Relationship between Measurement Error and Unit Nonresponse in Household Surveys: An Approach in the Absence of Validation Data

Andy Peytchev[1], Emilia Peytcheva[2]
[1]Program in Survey Research and Methodology, Research Triangle Institute, 3040 Cornwallis Rd, RTP, NC, 27709
[2]Program in Survey Methodology, University of Michigan, 426 Thompson St, Ann Arbor, MI, 48104

Abstract

Response rates in national probability surveys are falling despite higher cost of data collection from greater field efforts. The inherent threat of lower response rates is the increased potential for nonresponse bias. Some have argued that nonrespondents in a survey would be poor respondents if their cooperation is gained. If true, the higher cost per interview for cases that have already received a lot of effort could be better directed to other areas of the survey process.

Bringing nonresponse and measurement error combines two estimation problems in sample surveys. Nonresponse is a problem of missing survey data, while measurement error requires survey and validation data.

This study uses a model for nonresponse, and proposes a model for measurement error in the absence of auxiliary information. The technique involves the simultaneous estimation of means and variances in purposefully constructed models, and provides respondent-level estimates of measurement error. These estimates provide the potential for studying linkages between survey errors, including the identification of preferable measurement conditions such as particular interviewers inducing less measurement error.

1 Introduction

Nonrespondents would provide excessive measurement error if interviewed. This bold assertion can have great impact; if erroneously accepted, survey effort may be diverted away from minimizing nonresponse bias. There is limited empirical evidence that response propensity may be related to measurement error Cannell and Fowler, 1963; Assael and Keon, 1982; De Leeuw and Hox, 1988; Biemer, 2001; Voogt, 2005, but no published study examines the joint relationship between the nonresponse and measurement error. Yet, the notion of measurement error in responses provided by nonrespondents is a counterfactual one; by definition, these outcomes are never observed together.

Inherently there are two problems that need to be solved to test the counterfactual hypothesis, particularly in the common case where validation data is not available: identification of nonrespondents and estimation of measurement error.

Some studies have compared cases that have refused participation at some point but completed the interview (e.g., Billiet et al., 2007). Most authors are quick to acknowledge the obvious limitation: comparing those who hesitate to respond (coded differently by interviewers and organizations) to those who cooperate without hesitation is not equivalent to comparing respondents to nonrespondents. It involves the untested assumption that in each study, those who refused at some point to some degree are like those who never cooperated. A single binary indicator for whether the respondent ever refused reported by the interviewer is subjected to measurement error itself and asserts a deterministic equivalence between those who refused at some point and those who never cooperated based on a single behavior.

Instead, a model can be adopted that is both stochastic and is informed by multiple variables predicting membership in the respondent and nonrespondent group while also reflecting the likelihood of that membership, or response propensity (Rosenbaum and Rubin, 1983). This method reduces the extent of misclassification cases, and utilizes more information into the assignment of a value.

In the absence of validation data, a common proxy for measurement error is item nonresponse (e.g., Willimack et al., 1995). However, item nonresponse reflects data quality, which is distinct from measurement error - not providing a substantive response may be the most accurate response for some respondents and measurement error would have occurred had they provided a response.

A definition for measurement error is offered in Classical Test Theory (c.f. Novick, 1966). Under this psychometric theory, measurement error is defined as the difference between the obtained response and the underlying true value; Generalizability Theory (Cronbach, Rajaratnam and Gleser, 1963) extends this model by allowing different respondents to have different true values. The fundamental concept is that each person has a True Score that is imperfectly measured.

Unfortunately, true values for key survey variables are seldom available; they are the very purpose of the survey and when available, they are subject to measurement error themselves. Obtaining alternative measures of a variable of interest is usually costly and often impossible. Furthermore, measurement error is the property of a statistic and the relationship between unit nonresponse and measurement error may depend on both the particular survey and the statistic being examined. In order to

understand the relationship between nonresponse and measurement error, we need the ability to estimate and evaluate measurement error in each survey and for various estimates.

