A Comprehensive Assessment of Selection in a Major Internet Panel for the Case of Attitudes toward Government Regulation

by

T.A. Cameron, University of Oregon
J.R. DeShazo, UCLA

.

A Comprehensive Assessment of Selection in a Major Internet Panel for the Case of Attitudes toward Government Regulation

ABSTRACT

The rise of large internet consumer panels for survey research raises the question of whether samples of respondents drawn from these panels are representative of the underlying population. To assess this question we model the attrition/selection process for one major consumer panel, maintained by Knowledge Networks, Inc (KN). Starting from KN's over 525,000 random-digit-dialed (RDD) panel-recruitment telephone contact attempts, and ending with a sample of respondents to an actual online survey, we span all junctures at which systematic selection could occur. Our analysis begins by matching addresses or telephone exchanges to the appropriate census tract for the initial half-million residential telephone numbers in the RDD contact-attempt pool. This permits us to use a set of fifteen orthogonal factors based on census tract characteristics, plus county voting percentages in the 2000 Presidential election, to look for neighborhood characteristics that influence whether an initial RDD contact attempt eventually results in a usable response to a specific survey with a sample size of 2,911. We then examine how non-random selection into the estimating sample affects respondents' answers to one specific survey question about the proper role of government in environmental, health and safety regulation. Using two distinct approaches, we do find evidence of modest sample selectivity. However, we find that these selection effects are not statistically significant in explaining respondents' attitudes about the proper role of government in society.

Keywords: sample selection, consumer panel, census tracts, voting percentages, GIS, attitudes toward government regulation

1 Introduction

In recent years, online survey methods have made rapid gains in popularity among researchers. Deutskens, et al. (2006) note that by 2004 about 35% of the U.S. survey research market consisted of online surveys.[1] A large number of survey research firms now offer this mode of delivery (see Evans and Mathur (2005) and Wright (2005)). The online survey mode is attractive because it allows researchers to reduce field costs and improve response and data processing times. Despite these advantages, the sampling properties of these surveys can be less than ideal. As Best, et al. (2001) note, most Internet sampling procedures "only permit the generation of diverse, not representative, samples." Much recent effort has been devoted to assessing the representativeness of online surveys as compared to traditional random-digit-dialed (RDD) telephone surveys or mail surveys.[2]

Two of the leading U.S. online survey research firms are Knowledge Networks, Inc. (KN, formerly Intersurvey) and Harris Interactive, Inc. (HI, formerly Harris Black International).[3] There are a variety of ways to recruit members for an online survey panel.[4] Berrens, et al. (2003) provide a detailed description of the recruitment methods used by each firm.[5]

KN endeavors to build a panel which is representative to begin with. This company recruits its panelists via an initial attempt at RDD telephone contact. People who are willing to join the panel are equipped with Web-TV hardware and Internet access if they do not have their own computer and access to the Internet. However, members of households contacted by RDD become KN panelists only after they have survived several types of attrition.[6] Our study explores the representativeness of a KN survey sample, compared to the company's initial set of RDD telephone contact attempts.

In contrast, HI uses a wide range of recruitment methods, but panel membership is conditional on the panelist already having web access capability. Our study does not employ HI data, but this company's strategy involves ex post methods to correct for non-representativeness. While its recruitment methods cannot be expected to yield a representative panel, the company has developed a method using "propensity scores" to construct post-stratification weights to adjust the relative influence of different panelists.[7] These propensity scores are based on an array of benchmarking attitudinal questions posed both in each online survey and in Harris' regular RDD "reference surveys." Berrens, et al. (2003), Schonlau, et al. (2004), and Duffy, et al. (2005) describe how HI pools the data on these attitudinal questions across an online survey and their most current reference survey, using an indicator for the source of the data as the dependent variable in a logistic regression. The fitted values for the systematic portion of this regression (the propensity scores, or the associated conditional probabilities) are sorted into quintile or decile bins. These bins constitute an additional dimension (along with a number of study-specific observable sample characteristics such as race, gender, age, and income that may be used separately, or as part of the same logistic regression) to construct weights for each online survey observation that render its influence comparable to the likely influence of the same category of individual in the general population.[8]

The U.S. Office of Management and Budget has recommended data quality standards for survey research when that research is intended to be used as the basis for policy decisions.[9] One specific dimension of these standards concerns the representativeness of survey samples, which in some cases has been translated as expectations with respect to survey response rates. Most researchers understand, however, that even a very low response rate can yield data which are representative of the intended population if response patterns are independent of the issues being

examined in the survey. Response rates alone are not an adequate indicator of whether the findings from a survey study reflect the attitudes or opinions in the general population.

Earlier generations of researchers typically resorted to a simple descriptive assessments of the representativeness of survey samples. These consisted of side-by-side comparisons of the marginal distributions of key variables (such as age, income, and gender) for both the estimating sample and the relevant population. It is often straightforward to draw a sample in a manner that will ensure that the sample more or less matches the intended population in terms of the marginal distributions, and even the joint distribution, of common observable sociodemographic variables.

However, as most survey researchers now appreciate, there is a more subtle challenge. A sample that mimics the population in terms of the marginal distributions of a few observable variables may still be non-representative if the sample and the population differ in terms of unmeasured or unobserved characteristics. Correction methods such as the weights based on propensity score quantiles (as used by HI) still rely entirely on observed characteristics.[10] The effect of unobserved characteristics is especially relevant when the subject matter of the survey is more salient to some contacted households and less salient to others. Not all households, even in a group which is identical on some set of *observable* characteristics, will be equally inclined to participate in the survey.

Furthermore, when using a standing consumer panel for survey research, it is not sufficient merely to compare those panelists who were invited to participate in a particular survey with those who actually chose to participate (this might be called “end-stage” sample selection). The standing panel itself may already be self-selected. One really needs to reach all the way back to the random-digit-dialed recruiting contacts to assess representativeness. Most studies using consumer panels seem to report only these end-stage response rates, which can be

impressively high. For the sample we use in this study, the end-stage response rate is about seventy percent (for usable responses). However, as a fraction of the initial sample of RDD contact attempts, the overall "comprehensive" response rate is only slightly over one-half of one percent.[11]

In this paper, we assess the potential for sample selection bias in the empirical results from one survey sample drawn from the consumer panel maintained by Knowledge Networks, Inc. (KN). This survey research firm has undertaken to assemble the most representative consumer panel currently available. Our research goal is to determine whether the representativeness sought through the company's investment in over half a million random-digit-dialed recruitment attempts appears to be adequately maintained—through the attrition, selection, and response processes. Are models based on just a single estimating sample from the current active panel likely to produce inferences that can be considered valid for the entire U.S. population?

We conjecture that the policy implications of research projects which use the KN panel may be most widely accepted if it can be shown that there is no significant "liberal" or "conservative" bias among samples drawn randomly from the KN panel. If there is any significant liberal or conservative bias, then the panel's value in policy-related research may be compromised unless this bias is identified and corrected. We are able to explore this issue of bias in the political preferences of KN panelists because respondents to our particular survey were posed a specific auxiliary question about the proper role of government in regulating environmental, health, and safety risks. We do not pretend that this single attitudinal question gives a comprehensive picture of political ideologies, but we use this variable as an illustration of the types of selectivity assessments which might be performed.

