristics of the survey design.

A more structural method proceeds by specifying a set of functions and distributions as known up to a few parameters and by modeling the dependence among the variables. For example, in a yes or no answer, this approach would proceed by first specifying that

a particular individual will answer yes if the value of an unobservable variable is above some threshold; where the value of this variable is a linear function of some observable and unobservable characteristics of the respondent. Using data on each individual's responses and observable characteristics, this method provides numbers for the coefficients of the linear function, which could then be used to analyze, predict, and correct errors in the response of individuals. This analysis has been typically used to measure the effect of a variable in a certain response, while controlling for other variables that may also affect this response, and to uncover the distributions of relevant unobservable variables. The analysis of survey response biases in Hurd, McFadden, et al. (1998) and McFadden, Schwarz, and Winter (2003) are representatives of such type of analysis.

The methodology that we present in this paper provides a way of incorporating elements of both approaches of analysis described above. As with the latter method, it allows one to estimate underlying functions and distributions of key unobservable variables, in different stages of the response process. As with the former method, it does not require specifying a-priori parametric structures for the underlying functions and distributions. The new methods can possess as few or as many levels of complexity as one may be interested in analyzing.

As an example of how one can use the new methods to add a minimal amount of structure to a descriptive model of the former type described, suppose that one is interested in understanding the variation in the response of individuals that are otherwise equal in their relevant observable characteristics. For example, in a situation closely related to one analyzed in McFadden, Schwarz, and Winter (2003), suppose that individuals are asked their perceived probability of an end-of-life health hazard, such as needing nursing home care. Health status and family status will typically be the observable characteristics that one may

plot this answer against. However, another important variable, which is unobservable but can explain the variation in the response of individuals within a common health and family status, is the attitude of the respondent towards living in a nursing home. Being able to identify and estimate the distribution of attitudes towards living in a nursing home, and the variables that affect this taste distribution, is very important to predict future demand for nursing homes and to measure the well-being of nursing home users. Moreover, this distribution of attitudes towards living in a nursing home will also influence the variation in the response to other related questions, such as "Have you purchased insurance for nursing home care?" The response to the latter question may depend on a larger set of observable and unobservable characteristics than the former question, which may include the income and the unobservable attitude toward risk of the respondent. Identifying the distribution of attitudes towards living in a nursing home, from the response to the first question, will make it feasible to identify the distribution of the unobservable attitude towards risk, from the responses to the second question. Analyzing the latter distribution will have important implications to predict demand for various types of insurance, and its identification will allow to pursue the identification of other important unobservable variables as well as other further studies in which knowing this distribution is important.

The nonparametric, structural approach provides many benefits. First, it provides a bridge among the two different methods described above, which have been used to analyze behavior in survey response. Second, it provides more "trusted" predictions than the parametric analysis, because its conclusions do not depend on ad-hoc parametric specifications for the underlying functions and distributions. Third, it provides a method to test particular parametric assumptions, by evaluating how close the nonparametric estimates are from the

parametric ones. Fourth, it allows to infer all types of shapes for the structural anomalies, because the underlying functions will be estimated without imposing on them particular shapes. And, fifth, it allows one to infer the unobserved heterogeneity across otherwise observable equal individuals, which has implications towards their heterogenous responses in other questions and towards predicted behavior by these individuals.

When used in estimation, reported values of variables will suffer from measurement error, due to survey response errors. A large literature exists on measurement errors. Aigner (1984), Fuller (1987), Carroll, Ruppert, and Stefanski (1995), Wansbeek and Meijer (2000), Bound, Brown, and Mathiowetz (2001), Hausman (2001), Moffit and Ridder (2007), and Chen, Hong, and Nekipelov (2007) provide excellent surveys of the topic. This paper contributes to this literature by specifying structures that could be useful in the design of surveys and the analysis of data. Other recent papers within this realm, which also consider nonparametric methods, but make use of different structures, are McFadden (2006) and Hoderlein and Winter (2007). As in Philipson (1997, 2001), these papers incorporate a decision model for the respondent. McFadden (2006) applies incentive theory to the design and administration of economic surveys. Hoderlein and Winter (2007) analyze models of boundedly rational survey response behavior.

The outline of the paper is as follows. In the next section, we motivate our results making use of a well known empirical model. Section 3 deals with cases where the distribution of an unobservable variable of interest is known. Section 4 deals with cases where the distribution of the unobservable variable of interest is unknown. Section 5 provides conclusions, and describes extensions of the methods presented in the main sections. All the proofs are presented in the Appendix.

2. A motivating model

To provide motivation for our methods through a well known empirical problem, consider a model similar to that used in Bound (1991) and in Bound, Brown, and Mathiowetz (2001) to analyze the effect of health on the retirement decision of the elderly. Bound (1991) and Bound, Brown, and Mathiowetz (2001) discuss the different biases that may arise as a result of using various measures of health. Self reported health may generate biases not only because it is a subjective measure, but also because older workers may use poor health to justify decreasing the amount of hours worked, when their true reasons for decreasing hours of work are less sociably acceptable. In fact, Ettner (1997) provides evidence that women report health less frequently than men as a reason for retirement, which is consistent with the fact that it is more socially acceptable for women to retire due to reasons other than health.

