ELSEVIER

Journal of Econometrics 91 (1999) 273–298

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

Measurement errors: A principal investigator-agent approach

Tomas Philipson*, Anup Malani

Department of Economics, University of Chicago, Chicago, IL 60637, USA

Received 1 April 1997; received in revised form 1 July 1998

Abstract

This paper interprets the production of survey data as a labor market under asymmetric information. Measurement errors correspond to erroneous supply of observations by sample members. This paper discusses how to lower measurement errors using incentives that reward sample members for errorless supply and assesses the impact of such incentives on the sampling distributions of common estimators. When measurement errors are elastic with respect to such incentives, the latter can reduce bias even when information from the validation is not used and increase statistical efficiency when it is. We test our results on a survey of US physicians.

JEL classification: C1; C8; C9; I0; I1

Keywords: Measurement error; Survey design; Data collection

1. Introduction

The production of data and the functioning of the markets in which it takes place are concerns universal to all fields of positive economics and indeed to any field of empirical inquiry. However, greater emphasis in systematic analysis has been paid by economists to the *consumption*, rather than the *production*, of data. Much of economic data are obtained through surveys which record the behavior or experiences of included sample members or through social experiments in

* Corresponding author. E-mail: t-philipson@uchicago.edu

0304-4076/99/$ - see front matter
PII: S0304-4076(98)00078-5

which treatments are assigned by investigators. Whether the data of surveys or experiments are collected through mail, phone, interview, or direct observation, the difficulties introduced from not being able to produce data for all sample members (non-response) or the production of erroneous data (measurement errors) are well-known in applied work.

This paper interprets data production using surveys as a labor market under asymmetric information in which a principal investigator produces a data set by using the better informed agents that make up the sample. The information asymmetry is essential because the production of the data set would be unnecessary if the principal knew the information of sample members. However, the agency problem arises because the sample member has little stake in the final product made up of the data set, effort is unobservable, and erroneous production cannot be separated out from sampling variance. This agency problem suggests that contracts which align the sample member's incentives with those of the principal investigator may be of mutual interest to both parties.

We consider incentives affecting measurement errors that can be used in surveys or social experiments or any other context in which humans are observed. The basic way in which we suggest bias or efficiency may be improved is through so-called *validation incentives*. These incentives involve validating a very small part of the sample and compensating it if the produced outcomes are errorless. For example, in health surveys involving hospital experiences, patient self-reports could be verified most easily by using hospital records of the same events reported. This paper discusses how such incentives may be used to assess the impact of erroneous production on sampling distributions of common estimators. Although full-scale validation is often prohibitively costly, the advantage of our method is that it allows for less validation and may indeed only involve a single sample member being validated.

Although it may seem that validation incentives could be of theoretical interest, whether they affect the behavior of sample members is an empirical question. Many survey producers would argue that the effort required from sample members should be minimized, but such arguments concern behavior conditional on standard levels of compensation, which are often zero. We produced an experiment, *The Survey Supply Experiment* of The National Opinions Research Center (NORC), in which we tested our validation incentives on US physicians. This was part of larger survey sponsored by *US News and World Report* asking physicians about hospital quality. We found that the degree of error production may be highly elastic to validation incentives even among physicians – a group that a priori seemed likely to be unresponsive. Indeed, we found that monitoring 10% of the sample with an incentive of \$500 for errorless reply quadrupled the percentage of correct responses. We argue that these highly elastic measurement errors would enable one to separate out erroneous production from outcome distributions using the randomized incentives discussed.

The paper is briefly outlined as follows. Section 2 discusses validation incentives when the validation is small enough that the information it generates is not used in estimation. In this case, the bias may be reduced by incentives and we show how successful incentives are in achieving this, depending on what type of estimator is considered. Section 3 discusses the use of validation incentives when the amount of validation is large enough to be productive in the sense that errors may be estimated from the validation data. In this case, the data allow unbiased estimation so that the gains are in statistical efficiency. Section 4 provides our empirical estimates of measurement error elasticities. Lastly, the paper concludes in Section 5 by discussing what seems to be a rich set of issues raised by a labor economic approach to survey design.

The paper relates to several separate strands of work. There is, of course, a well-known literature devoted to principal-agent problems, and particularly the monitoring of agents in insurance, labor, financial and other market contexts. Our discussion differs from that literature in its main area of concern, econometrics, as well as in the fact that monitoring may be productive in the sense of being valued in itself for estimation purposes. The paper also relates to an extensive literature outside of economics on non-sampling errors in surveys and a large body of applied econometric work on measurement errors.[1] Our discussion complements this previous literature through a systematic analysis of the role of *exchange* in the production of data through the assumption of a rational supply-, and not just demand-, side for markets in observations. Although there exists an extensive *single-person* decision theoretic literature in statistics in general,[2] and survey sampling in particular, less emphasis has been put on the supply-side of observations as it interacts with the demand-side to make up the exchange of services involved in markets for observations. Finally, the paper relates to social scientific research on methods for improving surveys. This literature, however, contains little formal discussion of incentives to improve the statistical inferences made using the data produced.[3]

2. Bias effects of unproductive validation incentives

Consider measurement errors for a binary random variable. Following standard notation for measurement errors, let X denote the supplied outcome that is

[1] Representative of the literature on non-sampling errors in surveys are Bradburn and Sudman (1988), Beimer et al. (1991), and Lessler and Kalsbeek (1992). For a small subset of the work on measurement errors, see, e.g., Abowd and Zellner (1985), Anderson and Burkhauser (1984), Bound (1990), Butler et al. (1987), Manski (1990), and Rodgers et. al. (1993).

[2] As illustrated, for example, in the classics Savage (1977), Cochrane (1979), and Berger (1988).

[3] See Miller and Cannell (1982) and Cannell and Henson (1974). This research mainly focusses on nonmonetary incentives and only methods to raise response rates. There is no discussion of how incentives might be used to improve statistical inference, which is the focus here.

observed and X^* the true type of the supplier. The preferences of sample members to report correctly are captured in the utility function $u(x|x^*)$, where x and x^* equal 1 or 0. The knowledge of the respondent is represented by the subjective probability p of having the condition, $x^* = 1$.[4] We assume that the sources of erroneous supply by sample members are twofold preferences or knowledge. Fig. 1 presents the combinations of preferences and subjective knowledge that lead to erroneous reporting.

Preferences may cause erroneous supply for such variables as unemployment, abortion, or welfare participation, if sample members think they know but do not want to reveal their true status. Knowledge may be the cause of erroneous response for variables such as past visits to health care providers or any other variables for which sample members may themselves possess inaccurate retrospective information, even though they desire to respond truthfully. The basic agency problem stems from the fact that there is asymmetric information between the investigator and sample members who are assumed to possess, or at least have access at lower cost to, private information about their outcome.[5] So as to get the fundamental ideas of this paper across without confusion, we will focus on situations where either preference problems or knowledge problems arise, but not both.