Our KN sample was drawn for use in a stated-preference study concerning willingness to pay for public health policies that reduce the risk of illness and death (Bosworth, et al. (2005). Here, we develop a Heckman-type selectivity correction model for this specific sample of survey respondents from the KN panel. We assume that KN's over half-million initially attempted RDD telephone contacts for panel recruitment are adequately representative of the general population of the U.S. We build some *observed* neighborhood-level data (at the level of census tracts and counties) and link it to each one of these over half-million contact attempts. These half-million observations are then used to model the discrete outcome that is membership vs. non-membership in the 2,911-member estimating sample for our particular survey. We explore the extent to which *unobserved* factors that (a) make an individual more likely than one would expect to be present in our estimating sample also (b) make that person systematically more or less likely than expected (based on their observable characteristics) to prefer a greater role for government in environmental, health, and safety regulation. Answers to our "proper role of government" question are likely to be correlated with the respondent's position along an implicit conservative/liberal or anti-/pro-regulation spectrum. In preview, we find no conclusive evidence of bias along this spectrum in our data.

In Section 2, we outline the data construction procedures used to match each initial panel recruiting contact to both its county and an appropriate census tract. Section 3 describes results for a conventional selectivity-corrected model concerning the proper role of government in regulation of environmental, health, and safety risks, illustrated using our "public preferences" sample. Section 4 outlines some additional sensitivity tests, and Section 5 offers caveats and conclusions.

2 Data Construction

Sample selection correction algorithms generally require that the researcher know something about each member of the intended sample which might help explain whether each individual appears in the final estimating sample. For this work, the "intended sample" is the set of RDD panel recruitment attempts, which should be representative of the overall U.S. population.[12] To model the selection process, we need a lot of explanatory variables which are available (and conformable) for the entire "intended sample," not just the smaller final estimating sample. Ideally, we would like to have individual-specific data on a wide variety of household characteristics, but this is impossible. With random digit dialing, the only thing one truly knows about every RDD residential contact attempt is the telephone number itself. Therefore, we use proxy data in the form of neighborhood characteristics at the census tract level by linking census tract data from the 2000 census to each household in the original KN panel recruitment sample frame.

The KN panel recruitment sample frame includes all working residential RDD phone numbers that KN first sampled and called (using the proprietary MSG Genesys-ID sampling system). While recruitment at KN is ongoing, the relevant recruiting phone numbers for our particular study sample were dialed between 1999 (when panel recruitment began) and May 1, 2003 (the date when the particular survey samples to be investigated were drawn for the Cameron and DeShazo (2005) health risk study). KN retained for analysis all valid residential phone numbers which included all cases with a final recruitment disposition code of "answering machine," "call back," "interview," "no answer," "refusal," and "refusal - privacy manager." The only exclusions from the original RDD sample were phone numbers found to be non-residential or non-working. These phone numbers are excluded because they are not explicitly

associated with residential households. This recruiting strategy leaves more than 525,000 unique phone numbers in the sample frame.

2.1 Linking RDD Contacts to the Census Tracts

Of these over half-million phone numbers, roughly 400,000 had corresponding street addresses on file in the KN database (call this Subset 1). Some of these addresses came from reverse-address matching of just the phone numbers themselves, and others stemmed from telephone-based recruitment, where a telephone voice contact resulted in the contacted party providing a street address. Of these cases, about 80% had valid street addresses that could be successfully matched by ESRI's ArcView 3.3 and the ESRI StreetMap 2000 utility. These addresses were geocoded to identify approximate point locations (side of street and how far along block) for each residence. The approximate point locations of these residences were then overlaid with ESRI's census tract polygons, a standard GIS "theme" that is accompanied by an attribute file containing corresponding census tract data from the 2000 Census.

Of the remainder of the RDD telephone numbers with street addresses that could not be specifically matched by the StreetMap utility, most had usable zip code data (call this Subset 2). These cases were matched, albeit less accurately, to an approximate census tract FIPS code using the census tract corresponding to the geographical centroid of the zip code polygon.[13]

Finally, KN did not have either address or zip code information for the roughly 125,000 remaining RDD phone numbers (call this Subset 3). For these cases, the telephone exchange for each telephone number (i.e., the six digits making up the number's area code plus prefix) was used as the device for identifying an approximate census tract FIPS code. All of the census tracts overlapped by each active telephone exchange area—at the date of the recruitment attempt—were identified. (Directory-listed households in each identified census tract were enumerated

separately.) The census tract with the largest number of directory-listed households was then designated as the "majority" census tract for that exchange. Each telephone number without address information was assigned to an approximate census tract FIPS code in this manner.

There are thus three sources of data for this study. Knowledge Networks first provided to us just their proprietary identity-protected street addresses (Subsets 1 and 2), with no other associated data, for geocoding. These addresses were associated with their census tract FIPS codes and returned to KN to have (a) the addresses removed, and (b) the sampling status and attrition history of each contact appended. Proxy case identifiers were generated and the files were returned to us for subsequent analysis. For initial RDD contacts without address information (Subset 3), KN facilitated the task of matching each RDD telephone exchange with the census tract that best approximates the bulk of the telephone numbers in that exchange, delivering proxy identifiers and census tracts FIPS codes, along with sampling status and attrition history for each of these cases.[14] Subsets 1, 2 and 3, with their corresponding status and attrition histories, were then combined into one huge file. Each record contains an 11-character census tract FIPS code and a set of five indicator variables that identify whether each initial contact survived through five attrition processes:

a.) initially recruited to the Panel
b.) initial profile data collected
c.) still a part of the active Panel when a sample was drawn for the particular study in question
d.) drawn for our particular study
e.) responded to the invitation to participate in a sufficiently complete fashion to be included in the final estimating sample.

The proportions of the original RDD contacts surviving at each milepost are given in the top panel of Table 1.

2.2 Associating Census Tract Factors and 2000 Presidential Voting Patterns

We use the census tract FIPS codes for each tract to merge our data with the census tract factors resulting from the factor analysis described in Cameron and Crawford (2003). These factors capture variations in both short- and long-form census variables across tracts. These data consist of a set of 15 mutually orthogonal factors that capture approximately 88 percent of the variation, in a set of 95 variables, across the roughly 65,000 census tracts in the 2000 Census. Using census tract identifier (11-character FIPS code), we then merge the fifteen factor scores with the original RDD residential contact attempts.[15] Descriptive labels assigned to these census tract factors appear in Table 1. While the use of local averages or aggregates in lieu of household-specific data is always a compromise, we argue that models based on at least some information about possible systematic differences across RDD contacts in the original contact group are preferable to the alternative of ignoring the endogenous selection process altogether.

There is a clear reason for preferring census tract factor scores to the alternative of using a vastly larger number of raw census variables. Many census variables are highly collinear, making it extremely difficult to tease out the distinct incremental effect of a difference in any one variable upon the outcome of interest (e.g., sample membership/non-membership). Estimated factors produced by factor analysis have the attractive property of being orthogonal by design. The factor scores span the same space as the much larger number of correlated variables upon which they are based, but they are uncorrelated, so their distinct effects can be identified more easily (if such effects are indeed present). It is our goal merely to control for systematic variation in attrition propensities, rather than to quantify the specific causes of attrition. Thus factor scores can be particularly valuable in selection correction models.

However, the downside of using estimated factor scores as explanatory variables is that they must typically be considered to be "estimated" regressors (i.e. "generated" or "constructed" variables, as in Pagan (1984) or Pagan and Nicholls (1984)). Ordinarily, we are very concerned about this, since estimated quantities come with varying levels of precision. If we fail to recognize the estimated nature of factor scores, we will be understating the amount of noise in the overall model and distorting any hypothesis testing in any second-stage model which uses them. In this case, however, there is some basis for arguing that the estimated regressors problem is minimized. We are not using factor scores estimated for *just* the sample of census tracts represented in the RDD sample provided by Knowledge Networks. The factor scores used in this study are instead calculated for the complete set of all census tracts in the U.S. As such, our tract-level factor scores are technically not just estimates of the corresponding "population" values, but are the calculated population values themselves (although only for the 2000 Census).

