Limiting the Risk of Nonresponse Bias by using Regression Diagnostics as a Guide to Data Collection

James Wagner[*]

Abstract

Survey data collections have traditionally been judged by the response rate. In terms of improving the response rate, every case has the same value. An optimal approach to maximizing the response rate would be to always attempt to interview the "easiest" remaining case. However, this strategy might lead to interviewing very similar cases. In this way, as a guiding indicator, the response rate might distort the survey data. The ultimate goal is to control nonresponse bias, but this bias is never observed in practice. Researchers are investigating proxy indicators that may be used to control the risk of nonresponse bias. Unfortunately, the field of survey methodology knows very little about how other indicators might perform as guides to data collection. One possible set of indicators are drawn from regression diagnostics. These diagnostics can be used to identify influential data points. Exploring the covariate space near these influential points may reduce our uncertainty about these regions. We present preliminary results from experiments that target cases that are similar to influential cases for interview.

Key Words: Nonresponse bias, quality indicators, adaptive design

1. Introduction

Nonresponse is a growing problem faced by social surveys (de Leeuw and de Heer, 2002; Atrostic et al., 2001; Petroni et al., 2004; Curtin et al., 2005; Brick and Williams, 2013). This growth in nonresponse has led to increasing concerns about the possibility of nonresponse bias. Unfortunately, recent research has found that the response rate is not a good predictor for when nonresponse bias is likely to occur (Groves and Peytcheva, 2008). Therefore, maximizing the response rate may not lead to reduced bias. This leaves data collection agencies with a void. If maximizing the response rate is no longer the goal, then what should be the goal?

There have been several options proposed. One option is to attempt to balance the set of responding units on a set of covariates available for the whole sample or population. This approach was adopted by Wagner and colleagues (2012) for the National Survey of Family Growth (NSFG) Continuous 2006-2010. They attempted to reduce the variation of subgroup response rates and found several interventions that were effective in doing so. Others have attempted to reduce the variation in response probabilities in a multivariate model. The R-Indicator is a measure of this variability (Schouten et al., 2009). Several studies have attempted to minimize variation in response probabilities using the R-Indicator as a guide (Schouten et al., 2012; Luiten and Schouten, 2013; Finamore et al., 2013). Others have attempted to limit this variation by focusing on the cases with the lowest estimated response propensities (Rosen et al., 2014).

A key question for this approach is whether this kind of balancing during data collection is able to reduce nonresponse bias of adjusted estimates when the adjustments are based on the same variables used to measure the balance of response. An argument against attempting to balance response during data collection is that the same results can be obtained with

[*] Survey Research Center, University of Michigan, 426 Thompson St, Ann Arbor, MI 48014

a nonresponse weighting adjustment. The assumption of this argument is that responders and nonresponders are alike, irrespective of the achieved response rate. For example, if men have a response rate of 10% and are weighted up to 50% of the response, then the estimates will be the same as if we had balanced response such that 50% of the responders were men.

The argument for balancing on the covariates is that the MAR assumption becomes more plausible with improved balance. Of course, there may be confounding variables. That is also a problem with the assumption that everything can be fixed with weighting. The hope is that the effect of any confounding variables can be weakened when there is good balance on other covariates.

In any event, it is an empirical question whether either of these assumptions is true. Schouten and colleagues (Under Review) completed a simulation study – based upon real survey data from 16 different surveys – that showed that increasing the R-Indicator (i.e. increasing the balance of the responding set) led to decreases in the bias of adjusted estimates for a majority of the situations tested. These simulations used frame data as "Proxy Y" variables. They are nevertheless an empirical evaluation of this question.

An alternative to balancing response with respect to the $\mathbf{X}$ covariates is to use predicted Y values to guide the data collection. Presumably, such an approach would target cases with predicted Y values that would reduce the bias of the respondent mean. If we assume a model predicting Y based upon the $\mathbf{X}$ covariates, then this may be very similar to simply balancing upon the $\mathbf{X}$ covariates. The predicted Ys are a simple, scalar summary of $\mathbf{X}$ available for that case. In this way, balancing the Ys is likely to balance $\mathbf{X}$, apart from problems like interactions in the data that are not included in the model, confounding, and other similar mechanisms.