A *validation incentive* monitors the sample with probability π and pays wages $w = (w_1, w_0)$ to the two types for a correct report. The expected utility comparison that leads one to report having the condition is

$$X = 1 \Leftrightarrow \mathrm{E}[U(X = 1)] \geqslant \mathrm{E}[U(X = 0)], \tag{1}$$

where

$$\begin{aligned} \mathrm{E}[U(X = 1)] &= p[u(1|1) + \pi w_1] + (1 - p)u(1|0), \\ \mathrm{E}[U(X = 0)] &= pu(0|1) + (1 - p)[u(0|0) + \pi w_0]. \end{aligned} \tag{2}$$

The expected utility from reporting 1 is a probability-weighted average of the utility of truthfully reporting 1, thereby earning the expected validation wage πw_1, and the utility from falsely reporting 1. The expected utility from reporting 0 is similarly defined, but with a validation wage of w_0. If $F(p, u)$ is the

[4] Throughout, we assume that respondents with limiting prior beliefs ($p = 0$ or 1) correctly know their type.

[5] Sometimes incentive problems between asymmetrically informed sample members and investigators can possibly be overcome by information intermediaries – third parties (so called proxies in survey practice) who do not have an incentive to misreport. Our incentives would apply to such intermediaries as well.

		Does the respondent think she knows the truth?	
		Yes: $p = 0$ or 1	*No*: $p \in (0,1)$
Does the respondent want to tell the truth?	*Yes*: $u(i \mid i) \geq u(j \mid i)$	No Problem	Knowledge Problem
	No: $u(j \mid i) \geq u(i \mid i)$	Incentive Problem	Both Problems

Fig. 1. Sources of erroneous reporting.

distribution of subjective probability assessments and preferences in the population, the fractions of each type with errorless supply is

$$s_1(w) = \Pr\{x = 1 | x^* = 1, w\} = \int X \, dF(p, u | x^* = 1),$$
$$s_0(w) = \Pr\{x = 0 | x^* = 0, w\} = \int [1 - X] \, dF(p, u | x^* = 0). \tag{3}$$

One important feature of these supply functions is that each is increasing in its own wage

$$\frac{ds_1}{dw_1} \geq 0, \qquad \frac{ds_0}{dw_0} \geq 0. \tag{4}$$

Let supply, when conditioning the other way – the probability of true type conditional on supplied outcome, be defined by

$$s_1^*(w) = \Pr\{x^* = 1 | x = 1, w\} = s_1(w) \frac{P(X^* = 1)}{P(X = 1)}$$
$$s_0^*(w) = \Pr\{x^* = 0 | x = 0, w\} = s_0(w) \frac{P(X^* = 0)}{P(X = 0)} \tag{5}$$

Then if one type of supply rises in its wage, this implies the other will too. Regardless of the conditioning, such elastic supply may result from a straight incentive effect when there are no knowledge problems ($p = 0$ or 1). When such knowledge problems are present, it may occur through *self-validation* of sample members. Under such self-validation, an alternative to choosing between reporting one's current subjective assessment of type is to undertake some activity, at

cost c_a, which alters the beliefs from p to the correct status. For example, in an income survey, this may involve the agent's time and effort of validating his financial status before responding, while, in a health survey, this may entail consultation of medical records. The expected utility calculus respondents face in the absence of preference problems is therefore

$$V(w) = \max\{X\mathrm{E}[U(X=1)] + [1-X]\mathrm{E}[U(X=0),$$
$$p[u(1|1) + \pi w_1] + (1-p)[u(0|0) + \pi w_0] - c_a\}. \tag{6}$$

Respondents can either not self-validate or self-validate. If they do not self-validate, they get the appropriate expected utility in Eq. (2) from optimally reporting 1 ($X=1$) or 0 ($X=0$). If they self-validate, they get a weighted average of telling the truth and its validated rewards, where the weights are determined by the subjective probabilities of being 1 or 0, without any risk of reporting falsely – all at a cost of c_a.

Fig. 2 illustrates the demand for self-validation when there are no preference problems. Here validation occurs for those most unsure, with those more sure about their status gaining less from self-validating. More precisely, there exist prior beliefs $p \in [p_L, p_H]$, where

$$p_L = \frac{c_a}{u(1|1) + \pi w_1 - u(0|1)},$$
$$p_H = 1 - \frac{c_a}{u(0|0) + \pi w_0 - u(1|0)}, \tag{7}$$

for which the benefits of self-validation justify its expense, and increasing compensation w pushes these bounds toward more extreme prior beliefs.

The second important feature of these supply functions allows one to distinguish whether knowledge or preferences are causing erroneous supply. The key difference turns out to be the cross-elasticity with respect to the compensation of other types. If preferences are the cause of erroneous supply, then $p=1$ or $p=0$ and cross-reported type price effects are zero

$$\frac{\mathrm{d}s_1}{\mathrm{d}w_0} = 0, \qquad \frac{\mathrm{d}s_0}{\mathrm{d}w_1} = 0. \tag{8}$$

This occurs because incentives for the other type do not affect one's behavior when one knows one's type. On the other hand, if subjects are uncertain about their status, these cross reported-type effects are negative

$$\frac{\partial s_1}{\partial w_0} \leqslant 0, \qquad \frac{\partial s_0}{\partial w_1} \leqslant 0. \tag{9}$$

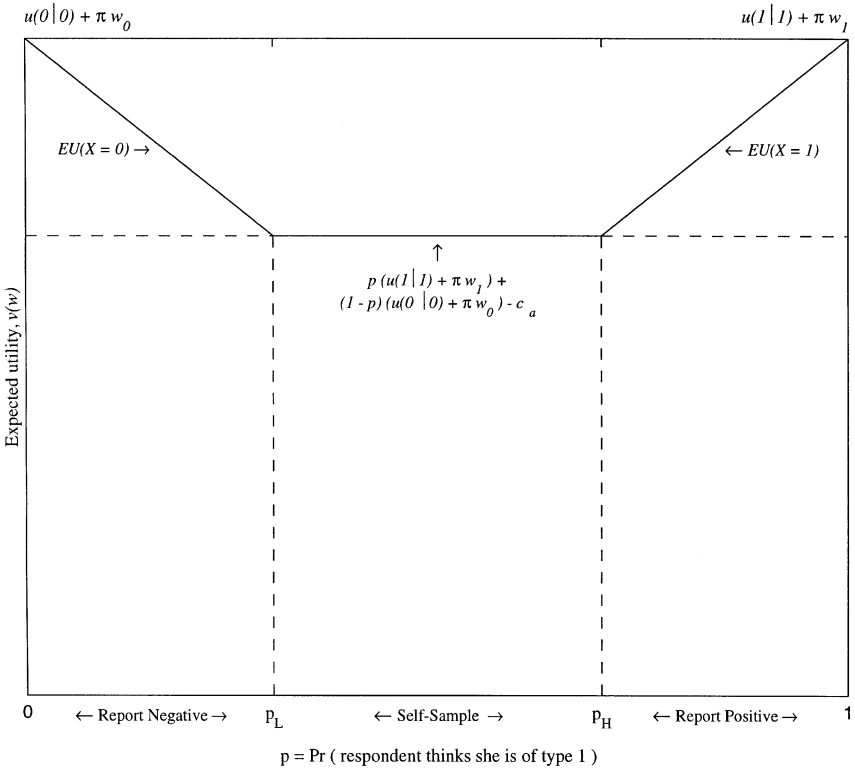

Fig. 2. Self-sampling of respondents.

This is because a sample member who does not know her type but is contemplating reporting type 1 is less likely to report 1 as the wage for 0 rises. These two patterns of the substitution matrix under imperfect knowledge, the positive own-reported type elasticity and the negative cross-reported type elasticity, impact the optimal incentive designs given an estimator.