While our census tract factor scores may approximate the true "population values" of the tract-level factors, they are not the attributes of the specific individual who was contacted in the RDD sample. The census tract factors will be a better estimate of the individual's characteristics, the more homogeneous the population of the census tract. However, we are not able to control for the amount of noise introduced by using census tract characteristics as proxies for the individual characteristics that we would prefer to use if they were available.[16]

In many survey applications, especially if the research is intended to inform policy-making, we are concerned not only whether sociodemographic groups are proportionately represented, but also whether political constituencies are proportionately represented. To allow this question to be addressed in at least a rudimentary fashion, we have also merged in, by county FIPS code, all of the available information at the county level about percentages of voters

who voted for the Democratic candidate (Al Gore) and for the Green Party Candidate (Ralph Nader) in the 2000 Presidential election.[17]

The lower panel of Table 1 includes descriptive statistics, for the roughly 525,000+ initial RDD contacts, for the merged-in census tract factors and county voting proportions variables. Across the universe of census tracts in the entire U.S., the mean and variance of the census tract factors should be zero and one, respectively, since the factor scores are standardized by the algorithm that calculates them. Departures from these standardized means and variances, for our half-million cases, reflect the slightly disproportionate presence of RDD contacts in some physical census tracts and also the approximations necessary to match telephone exchanges with the right census tracts.

We posed all respondents an auxiliary question: "People have different ideas about what their government should be doing. How involved do you feel the government should be in regulating environmental, health and safety hazards?" Answer options ranged from 1=minimally involved, up to 7=heavily involved. Ideally, one would prefer a continuous variable for this exercise, but we will treat this discrete ordinal variable as though it were a continuous and cardinal measure, and call it *govt* .[18] Of the individuals receiving our public health program survey instruments, a total of 2,911 individuals provided an answer to this question (see Table 2).

3 Naïve OLS versus selectivity-corrected models

We first estimate a "naïve" ordinary least squares (OLS) model that ignores possible systematic selection problems. This model explains the level of the *govt* variable using observed panelist attributes (age, gender, etc.) to identify sources of systematic variation in this opinion across the estimating sample (of size 2911) assuming the data consist of a truly random sample from the U.S. population. The respondent characteristics available for use in explaining this

rating are itemized in Table 3. For extended specifications with a rich set of linear, interaction, and non-linear effects in these variables, only a subset of slope coefficients were persistently significant. For our working model, therefore, we adopt a more parsimonious specification that retains only those terms that are robustly significant.[19]

Model 1 in Table 4 presents results for this first parsimonious model. Respondent characteristics with a positive effect on the level of the *govt* dependent variable are factors associated with a more favorable attitude towards government regulation of environmental, health and safety risks. Working down the list of explanatory variables in Table 4, our working specification includes baseline dummy variables for being female (with a positive effect), non-white (insignificant), less than high school education (positive), high school grad (insignificant), some college education (insignificant). Simple dummy variables without interaction terms are included for non-employed (positive), divorced (positive), and single (positive), and we use continuous variables for age (baseline insignificant) and income (baseline positive).

While a number of the baseline effects are statistically insignificant, several interaction terms do matter. The effect of having an education level less than high school is shifted systematically by gender and non-white status. For the baseline group of white males, having education less than high school significantly increases sentiment for these types of regulations, whereas for both females and non-whites, this less-than-high-school effect is decreased to the point where it becomes negative. For males, the baseline effects of the high school educational attainment category and the some-college category are statistically insignificant. For females, however, these lower levels of educational attainment decrease sentiment for regulation. The baseline age effect for whites is indistinguishable from zero, but for non-whites there is a small but statistically significant increase in demand for such regulations with each year of age.

Lastly, demand for environmental, health, and safety regulations increases with income for the baseline gender category (males), but the effect of higher income appears to be near zero for females. Being divorced, or being single, produces a statistically significant increase in demand for such regulation, relative to the baseline married group.

A Heckman selectivity-corrected model involves the joint estimation of two equations: the "selection equation" which explains each initial RDD contact's presence in the final estimating sample and the "outcome equation" which models respondents' answers to the *govt* question. The outcome portion of our Heckman selectivity-corrected model is reported as Model 2 in Table 4, adjacent to the outcome equation estimated by naïve ordinary least squares. The selection portion of the jointly estimated model, along with the estimated error correlation parameter, is reported as Model 2' in Table 5. These estimates are displayed next to sing-equation (independent probit) estimates of the selection equation alone. The parameters of the joint model (with results displayed as Models 2 and 2') are estimated by maximum likelihood.[20]

In our selection equation, we use as explanatory variables the fifteen census tract factors and the percents of each county voting for Gore or for Nader in 2000 (Cameron and Crawford, 2003). There are 525,139 observations in the selection model, and 2,911 observations in the outcome model (for the sample from the panel that provided answers to the *govt* question on our survey). The results for Model 2' in Table 5 reveal that a member of the 525,139-person RDD contact pool is *more* likely to show up in the 2911-person estimating sample if their census tract includes more "well-to-do seniors," more "rural farming self-employed," more "Native Americans," more "health-care workers," or a greater percentage of their county voted for Nader in 2000. In contrast, a member of the RDD contact pool is *less* likely to appear in the estimating sample if their census tract includes more "well-to-do prime aged," more "single renter

twenties," more "minority single moms," more "thirty-somethings," more "some college, no graduation," or more "Asian-Hispanic language-isolated," or if a greater percentage of their county voted for Gore in 2000.

There are several ways to evaluate the effectiveness of this Heckman model. The statistics of greatest interest from a Heckman model include the sign and statistical significance of the estimated correlation between the errors in the selection equation and the errors in the outcome equation. For the jointly estimated specification reported as Model 2 in Table 4 and Model 2' in Table 5, this correlation is displayed in the last row of Table 5. It is positive but relatively small, at 0.086, with an asymptotic t-test statistic of only 1.54. This Wald-type hypothesis test suggests that the error correlation is positive but statistically insignificant at the conventional 5% level (and even at the less-stringent 10% level).

A second way of evaluating the model is to evaluate difference in log likelihood of the the selection equation portion of the Heckman specification compared to simple probit results for an independently (rather than jointly) estimated selection equation as we do in Table 5. Recall that if the outcome and selection equations are estimated separately, the error correlation is implicitly constrained to be zero. The Heckman model frees up this single constraint and produces an improvement in the overall log-likelihood from -17458.33 + -5549.30 = -23007.63 for the separately estimated components, with uncorrelated errors, to -23006.49 for the jointly estimated specification. The likelihood ratio test statistic for the restriction that the error correlation is zero is only 2.28, whereas even the 10% critical value for the corresponding chi-squared distribution is 2.71. Thus we cannot reject the hypothesis of "no error correlation" by this test either.

These tests suggest that there may be some evidence of selection effects, but this evidence is not sufficiently strong to warrant concern if we subscribe to conventional standards for statistical significance. Even though the key error correlation is not statistically different from zero, it is relevant to consider the extent to which relaxation of the constraint (that it be zero) affects the estimated parameters of the outcome equation. We wish to know if the parameters of the outcome equation are appreciably different in the presence of the Heckman correction, even if the generalization it embodies does not seem entirely warranted. The intercept is 4.89 in the naïve model, but only 4.51 in the Heckman model. This is a little less than an 8 percent reduction. However, the various slope estimates change only minimally between the naïve and corrected specifications for *govt* in Models 1 and 2 in Table 4.