One of the models analyzed by Bound (1991) and Bound, Brown, and Mathiowetz (2001) is (after changing the notation to one that is more suitable for this paper)

$$H = \beta_1 X_1 + \lambda_1 Y^* + \varepsilon_1$$

where H denotes the choice of hours of work, X_1 denotes the benefit of working, Y^* denotes unobserved health status, and ε_1 denotes the effect of other random components. Letting Y denote self-reported health, these authors consider the equation

$$Y = \beta_2 \, X_1 + \lambda_2 Y^* + \varepsilon_2$$

They assumed that Y^* and X_1 are orthogonal to ε_1 and ε_2. In addition, the correlation

between benefits from work (X_1) and health (Y^*), is specified by

$$X_1 = \lambda_4 Y^* + \zeta$$

The models that we present is this paper can be used to estimate the function that determines self reported health, either as a function of X_1 or as a function of H, nonparametrically. This can then be used to identify the effect of health on hours of work. We consider different situations, starting from the simplest ones, and ending with the simultaneous model where the effect of desired hours of work on reported health is identified.

As above, let Y^* denote the true value of an unobservable variable, and let Y denote the response when asked the value of Y^*. Let $X = (X_1, X_2)$ denote a vector of observable characteristics of the respondent, let W denote observable characteristics of the survey, and let Z denote other observable external variables. In the next section, we first consider the identification of the nonparametric response function m, specified as

$$Y = m\left(Y^*, X, W\right)$$

where m is assumed to be strictly increasing in Y^* and the distribution of Y^* (health) conditional on X_1 (work benefits) is known. We then extend this model to include a response error, which depends on X_1 (work benefits, gender), directly added to Y^*, of the form

$$Y = m\left(Y^* + \eta, X_2, W_2\right)$$

$$\eta = v\left(X_1, W_1, \delta\right)$$

where δ is an unobservable random term that is distributed independently of (X, W, Y^*), and where v is an unknown function that is strictly increasing in δ. We then show that adding an error ε_2, can be done in a similar fashion. Hence, we consider

$$Y = m\left(Y^* + \eta, X_2, W_2\right) + \varepsilon_2$$

where now

$$\varepsilon_2 = s\left(Z, \omega\right)$$

In Section 4.1, we deal with the situation where the distribution of Y^* conditional on X_1 is unknown. We show how one may use the knowledge about an exogenous change in work benefits (e.g. a change in social security benefits) to identify this distribution. We also show how observable determinants of Y^* can be used to identify this distribution.

In Section 4.2, we deal with the simultaneous model where desired hours of work enter directly into the response function. We consider the model

$$\begin{aligned} H &= m_1\left(Y^*, X_1\right) + \varepsilon_1 \\ Y &= Y^* + m_2\left(H, X_2\right) + \mu_2 \end{aligned}$$

and describe what results can be used to identify and estimate the unknown functions and distributions in this model.

3. Known distribution of the true value Y^*

In many situations, one may be able to know the distribution of some variable of interest but not be able to know the particular value of such variable for any particular individual. For example, when Y^* denotes true income, the distribution of Y^* may be obtained from the Social Security administration records. In the model in Section 2, one may be able to know the distribution of health status, Y^*, conditional on work benefits, X_1, using medical records. Following the notation in Section 2, let Y denote the response when asked about Y^*, $X = (X_1, X_2)$ denote the observable characteristics of the worker, and W denote the vector of characteristics of the survey. Assume that Y^* is independent of (X_2, W) conditional on X_1, and that the model is given by

$$(3.1) \quad Y = m\left(Y^*, X, W\right)$$

Let $F_{Y|X,W}$ denote the joint distribution of the observable variables (Y, X, W). Let $F_{Y^*|X_1}$ denote the distribution of Y^* conditional on X_1. The following result follows by Matzkin (2003):

Theorem 1 (Matzkin (2003))**:** *Suppose that, conditional on* $X_1 = x_1$, Y^* *is distributed independently of* (X_2, W) *with an everywhere positive, known density. Suppose also that for each* (X, W), *the function* m *is strictly increasing in* Y^*. *Then, for all* y^*, x, w for which the values of $F_{Y^*|X_1=x_1}(y^*)$ and of $F^{-1}_{Y|X=x,W=w}\left(F_{Y^*|X_1=x_1}(y^*)\right)$ exist,

$$(3.2) \quad m\left(y^*, x, w\right) = F^{-1}_{Y|X_1=x_1,X_2=x_2,W=w}\left(F_{Y^*|X_1=x_1}(y^*)\right)$$

Hence, the function m is identified nonparametrically from the joint distribution of (Y, X, W).

Equivalently, we can state that, under the assumptions of Theorem 1, an individual for which $X_1 = x_1$ and who answers $Y = y$ when faced with $(X_2, W) = (x_2, w)$ has a value of the latent variable Y^* equal to

$$y^* = F^{-1}_{Y^*|X_1=x_1}(F_{Y|X_1=x_1,X_2=x_2,W=w}(y))$$

Replacing $F_{Y|X,W,Z=z}$ (and $F_{Y^*|Z=z}$) by a nonparametric estimator, one can obtain from the above equations estimators for the response function, $m(y^*, x, w)$, and for the value of the latent variable of any individual, given his response, y.