While the focus here is on the binary response case, this framework can easily be extended to a multinomial setting with M possible responses. In that case, respondents supply response i if

$$\mathrm{E}[U(X=i)] \geqslant \mathrm{E}[U(X=j)] \tag{10}$$

for all $j = 1, \ldots, M$, where

$$\mathrm{E}[U(X=k)] = p_k[u(k|k) + \pi w_k] + \sum_{k' \neq k} p_{k'} u(k|k'). \tag{11}$$

This yields supply functions for reports of type i given a true type j,

$$s_{ij}(w) = \Pr\{x = i | x^* = j, w\} = \int_p \int_u I(x = i) f(p, u | x^* = j)\, du\, dp \tag{12}$$

and supply functions for true types j given reported type i

$$s^*_{ji}(w) = \Pr\{x^* = j | x = i, w\} = s_{ij}(w) \frac{\Pr\{x^* = j\}}{\Pr\{x = i\}}. \tag{13}$$

Knowledge and preference problems are defined just as in Table 1, except that the latter are associated with $p_j = 1$ and $p_k = 0$, for all $k \neq j$, for some j. As before, with preference problems, own-reported type price effects are positive and cross-reported type price effects are negative but cross-true type price effects are zero:

$$\frac{\partial s_{ii}}{\partial w_i} \geqslant 0, \quad \frac{\partial s_{ij}}{\partial w_j} \leqslant 0, \quad j \neq i, \qquad \frac{\partial s_{ij}}{\partial w_k} = 0, \quad k \neq j. \tag{14}$$

When knowledge problems arise, these cross-true type price effects are positive and cross reported-type effects are negative

$$\frac{\partial s_{ij}}{\partial w_i} \geqslant 0, \quad \forall j, \qquad \frac{\partial s_{ij}}{\partial w_k} \leqslant 0, \quad k \neq i. \tag{15}$$

As an illustrative example, consider McFadden's (1973, 1974) choice-based multinomial logit model. Applied to our response-choice problem, it yields

$$\mathrm{E}[U(X = k)] = V_k + \varepsilon_k, \tag{16}$$

where

$$V_k = p_k[u(k|k) + \pi w_k] + \sum_{k' \neq k} p_{k'} u(k|k') \tag{17}$$

and ε_k are distributed i.i.d. Weibull. It is straightforward to show that, under preference problems, the wage effects are

$$\pi s_{ii} \geqslant \frac{\partial s_{ii}}{\partial w_i} \geqslant 0, \quad -\pi s_{ij} \leqslant \frac{\partial s_{ij}}{\partial w_j} \leqslant 0, \quad j \neq i, \quad \frac{\partial s_{ij}}{\partial w_k} = 0, \quad k \neq j. \tag{18}$$

When knowledge problems are present, we get

$$\pi p_i s_{ij} \geqslant \frac{\partial s_{ij}}{\partial w_i} \geqslant 0, \quad \forall j, \quad -\pi s_{ij} \leqslant \frac{\partial s_{ij}}{\partial w_k} \leqslant 0, \quad k \neq i. \tag{19}$$

These results are consistent with the theory above.

2.1. Validation incentives for a dependent variable

Now consider the bias effects of using validation incentives when X is a dependent variable such as an unemployment or uninsurance rate. The supplied proportion is then determined by those having the condition supplying the truth and those who do not supplying erroneously:

$$P(X = 1) = s_1(w)P(X^* = 1) + (1 - s_0(w)P(X^* = 0). \tag{20}$$

The absolute value of the bias when estimating $P(X^* = 1)$ is then monotonically related to

$$\begin{aligned} B^2(w) &\equiv [P(X = 1) - P(X^* = 1)]^2 \\ &= [(1 - s_0(w)P(X^* = 0) - (1 - s_1(w))P(X^* = 1)]^2. \end{aligned} \tag{21}$$

The effect of validation incentives on this bias under *preference* problems only (no cross-price effects) is

$$\begin{aligned} \frac{\mathrm{d}B^2}{\mathrm{d}w_0} &= -2[B]P(X^* = 0)\frac{\mathrm{d}s_0}{\mathrm{d}w_0}, \\ \frac{\mathrm{d}B^2}{\mathrm{d}w_1} &= 2[B]P(X^* = 1)\frac{\mathrm{d}s_1}{\mathrm{d}w_1}. \end{aligned} \tag{22}$$

In contrast to the effect of incentives in the case of measurement errors with binary independent variables, this implies that incentives may increase bias with binary dependent variables. In particular, if the dependent variable $(X = 1)$ is over-reported, then compensating those who have the condition to reduce error makes the bias even larger, i.e. increasing w_1 increases over-reporting. Similarly, if the variable $(X = 1)$ is under-reported, giving those who do not have the condition more incentive to report truthfully makes it even more under-reported, i.e. increasing w_0 increases under-reporting. This occurs because when errors cancel each other out, as when equal over and under-reporting by both sides implies no bias, incentive effects are *non-separable* in the sense that effects of one incentive depends on the level of the other. This differs from the effects of

validation incentives for an independent variable in which case error reduction across types was perfectly substitutable.

However, if systematic errors are known to be over-reported,[6] this implies that the effect of increasing the incentives for the two types can be signed according to

$$\frac{dB}{dw_0} \leqslant 0,$$

$$\frac{dB}{dw_1} \geqslant 0. \tag{23}$$

The first effect says that the bias is reduced by making those who do not have the condition report it better. The second effect, however, says that it is counter-productive to reduce errors for those that have the condition: it involves costs that increase bias. Therefore, it follows that under such a systematic error of dependent variables, a corner solution will be optimal in which those who are on the side of the bias receive no incentives. As opposed to the case of independent variable incentives, in which supply was perfectly substitutable across the two conditions, compensation should not necessarily be higher for the type with most elastic measurement errors.

These results extend directly to the multinomial case. There, under preference problems, validation incentives have the following effects on bias in estimating $P(X^* = i)$:

$$\frac{\partial B(i,w)^2}{\partial w_i} = 2B\frac{\partial s_{ii}}{\partial w_i}\Pr\{x^* = i\} \geqslant 0,$$

$$\frac{\partial B(i,w)^2}{\partial w_j} = 2B\frac{\partial s_{ij}}{\partial w_j}\Pr\{x^* = j\} \leqslant 0,\ i \neq j. \tag{24}$$

[6] A natural question is where researchers learn of systematic errors that lead one type to be over-reported, and how this additional information about the nature of systematic errors might be used more directly in the estimation process, e.g., to bound the true parameter? As previous studies such as Parry and Crossley (1950), Ferber et al. (1969), and Lansing (1961) have suggested, it is common knowledge that behavior perceived as desireable is often over-reported. Examples include voting, contributions to charity, and how often one takes one's medicine. Conversely, undesireable activities such as illegal drug use or indebtedness are often under-reported. The problem is that, here, only the sign, not the magnitude of the effect, is known. Such information is not very useful without further knowledge on the fraction of the sample reporting without error, although an interesting future research question is whether our type of incentives can be combined solely with information on the sign of errors to improve statistical inference. If the fraction with errorless supply is known, Horowitz and Manski (1995) suggest techniques to bound the distribution and certain moments of outcome variables that are reported with error.