This paper proposes an alternative approach to this problem and evaluates the results of the proposed method experimentally. The goal is to identify cases for which it is more difficult to predict the Y values. These cases would be difficult to impute. There would be more uncertainty about which values to impute and, as a result, the range of imputed values would be greater. These cases would also be difficult to adjust for using a weighting adjustment since it would be difficult to match them to cases with similar characteristics. The cases that are more difficult to predict would be the cases that are more difficult to adjust for if they are missing – either through imputation or weighting.

2. Proposed Procedure

The new approach that is proposed in this paper is to guide data collection using regression diagnostics. Many standard textbooks in regression, when discussing influential cases, recommend checking the data to see if there is an explanation for these data – data entry error, other measurement problems, or known confounding factors. It seems logical to extend this thinking to survey data collections.

The following figures show how a point may be influential in the regression setting. In each figure, only 1 point is moved and yet the estimated regression lines are vastly different. Since the other points form a sphere without an apparent relationship to X, the single point has as much information with respect to the line as all the other points.

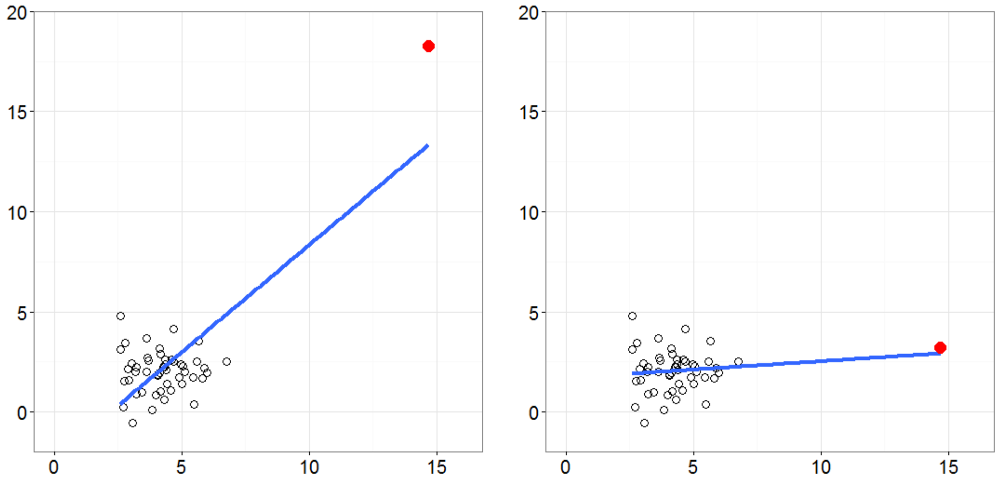

Figure 1: Example of Influence.

In the survey data collection setting, we might think of these data as the respondent data. The x-axis is the observed covariate and the y-axis is the reported survey variable. During the field period, we notice that the red point is influential. Assuming that is has been recorded correctly, we may want to explore the region of the x-axis where this point is observed (around $X = 15$). In this way, we can learn if this point is an outlier on X and Y (as in the figure on the left) or if it is just an outlier X (as in the figure on the right). This information will be useful as we predict the Ys for the missing Xs. Our uncertainty about imputing values in this region ($X = 15$) will be reduced with additional (X, Y) pairs observed in that region.

There are several kinds of influence diagnostics for regression models (Belsley et al., 1980). One important distinction is between the influence diagnostics available for linear models versus those available for generalized linear models (e.g. logistic regression). For both models, the leverages are the diagonal elements of the hat matrix. These are a measure of the distance from the center of the x-space. It is also possible to estimate the effect of each case on estimated regression coefficients using a statistics known as DFBETAS. For linear regression models, it is also possible to calculate Cooks Distance and DFFITS. The former is a measure of the influence each case is likely to have on the estimated coefficients, while the latter measures change in the fitted values when each case is dropped. Given the categorical nature of the outcome variable in logistic regression, the difference in the deviance (DIFDEV) or chi-square (DIFCHISQ) when each case is deleted are more appropriate measures of fit than the DFFITS.

