Validity of Survey Response Propensity Indicators: A Behavior Genetics Approach*

Levente Littvay, *Central European University*

Sebastian Adrian Popa, *Central European University and University of Mannheim*

Zoltán Fazekas, *University of Vienna*

Objectives. This study explains how behavior genetic analysis using a twin design can help us assess the validity of our measures. *Methods.* We test multiple indicators of response propensity, a measure used by survey researchers to better understand the similarities and differences between survey respondents and nonrespondents. The response propensity indicators evaluated include response to follow-up surveys and subsequent waves of a panel and the completion of a sensitive recontact information sheet to aid subsequent recontact efforts. *Results.* A classical and the newly proposed method of validation all point to insufficient validity of our response propensity measures. Construct validation using data from the National Survey of Midlife Development in the United States exhibited little correlation between indicators. Genetic analysis suggests that the success of subsequent data-collection efforts is predominantly driven by additive genetic effects, while nonresponse to inquiries for recontact information is influenced predominantly by familial environmental predictors. *Conclusion.* Our results indicate that different underlying constructs drive the response propensity indicators, suggesting that nonresponse is, at minimum, multidimensional.

This article aims to demonstrate how behavior genetic approaches in the social sciences can help us better understand various operationalizations of our constructs. In particular, we analyze the validity of proposed survey nonresponse proxies.

Researchers' inability to obtain data from a sampled individual (nonresponse) is a constant threat to data quality in all of the social sciences (Allison, 2001; Dillman et al., 2002; Little and Rubin, 2002). The problem is nonexistent if the sampled individuals who responded to the survey do not differ from the nonrespondents (Allison, 2001; Dillman et al., 2002; Little and Rubin, 2002; Tourangeau, Rips, and Rasinski, 2000). But since nonresponse,

*Direct correspondence to Levente Littvay, Central European University Department of Political Science, Nádor u. 9, Budapest 1051, Hungary ⟨littvayl@ceu-budapest.edu⟩.
The Web Appendix is available at ⟨http://levente.littvay.hu/appendices/ssq.pdf⟩.

SOCIAL SCIENCE QUARTERLY, Volume 94, Number 2, June 2013

DOI: 10.1111/j.1540-6237.2012.00938.x

by definition, means that the data we wish to collect will not be observed, we will never know if this assumption is reasonable or not.

Hence, several creative proxies were proposed to assess the determinants and effects of nonresponse (see, e.g., Brehm, 1993; Campanelli, Sturgis, and Purdon, 1997; Goyder, 1986; Groves, 2006; Groves and Couper, 1996, 1998) and many of these procedures attempt to measure the response propensity—a measure of a person's likelihood of responding to a survey—for each actual respondent (Groves, 2006; Groves and Couper, 1996, 1998; Olson, 2006; Rao, 1983; Sarndal and Swensson, 1987; Singh, 1983).

The assumption made by these studies is that people with low survey response propensity are similar to the people who do not respond to the survey so they can be used as a proxy to better understand the characteristics of nonrespondents (Currivan and Carley-Baxter, 2006; Etter and Perneger, 1997; Gmel, 2000; Groves, Singer, and Corning, 2000; Groves and Peytcheva, 2008; Hill et al., 1997; Lahaut et al., 2002; Lynn, 1998; Voigt, Koepsell, and Daling, 2003). But response propensity measures come in many forms (Currivan and Carley-Baxter, 2006; Etter and Perneger, 1997; Gmel, 2000; Groves and Peytcheva, 2008; Groves, Singer, and Corning, 2000; Hill et al., 1997; Lahaut et al., 2002; Lynn, 1998; Voigt, Koepsell, and Daling, 2003). To date, no study has evaluated the validity of any of these measures. We wish to start filling this gap in the literature.

A valid measure is one that measures what it is supposed to measure, in order to "meaningfully capture the ideas contained in the corresponding concept" (Adcock and Collier, 2001; see also Carmines and Zeller, 1979). In this study we use three response propensity measures. First, we carry out a test of construct validity by assessing correlations between our measures (Carmines and Zeller, 1979). The results suggest questionable construct validity. We then introduce a second assessment of validity capitalizing on a behavior genetics approach.