If we wish to be particularly thorough, a related question concerns what the selection correction implies for the fitted marginal distribution of the *govt* variable. We counterfactually simulate what would have been the distribution of the *govt* variable in the absence of systematic selection.[21] Jumping ahead, the first two columns of Table 8 display these simulation results. Whereas the naïve model yields a mean *govt* value of 5.17 on our 1 to 7 scale from "minimally involved" to "heavily involved," the Heckman selectivity corrected model yields a mean *govt* value of 4.78, a decrease of about 7.5 percent. This suggests that respondents to this survey may be inclined to give ratings that are less than 8 percent higher than the general population on the question of the desirability of government involvement in environmental, health, and safety regulation. However, in terms of the estimated error correlation and the likelihood ratio test, this effect is not statistically significant, even at the 10% level. On the basis of these results, the hypothesis that there is "no pro- or anti-regulation bias" in the estimating sample cannot unambiguously be rejected.

Of particular interest may be the estimated effects of the voting percentages and the question of "liberal" versus "conservative" bias in our sample. Consider Table 5, Model 2', where we already control for the sociodemographic characteristics of each census tract. The county percentage voting for Gore has a statistically significant negative marginal effect on the propensity for an initial RDD contact to appear in the estimating sample. While the percent voting for Nader has a statistically significant and positive effect on this propensity, the actual Nader vote percentages are typically very small (less than 3% on average, compared to about 50% for Gore). In the RDD sample, the mean value of the response propensity index is about -2.6. The product of the average percentage and the estimated coefficient is roughly -0.087 for the Gore vote. There is an offsetting effect, averaging +0.043, due to the Nader vote.[22]

4 Sensitivity Analysis Using Fitted Participation Propensities/Probabilities

While incorporating a full-fledged Heckman selectivity correction is straightforward for ordinary least squares models, it is rather unwieldy to employ for other types of statistical models. In the case of non-OLS models, it may still be illuminative to model the sample selection process first, and then to use the fitted response "propensities" or the fitted response probabilities as control variables in a second-stage, sequentially estimated non-OLS model. (The fitted response *propensities* are the unbounded fitted values for the linear-in-parameters "index" for that binary outcome model, while the fitted response *probabilities* consist of a specific transformation of this index that converts it into a variable on a 0,1 scale.) This approach is a different use for "propensity scores" than employing them as post-stratification weights, as seems to be the strategy adopted by Harris Interactive. Instead, we use them to look for direct evidence that the estimated parameters of the outcome equation differ systematically with the propensity for a given observation to be in the estimating sample. Although this approach does

not appear to have been used elsewhere in the literature, it is a logical step in making an ad hoc assessment of heterogeneity in the parameters of the outcome equation that may be related to response/nonresponse patterns.

For both fitted response propensities and probabilities, if the estimated model parameters within the sample appear sensitive to the odds of each observation appearing in the sample, then there is a greater chance that the model's parameters also differ between sampled and non-sampled individuals. Conversely, if the estimated model parameters within the sample appear to be insensitive to the odds of each observation appearing in the sample, then there may be a greater chance that this insensitivity extends through to non-sampled individuals as well. We emphasize, however, that such results can only be treated as suggestive, rather than conclusive, since the fitted response propensities or probabilities are estimated regressors. A potential drawback of this strategy is that it ignores correlations in the unobservable factors that simultaneously affect both survey participation and attitudes toward government involvement.

For purposes of illustration, we use the data employed in our earlier analyses. For our estimating sample, the mean value of the fitted selection index (response propensity) is -2.582, with a standard deviation of 0.1681 and a range from -3.301 to -0.6220. The associated fitted probabilities have a mean value of 0.005542, a standard deviation of 0.006233, and a range from 0.0004692 to 0.2673. These fitted response propensities and probabilities are functions of observable variables (or in this case, of our census tract proxies for individual characteristics, and county-level voting patterns).

Next we propose a sensitivity assessment using these estimated propensities and probabilities. We use fitted response propensities (or probabilities) in an ad hoc fashion to investigate the possible effects of non-random sampling on the estimated parameters in the

outcome equation. We wish to know what would have been the vector of model parameters if each original RDD panel recruitment contact was equally likely (according to our selection equation) to show up in this particular estimating sample. Thus it is helpful to express all of the estimated propensities or probabilities as deviations from the average propensity or probability in the population. These normalized fitted propensities/probabilities can then be allowed to shift either just the intercept, or every outcome-model parameter. The baseline outcome-model parameter estimates then represent the "simulated" parameters for the counterfactual case where every respondent's chance of participating is equal to the average (meaning that all deviations-from-the-average are zero). This allows key parameter estimates to be systematically larger or smaller for observations with higher propensities to appear in the estimating sample, relative to their frequencies in the initial RDD contact pool.

Turning to Table 6, containing Models 3 and 4, the relevant fitted selection measure is allowed to shift just the intercept of the *govt* equation. Model 3 includes an intercept shift variable in the form of the "fitted index." Model 4 uses the "fitted probability" analogously. In neither case is this intercept differential statistically significant, although the point estimate of the differential is negative in each case. If anything, this evidence leans towards a suggestion that for the individuals in the estimating sample, the greater the predicted odds of a respondent being in our estimating sample, the lower the expected value of the *govt* response. However, positive, negative, or zero values for this effect cannot be statistically rejected by the data; the 95% confidence intervals for the relevant parameters are (-0.27, 0.04) in the case of the fitted index and (-1.86, 0.51) in the case of the fitted probability.

Of course, the fitted index or probability variable used to shift the intercept in the *govt* model in each of Models 3 and 4 is again an "estimated regressor." This means that the standard

errors in this second stage are inaccurate. These two-stage strategies are therefore less reliable than a full-information maximum likelihood Heckman selection correction model. In cases with egregiously severe sample selection bias, however, these cruder methods may sometimes clearly reveal an underlying problem. If the sample is truly random, there should be no dependence of the fitted E[*govt*] on the propensity of the individual to show up in this sample.

For good measure, we also consider specification where the fitted index (Model 5) or fitted probability (Model 6) is allowed to shift not only the intercept of the outcome equation, but also the full set of slopes. These results are presented in Table 7. In the Model 5 "index" variant, a larger value of the fitted selection index renders more positive the slope on the interaction term between an education level of high school and the female dummy variable (i.e., "High school * Female"). It may also render more negative the slope coefficient on the non-white dummy (but only at the 10% significance level).

For Model 6, the fitted probabilities are generally very tiny (on the order of a half of a percent), so the estimated coefficients on the slope differentials for the *govt* equation have correspondingly larger magnitudes than for Model 5. In this specification, the effect of the fitted selection probability again seems to be most pronounced in the case of the interaction term between high school and female (i.e. "High school * Female"). There is also some evidence (significant at only the 10% level) for an effect on the apparent influence of the "Education = Less than high school" dummy and on its interaction with gender, as well as on the baseline "Education = High school grad" dummy coefficient. Overall, the low-education status and gender effects coefficients in the *govt* model may be the most likely to be distorted by differing selection probabilities.