1.1 Unit Nonresponse

Rather than assuming a deterministic model where some people are classified as respondents and others as nonrespondents, a more realistic view is stochastic - respondents vary in their likelihoods of being respondents or nonrespondents. Under this view, we can estimate these response propensities on a continuum. In order to do this, covariates need to be available that are related to the response outcome. In the case where the response outcome of interest is binary, a logistic regression can be fit:

$$\ln\left(\frac{p_i(r=1)}{1-p_i(r=1)}\right)=\alpha+\tilde{\beta}\tilde{w}_i \qquad [1]$$

and therefore,

$$p_i(r=1)=\frac{e^{\alpha+\tilde{\beta}\tilde{w}_i}}{1+e^{\alpha+\tilde{\beta}\tilde{w}_i}}$$

where p_i is the estimated probability that sample member i is a respondent in the survey and the vector $\tilde{w}$ are the variables predictive of the likelihood of responding.

The utility of these estimated propensities depends on the inclusion of strong correlates of unit nonresponse. While overlooked in model building, one of the primary goals of response propensities in surveys is to reduce nonresponse bias by weighting by their inverse. In this study the response propensities will not be used for weighting and we are interested in them even if they are not associated with survey variables – the association between the response propensities and measurement error will be tested. Although seemingly contradictory, another aspect of optimization of the response propensity models is to have a sufficient proportion of sample members with low response propensities who cooperated with the survey request and provided responses. The extreme separation of respondents from nonrespondents in response propensity models that perform well in predicting response outcome can be problematic as it creates few influential cases. To avoid having very influential cases, response propensities can be divided into a few groups and assigning a single response propensity to each group, and trimming extreme propensities. The present study retains as much information as possible without grouping, and trims the extreme tails of the distribution.

1.2 Measurement Error

In the True Score Model, the observed value is equal to the true score and a measurement error. The true score is a latent variable that can not be measured perfectly.

$$x_i = X_i + \varepsilon_i \qquad [2]$$

In this model, x_i is the observed variable for person i, X_i is the true score, the underlying true value that is measured by x, and ε_i is the measurement error. This model specification that allows the true score to differ for each individual is known as Generalizability Theory.

In order to improve the measurement of the latent construct of interest, researchers often use multiple measures to identify it. Thus, the true score can be replaced by multiple measures. We restate the true score model in Equation 2, in terms of a multiple regression of observables:

$$x_i = \alpha + \tilde{\beta}\tilde{x}_i + \tilde{\beta}\tilde{z}_i + \varepsilon_i \qquad [3]$$

To the extent that the vector of covariates $\tilde{x}$ is successful in identifying the true score, the error term ε becomes the measurement error in x. In the social sciences, the identification of X for each respondent will certainly be imperfect, so that the error term will be affected by measurement error and some true score variance that we fail to completely identify. The goal here is to select the covariates that best identify X. In addition, a vector of covariates $\tilde{z}$ that explain systematic variance in x but are not necessarily related to the true score X can be included to evaluate measurement error bias. For example, in surveys where sample cases are randomly assigned to interviewers, including an interviewer identifier in the model would provide estimates of differences in interviewer bias, just as in the Analysis of Variance based models for interviewer variance proposed by Kish (1962) and Hansen, Hurwitz, and Bershad (1961).

This model provides a respondent-level estimate of measurement error, albeit imperfect. This provides the opportunity to then model these errors as a function of likely causes, such as experimentally varied questionnaire design features, individual interviewers, interviewer and respondent characteristics, and mode of data collection. Of particular importance here is that measurement error can be evaluated as a function of other sources of survey error as well – such as the likelihood of being a respondent. This is done in a second model where the dependent variable is the measurement error from the first model:

$$\ln(\varepsilon_i^2) = \alpha + \tilde{\beta}\tilde{z}_i + \delta_i \qquad [4]$$

In this model the vector of covariates $\tilde{z}$ can include all the covariates that are in $\tilde{x}$ in Equation 3, although this is not necessary, and any other factors that are possible causes of measurement error.

The two models in Equations 3 and 4 are interdependent; parameters in the second model depend on the estimated

measurement errors in the first model, while the measurement errors in the first model are a function of the factors in the second model. In order to solve for this interdependency, the two models are fit iteratively until a convergence criterion is satisfied.

As noted earlier, the identification of measurement error in the first step would be imperfect when examining survey variables as compared to physical, chemical, or production processes where the variability in an outcome measure may be perfectly explained. Other than a threat that an unexplained part of the true score is also correlated with any of the factors in the measurement error model (step two), the effect of unexplained variance that is not measurement error is to attenuate the parameter estimates in the measurement error model. To the extent that a large proportion of the variability in ε is random variability that we have failed to "remove," it would produce downward-biased parameter estimates in step two, thus it would be harder to reject the null hypothesis when testing for significance.