The classical twin design (CTD) decomposes variance of the observed phenotypes into additive genetic, common, and unique environmental proportions of the variation (Medland and Hatemi, 2009). This is done through the comparison of how similar monozygotic (MZ) co-twins are to each other and how similar dizygotic (DZ) co-twins are to each other. We know that MZ twins share their genome; they are genetically identical. We also know that DZ twins share 50 percent of their genome just like all other siblings. Both MZ and DZ twins grow up in the same household and therefore share a sizable, but on average equal, portion of their environments. And every twin, as an individual, is also exposed to environmental stimuli that are unique to him or her.

We argue in this article that the CTD is suitable to test the validity of various measures of the same concept. More precisely, if the sources of variation (additive genetic, shared environment, unique environment) are markedly different for the three response propensity measures, we can conclude that these are not valid response propensity measures. The behavior genetic method

of validation proposed shows that some measures of response propensity appear to be driven by additive genetic factors, while others are driven by socialized environment, demonstrating little overlap between the driving forces of the response propensity indicators. Accordingly, this study casts doubt on the usefulness of the response propensity measures. To date only one study assessed heritability of a nonresponse propensity measure (Thompson et al., 2010). Our results, though corroborating their finding, also warn that different response propensity measures can yield vastly different results.

We begin this study with a review of the survey nonresponse literature both from the perspective of survey research and twin studies, and the review of validation techniques. We then empirically test the validity of available response propensity indicators for a large nationally representative household survey in the United States. We conclude the study with a discussion of the findings, which all point to the lack of validity of the classic response propensity measures.

Survey Nonresponse and Response Propensity

The failure to obtain responses from an individual who was originally sampled to be in a representative survey raises data quality issues in social science research (Allison, 2001; Dillman et al., 2002; Little and Rubin, 2002). Nonresponse bias can emerge in both survey and experimental methods (Little and Rubin, 2002). The data stemming from such research efforts are analyzed using statistical methods that assume random (or at least representative) samples of the populations of interest. But how realistic is this assumption in light of considerable nonresponse rates? If those who responded to the survey do not differ from the nonrespondents (Allison, 2001; Dillman et al., 2002; Little and Rubin, 2002; Tourangeau, Rips, and Rasinski, 2000), this problem threat becomes nil. However, we cannot reasonably assess this assumption, as we do not observe the data for those who are nonrespondents.

Consequently, it is unsurprising that the study of nonresponse has been a major issue for survey designs. This constant methodological concern has its roots in the fact that survey nonresponse, both item nonresponse (our inability to observe an answer to a single question) and unit nonresponse (our inability to observe the entire sampled individual), can bias statistical point estimates, inflate variances, and bias precision estimates (Dillman et al., 2002; Singer, 2006; Tourangeau, Groves, and Redline, 2010). Often, under realistic assumptions, the problem of item nonresponse can be efficiently handled using advanced statistical methods such as multiple imputation and direct estimates (e.g., full information maximum likelihood and Bayesian estimations (Allison, 2001; Little and Rubin, 2002)). This is because the presence of item nonresponse still presumes some observed information available on the sampled individual. If this information predicts the missingness of answers, estimates can remain unbiased.

This is rarely the case for unit nonresponse in surveys. The direction taken to overcome the obstacle of unit nonresponse is to put substantial effort into collecting information that nonrespondents are not disclosing in the context of a normal survey. Such information can include any observable information in a face-to-face situation, such as age and sex of person refusing to talk to the interviewer, visible details of the dwelling, the neighborhood, and so forth. In a non face-to-face interview situation information is less readily available, although the form of nonresponse and any information in the sampling frame used can be considered (such as zip code possibly matched with precinct information or property values for the area) (Groves and Couper, 1998).

The most popular method to analyze response propensity is through the direct analysis of nonrespondents. Unfortunately, this method only gives us, at best, a small amount of contextual information about who these nonrespondents are (Brehm, 1993; Campanelli, Sturgis, and Purdon, 1997; Groves, 2006; Groves and Couper, 1998; Heiskanen and Laaksonen, 1995; Olson, 2006).