We chose to use a measure of fit. This seemed more natural for our purpose since we are more concerned with the Y value itself than the regression coefficients which may predict Y. The regression coefficients are – for most purposes – a nuisance parameter in this setting.

For the responding units at any point in time, it is possible to fit a model using data that are available for all sampled units (i.e. sampling frame, paradata, and screening data) to predict the Y variable. It is possible to store the measures of influence from these models. Of course, these are only available for the respondents.

Once we have these influence measures stored, we can see if they are associated with any of the characteristics available for all sampled units. These measures are continuous. Therefore, they can be predicted with a linear regression model. For each nonresponding case, we can predict what its influence would be. It is also possible to dichotomize the observed influence measures for the responding cases, and then use the sampling frame and paradata to predict the probability of being influential using logistic regression. The latter procedure uses somewhat less information.

The predicted influences or predicted probabilities of being influential can then be used to target cases. Cases with high predicted influence would be made "high priority" in the sample management system. This sends a signal to the field interviewers that these cases are important and should be prioritized. Wagner et al. (2012) show that this often works to improve response rates for targeted cases. The result of this prioritization of the influential cases is that they are interviewed at higher rates.

3. Data and Methods

The data for this study come from experiments run on the National Survey of Family Growth (NSFG) Continuous 2011-2019. The NSFG is an ongoing cross-sectional, multistage area probability sample of householdsIn each quarter, a new sample of housing units is selected. At each sample household, interviewers administer a screener questionnaire to determine whether anyone 15-44 years of age who lives in the household. If so, interviewers make a roster of age-eligible members – obtaining sex, age, race, and ethnicity for each age-eligible member of the household – and select one at random for the main interview. The interviewer requests that the selected person complete a 60 minute (men) or 80 minute (women) main interview. The survey asks about sexual and fertility experiences of the respondent, with more sensitive items administered using Audio Computer-Assisted Self Interviewing (ACASI).

The auxiliary data used in this study include data from the sampling frame, paradata (Couper, 1998), and data from the screening interview. Paradata include variables generated from records of each call attempt and interviewer observations made about neighborhoods, households, sampled persons (including estimated characteristics of these persons along with any relevant questions or comments they may have made). Paradata elements included variables describing the housing unit structure (multi-unit structure vs single family home) and whether the household is in a locked building or gated community. Call record data and contact observations were continually collected through all stages of interviewing.

The auxiliary data were further tailored to the content of the survey by asking interviewers to make estimates or guesses about person-level or household-level characteristics thought to be related to fertility and sexual practices. One of these tailored interviewer observations was used in the present study. This is an observation that was recorded after selecting an eligible person, but before interviewing the selected person. The interviewer was asked to guess whether the selected person was in a sexually active relationship with a person of the opposite sex.

The sampling frame data include data about the neighborhood (Census Block, Block Group, or Tract) from the 2000 Decennial Census. These data are very general and have been shown to be not very useful for predicting survey variables (Biemer and Peytchev, 2013). Examples include the proportion in the ZIP Code Tabulation Area (ZCTA) that have never been married.

Since the screening interview collects the age, sex, race, and ethnicity of each person in the sampled household, these data are available as auxiliary data for households that completed the screening interview.

We also have data available from a commercial vendor. These data come from a variety of sources and are matched to the US Postal Services Delivery Sequence File. Match rates vary across the many variables available. These match rates are extensively explored elsewhere (West, et al., Under Review). In general, about 70% of households have information such as the name of an adult household member and some information about ages of adults 18+ in the household.

There are several key statistics measured by the NSFG. For this study, we chose to use a binary variable, "never been married." This statistic is collected for all respondents – men, women, and teenagers.