Again, increasing the incentive for type i respondents to report more truthfully when i is over-reported and increasing the incentive of any type but i to report correctly when i is under-reported only increases the bias.

2.2. Validation incentives for treatment effects

Instead of considering the level of a dependent variable, consider estimating the *difference* in the levels between two groups when the independent binary variables (indicating the respondent's group) are classified correctly but the dependent variable may not be. Such comparisons may be relevant for pre-post comparisons in which the periods themselves have no error but the supplied outcomes in both periods may have error. The difference in the proportions produced across the two covariates A and B is

$$\begin{aligned}\mathrm{E}[X|A] - \mathrm{E}[X|B] &= [P(X^* = 1|A)s_1(w) + P(X^* = 0|A)(1 - s_0(w))] \\ &\quad - [P(X^* = 1|B)s_1(w) + P(X^* = 0|B)(1 - s_0(w))] \\ &= [P(X^* = 1|A) - P(X^* = 1|B)](s_1(w) + s_0(w) - 1).\end{aligned} \tag{25}$$

This again implies that errorless production is perfectly substitutable across the two types.

In pre-post comparisons, many times the difference in outcome over the two periods reflects things other than the difference in treatments, such as time trends unrelated to the treatments. Therefore, one estimator that is often considered is the difference in pre-post effects across two treatments. Assume that sample members produce outcomes on treatments A and B at two times t and $t + 1$. Using the impact of errors on the difference estimator, the difference-in-differences estimator can be shown to satisfy

$$\begin{aligned}&(\mathrm{E}[X_{t+1}|A] - \mathrm{E}[X_t|A]) - (\mathrm{E}[X_{t+1}|B] - \mathrm{E}[X_t|B]) \\ &= \{[P(X^*_{t+1} = 1|A) - P(X^*_t = 1|A)] - [P(X^*_{t+1} = 1|B) \\ &\quad - P(X^*_t = 1|B)]\}\{s_1(w) + s_0(w)\}.\end{aligned} \tag{26}$$

Hence, as before, the two types are substitutable so that whether erroneous production is due to preferences or knowledge implies that bias is reduced with incentives.

The frequent occurrence of the perfect substitution case implies that incentives should be tailored towards the more elastic and that it is sensitive to these

elasticities. In particular, by standard arguments, the type that is more elastic to validation incentives should receive higher rewards.[7]

Unlike Section 2.1, these results extend only partially to the multinomial case. The difference in probability that $X = i$ across two groups A and B is a weighted sum of the differences in probability that $X^* = j$ across A and B, for $j = 1, \ldots, M$, where the weights do not necessarily sum to one

$$P(X = i|A) - P(X = i|B) = \sum_j \{P(X^* = j|A) - p(X^* = j|B)\}s_{ij} \tag{27}$$

This is not as clean a result as in the binary case: the bias is not mere attenuation. However, in the case of preference problems where respondents know their true type, increasing w_i raises the weight placed on the difference in true probability of being type i between the true groups

$$\frac{\partial P(X = i|A) - P(X = i|B)}{\partial w_i} = \{P(X^* = i|A) - p(X^* = i|B)\}\frac{\partial s_{ii}}{\partial w_i}. \tag{28}$$

When one faces knowledge problems, however, raising the incentive to report i truthfully simply increases the value of all the weights, not just s_{ii}. This may increase or decrease the observed difference.

2.3. Validation incentives for an independent variable

Suppose that Y is a correctly observed dependent variable, for which an independent variable X is to be measured. Consider the difference in measured means as it relates to the difference in actual means

$$\begin{aligned} \mathrm{E}[Y|X = 1] - \mathrm{E}[Y|X = 0] = {} & \mathrm{E}[Y|X^* = 1]s_1^*(w) \\ & + \mathrm{E}[Y|X^* = 0](1 - s_1^*(w)) \\ & - \mathrm{E}[Y|X^* = 1](1 - s_0^*(w)) \\ & - \mathrm{E}[Y|X^* = 0]s_0^*(w)]. \end{aligned} \tag{29}$$

[7] For example, if under incentive problems (no cross-elasticities) the wages are chosen to maximize total truth supplied the bias subject to a total cost constraint C we have $\max_w s_1(w) + s_0(w)$ subject to $\pi w_0 P(X^* = 0)s_0(w) + \pi w_1 P(X^* = 1)s_1(w) \leqslant C$ which implies that a necessary condition for an interior wage-choice is that

$$\frac{1/\varepsilon_1 + 1}{1/\varepsilon_0 + 1} = \frac{w_0 P(X^* = 0)}{w_1 P(X^* = 1)},$$

so elasticities and wages should be positively related at the optimum.

This can be rewritten as

$$\mathrm{E}[Y|X=1]-\mathrm{E}[Y|X=0]=(\mathrm{E}[Y|X^*=1]-\mathrm{E}[Y|X^*=0])(s_1^*(w)+s_0^*(w)-1). \tag{30}$$

In other words, the supplied effect of the covariate on the dependent variable is the true effect, diluted by the degree to which the independent variable is produced with error. With a binary independent variable, measurement error may produce bias in both the sign (if $s_1^*+s_0^*<1$) of the coefficient and its level.[8] This is because outcomes for those observed to have the condition are actually weighted averages of the outcomes of both types, making the difference between the two observed groups smaller than the true difference. Contrast this to the continuous independent variable case, where the sign of the coefficient is unaffected if measurement errors are independent of the dependent variables.

For the purpose of structuring incentives to reduce erroneous supply, the important result, however, is that truthful production is separable across the true conditions. In fact, truthful supply of each type are perfect substitutes in that only the *total* supply of truth across the two types, $s_0^*(w)+s_1^*(w)$, determines the bias. This implies that, regardless of whether erroneous supply is due to knowledge or preferences, i.e. whether there are always negative or there may be zero cross-reported type substitution effects, bias will always be reduced through more aggressive incentives.[9]

This result does not hold in general for the multinomial case with $M>2$. It can be shown that

$$\begin{aligned}\mathrm{E}(y|x=i)-\mathrm{E}(y|x=j)&=\sum_k\{\mathrm{E}[y|x^*=k]-\mathrm{E}[y|x^*=j]\}(s_{ki}^*-s_{kj}^*)\\&=\{\mathrm{E}[y|x^*=i]-\mathrm{E}[y|x^*=j]\}(s_{ii}^*-s_{ij}^*)\\&\quad+\sum_{k\neq i,j}\{\mathrm{E}[y|x^*=k]-\mathrm{E}[y|x^*=j]\}(s_{ki}^*-s_{kj}^*).\end{aligned} \tag{31}$$

[8] This stands in contrast to the continuous variable case when the measurement error is uncorrelated with the true value of the independent variable. There the sign of the observed value remains the same as that of the true value. But in the binary case, measurement errors are necessarily negatively correlated with the truth. As such, they can change the sign of the observed response. Similarly, in the continuous case, if measurement errors are negatively correlated with the true value, the sign of the observed value may differ.

[9] Note the differential impact incentives have on Type I and II errors when testing a null hypothesis of a zero effect. Incentives may increase the power to detect alternatives since alternative effects are magnified. However, incentives do not affect significance levels because the produced effect is still zero under the under the null of no treatment effect. For example, in a social experiment or clinical trial with self-reported compliance, incentives would increase the ability to detect an effective treatment but have little effect on the location of the sampling distribution under the null.