What are the implications of these additional selectivity-assessment models for the predicted values of *govt* in the estimating sample? For each of our models, we calculate and save the non-systematic part of *govt* (i.e., we save the estimated error terms). We then recalculate the systematic portion of *govt* under the assumption of no systematic selection--for Model 2, we employ the selectivity-corrected Heckman estimates; for Models 3 through 6, we use the baseline coefficients that apply when the fitted index (or fitted probability) is equal to the mean value in the RDD "population" (so that the shift variable equals zero in each case). To each of these adjusted fitted values, we add back in the fitted error term to produce a set of crudely "selectivity-corrected" values for *govt* . The distributions of these corrected fitted values, summarized in Table 8, tell the story. As noted above, the statistically insignificant Heckman correction does produce a less-than-8% decrease in the mean value of the *govt* variables. However, despite the presence of a few statistically significant slope-shift parameters in Models 5 and 6, the corrections based on Models 3 through 6 produce only minimal differences in the implied distribution of the *govt* variable.

5 Conclusions and Caveats

We have conducted a careful inquiry into the possibility of systematic selection in a sample drawn from the Knowledge Networks panel—between the original random-digit-dialed recruiting contact and a respondent's eventual participation in a particular research study sample. The most innovative feature of this sample-selectivity assessment/correction exercise is that we reach all the way back to the initial RDD recruiting contacts made to build the panel, rather than considering just the "end-stage" selectivity for the subset of panelists actually invited to participate in this particular survey. We consider many characteristics of these panelists (proxied

very loosely by the sociodemographic characteristics of the census tract where they live, or the voting patterns in their county).

It is worth reiterating that the use of census tract or county averages as proxies for individual values produces an obvious "errors-in-variables" problem in selection models where the researcher must rely on these averages in lieu of the specific characteristics of each individual. Errors in regressors are typically expected to produce "errors-in-variables attenuation." As a consequence, failure to find statistically significant effects in these types of models does not necessarily mean they would *not* materialize if analogous individual-specific regressors were available. However, success in finding statistically significant effects, even in the presence of errors-in-variables attenuation, should be considered a potentially noteworthy finding. Statistically significant slope estimates in our selection-assessment models using these data are significant *in spite of* the errors-in-variable attenuation that may make such relationships less easy to detect.

For the Knowledge Networks sample examined here, we find numerous statistically significant determinants of membership in the estimating sample, starting from the pool of over one-half million original RDD contacts. An empirical researcher is more likely to enjoy a sense of triumph upon finding evidence of some kind of mischief—i.e., extensive systematic selection or non-response in a survey sample (a "smoking gun") which has caused major damage to one's parameter estimates and the inferences from the survey. While there might be a smoking gun in this case, there appears to have been very little injury produced. We have examined an application of Heckman's selectivity correction where a small positive point estimate of the error correlation is statistically insignificant, even at the 10% level. There are modest differences in the parameter point estimates between the uncorrected and corrected models, but the importance

of the differences seems debatable (a difference in the means of only about 8%). We also consider a less-sophisticated strategy to determine whether there is any systematic variation in parameters according to the estimated participation propensity (or probability) that each potential panelist, as an initial RDD recruiting contact attempt, ends up in the estimating sample. We identify a limited number of systematic effects, but the effects of these tendencies on the outcome variable are minimal.

Finally, some audiences may be concerned that the widely used Knowledge Networks panel may have either a "liberal bias" or a "conservative bias," but the preliminary results described here do not really support such a conclusion. In particular, controlling for sociodemographics, there is a somewhat lower overall response probability for panelists from counties where a higher proportion of votes in the 2000 Presidential election went to Gore. However, this effect is offset to a considerable extent by a slightly higher overall response probability for panelists from counties where a higher proportion of votes went to Nader. Overall, our results can probably be characterized as reassuring news for researchers who have used (or who contemplate using) the Knowledge Networks panel for policy-oriented research. Our finding are also good news, presumably, for policy makers who need to rely on survey-based research to support their decisions.

BIOSCKETCHES

Trudy Ann Cameron (Ph.D. Princeton University, 1982) is the Raymond F. Mikesell Professor of Environmental and Resource Economics at the University of Oregon. She is currently President of the Association of Environmental and Resource Economics, and is a past Associate Editor for the *Journal of Environmental Economics and Management* and the *American Journal of Agricultural Economics.* She is also a member of the Science Advisory Board for the U.S. Environmental Protection Agency. (cameron@uoregon.edu; Department of Economics, 435 PLC; 1285 University of Oregon, Eugene, OR 97403-1285.

J.R. DeShazo is an Associate Professor in the School of Public Policy and Social Research at the University of California at Los Angeles. (B.A., College of William and Mary, M.Sc., Oxford University, Rhodes Scholar; Ph.D., Harvard University) He was a faculty associate at the Harvard Institute for International Development (1997-2000) and is currently Director of the Lewis Center for Regional Studies at UCLA. (deshazo@ucla.edu; Department of Public Policy, UCLA, Los Angeles, CA 90095-1656).

We acknowledge the generous cooperation of Knowledge Networks, Inc., especially Mike Dennis and Rick Li, in making it possible for us to address the crucial issue of potential non-response bias in our own Knowledge Networks research sample, used in work funded by the U.S. Environmental Protection Agency (R829485), and by Health Canada (Contract H5431-010041/001/SS). This work has not yet been formally reviewed by either agency. Tatiana Raterman provided helpful editorial comments. Any remaining errors are our own.

Table 1 – Descriptive statistics, leading to "public preferences" sample

Variable	Mean	Std. Dev.	Min	Max
Disposition (n=525,139)				
Recruited	0.3542			
Profiled	0.1833			
Active panel at sample time	0.07355			
Eligible at sample time (>24 years)	0.06788			
Drawn for sample	0.007891			
Estimating sample	0.005543			
Have census data? (1=yes)	0.9969			
Census Tract Factors [a] *(*n=523,506*)*				
"well-to-do prime aged"	0.2381	1.060	-3.035	6.695
"well-to-do seniors"	-0.01341	0.9776	-5.403	6.521
"single renter twenties"	0.03283	1.067	-2.875	5.419
"unemployed"	-0.1040	0.9899	-4.568	6.835
"minority single moms"	-0.002739	0.9109	-7.406	5.185
"thirty-somethings"	0.07765	0.8225	-13.30	4.01
"working-age disabled"	-0.08279	0.8406	-5.042	11.73
"some college, no graduation"	0.1101	0.9577	-5.869	8.428
"elderly disabled"	0.02486	0.9900	-4.067	14.05
"rural farm self-employed"	-0.1947	0.6110	-2.966	11.00
"low mobility stable neighborhd"	-0.04835	0.9457	-8.485	5.049
"Native American"	-0.1064	0.7882	-7.075	11.85
"female"	0.04068	0.8163	-14.90	5.847
"health-care workers"	-0.005511	0.8431	-4.979	9.475
"Asian-Hisp language-isolated"	0.06347	0.9582	-3.433	9.173
Have county voting data? (1=yes)	0.9983			
Voting Percentages (n=524,238)				
gore_pct	0.5042	0.1290	0.0847	0.868
nader_pct	0.02726	0.01812	0	0.172

[a] These factors are developed in Cameron and Crawford (2003)

Table 2 – "Public preferences" estimating sample: distribution of ratings on the question:

"How involved should government be in regulating environment, health, safety?"