This method is related to the Box-Cox (Box and Cox, 1964) transformation for heterogeneity of variance, only in that case it is merely accounting for regression residuals that do not have a constant variance. It is more related to work in (offline) quality control, where of interest is not just achieving, for example, manufactured components of a particular size, but doing so consistently. When the goal is minimizing residual variance through altering production process parameters, the identification of the factors associated with higher variance is of prime interest. Modeling residual variance rather than total variance allows for better identification of these factors. Such an approach was implemented by Taguchi and Wu (1980) and has received some attention since (Nair et al., 1992; Engel and Huele, 1996), yet social sciences and particularly survey methodology have maintained their traditional focus on differences in means, looking at simple response variance at best. While in this paper we are interested in estimation and evaluation of causes and correlates of measurement error in survey data, the additional benefit from this approach is a reduction in bias in the parameter estimates in regression—something that is addressed with the Box-Cox transformation but often overlooked. One example of this in the survey methodology literature is the estimation of interviewer variance, based on differences in interviewer means, under the assumption that interviewers induce the same simple response variance, as acknowledged by Groves and Magilavy (1980).

In the statistical literature these models are referred to as mean-variance models as the first step is estimation of parameters for predicting means, and the second step is estimation of parameters on residual variances. From here on we will alter the naming to fit the terminology used in surveys; we refer to the first step as the measurement error "Estimation" step, and the second model as the "Evaluation" step. Implicitly, when causes of measurement error are included in both steps, the first set of parameters can be interpreted as influences on measurement error bias, while the second set are the measurement error variance.

To evaluate the validity of this method, Peytchev (2006) examined the relationship between the estimated measurement error and known and expected correlates of measurement error. He used data from the 2002 National Election Studies (NES), with telephone interviews conducted before and after the presidential elections, minimizing the threat of correlated measurement error between the dependent variable and the predictors in the measurement error estimation model by regressing a post-election variable on pre-election covariates. For a thermometer rating on feelings towards blacks, race of interviewer was significantly associated with the obtained measurement error. For the same variable, respondent level of cooperation rated by the interviewer, was also associated with the measurement error. While the performance of this method depends on the variable of interest, and on the fit of the estimation part of the model (which was relatively poor in his study), the study presents some support for employing this method for estimation and modeling of measurement error.

In this study, we then turn to an empirical test of the association between response propensities and measurement error.

1.3 Relationship between Unit Nonresponse and Measurement Error

In order to address the counterfactual problem of relating unit nonresponse and measurement error, we would need to obtain responses from all nonrespondents under a particular survey protocol. While this is unrealistic, we could obtain responses from some of the nonrespondents and assume that the remaining nonrespondents are like them. A less deterministic approach is to assign stochastic response propensities as discussed above, while obtaining responses from some of the likely nonrespondents (initial refusals) will still be beneficial in providing responses from those with low response propensities. We can then estimate measurement error as described above, and the covariance between the response propensity and the measurement error, cov(p,ε). Entering the response propensity p in the first stage of the measurement error model estimates and controls for nonresponse bias in the dependent variable, conditional on the other predictors in the model, while entering it also in the second stage of the measurement error model provides the sought estimate of cov(p,ε):

$$x_i = \alpha + \tilde{\beta}\tilde{x}_i + \tilde{\beta}\tilde{z}_i + \beta p_i + \varepsilon_i \qquad [5]$$

and

$$\ln(\varepsilon_i^2) = \alpha + \beta p_i + \delta_i \qquad [6]$$

where the two equations are fit iteratively.