In the binary case, because outcomes are weighted averages of the outcome of the two true types, differencing does not involve any terms for types other than those differenced. If $M \geqslant 3$, however, differencing will involve true types other than the two differenced. Hence, bias may change both the sign and level of the coefficient in a more complicated way. Moreover, there is no true substitutability of truthful supplies: as written above only the truthful supply of i for sure lowers the bias in the observed difference between i and j. Increasing the truthful supply of other types (even j) simply alters the relative weights of the true differences of other types that comprise the observed difference in i and j, but does not necessarily lower them.

3. Efficiency effects of productive validation incentives

The previous bias effects assumed that the validation itself was not on a large enough scale to enter into the estimation of the parameters of interest. When the validation is productive in the sense that the information it generates is used in the estimation, then not only is there no bias, but statistical efficiency gains may be had. When the monitoring is productive in this sense, the full estimation problem is one of multi-stage sampling.[10] Consider our simple case of producing estimates of the true and reported population fractions for a binary dependent variable. If N sample members supply observations, the data may be described as $(X,\{Z_{x,x^*}\})$, where X are the number of positive reports as before of the $N(1-\pi)$ sample members not monitored and Z_{x,x^*} are the data of the πN validated sample members that report x and are of type x^*. The variable X is binomial and the variable Z_{x,x^*} is multinomial.

The likelihood function for the data given the wages is proportional to

$$f(\theta|\pi, w) = (s_1^*)^{Z_{11}}(1-s_1^*)^{Z_{10}}(1-s_0^*)^{Z_{01}}(s_0^*)^{Z_{00}} \times P(X=1)^{X+(Z_{11}+Z_{10})}(1-P(X=1))^{N(1-\pi)-X+(Z_{00}+Z_{01})}, \quad (32)$$

where the coefficients of interest, $\theta = (s_1^*, s_0^*, P(X=1))$, are the conditional probabilities of having and not having the condition given that you do and do not report it, respectively, and the probability of reporting type 1. Maximizing

[10] This problem has been worked out by Tenenbein (1970, 1971, 1972), Hochberg and Tenenbein (1983) and Selen (1986).

and rearranging yields the estimators

$$\hat{P}(X=1)=\frac{X+Z_{11}+Z_{10}}{N},$$

$$\hat{s}_1^*=\frac{Z_{11}}{Z_{10}+Z_{11}},$$

$$\hat{s}_0^*=\frac{Z_{00}}{Z_{01}+Z_{00}}. \tag{33}$$

This suggest the estimator

$$\hat{P}(X^*=1)=\hat{s}_1^*\hat{P}(X=1)+(1-\hat{s}_0^*)\hat{P}(X=0) \tag{34}$$

for the true probability of having the condition.[11] This estimator is simply the sample analog of the population relation: the sum of the fraction of the population that reported positive, corrected by the conditional probability that they reported correctly, and the fraction that reported negatively, corrected by the conditional probability of reporting incorrectly, with the corrections based on the validated subsample.

The delta method can be shown to imply that the estimator is asymptotically normal with mean and variance

$$\text{Plim } \hat{P}(X^*=1)=s_1^*P(X=1)+(1-s_0^*)P(X=0)=P(X^*=1), \tag{35}$$

$$\text{Asy.Var}\,[\hat{P}(X^*=1)]=Q\frac{P(X^*=1)P(X^*=0)}{N}+(1-Q)\frac{P(X^*=1)P(X^*=0)}{\pi N},$$

where

$$Q\equiv\frac{P(X^*=1)P(X^*=0)}{P(X=1)P(X=0)}[s_1-(1-s_0)]^2 \tag{36}$$

[11] This is very much in the spirit of Imbens and Lancaster (1994) and Imbens and Hellerstein (1998) which seek to combine information on moments of certain economic variables from macroeconomic data with micro data on these and other variables to improve the efficiency of estimating microeconometric models involving these variables. Here we seek to combine information on $P(X=1)$ from the entire sample with information on (s_1^*, s_0^*) from just the validated subsample to estimate a simple function of these estimators. The major difference is that sampling sampling variance is still present for the marginal distribution estimated.

which is simply the square of the correlation between the reported type and true type. It measures how well the true type can be predicted from the reported type and, as such, is termed the coefficient of reliability between supplied and actual status. In other words, the estimator is asymptotically unbiased with a variance that is a reliability-weighted average of the variance without any reliability and that with full reliability.

This result holds with few modifications in the multinomial setting. It can easily be shown that the estimator

$$\hat{P}(X^* = i) = \sum_j \hat{s}^*_{ij} \hat{P}(X = j), \tag{37}$$

where $\hat{s}^*_{ij}$ and $\hat{P}(X = j)$ are estimated in the obvious manner, is asymptotically normal with mean $P(X^* = i)$ and asymptotic variance

$$\text{Asy.Var}\,[\hat{P}(X^* = i)] = Q_i \frac{P(X^* = i)P(X^* \neq i)}{N} + (1 - Q_i)\frac{P(X^* = i)P(X^* \neq i)}{\pi N}, \tag{38}$$

where

$$Q_i \equiv \frac{P(X^* = i)}{P(X^* \neq i)}\left[\sum_j \frac{s^*_{ji}}{P(X = j)} - 1\right]. \tag{39}$$

The definition given to Q_i here is the same as that for Q above.

Returning to the binary response setting, one may interpret the reliability coefficient Q as the *quality* of labor supplied, in the sense that, as the fractions that respond truthfully (s_1, s_0) rise, so does Q. The sample size, N, on the other hand, measures the *quantity* of labor supplied. Under productive monitoring, therefore, there is a cost-based trade-off between the quality of production, which may be increased by more aggressive incentives, and the quantity of production which may be increased by a larger labor force. Indeed, assume that the quality and quantity are related through a budget constraint defined by the implicit function

$$\Phi(N, w, \pi) = 0 \tag{40}$$

such that $\Phi_N, \Phi_w, \Phi_\pi \geqslant 0$. Then the sample size afforded under a given wage and monitoring probability $N(w, \pi)$ is decreasing in each of its arguments. In this case we see that the mean-squared error $MSE = B^2 + V$ of the estimator for two

wages w and w' may be written

$$MSE(w, \pi) \geqslant MSE(w', \pi') \Leftrightarrow \frac{Q(w, \pi)}{N(w, \pi)} + \frac{1 - Q(w, \pi)}{N(w, \pi)\pi} \leqslant \frac{Q(w', \pi')}{N(w', \pi')} + \frac{1 - Q(w', \pi')}{N(w', \pi')\pi'}. \quad (41)$$