The assumptions in the above model are such that there is a 1-1 relationship between the true value, Y^*, of the response, and the response, Y. The distributions of Y^* and of Y conditional on (X, W) allow one to determine uniquely such a 1-1 relationship. From this, one can map any response with its true corresponding value, without any error. In some cases, this situation may be unrealistic. The same value of the true latent variable Y^* may generate different responses even from individuals that possess the same observable characteristics, X, and are asked the same question, characterized by W. In the model considered in Bound (1991), an unobservable variable, μ_1, added to the response model allowed this effect. Another alternative is to model an added response error to Y^*. In the model in Section 2, we may specify that the response error is

$$(3.3) \quad \eta = v(X_1, W_1, \delta)$$

where X_1 denotes work benefits and gender, δ is an unobservable random term that is distributed independently of (X, W, Y^*), and where v is an unknown function that is strictly increasing in δ. The inclusion of W_1 allows for this response error to be affected by characteristics of the survey. The model that determines the response Y as a function of Y^*, η, X and W may then be given by

$$(3.4) \quad Y = m\left(Y^* + \eta, X_2, W_2\right)$$

An added term, μ_1, as in the model in Bound (1991), can be handled similarly, by specifying that for some unknown function s, a vector of observable variables, Z, and an unobservable ω

$$(3.5) \quad \mu_1 = s\left(Z, \omega\right)$$

where s is strictly increasing in ω, and ω has an everywhere positive density. The models above then become

$$(3.6) \quad Y = m\left(Y^*, X_2, W_2\right) + \mu_1$$

and

$$(3.7) \quad Y = m\left(Y^* + \eta, X_2, W_2\right) + \mu_1$$

where

$$(3.8) \quad \mu_1 = s\left(Z, \omega\right)$$

The critical assumptions we make to deal with the multiple unobservables are that for at least one value, $\widetilde{z}$, of Z, $s\left(\widetilde{z}, \omega\right) = 0$ for all ω, and for at least one value $(\widetilde{x}_1, \widetilde{w}_1)$ of (X_1, W_1), $v\left(\widetilde{x}_1, \widetilde{w}_1, \delta\right) = 0$ for all δ, in addition to assuming that δ is independent of (X, W, Z, ω, Y^*)

and ω is independent of (X, W, Z, δ, Y^*). In the model of Section 2, X_1 may denote work benefits from working and gender. Since there exists evidence that women self reported health is much less influenced by work benefits than that of men, one may assume that $(\widetilde{x}_1, \widetilde{w}_1)$ denotes the value of a vector for which a coordinate of X_1 corresponds to women. Another obvious example is where η denotes recall error, on a question that refers to an event in the past, and W_1 denotes how long ago the event refers to. When $W_1 = 0$, one may assume that the value of v is zero for all values of X_1 and δ.

Since the analysis is similar in either case, with η and μ_1, or with only η, we next consider the case with η only. Hence, the model we consider is (3.4) with (3.3). In such a model, the following theorem establishes the identification of the functions v and m and of the distributions of δ and η, under some assumptions.

Theorem 2: *Suppose that* (Y^*, δ) *is distributed independently of* (X_2, W) *conditional on* X_1, *with an everywhere positive density,* δ *is distributed independently of* (X, W, Y^*) *with an everywhere positive unknown density, the function* m *is strictly increasing in* $Y^* + \eta$, *the function* v *is strictly increasing in* δ, *the distribution of* Y^* *conditional on* X_1 *is known, and its characteristic function is everywhere different zero. Restrict the function* v *to satisfy at one point* $(\overline{x}_1, \overline{w}_1)$ *of* (X_1, W_1) *the condition:* $v(\overline{x}_1, \overline{w}_1, \delta) = 0$ *for all* δ, a*nd at another point* $(\widetilde{x}_1, \widetilde{w}_1)$ *of* (X_1, W_1) *the condition:* $v(\widetilde{x}_1, \widetilde{w}_1, \delta) = \delta$. *Then, the function* m, *the function* v, *and the distributions of* δ *and of* η *conditional on* (X_1, W_1) *are identified nonparametrically from the joint distribution of* (Y, X, W).

Theorem 2 establishes the nonparametric identification of the response error, $v(x_1, w_1, \delta)$,

and the response function m, when the distribution of the variable of interest, Y^* is known. Since the proof is constructive, one can use the proof to derive nonparametric estimators for these functions. Note that if the function v were specified as: $v(x_1, w_1, \delta) = \delta x_1 + \delta w_1$, then the restrictions on the function v would be satisfied for $\widetilde{x}_1 = \widetilde{w}_1 = 0$, and for $(\overline{x}_1, \overline{w}_1) = (1, 0)$ or $(\overline{x}_1, \overline{w}_1) = (0, 1)$.

4. Unknown distribution of the true value Y^*

4.1. Y^* conditionally independent of the explanatory variables.

The analysis in Section 3 rested on the assumption that the distribution of Y^* conditional on X_1 was either known or could be estimated. Often, one may need to impose additional restrictions or to augment the data to be able to identify, and therefore estimate, this distribution. We next analyze how this can be done by augmenting the data. Consider again the model in Section 2. Suppose that in some period there was a change in the laws regarding work benefits. Let Z_1 denote work benefits for a worker before the change and X_1 denote the work benefits after the change. The exogeneity of the law change may allow one to assume that for some unknown function s,

$$X_1 = s\left(Z_1, \delta\right)$$

where δ is independent of the vector of all the observable and other unobservable variables in the model and s is strictly increasing in δ. Using the results in Matzkin (2004), one can then show that, under some monotonicity and support conditions, one can identify the

distribution of Y^* conditional on X_1.

Alternatively, we may have information about some of the determinants of Y^*. Consider, for example, the model where

$$Y = m\left(Y^* + \eta, X_2, W_2\right), \text{ and}$$

$$\eta = v\left(X_1, W_1, \delta\right)$$

Suppose that for some unknown function s, some vector, Z, of observable variables, and an unobservable variable, ξ,

$$Y^* = s\left(Z, \xi\right)$$

Under assumptions similar to those made in Theorem 2, and some additional assumptions on s and ξ, one can show that the functions m, v, and s are identified. The later set of assumptions may be imposed on the function s, similarly to those imposed in Theorem 2 on the function v and δ, or on the distribution of ξ. We consider the former.