It is possible to use the methods described in this paper to target influential cases across a range of key outcome variables. The methods we would propose are akin to methods for sample design for multipurpose surveys (Kish, 1988; Valliant and Gentle, 1997). We can expand our notation to include key stastistics. For each statistic, we have $\pi_{ij} = PR(INFLUENTIAL_{ij} = 1)$ where $j = 1, \ldots, P$, and j indexes the key statistics. We could take the average value $\bar{\pi}_i = \sum_{j=1}^{P} \pi_{ji}/P$ or even a weighted average and establish a cutoff for that value. Anything above that cutoff is targeted for data collection.

We tested the approach experimentally with a single outcome Y. The experiment was designed to be implemented in Week 6 of the 12 week data collection. At the end of Week 5, we estimated the models described earlier. The first model is using the respondents and predicting their reported Y using the available sampling frame, paradata, and screening data. The screening data are only available for households that have completed a screening interview. There was one set of models estimated for cases that had not been screened, and another set of models fit for cases that had been screened. The latter models included data from the screening interview. We stored the predicted probability of being influential for each nonresponding case.

We then randomly divided the remaining active cases into two groups – experimental and control treatments. The experimental group had the 30% of cases with the highest predicted probability of being influential flagged. The flag was displayed in the sample management system. Interviewers were asked to make these flagged cases a high priority for interviewing. In other research (Wagner et al., 2012), we have found this prioritization scheme to be effective in increasing the effort on the flagged cases. This extra effort frequently leads to higher response rates for the flagged cases. The control group had a random 30% of cases flagged.

4. Results

4.1 Effort by Treatment

Table 1 shows the results from a model testing whether the number of calls on a case was different for experimental vs control cases, flagged vs not, and an interaction term for flagged by experimental status. We want flagged cases to be called more, but we do not want experimental vs control to differ. Nor do we want the interaction to be significant as that would indicate that flagged cases in one treatment group had more or fewer than the average number of calls. The model results indicate that the effort applied to the flagged cases was more, but not different across the two treatment groups.

Table 1: Coefficients for Regression Model Predicting the Number of Calls during the Intervention

Parameter	**Estimate**	**Standard Error**	**t-Value**	$\mathbf{Pr} > \mathbf{|t|}$
Intercept	1.363296	0.043684	31.21	$< .0001$
Experimental Group	0.049625	0.061778	0.8	0.4218
Not Flagged	-0.22999	0.048994	-4.69	$< .0001$
Experimental Group * Not Flagged	-0.07165	0.069286	-1.03	0.3011

4.2 Which Cases are Influential?

Each quarter, a model was run predicting who would be influential. Table 2 shows the odds ratios estimated for such a model from Q8.

Table 2: Odds Ratios for Predictors from Q8 Model for Which Cases will be Influential

Effect	$\mathbf{\hat{OR}}$	**95% CI**	
Commercial Data: Someone 18-24	0.943	0.719	1.238
Commercial Data: Someone 46+	1.282	1.046	1.572
Commercial Data: Someone Married	0.728	0.578	0.917
Commercial Data: 3+ Adults in HH	0.985	0.759	1.278
Commercial Data: No One 18-24	0.889	0.689	1.148
Commercial Data: No One 35-64	1.194	1.004	1.419
Commercial Data: No One 65+	0.94	0.741	1.192
PSU Not Self Representing	1.085	0.937	1.257
Single Family Home	0.757	0.638	0.898
Multi-Unit Structure	0.953	0.786	1.155
Selected R Sex Active	1.355	1.069	1.717
Selected R Age	0.998	0.988	1.008
Selected R Hispanic	0.957	0.802	1.141
Selected R Black	1.344	1.147	1.573
Selected R Male	0.977	0.858	1.113

It appears that persons who are in households that may not have someone 35-64, may have someone who is 46+, where a selected respondent is judged to be sexually active, or the selected respondent is black have a higher probability of being influential. These odds ratios are interesting, but the probability of being influential is predicted using a multivariate model. The next section looks at the results of this model.