The essential trade-offs that determine whether larger compensation is preferred is then how wage-elastic the quality is compared to how large the foregone sample size is. To see how this compares to the inelastic case that may be considered the statistical benchmark, consider the budget constraint

$$\Phi(N, w, \pi) = N[c + \pi[c_v + w_1 P(X^* = 1)s_1 + w_0 P(X^* = 0)s_0]] - C, \quad (42)$$

where c is the cost of sampling, c_v[12] the cost of validation by the principal, and C the size of the total budget. Thus, the total cost is the sampling cost plus validation costs and wage expenditures for those who produce correct outcomes. It can easily be shown that a necessary condition for an interior solution when errors are inelastic has no compensation ($w = 0$) and the optimal monitoring probability

$$\pi_0 \equiv \sqrt{\left(\frac{1 - Q_0}{Q_0}\right)\left(\frac{c}{c_v}\right)}, \quad (43)$$

where Q_0 is the level of the inelastic quality. In other words, the percentage of the sample monitored rises in the relative costs of monitoring to sampling and falls in the quality of production. For example, if the supplied outcomes are fully reliable without incentives, $Q_0 = 1$, then, obviously, no validation by the principal takes place, $\pi_0 = 0$. The fraction of the total budget spent on monitoring is

$$\frac{N\pi c_v}{C} = \frac{1}{\frac{c}{c_v}\frac{1}{\pi_0} + 1}. \quad (44)$$

As the cost of direct sampling converges to the cost of monitoring (c_v/c goes to unity) and as the quality of the observations goes to zero (π_0 goes to unity), this

[12] This is the analog quantity for the principal to the self-validation cost c_a for the agents (sample members) discussed before.

budget fraction goes to one-half – the budget share when everyone is validated at the same cost of sampling.

This degree of principal validation and the implied budget fractions are desired for statistical purposes but is excessive when measurement errors are elastic. If η is the common and constant elasticity of the supply schedules s_1 and s_0, then it is straightforward to show that the optimal monitoring probability starts at the statistical levels but falls with more elastic supply while the optimal validation wage starts at zero and rises as supply becomes more elastic

$$\pi(\eta = 0) = \pi_0, \quad \frac{\partial \pi}{\partial \eta} \leqslant 0,$$

$$w(\eta = 0) = 0, \quad \frac{\partial w}{\partial \eta} \geqslant 0. \tag{45}$$

Therefore, the more elastic are measurement errors, the larger is the *over-validation* and the less sufficient is the validation wage of the statistical solution under inelastic supply since resources can be used to improve quality by raising validation wages instead of monitoring the sample. Note that since monitoring is productive in the inelastic case, monitoring takes place, in contrast to most principal-agent problems where monitoring is unproductive.[13]

4. Empirical estimates of measurement error elasticities

This section provides illustrative estimates of how elastic measurement errors are to validation incentives using an incentive experiment on a random sample of US physicians. Even among this group that one may expect a priori to be inelastic to survey incentives, we find strong evidence of elastic measurement errors. This was investigated in a module we designed called the *Survey Supply Experiment* (SSE) of the physician survey *The Index of Hospital Quality* (IHQ). The main survey IHQ was produced by The National Opinion Research Center (NORC) at The University of Chicago during October–December 1995 and the SSE was added to the main survey on a subsample of the entire sample.[14] The survey was a mail survey privately sponsored by the *US News and World Report*. The private funding is important for the type of incentives used to identify production bias because both the levels and variation in survey wages are

[13] See, e.g., Becker (1968) or Polinsky and Shavell (1979).

[14] See Philipson and Grabowski (1996) for a more complete description of the experiment and its design.

regulated in publicly financed surveys in the US under the wage regulations imposed by the Office and Management and Budget (OMB).[15]

The sample consisted of a total of 2550 physicians, with 150 individuals in each of 17 specialties. The survey was very short in length and in the amount of time required to complete it. It contained a total of 34 items and took about 5 minutes to complete. The survey frame of the sample was the physician directory of The American Medical Association (AMA). This frame has been produced by NORC for other physician surveys and includes members as well as non-members and is the primary frame used for physician surveys in the US. This frame is typical in that it contains information on the frame members prior to sampling. We used this frame information to test the validation incentives by including two questions on the survey about the values of two frame variables which we already knew. In this respect we could, without additional costs, assess the measurement errors, i.e., whether supplied observations matched the frame, for the *entire* sample not only those validated. The objective and factual frame-variables to which the validation incentive were tied concerned the exact date of graduation from medical school as well as the number representing the AMA Code for the physicians medical specialty.[16]

The design of the incentive experiment was a randomized block design involving several types of incentives of which we only focus on the validation incentives. This validation incentive was assigned in equal proportions to a set of 6 specialties allocated to the experiment (out of a total of 17), to neutralize unobservables that were not independently distributed with specialties. In particular, the incentive was allocated to 120 individuals across these six specialties which we compare with two types of control groups. The first contains sample members in the six specialties that did not receive any validation incentive. The second group consists of the entire set of sample members in the survey not receiving the incentive.

The sample members receiving the validation incentive were given a simple set of instructions explaining how the incentive worked (see Appendix A1). The amount of the incentive was \$50 in expected value terms, with the monitoring probability being ten percent of the treatment group ($\pi = 0.10$) and the award conditional on having been monitored and giving a truthful response being five hundred dollars ($w = 500$). We did not vary compensation across truthfully reported types for three reasons. One, the frame variables we had access to were not of substantive interest in themselves for this particular experiment. For

[15] The total operating budget of the survey was \$170000 with an average cost per observation of \$170000/2550 = \$67 (throughout the paper we use 1996 dollars).

[16] The exact wording of the two questions were (italic not added); 1. What is the *exact date* when you received your medical degree? 2. What is the (3 digit) AMA code of your primary specialty?

example, we did not care what share were really cardiologists rather than pediatricians. Two, there were too many specialty types and dates of graduation to vary wages across. Such variations would have made the survey far too complicated. Three, we had no a priori reason to suspect that there was over-reporting of certain types that needed to be corrected. Nor did we have prior information on the relative elasticities of each type to help us decide on how much incentives for under-reported types would have to be raised. Moreover, we didn't attempt to break the sample into a control and multiple treatment groups with different levels of incentives because we felt that the sample size of each group would be too small, especially after validation. Also, such a breakdown would have substantially increased the complexity of the experiment beyond what the sponsor deemed worthwhile given its budget constraint. The sample members were not informed about the fact that they participated in an experiment in which their incentives did not necessarily equal those of others and in which pay was randomly assigned.

We calculated the mean errorless supply rate of specialty codes and graduation dates for the subsample with and the subsample without validation incentives. Our basic finding is that sample members that faced validation incentives were much more likely to supply their true specialty code (27% versus 11%). The difference in their truthful supply of graduation dates was *not*, however, found to be significant (83% for those with incentives versus 87% for those without). We further broke down the sample by gender, income, age and medical specialty and found those conditional results were consistent with the unconditional results.

The poor quality of the supply of AMA medical specialty codes among sample members is a product of two factors: (1) doctors generally do not memorize the AMA code for their medical specialty and (2) doctors did not receive a table of the codes with the survey. This suggests that doctors were unsure of their specialty code (p_i for specialty code i is not very high). In fact, only 11% correctly stated their code in the control group. Assuming doctors wanted to tell the truth, this suggests that their prior beliefs regarding their specialty *code* fell in the interval (p_L, p_H) for which self-validation was profitable. On the other hand, doctors generally think they know their graduation date, i.e., they posses extreme prior beliefs regarding this outcome. The data suggest 87% actually do. As such their beliefs are more likely to fall outside the same interval, making self-sampling less productive. So validation incentives and knowledge were such that doctors were induced to self-sample and improve their estimates only for specialty codes, but not for graduation dates.