Theorem 3: *Suppose that (ξ, δ) is distributed independently of (X, W, Z) with an everywhere positive unknown density, δ is distributed independently of (X, W, Z, ξ) with an everywhere positive unknown density, the function m is strictly increasing in $Y^* + \eta$, the function v is strictly increasing in δ, the function s is strictly increasing in ξ, and the characteristic function of ξ is everywhere different from 0. Restrict the function v to satisfy at one point $(\overline{x}_1, \overline{w}_1)$ of (X_1, W_1) the condition: $v\left(\overline{x}_1, \overline{w}_1, \delta\right) = 0$ and at another point $(\widetilde{x}_1, \widetilde{w}_1)$ of (X_1, W_1) the condition: $v\left(\widetilde{x}_1, \widetilde{w}_1, \delta\right) = \delta$. Restrict the function s to satisfy at one point $\overline{z}$*

of Z the condition: $s(\overline{z}, \xi) = \xi$. Restrict the function m to satisfy at one point $(\overline{x}_2, \overline{w}_2)$ of (X_2, W_2) the condition: $m(t, \overline{x}_2, \overline{w}_2) = t$. Then, the functions m, v, and s as well as the distributions of δ, η, ξ, and Y^* conditional on (X, W, Z) are identified nonparametrically from the joint distribution of (Y, X, W, Z).

4.2. Y^* not conditionally independent of the explanatory variables.

The model where work benefits are included directly in the response function does not allow one to infer directly the effect of the desired hours of work on self reported health. A more structural model would be

$$\begin{aligned} H &= m_1(Y^*, X_1) + \varepsilon_1 \\ Y &= m_2(Y^*, H, X_2, W) + \mu_2 \end{aligned}$$

The identification and estimation of the unknown functions and distributions in this model can be analyzed using the results in Matzkin (2007) (See also Matzkin (2005).) In particular, suppose that for unknown functions m_1, m_2, s_1, s_2,

$$\begin{aligned} H &= m_1(Y^*, X_1) + \varepsilon_1 \\ Y &= Y^* + m_2(H, X_2) + \mu_2 \end{aligned}$$

$$X_1 = s_1(Z_1, \delta)$$

and

$$\mu_2 = s_2\left(Z_2, \omega_2\right)$$

where X_2 may denote here a characteristic of either the respondent or the survey. Let $f_{\varepsilon_1,y^*}(\varepsilon_1, y^*)$ denote the joint density of (ε_1, y^*). Let $\overline{z}_1$ and $\overline{z}_2$ denote particular given values of Z_1 and Z_2. Let $\widetilde{x}_1$ denote a given value of X_1. Matzkin (2007) makes use of the following assumptions to show that the derivatives of m_1 and m_2, as well as the derivatives of the density $f_{\varepsilon_1,y^*}(\varepsilon_1, y^*)$ are identified:

(i) m_1 is strictly increasing in Y^*, s_1 is strictly increasing in δ, and s_2 is strictly increasing in ω_2,

(ii) the vector (X, W, Z, H, Y) and the vector $(X, W, Z, \varepsilon_1, \omega_2, \delta, Y^*)$ have full support,

(iii) ω_2 is independent of (Y^*, X_1, X_2, δ) conditional on $Z_1 = \overline{z}_1$, $Z_2 = \overline{z}_2$;

(iv) δ is independent of $(Y^*, \varepsilon_1, X_1, X_2, \delta)$ conditional on $Z_1 = \overline{z}_1$, $Z_2 = \overline{z}_2$,

(iv) for all ω_2, $s_2\left(\overline{z}_2, \omega_2\right)$,

(v) at $X_1 = \widetilde{x}_1$, $m_1\left(y^*, \widetilde{x}_1\right) = y^*$;

(vi) (ε_1, Y^*) is independent of (Z_2, X_2, δ), and

(vii) for all values of y^*, there exists a value of ε_1 such that

$$\frac{\partial \log f_{\varepsilon_1,y^*}(\varepsilon_1, y^*)}{\partial \varepsilon_1} = 0 \quad \text{and} \quad \frac{\partial \log f_{\varepsilon_1,y^*}(\varepsilon_1, y^*)}{\partial y^*} \neq 0$$

In Matzkin (2007) a method to directly calculate, from the density of the observable variables, the derivatives of m_1 and m_2, as well as the derivatives of the density $f_{\varepsilon_1,y^*}(\varepsilon_1, y^*)$, is presented.

5. Extensions

The results that have been presented in the previous sections can be extended to analyze models with nested response errors and models where the responses are discrete.

Answering a survey question involves the stage of comprehension of the question, the stage of retrieving and assembling relevant information, the stage of filtering the information, and the stage of actually responding to the question with an answer which may or may not be the one that the respondent have come up with prior to responding (see Tourangeau, Rips, and Rasinski (2000)). Each of these stages adds a level of noise to the response, which will typically be different across respondents. Variation in the noise may depend on observable characteristics, but, it will typically also depend on unobservable characteristics. The design of the survey and of the particular question being asked add additional layers of possible noise, which interact and affect the magnitudes of the processing noise. A model that allows for all these errors at the different stages may take, for example, the form

$$Y = m\left(s\left(\left(v\left(\delta_3, X_3, W_3\right) + \delta_2\right), X_2, W_2\right) + \delta_1, X_1, W_1\right)$$

or the form

$$Y = m\left(s\left(X_2, W_2, \delta_2\right) + v\left(X_3, W_3, \delta_3\right) + \delta_1, X_1, W_1\right).$$

The identification and estimation of the unknown functions and distributions in these models can be achieved by following a similar analysis to that described in the previous sections.