Yet we are interested not only in the possible association between response propensity and measurement error, but also in the factors that explain it. This can be achieved by changing the error variance equation in [6] to include correlates of measurement error; to the extent that they are also related to the response propensities, this will reduce the association between the response propensity and measurement error, if such had been found. The altered equation adds the vector of measurement error covariates, $\tilde{z}$, such as interviewer observations about the interview process:

$$\ln(\varepsilon_i^2) = \alpha + \tilde{\beta}\tilde{z}_i + \beta p_i + \delta_i \qquad [7]$$

2 Data and Methods

The ideal data set to evaluate the estimation of measurement error and test the association between response propensities and measurement error would contain true values for all respondents, implement random assignment of sample cases to interviewers (and random assignment of other known causes of measurement error, such as mode if multiple modes are implemented), would minimize the possibility of correlated errors between the variable of interest, x, and the variables explaining it, $\tilde{x}$, and include measures for correlates of response propensity, $\tilde{w}$, and measurement error, $\tilde{z}$.

We do not have the ideal data set and it may not exist, but we can meet different subsets of these beneficial conditions through studies with different designs that vary in the assumptions that need to be made. Peytchev (2006) used the 2002 NES that does not have the degree of clustering of respondents by interviewer that occurs in area probability surveys, but had a substantive model for estimating measurement error with relatively poor fit, and lacked information on unit nonresponse.

We turn to the National Comorbidity Survey 2001-2003 Replication (NCS-R) a multi-stage area-probability face-to-face survey of adults 18 and over. This survey has a focus on prevalence of mental disorders, as well as their correlates, providing the ability of building OLS models (for Equation 5) with relatively high explanatory power that is necessary for this approach.

Ideally, variance estimation would reflect the complex sample design. Due to software limitations, we included cluster indicators in the first stage; hence, the standard errors of coefficients may be underestimated. Sampling weights were used in all models to account for subsampling of locked status buildings and within household selection.

We would expect to find a relationship between unit nonresponse and measurement error when there is a common cause for both outcomes. We selected a dependent variable measuring the degree to which depression interferes with the respondent's work. Depression is a sensitive issue to many respondents that is likely to induce measurement error, but it may also be associated with unit nonresponse as it is part of the survey topic. This variable is measured on an 11-point scale, where 0 is "No interference" and 10 is "Very severe interference." It also did not have a skewed distribution, although as often found in telephone surveys, slightly higher frequencies were observed at 0, 5, and 10.

This study made exceptional efforts on nonresponse, in terms of reduction in nonresponse rate, creating weighting adjustments, and juxtaposing alternative methods for addressing nonresponse bias to examine sensitivity to the methods used (see Kessler et al., 2004). The study implemented a phased design; one month before the end of the field period remaining nonrespondents were subjected to a change in study protocol, increasing incentives from $50 to $100, shortening the survey to about a third of the full instrument, and providing additional incentives to interviewers for completing any of the remaining cases. One method for evaluating differences between respondents and nonrespondents in a phased design is to make the assumption that the remaining nonrespondents are like the respondents in the second phase. There is likely a large similarity between the two, but this is a very stringent assumption of homogeneity in a group that includes reluctant and difficult to contact sample members. Indeed, Kessler and his colleagues (2004) found some similarity between these subgroups, but also some differences. They weighted all cases to the full sample and gave an additional weight to the respondents in the second phase to compensate for the remaining nonrespondents under the above homogeneity assumption. However, instead of using this as a nonresponse weight, they used the weight in the estimation of a response propensity model, essentially giving more influence to cases in the second phase. The constructed response propensity model included Census segment level covariates (region and urbanicity), block group level measures (average household income, average number of adults per household, proportion not in the labor force, proportion Hispanic, etc.), individual level demographics (age, sex, marital status, employment status, etc.), and individual level substantive variables, diagnostic questions in the screening section (various questions measuring mood disturbance, anxiety, substance use, and impulse control problems). From the four sets, the individual level substantive variables were not significant predictors of response outcome and were excluded. The inverse of these response propensities became the nonresponse weight component, and the weights were trimmed at the extreme 2% at either end of the distribution. If we take the inverse of these weights, we essentially reproduce a trimmed distribution of the response propensities that can be used in relating response propensity to measurement error.

While we are interested in the association between measurement error and unit nonresponse, we also need to evaluate the estimated measurement error. These data include interviewer observations about the interview. Interviewer ratings of respondent understanding of the survey questions, respondent level of effort in answering questions, respondent cooperativeness during the interview, and whether someone else was present during the interview, should be related to measurement error. The interviewer observations may not be as valid or reliable as we would like. As a result, the parameter estimates from these associations are likely downward biased, or an underpowered test. The properties of different interviewer observations can also vary, so we may find only some variables to exhibit the expected relationship.