Facing this difficulty, survey researchers have turned to indirect indicators of response propensity to assess the causes of nonresponse and nonresponse bias (Bergman, Hanve, and Rapp, 1978; Brehm, 1993; Campanelli, Sturgis, and Purdon, 1997; Goyder, 1986; Groves, 2006; Groves and Couper, 1996, 1998; Heiskanen and Laaksonen, 1995; Olson, 2006; Rao, 1983; Sarndal and Swensson, 1987; Smith, 1984; Singh, 1983). Many of these attempt to measure the response propensity for each actual respondent (Groves, 2006; Groves and Couper, 1996, 1998; Olson, 2006; Rao, 1983; Sarndal and Swensson, 1987; Singh, 1983). Having nonresponse proxy measures we can see if a correlation exists between the traits of interest and the likelihood of nonresponse to better understand the bias associated with nonresponse (Groves and Couper, 1998; Olson, 2006; Smith, 1984). Even though a myriad of response propensity measures exist in the literature (Currivan and Carley-Baxter, 2006; Etter and Perneger, 1997; Gmel, 2000; Groves and Peytcheva, 2008; Groves, Singer, and Corning, 2000; Hill et al., 1997; Lahaut et al., 2002; Lynn, 1998; Voigt, Koepsell, and Daling, 2003), surprisingly enough, to date, no study has evaluated the validity of any of these measures. Our first contribution lies in starting to fill in this lacuna.

Our study utilizes one new and one common method of validation to assess the quality of response propensity measures. Construct validation searched for the presence of a latent continuous response propensity trait underlying the indicators (Carmines and Zeller, 1979). Second, we propose that genetic analysis using a twin sample can further aid our validation attempts by focusing on the underlying sources of variation for the proposed indicators. We present evidence through both methods of validation that the indicators cannot be used as a general measure of response propensity. In the next section we introduce the response propensity measures to be analyzed.

Dealing with Survey Nonresponse Bias

In order to tackle this problem, multiple approaches have been proposed: collecting additional information on nonrespondents based on the sampling frame with useful information available for everyone who could end up in the sample[1] (Groves and Couper, 1998; Heiskanen and Laaksonen, 1995); studying reluctant respondents (Groves and Couper, 1998; Olson, 2006; Smith, 1984); collecting additional observational data about households of nonrespondents often used in conjunction with information derived from the sampling frame (Brehm, 1993; Campanelli, Sturgis, and Purdon, 1997; Groves and Couper, 1998); recruiting and assessing nonrespondents of a prior survey (Rao, 1983; Sarndal and Swensson, 1987; Singh, 1983); designing special surveys about survey participation, which attempt to ask respondents about their past experiences with surveys in order to estimate the probability of acceptance and refusal (Goyder, 1986); or experimental designs involving the alternation of survey design (Groves and Couper, 1998).

The approach we investigate further in this study involves the study of nonrespondents in a panel setting using the characteristics of those who responded to the initial survey request but dropped out in subsequent follow-up efforts (Groves and Couper, 1998). Using the National Survey of Midlife Development in the United States (MIDUS) we compile measures of response propensity through the assessment of refusals to follow-up survey efforts.

Needless to say, all of these methods face major obstacles. No contextual information can totally compensate for the lack of response; using contextual data clearly gives different estimates but no evidence was found that they are more accurate (Johnson et al., 2006; Olson, 2006). Alternative methods of data collection (e.g., extensive study of reluctant responses, offering financial incentives, extensive study of subsamples) do not successfully overcome the problem of biased results caused by the fact that those who still do not respond are essentially different from the rest of the population (Groves and Couper, 1998; Groves et al., 2006; Singer, 2002). While methods are available to correct for nonresponse bias (Abraham, Maitland, and Bianchi, 2006; Groves, 2006; Groves et al., 2006; Groves and Peytcheva, 2008; Singer, 2006; Wagner, 2010), these methods depend heavily on the availability of relevant information to correct for (Shadish, Clark, and Steiner, 2008). Unfortunately in most survey contexts it is unlikely that such relevant information can be collected about the nonrespondents.