Discrete responses are also very common. Some questions ask for a 0 or 1 answer.

Some respondents prefer not to answer some questions. The identification and estimation of such models can be performed also by extending the above results. Consider for example a situation where an individual is asked whether he thinks that the probability that he will ever need a nursing home is above or below .5. The individual may decide jointly the answer to the question and whether or not to answer. Let $R = 1$ if he answers and $R = 0$ otherwise. Let $Z = 1$ if the response is that the probability is above .5; $Z = 0$ otherwise. Some of the variables affecting the respondent's behavior may be observable and some may be unobservable. Suppose, in particular, that $R = 1$ if $W_1 + v(X, W_2, \eta) \geq \varepsilon_R$, where W_1, W_2, and X are observable and η and ε_R are unobservable, and the value of v is known at one point. (This setup generalizes a model of discrete response to survey treatments W_1 analyzed by McFadden (1994).) Suppose that $Z = 1$ if $\widetilde{W}_1 + r(\widetilde{X}, \widetilde{W}_2, \eta) \geq \varepsilon_Z$, where $\widetilde{W}_1$, $\widetilde{W}_2$, and $\widetilde{X}$ are observable, ε_Z is unobservable, and the function r is known at one point. Then

$$\Pr(R = 1, Z = 1 | X, W, \widetilde{X}, \widetilde{W}, \eta) = F_{\varepsilon_R, \varepsilon_Z}\left(W_1 + v(X, W_2, \eta), W_1 + r(\widetilde{X}, \widetilde{W}_2, \eta)\right)$$

Under strong independence and support conditions, and some normalizations on the functions v and r, one can identify the nonparametric distribution of $(\varepsilon_R, \varepsilon_Z, \eta)$ and the nonparametric functions v and r. (See Matzkin (1993, 1994, 2005, 2006), Briesch, Chintagunta, and Matzkin (1997, 2007), and Lewbel (2000).)

6. Appendix

Proof of Theorem 1: By conditional independence between Y^* and (X_2, W) given X_1, and strict monotonicity, it follows that for all (x, w)

$$\begin{aligned} F_{Y^*|X_1=x_1}(y^*) &= \Pr(Y^* \leq y^* | X_1 = x_1) = \Pr(Y^* \leq y^* | X = x, W = w) \\ &= \Pr(m(Y^*, X, W) \leq m(y^*, x, w) | X = x, W = w) \\ &= F_{Y|(X,W)=(x,w)}(m(y^*, x, w)) \end{aligned}$$

Since the assumptions imply that $F_{Y|(X,W)=(x,w)}$ is strictly increasing, it follows that

$$m(y^*, x, w) = F^{-1}_{Y|X=x,W=w}\left(F_{Y^*|X_1=x_1}(y^*)\right)$$

Hence, m is identified.

Proof of Theorem 2: Since (Y^*, δ) is distributed independently of (X_2, W), conditional on X_1, Y^* is distributed independently of (X_2, W_2) conditional on any values of (X_1, W_1), $Y^* + \delta$ is distributed independently of (X_2, W_2) conditional on any values of (X_1, W_1), and $Y^* + \eta = Y^* + v(X_1, W_1, \delta)$ is distributed independently of (X_2, W_2) conditional on any values of (X_1, W_1). Moreover, when $(X_1, W_1) = (\overline{x}_1, \overline{w}_1)$, $\eta = v(\overline{x}_1, \overline{w}_1, \delta) = 0$ and $Y^* + \eta = Y^*$; and when $(X_1, W_1) = (\widetilde{x}_1, \widetilde{w}_1)$, $\eta = v(\widetilde{x}_1, \widetilde{w}_1, \delta) = \delta$ and $Y^* + \eta = Y^* + \delta$. Hence, for any

(x_2, w_2)

$$
\begin{aligned}
& (T2.1) \quad F_{Y^*|X_1=\overline{x}_1}(y^*) \\
= \; & F_{Y^*+\eta|(X_1,W_1)=(\overline{x}_1,\overline{w}_1)}(y^*) \\
= \; & F_{Y^*+\eta}\left(y^* \middle| \left(X_1, W_1\right) = \left(\overline{x}_1, \overline{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & \Pr\left(Y^* + \eta \leq y^* \middle| \left(X_1, W_1\right) = \left(\overline{x}_1, \overline{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & \Pr\left(m\left(Y^* + \eta, X_2, W_2\right) \leq m\left(y^*, x_2, w_2\right) \middle| \left(X_1, W_1\right) = \left(\overline{x}_1, \overline{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(m\left(y^*, x_2, w_2\right)\right)
\end{aligned}
$$

Since our assumptions imply that $F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}$ is invertible, this implies that for any t and any (x_2, w_2)

$$
(T2.2) \quad m\left(t, x_2, w_2\right) = F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(F_{Y^*|(X_1,W_1)=(\overline{x}_1,\overline{w}_1)}(t)\right)
$$

Hence, m is identified nonparametrically.