In addition to interviewer observations, we enter respondent demographic characteristics that should be related to measurement error. Respondents with lower levels of education have been found to elicit higher measurement error (proxied by response order effects) under the hypothesis of cognitive sophistication (Krosnick and Alwin, 1987; Knauper, 1999). Same was observed for older respondents (Andrews and Herzog, 1986), explained by a cognitive aging mechanism (Hertzog and Bleckley, 2001). Unlike the interviewer observations, these effects should be resilient to measurement error from the interviewers themselves.

We first estimate an OLS model that does not model the measurement error. We then estimate the two-stage model with response propensity in the measurement error equation. Correlates of measurement error are then entered, which serves the dual purpose of providing another evaluation of how well we do indeed obtain estimates of measurement error, and of testing whether the correlates we have also manage to explain any association between the response propensities and measurement error – identifying potential common causes.

One example of a common cause of measurement error and unit nonresponse that Peytchev (2006) examined are the interviewers themselves; they can induce measurement error and they vary in their response rates. Of particular interest is the direction of each association. We may expect that interviewers who do one task well, that is, obtaining respondent cooperation (thus., minimizing nonresponse), are also likely to do other tasks well - for example, conducting standardized interviews without influencing respondent answers (thus., minimizing measurement error). Yet one task requires the interviewer to influence the respondent, while the other to minimize influence on the respondent. Some evidence suggests that the more experienced interviewers (who tend to produce higher response rates) may also elicit higher measurement error through lower reports of sensitive behaviors (Hughes et al., 2002; Chromy et al., 2005).

3 Results

The model in the NCS-R for measurement error in the degree to which depression interfered with the respondent's work, fit the data acceptably well; the proportion of variance explained in the measurement error estimation model without accounting for the heterogeneous variance was 0.3.

Model 1, the Means Only Model, in Table 1 is an OLS regression with no modeling of the measurement error variance. The parameter estimate for response propensity in this mean-only model is not significant; there is no nonresponse bias in the mean for this dependent variable, when estimating it as the covariance between the response propensity and the survey variable (see Bethlehem, 2002). Among the interviewer observations, lower degree of interference by depression was reported when someone was present at some point during the interview, to which we return shortly, and also when the respondent was rated as putting a lot of effort in answering the questions.

Model 2, the Response Propensity Model, has the error variance as a function of the response propensity. We find no association between the measurement error and the response propensity.

Interviewer observations were added to the error variance model in Model 3, the Response Propensity and Interviewer Observations Model. Respondents who were rated by the interviewer as having very good question understanding, produced significantly less measurement error (about 79% of the measurement error estimated for those who were not rated as having very good question understanding).

Finally, education and age are added to the error variance model in Model 4, the Response Propensity, Interviewer Observations, and Respondent Characteristics Model, as proxies for cognitive sophistication and cognitive aging. Education was associated with measurement error in the expected direction – every additional year of education resulted in 5% decrease in measurement error. Older respondents produced significantly more measurement error - a two percent increase in measurement error for every year of age.

Accounting for measurement error as a function of the covariates in Model 4 also improves the substantive part of the model. Recall that the measurement error estimates are dependent on the parameter estimates in the mean part of the model, but the coefficients for the means are in turn dependent on the error variances; this interdependency is the reason why iterative model fitting is needed. In the simple OLS regression in Model 1, only 3 of the 13 substantive predictors were statistically significant (Lose interest in doing things; Unable to make up mind; and Can't cope with responsibilities); after modeling the error variance in Model 4, 6 of the substantive predictors had significant coefficients. This difference was driven by the increased magnitude of the coefficients, rather than reduction in standard errors.

When accounting for differential measurement error (Model 4), the "bias" coefficient for the interviewer observation whether someone else was present during the interview (suggesting that presence of others leads to lower reports of interference by depression), was no longer significant. This relationship between measurement error bias and measurement error variance calls for further attention.