4.3 Characteristics of Flagged and Interviewed Cases by Treatment.

Next, we look at who is being flagged and interviewed across the two treatment groups. Table 3 includes the variables used in the model predicting who would be influential. The asterisks in the significance tests column indicate $p < 0.05$. The flagged cases in the control group are a random sample of the active cases. Therefore, the differences between the experimental and control group cases that are flagged gives an indication of which cases are predicted by the multivariate model to be influential. For instance, there are significantly more households where the commercial data indicate that someone 46+ is in the household that are flagged for interviewing in the experimental group. This indicates that households with that characteristic are more likely to be influential.

Table 3: Flagged Cases By Treatment (Experiment and Control); Flagged and Interviewed Cases by Treatment

	Flagged			**Flagged and Iw'd**		
Variable	**Exp**	**Con**	**Sig.**	**Exp**	**Con**	**Sig.**
Never Been Married				55.60%	53.80%	
Commercial Data: Someone 18-24	3.50%	3.10%		7.80%	7.30%	
Commercial Data: Someone 46+	43.30%	23.70%	*	31.20%	29.30%	
Commercial Data: Someone Married	21.10%	19.90%		5.20%	19.50%	*
Commercial Data: 3+ Adults in HH	13.90%	11.40%	0.09	7.80%	11.00%	
Commercial Data: No One 18-24	61.90%	54.80%	*	74.00%	65.90%	
Commercial Data: No One 35-64	24.80%	25.90%		40.30%	36.60%	
Commercial Data: No One 65+	53.60%	51.50%		67.50%	74.40%	
PSU Not Self Representing	63.40%	63.60%		72.70%	73.20%	
Single Family Home	56.60%	65.70%	*	45.50%	61.00%	*
Multi-Unit Structure	19.90%	15.80%	*	29.90%	15.90%	*
Selected R Sex Active	97.90%	94.20%	*	92.20%	76.80%	*
Selected R Hispanic	26.10%	21.60%		13.00%	14.60%	
Selected R White/Other	65.00%	47.30%	*	58.40%	47.60%	
Selected R Male	52.40%	49.20%		53.30%	47.60%	
Selected R Only Adult in HH	30.30%	9.20%	*	16.90%	12.20%	
Selected R 15-19	0.30%	21.20%	*	2.60%	29.30%	*
Segment Obs: No Access Problems	64.00%	63.10%		68.80%	63.40%	
Segment Obs: Unimproved Roads	11.30%	12.40%		5.20%	12.20%	
Selected R Age (Mean)	28.8	28.8		27.5	27.1	
ZCTA Never Married Rate	55.6	55.6		56.6	57.5	
ZCTA Elig Rate	0.4	0.4		0.4	0.4	

It is interesting to note that teens who are sampled to be interviewed are flagged at a much lower rate. This is apparently because most teens have never been married and their value for this variable is more easily predicted for this group. On the other hand, when the selected respondent is the only adult in the household, it is more difficult to predict whether they have never been married. This makes sense as they are apparently not married at the