Using a similar reasoning as above, we get that for any (x_2, w_2)

$$
\begin{aligned}
& (T2.3) \quad F_{Y^*+\delta|X_1=\widetilde{x}_1}(y^* + \delta) \\
= \; & F_{Y^*+\eta}(y^* + \delta| \left(X_1, W_1\right) = \left(\widetilde{x}_1, \widetilde{w}_1\right)) \\
= \; & F_{Y^*+\eta}\left(y^* + \delta \middle| \left(X_1, W_1\right) = \left(\widetilde{x}_1, \widetilde{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & \Pr\left(Y^* + \eta \leq y^* + \delta \middle| \left(X_1, W_1\right) = \left(\widetilde{x}_1, \widetilde{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & \Pr\left(m\left(Y^* + \eta, X_2, W_2\right) \leq m\left(y^* + \delta, x_2, w_2\right) \middle| \left(X_1, W_1\right) = \left(\widetilde{x}_1, \widetilde{w}_1\right), \left(X_2, W_2\right) = \left(x_2, w_2\right)\right) \\
= \; & F_{Y|X=(\widetilde{x}_1,x_2),W=(\widetilde{w}_1,w_2)}\left(m\left(y^* + \delta, x_2, w_2\right)\right)
\end{aligned}
$$

Using (T2.1), this implies that

$$F_{Y^*+\delta|X_1=\widetilde{x}_1}(y^*+\delta) = F_{Y|X=(\widetilde{x}_1,x_2),W=(\widetilde{w}_1,w_2)}\left(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(F_{Y^*|X_1=\overline{x}_1}(y^*+\delta)\right)\right)$$

Hence, for any t

$$(T2.4) \quad F_{Y^*+\delta||X_1=\widetilde{x}_1}(t) = F_{Y|X=(\widetilde{x}_1,x_2),W=(\widetilde{w}_1,w_2)}\left(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(F_{Y^*|X_1=\overline{x}_1}(t)\right)\right)$$

This implies that the distribution of $Y^*+\delta$ conditional on $X_1=\widetilde{x}_1$, is identified. Since, by assumption, the distribution of Y^* conditional on $X_1=\widetilde{x}_1$ is known, and δ is independent of X_1, one can obtain the distribution of δ by deconvolution. Hence, the distribution of δ is identified.

Next, we derive an expression for the distribution of $Y^*+\eta$ conditional on any (X_1,W_1). Similarly to above,

$$\begin{aligned}
&(T2.5) \quad F_{Y^*+\eta|(X_1,W_1)=(x_1,w_1)}(y^*+\eta) \\
= \;& F_{Y^*+\eta}\left(y^*+\eta|\left(X_1,W_1\right)=\left(x_1,w_1\right),\left(X_2,W_2\right)=\left(x_2,w_2\right)\right) \\
= \;& \Pr\left(Y^*+\eta\leq y^*+\eta|\left(X_1,W_1\right)=\left(x_1,w_1\right),\left(X_2,W_2\right)=\left(x_2,w_2\right)\right) \\
= \;& \Pr\left(m\left(Y^*+\eta,X_2,W_2\right)\leq m\left(y^*+\eta,x_2,w_2\right)|\left(X_1,W_1\right)=\left(x_1,w_1\right),\left(X_2,W_2\right)=\left(x_2,w_2\right)\right) \\
= \;& F_{Y|X=(x_1,x_2),W=(w_1,w_2)}\left(m\left(y^*+\eta,x_2,w_2\right)\right) \\
= \;& F_{Y|X=(x_1,x_2),W=(w_1,w_2)}\left(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(F_{Y^*|X_1=\overline{x}_1}(y^*+\eta)\right)\right)
\end{aligned}$$

where the last equality follows by (T2.2). Hence, for any t and any (x_1, w_1),

$$(T2.6) \quad F_{Y^*+\eta|(X_1,W_1)=(x_1,w_1)}(t) = F_{Y|X=(x_1,x_2),W=(w_1,w_2)}\left(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2)}\left(F_{Y^*|X_1=\overline{x}_1}(t)\right)\right)$$

This implies that the distribution of $Y^* + \eta$ conditional on $(X_1, W_1) = (x_1, w_1)$ is identified. Since, by our assumptions, Y^* is distributed independently of η conditional on (X_1, W_1), Y^* is distributed independently of W_1, conditional on X_1, and the distribution of Y^* conditional on X_1 is known, we can get from the distribution of $Y^* + \eta$ conditional on (X_1, W_1) the distribution of η conditional on (X_1, W_1), by deconvolution. Hence, the distribution of η conditional on (X_1, W_1) is identified.

Last, to show that the function v is identified, we use the strictly monotonicity of v in δ and the independence between δ and (X_1, W_1), to establish as in the proof of Theorem 1 that for any t and any (x_1, w_1)

$$F_\delta(t) = F_{\eta|(X_1,W_1)=(x_1,w_1)}(v(x_1, w_1, t))$$

This implies that

$$(T2.7) \quad v(x_1, w_1, t) = F^{-1}_{\eta|(X_1,W_1)=(x_1,w_1)}(F_\delta(t)).$$

Since F_δ and $F_{\eta|(X_1,W_1)=(x_1,w_1)}$ are identified, v is identified. This completes the proof.