4 Discussion and Conclusions

This study aims to contribute to two lines of research: to the investigation of the link between unit nonresponse and measurement error, and to solutions to the problem of estimation of measurement error in the absence of validation data. We present four important results:

1. Response propensity was not associated with measurement error.

While this is good news for survey practitioners, it should be taken with caution. The association between unit nonresponse and measurement error is at the statistic level; therefore, it may exist for other statistics. More importantly, it may appear only in some surveys, depending on the protocol used. The combination of these two arguments leads to a critical point: nonresponse and measurement error are going to be associated when they have common causes. In this case, they were either not present, or not captured in the response propensity model. Future research is much needed in the identification of common causes to allow anticipation of when minimizing one source of survey error may increase another, or anticipation of the desired situation of decreasing multiple sources of error. One such mechanism is how sensitive the topic is to some of the sample members. The current study tried to explore this mechanism, as depression was the topic of the survey and of the question examined, but there was no association between the response propensity and the depression variable.

2. Those rated as having very good question understanding answered with less measurement error.

While this is expected, it provides support for the validity of the measurement error estimates. The magnitude of this association is likely attenuated by measurement error in the interviewer observations themselves, a potential explanation why the other three interviewer observations were not significantly associated with the estimated measurement error.

3. Older and less educated respondents provided more measurement error.

These effects and their direction are inline with causal arguments and empirical findings on cognitive aging and cognitive sophistication, albeit like any variables, they are imperfect proxies of the constructs. The magnitudes of these effects is noteworthy – every year in age is associated with a 2% increase in measurement error, and a year of education associated with 5% decrease in measurement error.

4. Accounting for differences in measurement error improves substantive models.

Half of the substantive associations in the model appeared only after accounting for differential measurement error, particularly due to age and education. A naïve OLS model assumes constant error variance, an assumption that is most likely seldom tested. Ignoring it would be wrong as the parameter estimates and standard errors could be biased, as found here. Using econometric models for heterogeneity of variances may be complicated to perform and interpret for data users. This simultaneous (through iterative fitting) explicit modeling of error variances allows accounting for heterogeneity, while employing social/psychological theories by controlling the factors in the measurement error model, and provides separate interpretation of the parameter estimates for these factors.

This two-stage model allows accounting for differential measurement error in analysis of survey data, which even if not done on the basis of cognitive theories, should be done to address violations of assumptions in common statistical models. There was supporting evidence that using the proposed model for estimation of measurement error provided valid estimates. This is one method for obtaining measurement error; other approaches that make different assumptions would improve inferences drawn about the causes and correlates of measurement error. The observational approach to identify common causes of measurement error and unit nonresponse here was to select a question that was on the same sensitive topic as the survey; hence, a likely common cause. Another approach to addressing causal questions would have experimental manipulations of known causes of measurement error and unit nonresponse.

There are other error sources not examined here. Survey practitioners need to consider all of them. Different sources of error can have common causes, creating an association between them. When this occurs, it would be advantageous if decreasing the impact of one source of error in a survey estimate also decreases another; yet the opposite effect may occur. We need to know for which statistics this could occur, and identify any manipulable features in the data collection protocol that would decrease multiple error sources. Future research is needed to examine multiple sources of survey error in various statistics.

Acknowledgment

The authors are grateful to Patricia Berglund for recreating and providing the NCS-R weight components.

References

Andrews, F. M. and A. R. Herzog (1986). "Respondent Age and Survey Measurement Error." Journal of the American Statistical Association **81**: 403-410.

Assael, H. and J. Keon (1982). "Nonsampling Vs. Sampling Errors in Survey Research." Journal of Marketing **46**(2): 114-123.

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information. Survey Nonresponse. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York, Wiley**:** 275-288.

Biemer, P. P. (2001). "Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing." Journal of Official Statistics **17**(2): 295-320.

Billiet, J., M. Philippens, R. Fitzgerald and I. Stoop (2007). "Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents." Journal of Official Statistics **23**(2): 135-162.

Box, G. E. P. and D. R. Cox (1964). "An Analysis of Transformations." Journal of Royal Statistical Society, Series B **26**: 211-246.

Cannell, C. F. and F. J. Fowler (1963). "Comparison of a Self-Enumerative Procedure and a Personal Interview: A Validity Study." Public Opinion Quarterly **27**(2): 250-264.

Chromy, J. R., J. Eyerman, D. Odom, M. E. McNeeley and A. Hughes (2005). Association between Interviewer Experience and Substance Use Prevalence Rates in Nsduh. Evaluating and Improving Methods Used in the National Survey on Drug Use and Health. J. Kennet and J. Gfroerer. Washington, DC, Substance Abuse and Mental Health Services Administration**:** 59-86.