Rating	Frequency	Percent	Cum. Percent
1 – minimally involved	134	4.60	4.60
2	90	3.09	7.69
3	199	6.84	14.53
4	493	16.94	31.47
5	609	20.92	52.39
6	591	20.30	72.69
7 – heavily involved	795	27.31	100.00
Total	2,911	100.00	

Table 3 – Attributes of "public preferences" survey respondents (n = 2,911)

Variable	Description	Mean	Std.Dev.	Min	Max
Female	=1 if female	.5211			
Non-white	=1 if non-white	.2188			
Not employed	=1 if not employed	.3858			
Less than high school	=1 if education < high school	.1268			
High school	=1 if education = high school	.3274			
Some college	=1 if education = some college	.2728			
Age of respondent	age in years	49.76	15.15	24	93
Income	income in $10,000	4.914	3.463	.25	20
Divorced	=1 if divorced	.1244			
Single	=1 if single	.1546			

Table 4 –Naïve OLS versus Heckman FIML selection-correction model
(n=2911, asymptotic t-test statistics in parentheses)

Dependent variable: *govt* Panelist attributes	**Model 1** OLS	**Model 2** Heckman[a] "outcome" equation
Female	0.5074 (2.92)***	0.5079 (2.93)***
Non-white	-0.2250 (0.87)	-0.2406 (0.93)
Not employed	0.2358 (3.16)***	0.2315 (3.12)***
Education = Less than high school	0.3860 (2.24)**	0.3931 (2.29)**
Less than high school * Female	-0.5311 (2.43)**	-0.5268 (2.41)**
Less than high school * Non-white	-0.9404 (4.76)***	-0.9401 (4.78)***
Education = High school grad	-0.0027 (0.02)	0.0069 (0.06)
High school grad* Female	-0.4171 (2.52)**	-0.4167 (2.52)**
Education = Some college	0.0124 (0.10)	0.0147 (0.12)
Some college*Female	-0.3519 (2.08)**	-0.3477 (2.06)**
Age of respondent	-0.0033 (1.24)	-0.0033 (1.27)
Age of respondent*Non-white	0.0142 (2.60)***	0.0143 (2.62)***
Income (measured in $10,000)	0.0346 (2.56)**	0.0334 (2.46)**
Income*Female	-0.0346 (1.83)*	-0.0351 (1.86)*
Marital status = Divorced	0.2421 (2.55)**	0.2377 (2.51)**
Marital status = Single	0.2096 (2.27)**	0.2024 (2.20)**
Constant	4.8861 (26.66)***	4.5131 (14.71)***
Log L	-5549.30	-23006.49

[a] Selection-equation portion of estimates for Heckman model are contained in Table 5, in comparison with simple probit model.
* Significant at the 10% level; ** Significant at the 5% level; ***Significant at the 1% level

Table 5 –Selection equation for Heckman model and analogous simple probit. (n=525,139)

Explanatory variables:	Independent single-eqn probit model	**Model 2'** Heckman "selection" eqn
Census tract factors avail. [a]	-0.7426 (12.50)***	-0.7427 (12.50)***
"well-to-do prime aged"	-0.1185 (15.00)***	-0.1185 (15.00)***
"well-to-do seniors"	0.0327 (4.66)***	0.0324 (4.62)***
"single renter twenties"	-0.0323 (4.32)***	-0.0324 (4.34)***
"unemployed"	0.0063 (0.86)	0.0062 (0.85)
"minority single moms"	-0.0191 (2.33)**	-0.0194 (2.37)**
"thirty-somethings"	-0.0185 (2.18)**	-0.0181 (2.13)**
"working-age disabled"	-0.0009 (0.12)	-0.0013 (0.17)
"some college, no graduation"	-0.0190 (2.53)**	-0.0183 (2.44)**
"elderly disabled"	0.0056 (0.78)	0.0055 (0.76)
"rural farm self-employed"	0.0421 (4.05)***	0.0425 (4.09)***
"low mobility stable neighborhood"	-0.0097 (1.22)	-0.0094 (1.18)
"Native American"	0.0407 (4.60)***	0.0407 (4.59)***
"female"	0.0112 (1.29)	0.0112 (1.29)
"health-care workers"	0.0191 (2.32)**	0.0192 (2.33)**

"Asian-Hisp language isolated"	-0.0664 (7.74)***	-0.0666 (7.75)***
2000 vote percentage avail.	-1.1297 (15.38)***	-1.1264 (15.32)***
Gore percent (county)	-0.1729 (2.39)**	-0.1769 (2.45)**
Nader percent (county)	1.6050 (3.90)***	1.6150 (3.93)***
Constant	-0.6208 (9.00)***	-0.6224 (9.02)***
ρ (implied Heckman error correlation)	-	0.08462[b] (1.52)
Log-Likelihood	-17458.33	-23006.49

[a] These factor analysis scores are developed and explained in Cameron and Crawford (2003)
[b] Point estimate for correlation parameter is de-transformed, but t-test statistic is not.
** Significant at the 5% level; *** Significant at the 1% level.

Table 6 – Models using just an intercept shifter
(n = 2911, asymptotic t-test statistics in parentheses)

Dependent var: *govt* Panelist attributes	**Model 3** Fitted selection "index"	**Model 4** Fitted selection "probability"
Female	0.5078 (2.92)***	0.5068 (2.91)***
Non-white	-0.2395 (0.92)	-0.2284 (0.88)
Not employed	0.2317 (3.11)***	0.2332 (3.13)***
Education = Less than high school	0.3923 (2.28)**	0.3854 (2.24)**
Less than high school * Female	-0.5267 (2.41)**	-0.5264 (2.40)**
Less than high school * Non-white	-0.9399 (4.76)***	-0.9381 (4.75)***
Education = High school grad	0.0062 (0.05)	-0.0005 (0.00)
High school grad* Female	-0.4166 (2.51)**	-0.4168 (2.51)**
Education = Some college	0.0144 (0.12)	0.0117 (0.10)
Some college*Female	-0.3477 (2.06)**	-0.3485 (2.06)**
Age of respondent	-0.0033 (1.26)	-0.0032 (1.22)
Age of respondent*Non-white	0.0143 (2.61)***	0.0142 (2.60)***
Income (measured in $10,000)	0.0335 (2.46)**	0.0344 (2.54)**
Income*Female	-0.0351 (1.85)*	-0.0349 (1.84)*
Marital status = Divorced	0.2380 (2.51)**	0.2407 (2.54)**
Marital status = Single	0.2031 (2.20)**	0.2086 (2.26)**
Constant	4.9107 (26.69)***	4.8940 (26.68)***
Fitted index (or probability) Intercept shifter	-0.1146 (1.47)	-0.6733 (1.12)
Log L	-5548.21	-5548.67

* Significant at the 10% level; ** Significant at the 5% level; ***Significant at the 1% level

Table 7 – Models using both intercept and slope shifters
(n = 2911, asymptotic t-test statistics in parentheses)