Proof of Theorem 3: Following arguments as in the proof of Theorem 2, we have

that since (ξ, δ) is distributed independently of (X, W, Z), ξ is distributed independently of (X_2, W_2, Z) conditional on any values of (X_1, W_1), $\xi + \delta$ is distributed independently of (X_2, W_2, Z) conditional on any values of (X_1, W_1), and $Y^* + \eta = s(Z, \xi) + v(X_1, W_1, \delta)$ is distributed independently of (X_2, W_2) conditional on any values of (X_1, W_1, Z). Moreover, when $(X_1, W_1, Z) = (\overline{x}_1, \overline{w}_1, \overline{z})$, $Y^* + \eta = \xi$; and when $(X_1, W_1) = (\widetilde{x}_1, \widetilde{w}_1)$, $\eta = v(\widetilde{x}_1, \widetilde{w}_1, \delta) = \delta$ and $Y^* + \eta = Y^* + \delta$. Hence, for any e and (x_2, w_2)

$$
\begin{aligned}
& (T5.1) \quad F_\xi(e) \\
= \; & F_{\xi|X=(x_1,x_2),W=(w_1,w_2),Z=z}(e) \\
= \; & F_{Y^*+\eta|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}(e) \\
= \; & \Pr\left(Y^* + \eta \leq e \right| \left.(X_1, W_1, Z) = (\overline{x}_1, \overline{w}_1, \overline{z}), (X_2, W_2) = (x_2, w_2)\right) \\
= \; & \Pr\left(m(Y^* + \eta, X_2, W_2) \leq m(e, x_2, w_2) \,|\, (X_1, W_1, Z) = (\overline{x}_1, \overline{w}_1, \overline{z}), (X_2, W_2) = (x_2, w_2)\right) \\
= \; & F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}\left(m(e, x_2, w_2)\right)
\end{aligned}
$$

Since for all t, $m(t, \overline{x}_2, \overline{w}_2) = t$

$$
(T5.2) \quad F_\xi(e) = F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}(e)
$$

Hence, F_ξ is identified. Using this in (T5.1), we get that

$$
F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}(e) = F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}\left(m(e, x_2, w_2)\right)
$$

Since our assumption imply that $F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),,Z=\overline{z}}$ is invertible, this implies that for

any t and any (x_2, w_2)

$$(T5.3) \quad m\left(t, x_2, w_2\right) = F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}\left(F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}\left(t\right)\right)$$

Hence, m is identified nonparametrically.

Using a similar reasoning as in (T5.1), we get that for any (x_2, w_2) and any z

$$\begin{aligned} & (T5.4) \quad F_{\xi}(e) \\ = \; & F_{\xi|X=(x_1,x_2),W=(w_1,w_2),Z=z}(e) \\ = \; & F_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=z}(s\left(z, e\right)) \end{aligned}$$

Equation (T5.2) together with the strict monotonicity of $F_{Y^*|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=z}$ imply then that

$$(T5.5) \quad s\left(z, e\right) = F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=z}\left(F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}\left(e\right)\right)$$

Hence, the function s is identified. Since the distribution of ξ is also identified, this implies that the distribution of Y^* conditional on Z is also identified.

The distribution of δ is identified because for any t

$$\begin{aligned} & F_{\xi+\delta}(t) \\ = \; & F_{Y|(X_1,W_1)=(\widetilde{x}_1,\widetilde{w}_1),(X_2,W_2)=(x_2,w_2),Z=\overline{z}}(m\left(t, x_2, w_2\right)) \end{aligned}$$

Hence, using (T5.3), it follows that for any x_2, w_2

$$F_{\xi+\delta}(t) = F_{Y|(X_1,W_1)=(\widetilde{x}_1,\widetilde{w}_1),(X_2,W_2)=(x_2,w_2),Z=\overline{z}}(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}\left(F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}\left(t\right)\right))$$

This implies that the distribution of $\xi + \delta$ is identified. Hence, by deconvolution, we can obtain the distribution of δ, using the already identified distribution of ξ.

The distribution of η conditional on (X_1, W_1) is identified because for any t, x_2, w_2

$$\begin{aligned}
& F_{\xi+\eta|(X_1,W_1)=(x_1,w_1)}(t) \\
= \; & (X_1,W_1) = (x_1,w_1), (X_2,W_2) = (x_2,w_2), Z = \overline{z} \\
= \; & F_{Y|(X_1,W_1)=(x_1,w_1),(X_2,W_2)=(x_2,w_2),Z=\overline{z}}(m\left(t,x_2,w_2\right)) \\
= \; & F_{Y|(X_1,W_1)=(x_1,w_1),(X_2,W_2)=(x_2,w_2),Z=\overline{z}}(F^{-1}_{Y|X=(\overline{x}_1,x_2),W=(\overline{w}_1,w_2),Z=\overline{z}}\left(F_{Y|X=(\overline{x}_1,\overline{x}_2),W=(\overline{w}_1,\overline{w}_2),Z=\overline{z}}\left(t\right)\right))
\end{aligned}$$

Hence, since the distribution of ξ is known, and ξ is distributed independently of (X_1, W_1), the distribution of η conditional on (X_1, W_1) is identified. From this conditional distribution and the distribution of δ, we can identify $v\left(x_1, w_1, \delta\right)$, by the arguments in Theorem 1, as

$$v\left(x_1,w_1,\delta\right) = F^{-1}_{\eta|(X_1,W_1)=(x_1,w_1)}\left(F_\delta(\delta)\right)$$

This completes the proof.

7. References

BATTISTIN, E. (2003) "Errors in Survey Reports of Consumption Expenditures," working paper # 0307, *Institute for Fiscal Studies*, London.

BOLLINGER, C.R. (1998) "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics,* Vol. 16, No. 3, pp. 576-594.

BOUND, J., C. BROWN, and N. MATHIOWETZ (2001) "Measurement Error in Survey Data," in *Handbook of Econometrics,* Vol. 5, edited by J.J. Heckman and E. Leamer, 3705-3843. Amsterdam: Elsevier.

BOUND, J. and A. KRUEGER (1991) "The Extent of Measurement Error in Longitudinal Earning Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 16, 576-94.