Table 1 reports the marginal effects from a probit regression estimating how elastic the supply functions, $s(w)$, were to the validation incentive which enters in as randomized treatment dummy in all the specifications displayed. The table is for the specialty code outcome. The table shows that for successively larger specifications, taking into account differences in other determinants of errorless

Table 1
Marginal effect estimates from probit on measurement error effects. Dependent variable: MtchCode (supplied errorless speciality code). Standard error estiamtes for marginal effects are in parentheses

	Equation			
Independent variable[a]	(a)	(b)	(c)	(d)
Validation incentive	0.163**	0.170**	0.195**	0.241**
	(0.064)	(0.069)	(0.071)	(0.085)
Income < 50K		0.055	0.033	0.165
		(0.097)	(0.090)	(0.162)
Income 50–100K		− 0.068	− 0.061	− 0.039
		(0.050)	(0.047)	(0.047)
Income 100–150K		− 0.010	− 0.016	0.031
		(0.045)	(0.041)	(0.048)
Income 150–175K		0.020	0.010	0.093
		(0.061)	(0.056)	(0.087)
Income 175–200K		− 0.019	− 0.004	0.015
		(0.051)	(0.051)	(0.051)
West			− 0.098**	− (0.072)**
			(0.030)	(0.029)
North			− 0.083**	− 0.056
			(0.031)	(0.030)
North-central			− 0.017	− 0.011
			(0.038)	(0.036)
Age 25–39			0.000	0.010
			(0.052)	(0.051)
Age > 55			0.097**	0.056
			(0.045)	(0.042)
Male			0.087**	0.063*
			(0.036)	(0.034)
Specialty cardiology[b]				
Specialty cancer				0.112
				(0.110)
Specialty neurology				0.031
				(0.082)
Specialty opthamology				0.427**
				(0.130)
Specialty orthopedics				0.469**
				(0.123)
Number of obs	429	368	368	311
Log likelihood	− 157.823	− 137.217	− 127.152	− 93.038
Pseudo-*R*-squared	0.028	0.037	0.108	0.305

[a]Marginal effects are for discrete changes of a dummy variable from 0 to 1.
[b]Specialty cardiology was dropped because it predicts perfectly. 57 observations were lost.
*Indicates that the marginal effect is statistically different from zero at the 10% level.
**Indicates that the marginal effect is statistically different from zero at the 5% level.

production that may not be balanced across treatment groups, the incentive effect remains fairly robust and highly significant. Indeed, the incentive effect is the most significant determinant of errorless production throughout the specifications considered.

Throughout the specifications, the effect of the validation incentive on truthful supply is positive and highly significant with the relationship being stronger with more elaborate controls for truthful supply. The income bracketing has been refined to cover five as opposed to two income categories. The evidence above suggests that among US physicians, a group of individuals who a priori would seem unlikely to respond to survey incentives, there is substantial evidence that modest (average) validation incentives can have strong effects on the effort put into the production of observations to lower measurement errors.

To show the effect of these incentives on the bias, in Fig. 3, we plot the change, when the validation wage rises by \$10 in the absolute ($B(i, w)$) and relative

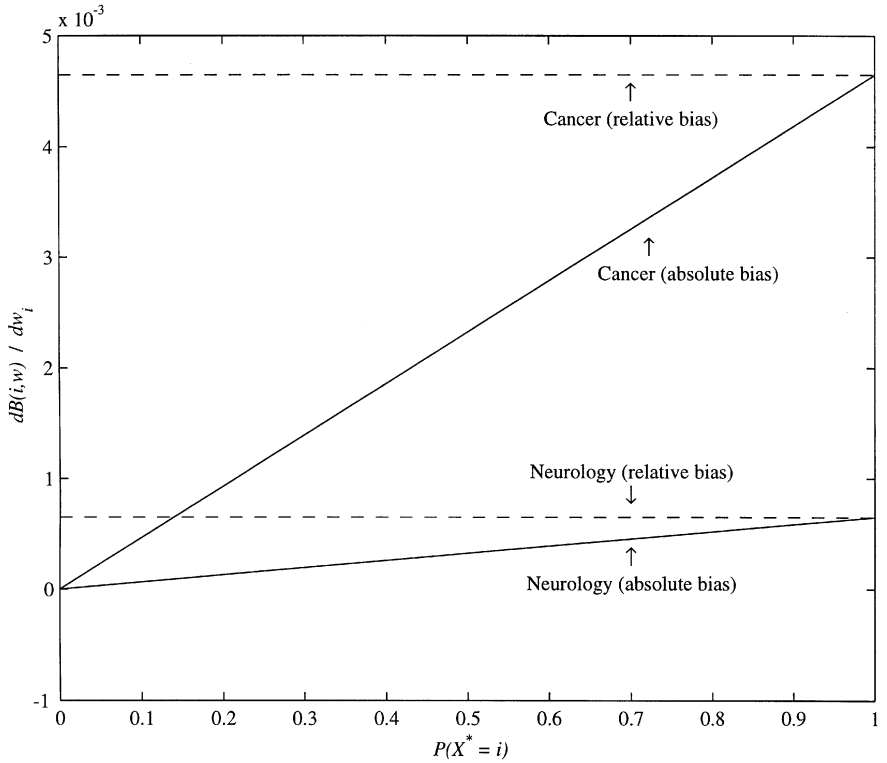

Fig. 3. Change in absolute and relative bias due to a \$10 increase in w_i when estiamting $P(X^* = i)$, with fixed $\pi = 0.1$.

$(B(i, w)/P(X^* = i))$ biases in estimating true proportions from reported ones. As suggested in Eq. (22), these changes depend on the true proportion of the type in question as well as the elasticity of errorless supply for that type, so we present the results for two types : cancer specialists and neurology specialists. Estimates of supply elasticities for each type are drawn from probits of errorless supply on dummies for validation incentives run on subsamples comprised only of cancer specialists and neurology specialists. Because $\pi = 0.1$, a \$10 increase in the validation wage is equivalent to a \$1 increase in the expect wage from truth-telling. The absolute bias (which is negative in the case of under-reporting) rises as the true proportion rises, suggesting that validation wages are more effective for correcting under-reporting the higher is the true proportion. The relative bias, on the other hand, is constant across true proportions. This is a product of the fact that relative bias is simply the elasticity of errorless supply and that we assume here that our estimate of supply elasticity holds for all true proportions. Notice that since cancer specialists are more responsive to validation wages, the bias in estimating their true proportions from their reported proportions changes more for each incremental change in the validation wage than does bias for neurologists.

5. Concluding remarks

This paper showed the benefit of using validation incentives to reduce bias when the fraction validated is small or to increase efficiency when the fraction is large so that it provides information that can be used in estimation. The paper documented the empirical relevance of elastic measurement errors with respect to such incentives. We conclude by discussing some of the shortcomings of the analysis as well as future questions raised by a labor economic approach to survey design.[17]

Although we were able to document that validation wages were able to lower the number of people reporting erroneously, we were unable document the effects of these same wages on the extent of measurements errors among those that reported erroneously. Our contractual agreement with *US News and World Report*, the sponsor for *Survey Supply Experiment* of the larger NORC physician survey, did not provide us with the actual and validated responses of physicians in the survey. Instead, we only received dummy variables that indicated whether each physician was part of the control or treatment, and if part of the latter, whether she gave the correct answer to each of the validated questions. As such we could not provide measures of differences in the extent of measurement

[17] See also Philipson (1997), Philipson and Desimone (1997), and Philipson and Hedges (1998).

errors in erroneous responses from the treatment and control groups. Consequently, we focused on the external margin of truthful response rather than the internal margin that governs the extent of errors in non-truthful response.