Dependent var: *govt*	**Model 5**		**Model 6**	
Panelist attributes:	Linear term in variable	* fitted selection index	Linear term in variable	* fitted selection prob.
Female	0.5336 (2.97)***	-0.3979 (0.86)	0.5360 (3.03)***	-4.5286 (1.18)
Non-white	-0.1586 (0.60)	-1.5744 (1.76)*	-0.1457 (0.55)	-11.8682 (1.48)
Not employed	0.2565 (3.26)***	-0.2278 (1.13)	0.2471 (3.25)***	-1.7786 (1.09)
Education = Less than high school	0.4566 (2.48)**	-0.5223 (1.05)	0.4540 (2.59)***	-7.0341 (1.79)*
Less than high school * Female	-0.5972 (2.58)***	0.8584 (1.42)	-0.5841 (2.61)***	8.9739 (1.80)*
Less than high school * Non-white	-0.9144 (4.40)***	-0.2127 (0.40)	-0.9369 (4.66)***	-0.0552 (0.01)
Education = High school grad	0.0441 (0.35)	-0.3600 (1.21)	0.0394 (0.32)	-4.3934 (1.85)*
High school grad* Female	-0.4999 (2.87)***	0.9393 (2.10)**	-0.4906 (2.90)***	9.1203 (2.49)**
Education = Some college	0.0007 (0.01)	-0.0283 (0.09)	0.0158 (0.13)	-1.5325 (0.61)
Some college*Female	-0.3449 (1.97)**	0.4970 (1.07)	-0.3562 (2.07)**	5.1662 (1.33)
Age of respondent	-0.0044 (1.59)	0.0089 (1.26)	-0.0040 (1.49)	0.0905 (1.61)
Age of respondent*Non-white	0.0129 (2.33)**	0.0295 (1.55)	0.0130 (2.35)**	0.2086 (1.21)
Income (measured in $10,000)	0.0350 (2.52)**	-0.0191 (0.56)	0.0357 (2.58)***	-0.2511 (0.90)
Income*Female	-0.0345 (1.78)*	-0.0649 (1.21)	-0.0323 (1.68)*	-0.6142 (1.30)
Marital status = Divorced	0.2623 (2.63)***	-0.2444 (0.97)	0.2456 (2.54)**	-1.6382 (0.82)
Marital status = Single	0.1685 (1.77)*	0.2950 (1.17)	0.1837 (1.96)*	2.8594 (1.36)
Constant	4.9291 (25.67)***	-0.1276 (0.27)	4.8920 (26.17)***	-0.0776 (0.02)
Log L	-5535.89		-5533.65	

* Significant at the 10% level; ** Significant at the 5% level; ***Significant at the 1% level

Table 8 – Distributions of actual and "corrected" dependent (*govt*) variable (n=2911)

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
	OLS	Heckman selectivity-corrected model (MLE)	OLS-fitted index shifting intercept	OLS-fitted prob. shifting intercept	OLS-fitted index shifting all parameters	OLS-fitted prob. shifting all parameters
Mean	5.166 [a]	4.783	5.181	5.174	5.177	5.173
5%	2.000	1.608	1.991	1.999	1.976	1.997
25%	4.000	3.622	4.015	4.001	4.011	4.001
50%	5.000	4.627	5.026	5.003	5.074	5.010
75%	7.000	6.601	6.981	6.998	6.909	6.987
95%	7.000	6.625	7.022	7.002	7.053	7.007

[a] Actual distribution of dependent variable for Model 1; simulated distribution after correction for Models 2-6.

References

Bandilla, W., Bosnjak, M., Altdorfer, P. (2003). "Survey administration effects? A comparison of web-based and traditional written self-administered surveys using the ISSP environment module", *Social Science Computer Review* 21, 235-243.

Berrens, R.P., Bohara, A.K., Jenkins-Smith, H., Silva, C., Weimer, D.L. (2003). "The advent of internet surveys for political research: A comparison of telephone and internet samples", *Political Analysis* 11, 1-22.

Best, S.J., Krueger, B. (2002). "New approaches to assessing opinion: The prospects for electronic mail surveys", *International Journal of Public Opinion Research* 14, 73-92.

Best, S.J., Krueger, B., Hubbard, C., Smith, A. (2001). "An assessment of the generalizability of internet surveys", *Social Science Computer Review* 19, 131-145.

Birnbaum, M.H. (2004). "Human research and data collection via the internet", *Annual Review of Psychology* 55, 803-832.

Bosworth, R., Cameron, T.A., DeShazo, J.R. (2005). Advances in evaluating the demand for risk prevention policies. Department of Economics, University of Oregon. Eugene, OR.

Cameron, T.A., Crawford, G.D. (2003). Note: Independent dimensions of sociodemographic variability in neighborhood characteristics at the tract level of the 2000 census. University of Oregon; http://economics.uoregon.edu/papers/UO-2004-10_Cameron_Crawford_Census_Factors.pdf. Eugene, OR.

Cameron, T.A., DeShazo, J.R. (2005). Valuing health risk reductions: Sick-years, lost life-years, and latency. University of Oregon. Eugene, OR.

Cuddeback, G., Wilson, E., Orme, J.G., Combs-Orme, T. (2004). "Detecting and statistically correcting sample selection bias", *Journal of Social Service Research* 30, 19-33.

Das, M., Newey, W.K., Vella, F. (2003). "Nonparametric estimation of sample selection models", *Review of Economic Studies* 70, 33-58.

Deutskens, E., de Jong, A., de Ruyter, K., Wetzels, M. (2006). "Comparing the generalizability of online and mail surveys in cross-national service quality research", *Marketing Letters* 17, 119-136.

Duffy, B., Smith, K., Terhanian, G., Bremer, J. (2005). "Comparing data from online and face-to-face surveys", *International Journal of Market Research* 47, 615-639.

Evans, J.R., Mathur, A. (2005). "The value of online surveys", *Internet Research-Electronic Networking Applications and Policy* 15, 195-219.

Eysenbach, G. (2004). "Improving the quality of web surveys: The checklist for reporting results of internet e-surveys (cherries)", *Journal of Medical Internet Research* 6, 12-16.

Fitzgerald, J., Gottschalk, P., Moffitt, R. (1998a). "An analysis of sample attrition in panel data - the Michigan Panel Study of Income Dynamics", *Journal of Human Resources* 33, 251-299.

Fitzgerald, J., Gottschalk, P., Moffitt, R. (1998b). "An analysis of the impact of sample attrition on the second generation of respondents in the Michigan Panel Study of Income Dynamics", *Journal of Human Resources* 33, 300-344.

Goritz, A.S. (2004). "Recruitment for online access panels", *International Journal of Market Research* 46, 411-425.

Hausman, J.A., Wise, D.A. (1979). "Attrition bias in experimental and panel data - gary income-maintenance experiment", *Econometrica* 47, 455-473.

Heckman, J.J. (1979). "Sample selection bias as a specification error", *Econometrica* 47, 153-161.

Ilieva, J., Baron, S., Healey, N.M. (2002). "Online surveys in marketing research: Pros and cons", *International Journal of Market Research* 44, 361-376.

Lee, S. (2006). "An evaluation of nonresponse and coverage errors in a prerecruited probability web panel survey", *Social Science Computer Review* 24, 460-475.

Leip, D. (2003). Atlas of u.S. Presidential elections. http://www.uselectionatlas.org/

Lillard, L.A., Panis, C.W.A. (1998). "Panel attrition from the Panel Study of Income Dynamics - household income, marital status, and mortality", *Journal of Human Resources* 33, 437-457.

Nevo, A. (2003). "Using weights to adjust for sample selection when auxiliary information is available", *Journal of Business & Economic Statistics* 21, 43-52.

Nicoletti, C., Peracchi, F. (2005). "Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel", *Journal of the Royal Statistical Society Series a-Statistics in Society* 168, 763-781.

Pagan, A. (1984). "Econometric issues in the analysis of regressions with generated regressors", *International Economic Review* 25, 221-247.

Pagan, A.R., Nicholls, D.F. (1984). "Estimating predictions, prediction errors and their standard deviations using constructed variables", *Journal of Econometrics* 24, 293-310.

Ridder, G. (1990). "Attrition in multi-wave panel data". In J. Hartog, G. Ridder, J. Theeuwes (eds.), *Panel data and labor market studies*. Amsterdam: North-Holland.