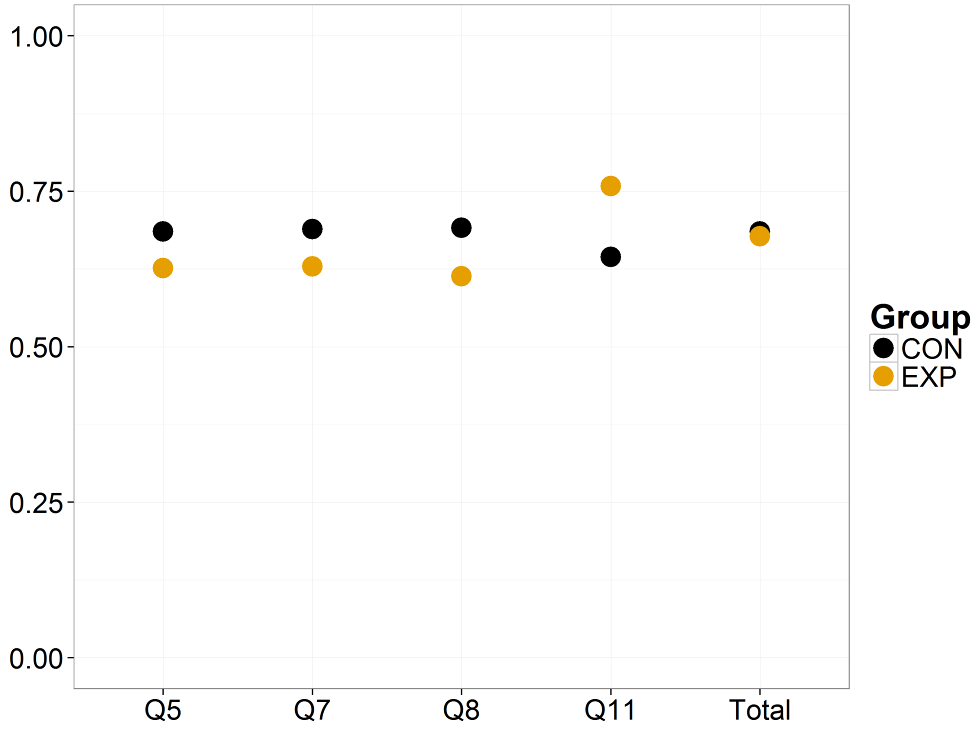

Figure 2: Proportion of AIC relative to Null Model by Treatment

moment, but may have been married in the past.

The difference in the outcome variable "Never been married" between the two treatments is small and not statistically significant. But that does not necessarily mean that the two sets of respondents are equivalent. It may be that one set of respondents is better distributed across the observed sampling frame and screening interview characteristics, where "better" indicates more accurate predictions for nonresponders.

4.4 Model Fit

Finally, we look at whether the flagging and interviewing of cases that are predicted to be influential made any difference in the final results. In 3 of the 4 quarters, the experimental group had more of a reduction in AIC than the control group with the full model. This indicates better model fit and is analogous to the inverse of the Pseudo-R2 statistic. However, the difference in Q11 is larger than that of any other quarter, and runs in the opposite direction. Therefore, the overall difference is 0.

We also looked at a model predicting marital status using the same predictors as earlier models, but now with each predictor also interacted with an indicator for the experimental treatment. There were 21 predictors in the original model (plus an intercept). Of the interactions of the experimental treatment with these 21 predictors, 7 were significant at the $p < 0.05$ level. These interactions are shown in Table 4.

Table 4: Statistically Significant Interactions of Predictors with Experimental Treatment Group Indicator

Parameter Interacted with Experimental Treatment Indicator	**Estimate**	**SE**	$\mathbf{Pr > ChiSq}$
Commercial Data: 3+ Adults in HH	-3.5534	1.7978	0.0481
Commercial Data: No One 18-24	1.7128	0.625	0.0061
Commercial Data: No One 65+	-1.4916	0.5703	0.0089
Screening Data: Race=Black	-0.9979	0.4139	0.0159
Screening Data: Sex=Male	0.9395	0.3412	0.0059
Screening Data: Single Adult in HH	-1.1704	0.5208	0.0246
Segment Obs: No Access Problems	0.9267	0.4167	0.0262

5. Discussion

The experiment produced some interesting results. The method could identify cases to be flagged that were different in ways that made sense. In other words, flagging persons who live in single adult households makes sense if our goal is to predict which persons have never been married. It is more difficult to predict this characteristic for such persons than persons from households with two adults. It may be, however, that some of the characteristics in the model can work in a manner that they offset each other. For instance, younger persons who live in single person households may be easier to predict whether they have never been married than older persons living in households with more than one adult. This highlights the importance of the modeling step in setting up interventions based on influence diagnostics.

There was an apparent pattern that use of the method frequently led to collecting data that had a better model fit. The exception was Q11, where the reverse pattern emerged. We also found that the experiment flag, when interacted with some characteristics from the original model, were significant predictors of survey outcome variable. This provides evidence that we are not getting the same kind of persons by flagging and interviewing cases based on the observed data. In other words, it seems that an adjustment strategy based on the observed data would not necessarily produce the same results as our data collection strategy.