There is, of course, a whole research agenda concerned with making survey instruments better and more user-friendly. In many cases, it becomes infeasible to ask questions that are of interest because they are ruled out as unlikely to be producible by sample members that are not willing to undertake time and effort to supply better responses. This *is* the principal-agent aspect of survey design. The problem is that this research agenda is motivated by the fact that sample members need user-friendly instruments *conditional* on the low-powered incentives currently observed. Answers to better questions, as well as more extensive efforts on the part of sample members, may be obtained by use of incentives whether aimed at measurement errors, as discussed here, or other aspects of data production such as unit- and item-response.

Our discussion highlights some of the potential drawbacks to wage regulations, both in maximum wages imposed as well as the restrictions on wage discrimination, of publicly funded surveys in the US. Survey producers in the US are affected by wage regulations set by federal agencies, such as the Office of Management and Budget (OMB). In particular, maximum wage policies are often invoked. These policies are often justified, by economists as well as other survey producers, by the argument that compensating sample members would unjustifiably inflate survey budgets. However, such restrictions on the feasible set of survey production inputs may increase, rather than decrease, the cost of production. A more sensible argument against compensation may be that when compensation is used, income effects arise which imply that behavior is observed that would not be otherwise. This argument does not apply to most surveys since they are *retrospective*, and past behavior is presumably inelastic to unexpected current compensation. Such concerns may be relevant for panel surveys, however, if back-loaded compensation is used.

Indeed, panels raise a large set of issues abstracted from here. They correspond to long-term contracts between the principal and agent as opposed to cross-sections which involve spot markets. Of particular interest is the fact that the production of panels may often be inconsistent with perfect recall on the part of sample members, a model of knowledge which dominates current economic models of dynamic choice. There is a tension therefore between models of the behavior of agents sampled and the survey design since few panels should be produced if sample members behaved according to current dynamic models. The interaction between memory and compensation in panel production are non-trivial, since with perfect recall the production of repeated measurement is not optimal relative to a single retrospective study asking the sample member to recall long histories.

Acknowledgements

We thank John Cawley, Tom Lawless, and Diana Seger for their research assistance and two anonymous referees of the Journal for recommendations that helped improve the paper. Gary Becker, Norman Bradburn, Phil DePoy, James Heckman, Joseph Hotz, Edward Kaplan, Michael Kremer, Casey Mulligan, Edward Prescott, Kirk Wolters, Arnold Zellner, and seminar participants at University of Chicago, Yale University, the 1994 Econometric Society Meetings, and the 1995 NSF/CEME Micro Econometrics Conference at the University of Wisconsin provided useful comments. We are especially thankful to many members of the National Opinion Research Center (NORC) for many discussions, especially Craig Hill and Krishna Winfrey at the Health Section of NORC for helping us produce the data used. Philipson is grateful for support from the NSF (SBR-9709635) and the Alfred P. Sloan Foundation Faculty Research Fellowship.

References

Abowd, J., Zellner, A., 1985. Estimating gross labor force flows. Journal of Business and Economic Statistics 3, 254–283.

Anderson, K., Burkhauser, R., 1984. The importance of the measure of health in empirical estimates of the labor supply of older men. Economic Letters 16, 375–380.

Becker, G., 1968. Crime and punishment: an economic approach. Journal of Political Economy 76, 169–217.

Beimer, P., Groves, A., Lyberg, L., Mathiowetz, N., Sudman, S., 1991. Measurement Errors in Surveys. Wiley, New York, NY.

Berger, J., 1988. Statistical Decision Theory and Bayesian Analysis. Springer, New York, NY.

Bound, J., 1990. Self-reported versus objective measures of health in retirement models. Journal of Human Resources 26, 106–138.

Bradburn, N., Sudman, S., 1988. Polls and Surveys. Jossey-Bass, San Francisco, CA.

Butler, J., Burkhauser, R., Mitchell, J., Pincus, T., 1987. Measurement error in self-reported health variables. Review of Economics and Statistics 69, 644–650.

Cannell, C., Henson, R., 1974. Incentives, motives, and response bias. Annals of Economic and Social Measurement 3, 307–318.

Cochrane, W., 1979. Survey Sampling. Wiley, New York, NY.

Ferber, R., Forsythe, J., Guthrie, H., Maynes, E.S., 1969. Validation of a national survey of consumer financial characteristics: savings accounts. Review of Economics and Statistics 51, 436–444.

Hochberg, Y., Tenenbein, A., 1983. On triple sampling schemes for estimating from binomial data with misclassification errors. Communications in Statistics 12, 1523–1533.

Horowitz, J., Manski, C., 1995. Identification and robustness with contaminated and corrupted data. Econometrica 63, 281–302.

Imbens, G., Hellerstein, D., 1998. Imposing moment restrictions from auxiliary data by weighting. Review of Economics and Statistics, forthcoming.

Imbens, G., Lancaster, T., 1994. Combining micro and macro data in microeconometric models. Review of Economic Studies 61, 655–680.

Lansing, J.B., Ginsburg, G., Braaten, K., 1961. An investigation of response error. Bureau of Economic and Business Research, Urbana, IL.

Lessler, J., Kalsbeek, W., 1992. Nonsampling Error in Surveys. Wiley, New York, NY.
Manski, C., 1990. The use of intensions data to predict behavior: a best case analysis. Journal of the American Statistical Association 85, 934–940.
McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, NY.
McFadden, D., 1974. The measurement of urban travel demand. Journal of Public Economics 3, 303–328.
Miller, P., Cannell, C., 1982. A study of experimental techniques for telephone interviewing. In: Singer, E., Presser, S. (Eds.), Survey Research Methods. University of Chicago Press, Chicago, IL.
Parry, H., Crossley, H., 1950. Validity of response to survey questions. Public Opinion Quarterly 14, 61–80.
Philipson, T., 1997. Data markets and the production of surveys. Review of Economic Studies 64, 47–72.
Philipson, T., Desimone, J., 1997. Subject sampling and experiments. Biometrika 84, 618–632.
Philipson, T., Grabowski, D., 1996. The survey supply experiment: design and summary statistics. Mimeo, Department of Economics, University of Chicago, Chicago, IL.
Philipson, T., Hedges, L., 1998. Subject evaluation in social experiments. Econometrica 66, 381–408.
Polinsky, M., Shavell, S., 1979. The optimal tradeoff between the probability and magnitude of fines. American Economic Review 69, 880–891.
Rodgers, W.L., Brown, C., Duncan, G.J., 1993. Errors in survey reports of earnings, hours worked, and hourly wages. Journal of the American Statistical Society 88, 1208–1218.
Savage, L., 1977. The Foundations of Statistics. Dover, New York, NY.
Selen, J., 1986. Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. Journal of the American Statistical Association 81, 75–81.
Tenenbein, A., 1970. A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Society 65, 1350–1361.
Tenenbein, A., 1971. A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. Biometrics 27, 935–944.
Tenenbein, A., 1972. A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. Technometrics 14, 187–202.