BRIESCH, R., P. CHINTAGUNTA, and R.L. MATZKIN (2005) "Nonparametric Discrete Choice Models with Unobserved Heterogeneity," mimeo, Northwestern University.

CANNELL, C., G. FISHER and T. BAKKER (1965) "Reporting of Hospitalizations in the Health Interview Survey," *Vital and Health Statitics,* Series 2, Number 6 (Public Health Service, Washington)

CARROLL, R.J., D. RUPPERT, and D. STEFANSKI (1995) *Measurement Error in Nonlinear Models,* New York: Chapman and Hall.

CHASE, D.R and M. HARADA (1984) "Response Error in Self-Reported Recreation Participation," *Journal of Leisure Research, 16: 322329.*

CHEN, X, H. HONG and D. NEKIPELOV (2007) "Measurement Error Models," mimeo, NYU.

ETTNER, S. (1997) "Is Working Good for You? Evidence on the Endogeneity of Mental and Physical Health to Female Employment," Unpublished paper (Harvard School of Public

Health).

FULLER, W. (1987) *Measurement Error Models (*Wiley, New York).

HAUSMAN, J., W. NEWEY, H. ICHIMURA, and J. POWELL (1991) "Measurement Errors in Polynomial Regression Models," *Journal of Econometrics,* 50, 273-295.

GEMS, B., D. GHAOSH, and R. HITLIN (1982) "A Recall Experiment: Impact of Time on Recall of Recreational Fishing Trips," *Proceedings of the Section on Survey Research Methods (*American Statistical Association, Alexandria, VA) 168-173.

GROVES, R.M. (1989) *Survey Errors and Survey Costs* (Wiley, New York).

HAUSMAN, J. (2001) "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *The Journal of Economic Perspectives,* 15, 57-67.

HODERLEIN, S. and J. WINTER (2007) "Recall Errors in Surveys," mimeo, Mannheim University.

HURD, M.D., D. McFADDEN, H. CHAND, L. GAN, A. RMERRILLL and M. ROBERTS (1998) "Consumption and Saving Balances of the Elderly: Experimental Evidence on Survey Response Bias", in *Frontiers in the Economics of Aging,* edited by D. Wise, 353-387, Chicago, IL: University of Chicago Press.

LEWBEL, A. (2000), "Semiparametric Qualitative Response Model Estimation with Unknown Hetersokedasticity and Instrumental Variables," *Journal of Econometrics*, 97, 145-177.

LOFTUS, E.F. (1975) "Leading Questions and the Eyewitnes report," *Cognitive Psychology,* 7:560-572.

MATZKIN, R.L. (1993) "Nonparametric Identification and Estimation of Polychotomous Choice Models," *Journal of Econometics,* 58.

MATZKIN, R.L. (1994) "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of Econometrics,* Vol. 4, edited by R.F. Engel and D. McFadden.

MATZKIN, R.L. (2003) "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

MATZKIN, R.L. (2004) "Unobservable Instruments," mimeo, Northwestern University.

MATZKIN, R.L. (2005) "Identification in Nonparametric Simultaneous Equations," mimeo, Northwestern University.

MATZKIN, R.L. (2006) "Heterogeneous Choice," in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress,* edited by R. Blundell, W. Newey, and T. Persson, Cambridge University Press.

MATZKIN, R.L. (2007) "Estimation in Nonparametric Simultaneous Equations," mimeo, Northwestern University.

McFADDEN, D.L. (1994) "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics,* 76, 689-708.

McFADDEN, D.L. (2006) "How Consumers Respond to Incentives," Jean-Jacques Laffont Lecture, mimeo, University of California at Berkeley.

McFADDEN, D.L. (2007) *Foundations of Economic Survey Research,* Princeton University Press, forthcoming.

McFADDEN, D.L. et al. (2005) "Statistical Analysis of Choice Experiments and Surveys," Marketing Letters, Volume 16.

McFADDEN, D.L., N. SCHWARZ, and J. WINTER (2003) "Measuring Perceptions and behavior in Household Surveys," mimeo, Mannheim Research Institute for the Economics of Aging.

MOFFIT, R. and G. RIDDER (2007) "The Econometrics of Data Combination," in *Handbook of Econometrics,* Vol. 6, edited by J.J. Heckman and E.E. Leamer, Amsterdam: Elsevier.

PHILIPSON, T. (1997) "Data Markets and the Production of Surveys," *Review of Economic Studies,* 64 (1), 47-72.

PHILIPSON, T. (2001) "Data Markets, Missing Data, and Incentive Pay," *Econometrica,* 69 (4), 1099-1111.

POTERBA, J. and L. SUMMERS (1986) "Reporting Errors and Labor Market Dynamics," *Econometrica,* 54 (6), 1319-1338.

SCHWARZ, N., H.J. HIPPPLER, B. DEUTSCH, and F. STRACK (1985) "Response Categories: Effects on Behavioral Reports and Comparative Judgements," *Public Opinion Quarterly*, 49, 388-395.

SCHWARZ, N., B. KNAUPER, H.J. HIPPLER, E. NOELLE-NEUMANN & F. CLARK (1991) Rating Scales: Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly*, 55, 618-630.

TOURANGEAU, R., L.J. RIPS, and K. RASINSKI (2000) *The Psychology of Survey Response.* New York, NY and Cambridge, UK: Cambridge University Press.

WANSBEEK, T. and E. MEIJER (2000) "Measurement Error and Latent Variables in Econometrics," New York: North Holland.