A key question is why the experiments did not yield stronger results? We have several hypotheses. First, the experimental strategy influenced only 1 of 12 weeks of data collection. Most of the data was actually collected without guidance from influence diagnostics. It could be that the interviewers went back to what they had been doing before the experiment and, by the end of 12 weeks, had overcome the results of the experiment. It might be the case that running the experimental method for a longer period of time would produce stronger results.

A second hypothesis is that there were problems with the model used to predict never been married and then which cases are influential. The model fitting is an important step. It might be good to test the method across a range of variables and even with different models. In this experiment, the model did not vary. Varying models would be a way to test the importance of this dimension.

A third hypothesis is that the predictors in the model are too weak to allow for this strategy to be effective. If the associations with the outcome are small, then it may be that the influence diagnostics will not be very informative.

There are several next steps implied by these hypotheses. The first is to run the experiment for a larger proportion of the field period. The second is to try the method on different survey outcome variables. The third is to obtain better data to be used as predictors. These modifications should assist in further evaluation of this approach.

References

Atrostic, B. K., N. Bates, G. Burt, and A. Silberstein (2001). Nonresponse in us government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics 17*(2), 209–226.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression diagnostics: Identifying influential observations and sources of collinearity*, Volume 101 of *John Wiley and Sons*.

Biemer, P. and A. Peytchev (2013). Using geocoded census data for nonresponse bias correction: An assessment. *Journal of Survey Statistics and Methodology 1*(1), 24–44.

Brick, J. M. and D. Williams (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science 645*(1), 36–59.

Couper, M. P. (1998). Measuring survey quality in a casic environment. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 41–49.

Curtin, R., S. Presser, and E. Singer (2005). Changes in telephone survey nonresponse over the past quarter century. *Public Opinion Quarterly 69*(1), 87–98.

de Leeuw, E. and W. de Heer (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves (Ed.), *Survey Nonresponse*, pp. 41–54. New York: John Wiley & Sons.

Finamore, J., S. Coffey, and B. Reist (2013). 2013 national survey of college graduates: A practice-based investigation of adaptive design. In *Paper presented at the Annual AAPOR Conference*.

Groves, R. M. and E. Peytcheva (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly 72*(2), 167–189.

Kish, L. (1988). Multipurpose sample designs. *Survey Methodology 14*(1), 19–32.

Luiten, A. and B. Schouten (2013). Tailored fieldwork design to increase representative household survey response: an experiment in the survey of consumer satisfaction. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(1), 169–189. 1467-985X.

Petroni, R., R. Sigman, D. Willimack, S. Cohen, and C. Tucker (2004). Response rates and nonresponse in establishment surveys bls and census bureau. *Presented to the Federal Economic Statistics Advisory Committee*, 1–50.

Rosen, J. A., J. Murphy, A. Peytchev, T. Holder, J. Dever, D. Herget, and D. Pratt (2014). Prioritizing low propensity sample members in a survey: Implications for nonresponse bias. *Survey Practice 7*(1).

Schouten, B., J. Bethlehem, K. Beullens, y. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through r-indicators and partial r-indicators. *International Statistical Review 80*(3), 382–399. 1751-5823.

Schouten, B., F. Cobben, and J. G. Bethlehem (2009). Indicators for the representativeness of survey response. *Survey Methodology 35*(1), 101–113.

Schouten, B., F. Cobben, P. Lundquist, and J. Wagner (Under Review). Does balancing response reduce nonresponse bias?

Valliant, R. and J. E. Gentle (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis 25*(3), 337–360. 0167-9473 doi: DOI: 10.1016/S0167-9473(97)00007-8.

Wagner, J., B. T. West, N. Kirgis, J. M. Lepkowski, W. G. Axinn, and S. K. Ndiaye (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics 28*(4), 477–499.

West, B. T., J. Wagner, H. Gu, and F. Hubbard (Under Review). The utility of alternative commercial data sources for survey operations and estimation: Evidence from the national survey of family growth.