Cronbach, L. J., N. Rajaratnam and G. C. Gleser (1963). "Theory of Generalizability - a Liberalization of Reliability Theory." British Journal Of Statistical Psychology **16**(2): 137-163.

De Leeuw, E. D. and J. J. Hox (1988). "The Effects of Response Stimulating Factors on Response Rates and Data Quality in Mail Surveys." Journal of Official Statistics **4**(3): 241-249.

Engel, J. and A. F. Huele (1996). "Taguchi Parameter Design by Second-Order Response Surfaces." Quality And Reliability Engineering International **12**(2): 95-100.

Groves, R. M. and L. Magilavy (1980). Estimates of Interviewer Variance for Telephone Surveys. Proceedings of the Proceedings of the Survey Research Methods Section of the American Statistical Association.

Hansen, M. H., W. N. Hurwitz and M. A. Bershad (1961). "Measurement Errors in Censuses and Surveys." Bulletin of the ISI **38**: 351-374.

Hertzog, C. and M. K. Bleckley (2001). "Age Differences in the Structure of Intelligence: Influences of Information Processing Speed." Intelligence **29**(3): 191-217.

Hughes, A., J. Chromy, K. Giacoletti and D. Odom (2002). Impact of Interviewer Experience on Respondent Reports of Substance Use. Redesigning an Ongoing National Household Survey. J. Gfroerer, J. Eyerman and J. Chromy. Washington, DC, Substance Abuse and Mental Health Services Administration**:** 161-184.

Kessler, R. C., P. Berglund, W. T. Chiu, O. Demler, S. Heeringa, E. Hiripi, R. Jin, B. E. Pennell, E. E. Walters, A. Zaslavsky and H. Zheng (2004). "The Us National Comorbidity Survey Replication (Ncs-R) Design and Field Procedures." International Journal of Methods in Psychiatric Research **13**: 69-92.

Kish, L. (1962). "Studies of Interviewer Variance for Attitudinal Variables." Journal of the American Statistical Association **57**: 91-115.

Knauper, B. (1999). "The Impact of Age and Education on Response Order Effects in Attitude Measurement." Public Opinion Quarterly **63**(3): 347-370.

Krosnick, J. A. and D. F. Alwin (1987). "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement." Public Opinion Quarterly **51**(2): 201-219.

Nair, V. N., B. Abraham, J. Mackay, G. Box, R. N. Kacker, T. J. Lorenzen, J. M. Lucas, R. H. Myers, G. G. Vining, J. A. Nelder, M. S. Phadke, J. Sacks, W. J. Welch, A. C. Shoemaker, K. L. Tsui, S. Taguchi and C. F. J. Wu (1992). "Taguchi Parameter Design - a Panel Discussion." Technometrics **34**(2): 127-161.

Novick, M. R. (1966). "The Axioms and Principal Results of Classical Test Theory." Journal of Mathematical Psychology **3**: 1-18.

Peytchev, A. (2006). Estimation of Measurement Error and Identification of Causes: Linking Measurement Error to Nonresponse, Interviewers, and Interviewer Characteristics. Proceedings of the Joint Statistical Meetings of the American Statistical Association.

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika **70**: 41-55.

Taguchi, G. and Y. I. Wu (1980). Introduction to Off-Line Quality Control. Central Japan Quality Control Association. Nagoya, Japan.

Voogt, R. J. J. (2005). "An Alternative Approach to Correcting Response and Nonresponse Bias in Election Research." Acta Politica **40**: 94-116.

Willimack, D. K., H. Schuman, B.-E. Pennell and J. M. Lepkowski (1995). "Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey." Public Opinion Quarterly **59**(1): 78-92.

Table 1. Bias and Measurement Error Parameter Estimates in Degree Depression Interferes with Respondent's Work (0=No Interference, 10=Very Severe Interference) in the NCS-R, from four models: (1) Mean Only Model, (2) Response Propensities, (3) Response Propensities and Survey Design, and (4) Response Propensities, Survey Design, and Respondent Characteristics.