Schillewaert, N., Meulemeester, P. (2005). "Comparing response distributions of off line and online data collection methods", *International Journal of Market Research* 47, 163-178.

Schonlau, M. (2004). "Will web surveys ever become part of mainstream research?" *Journal of Medical Internet Research* 6, 10-11.

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K.H., Marcus, S.M., Adams, J., Spranca, M., Kan, H.J., Turner, R., Berry, S.H. (2004). "A comparison between responses from a propensity-weighted web survey and an identical RDD survey", *Social Science Computer Review* 22, 128-138.

Smith, T.W. (2003). "An experimental comparison of Knowledge Networks and the GSS", *International Journal of Public Opinion Research* 15, 167-179.

Tourangeau, R. (2004). "Survey research and societal change", *Annual Review of Psychology* 55, 775-801.

Vella, F. (1998). "Estimating models with sample selection bias: A survey", *Journal of Human Resources* 33, 127-169.

Winship, C., Mare, R.D. (1992). "Models for sample selection bias", *Annual Review of Sociology* 18, 327-350.

Wright, K.B. (2005). "Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services", *Journal of Computer-Mediated Communication* 10,

ENDNOTES

[1] See discussion of the state-of-the-science in survey research in Tourangeau (2004). A number of relevant concerns are also outlined in Birnbaum (2004).

[2] Ilieva, et al. (2002), Schonlau (2004), Schillewaert and Meulemeester (2005) address the (relative) sampling properties of web-based or email surveys. Some social science disciplines, such as economics, have struggled with sample selection bias detection and correction for decades (e.g. going back to early work by Heckman (1979), with the broader scope of the early work surveyed by Vella (1998)). Winship and Mare (1992) summarized the issue for sociologists. In other social science disciplines, these issues have been addressed routinely only in more recent years (e.g., Cuddeback, et al. (2004) describe the state of practice in social work research).

[3] See http://www.knowledgenetworks.com and http://www.harrisinteractive.com.

[4] Some of the possibilities have been explored systematically by Goritz (2004), for example.

[5] In reaction to concerns about validity of the inferences from online surveys, the *Journal of Medical Internet Research* has proposed a checklist of recommendations for authors in an effort to ensure complete descriptions of Web surveys (Eysenbach (2004).

[6] Smith (2003) compares answers from the General Social Survey (GSS) with answers to the same questions by a sample of KN panel members. They find many similarities, but a few differences.

[7] Simple post-stratification weights, based upon the relative frequencies of types of respondents in the sample versus the population (say, according to a recent census data) have been discussed in many studies. The viability of this strategy has been assessed for email-based surveys by Best and Krueger (2002), and for web-based surveys by Bandilla, et al. (2003).

[8] A similar technique, based on method of moments estimation, has been demonstrated by Nevo (2003).

[9] The Data Quality Act amends the Paperwork Reduction Act (44 U.S.C. 3501 et seq.) The DQA was enacted in December 2000 as a two-paragraph provision within an appropriations bill (see the Treasury and General Government Appropriation Act for Fiscal Year 2001, Pub. L. No. 106-554, § 515 Appendix C, 114 Stat. 2763A-153 (2000)). The DQA went into effect on October 1, 2002, which was a deadline for federal agencies to issue their final information quality guidelines. It is intended to apply to "influential scientific, financial, or statistical information," consisting of any data that will have an impact on significant public policies or major private sector decisions.

[10] Schonlau, et al. (2004) acknowledge this limitation: “Propensity scoring balances observed covariates. Propensity scoring balances unobserved covariates only to the extent that they are correlated with observed covariates. The assumption that unobserved variables can be ignored with respect to selection bias is called ignorability.” These authors also concede that their weighting scheme adjusts their California sample to match the national distribution for the attitudes in the reference survey, but that “the additional assumption that the California population answers attitudinal questions just like the U.S. population…is not verifiable.”

[11] An example of research which explores more than just end-stage selection is Lee (2006), who begins with the set of panelists who made it to the intermediate stage of being profiled for the panel (so that relatively complete sociodemographic information is known for each). She does not model selectivity formally, however, and there is no discussion of selection on unobservables. She emphasizes the differences in distributions of specific characteristics between the “profiled” and “final” samples. In our data, only 18% of the original RDD contact attempts produced profiled panelists, so Lee’s analysis misses whatever systematic selection may have affected the other 82% of the ostensibly random RDD contact attempts.

[12] We must assume that households without telephone numbers are a sufficiently tiny fraction of the population that they can be ignored for most purposes. For studies targeting certain specialized groups, of course, this underlying selection problem could not be ignored (e.g., studies concerning the homeless).

[13] These links were accomplished using utilities provided within ArcView.

[14] Dale Kulp of Marketing Systems Group (MSG) generously provided the exchange/census-tract matching for this subsample. We note also that a very limited set of initial RDD contacts were lost from KN's archival records. However, we are confident that this block of lost data occurred essentially randomly. We have no recourse but to assume this loss was independent of any of the other general processes modeled here and to proceed without those data.

[15] For only a tiny minority of census tracts (i.e., less than 0.4%), it was not possible to construct a set of census tract factors. Thus we include an indicator variable, *census factors available*, that takes on a value of 1 if the census tract factors are available, and is zero otherwise. See Table 1.

[16] Sworn Census employees can gain access to much more of the individual household data underlying the census tract totals. These data would allow the researcher to estimate the variance-covariance matrix for census variables within each census tract and would allow more rigorous corrections for this type of measurement error. This strategy, however, is still prohibitively difficult with current technologies and we do not have sworn Census employee status.

[17] Proportions in the "omitted category" voting for candidates other than the Republican candidate (George W. Bush) are assumed to be sufficiently small that little generality is lost by neglecting them. Presidential voting data are available in spreadsheet format from Leip (2003). However, the breakdown in votes for Alaska counties is not available. We thus include an indicator for election data availability, *vote percentage available*.

[18] Note that we have also explored ordered probit models for the "outcome" model in a full-information maximum likelihood selectivity-correction model. The same signs and levels of significance result for each of the coefficients in the outcome equation (although the parameter point estimates themselves are not comparable because of normalization in the ordered probit specification). Since the OLS estimates for the outcome equation are qualitatively the same, and will be more familiar to most readers, we present this simpler model in the paper.

[19] We do, however, retain statistically lower-order terms when a higher-order or interaction term is statistically significant.

[20] Sample selection models have been researched extensively. The seminal paper is Heckman (1979), and subsequent surveys of basic and alternative models have been provided by Vella (1998) and Das, et al. (2003). The role of sociodemographic characteristics on response propensities has been considered by Hausman and Wise (1979), Ridder (1990), Lillard and Panis (1998), Fitzgerald, et al. (1998a), Fitzgerald, et al. (1998b), and Nicoletti and Peracchi (2005), among others.

[21] In simulating corrected distributions for the *govt* variable, we use the corrected coefficients to form a linear combination with the explanatory variables to produce a predicted value for *govt* in the absence of selection. We then add back in the error term to produce a corrected distribution that mimics the distribution for the raw *govt* variable, but without the bias attributed to selection.

[22] Note that an ordered probit model with sample selection correction yields an almost identical estimate for the error correlation parameter (0.082, with an asymptotic t-test statistic of 1.49) We used NLOGIT 3.0 to estimate the ordered probit model with selection, rather than Stata SE 8.0, which we used for all of the other estimates presented here. Given the size of our data set (525,139 observations), NLOGIT's much longer loading time was a considerable disadvantage.