	Model 1: Means Only Model			**Model 2: Response Propensity**			**Model 3: Response Propensity & Interviewer Observations**			**Model 4: Response Propensity, Interviewer Observations, & Respondent Characteristics**		
Variable	**Param. Estimate**	**Std. Error**	**Error Variance Ratio**	**Param. Estimate**	**Std. Error**	**Error Variance Ratio**	**Param. Estimate**	**Std. Error**	**Error Variance Ratio**	**Param. Estimate**	**Std. Error**	**Error Variance Ratio**
Mean Model												
Lose interest in doing things	0.81***	(0.25)	--	0.81***	(0.25)	--	0.73***	(0.25)	--	0.75***	(0.24)	--
Trouble sleeping	0.43†	(0.23)	--	0.42†	(0.23)	--	0.41†	(0.23)	--	0.56*	(0.22)	--
Feel tired	0.51†	(0.29)	--	0.49†	(0.29)	--	0.56†	(0.29)	--	0.63*	(0.28)	--
Thoughts come slowly	0.27	(0.22)	--	0.27	(0.22)	--	0.28	(0.21)	--	0.23	(0.21)	--
Trouble concentrating	-0.19	(0.26)	--	-0.19	(0.26)	--	-0.19	(0.26)	--	-0.16	(0.25)	--
Unable to make up mind	1.10***	(0.23)	--	1.11***	(0.23)	--	1.12***	(0.22)	--	1.12***	(0.22)	--
Lose self-confidence	-0.42†	(0.25)	--	-0.44†	(0.25)	--	-0.48†	(0.25)	--	-0.56*	(0.25)	--
Feel irritable, bad mood	0.09	(0.21)	--	0.09	(0.21)	--	0.07	(0.21)	--	0.00	(0.21)	--
Feel nervous, anxious	0.16	(0.21)	--	0.16	(0.21)	--	0.17	(0.21)	--	0.13	(0.21)	--
Can't cope with responsibilities	1.66***	(0.22)	--	1.67***	(0.22)	--	1.74***	(0.22)	--	1.82***	(0.22)	--
Want to be alone	-0.21	(0.24)	--	-0.20	(0.24)	--	-0.21	(0.24)	--	-0.19	(0.23)	--
Less talkative	0.58*	(0.26)	--	0.57*	(0.26)	--	0.68**	(0.26)	--	0.68**	(0.26)	--
Often in tears	-0.18	(0.21)	--	-0.19	(0.21)	--	-0.21	(0.21)	--	-0.13	(0.20)	--
Response propensity (prob.)	-0.43	(0.80)	--	-0.42	(0.80)	--	0.00	(0.82)	--	-0.25	(0.80)	--
Anyone present during int'w	-0.47*	(0.20)	--	-0.46*	(0.20)	--	-0.48*	(0.20)	--	-0.33†	(0.19)	--
Very good question understanding	0.00	(0.22)	--	0.01	(0.22)	--	0.00	(0.22)	--	0.01	(0.22)	--
Excellent cooperation with int'w	0.41	(0.25)	--	0.40	(0.25)	--	0.40	(0.26)	--	0.28	(0.26)	--
A lot of effort in answering qns	-0.76**	(0.29)	--	-0.76**	(0.29)	--	-0.73*	(0.31)	--	-0.63*	(0.30)	--
Education (in years)	-0.03	(0.04)	--	-0.03	(0.04)	--	-0.02	(0.04)	--	0.01	(0.04)	--
Age (years)	0.002	(0.007)	--	0.002	(0.007)	--	0.000	(0.01)	--	0.002	(0.007)	--
Error Variance Model												
Response propensity				-0.17	(0.35)	0.84	-0.19	(0.35)	0.83	-0.06	(0.35)	0.94
Anyone present during int'w							-0.15	(0.10)	0.86	-0.10	(0.10)	0.91
Very good question understanding							-0.24*	(0.11)	0.79	-0.12	(0.12)	0.89
Excellent cooperation with int'w							-0.02	(0.14)	0.98	-0.09	(0.13)	0.91
A lot of effort in answering qns							-0.06	(0.15)	0.94	-0.02	(0.15)	0.98
Education (in years)										-0.05*	(0.02)	0.95
Age (years)										0.019***	(0.004)	1.019

Unless otherwise noted, variables are coded as indicators (1=yes, 0=no)
Parameter estimates for clusters and for model intercepts are omitted.
† p<.1; * p<.05; ** p<.01; *** p<.001