As discussed above, a wide range of approaches have been employed to gain some leverage on the direction and magnitude of nonresponse bias. These methods either provide very little information on nonrespondents or make assumptions that some respondents are much like nonrespondents. The

[1]A broader operationalization of the nonresponse construct could also include noncontacts, those who could not be located or reached by phone (Abraham et al., 2006; Groves, 2006; Groves and Couper, 1998; Johnson et al., 2006; Olson, 2006).

appeal of the latter approach is twofold. The survey did observe all data on these respondents with low response propensity so they can be compared directly to respondents with high response propensity on all characteristics. Second, it makes sense that people who exhibit reluctance should be similar to respondents who do not respond at all. To date, the authors know of no study that tested, in any form, if the common indicators of response propensity have validity. We seize the opportunity to test the validity of the available indicators in our data set and demonstrate how behavior genetic analysis using twin data can be used for validation analysis.

Validity of Nonresponse Indicators

A valid measure is one that measures what it is supposed to measure, in order to "meaningfully capture the ideas contained in the corresponding concept" (Adcock and Collier, 2001). When a construct is difficult or expensive to measure directly, but indirect measures are available, the best method of validation is a comparison. This approach is common when researchers attempt to develop less expensive or easier to use measures of an already established measure of the construct. In psychology, it is not uncommon to find reduced psychometric scales of various constructs that claim to be valid measures of something that has been measured with a longer inventory of questions in the past.

Finding the most appropriate "golden standard" direct measure to compare to is straightforward in some cases and impossible in others. The golden standard of nonresponse is easy to identify, it is nonresponse itself. The problem is that by definition we are not measuring anything self-reported on nonrespondents, and therefore response propensity measures cannot be observed for the true nonrespondents. For this reason, we are ruling out this type of validation.

On the other hand, validity could be assessed through two different approaches. Let us assume that our proxy of response propensity measures an unobservable (latent) underlying construct, a continuum that decides the likelihood of response for a person and a threshold that flips the scale between the decision to respond or not.[2] If we have multiple good measures that indirectly proxy this underlying response propensity scale, it is reasonable to assume that they are intercorrelated. Assessing these intercorrelations is the most basic test of construct validity (Carmines and Zeller, 1979). This will be the first method of validation presented for the available measures of response propensity.

Second, in line with traditional convergent and discriminant methods of validation, we argue that behavior genetics models utilizing twin data also have the ability to validate through the identification of the causes of variation in a trait. Assuming the indicators tap the same underlying construct, if different forces drive variance in one indicator than the rest, that indicator cannot be

[2]The most prominent theory of nonresponse, Groves, Singer, and Corning's (2000) leverage-salience theory of survey participation, uses the same scale analogy.

tapping the same variance as the others. We describe the behavior genetic analysis in the next section, pointing out how it helps assessing claims of validity or invalidity for our measures.

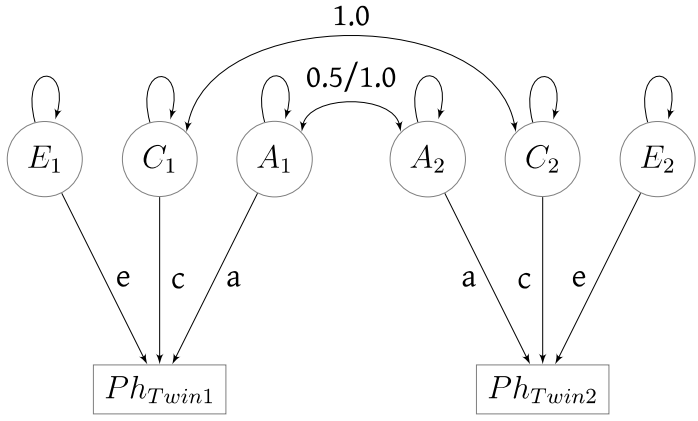
CTD and Behavior Genetics Analysis

Twin studies can contribute to our understanding of a construct as much as they can attribute the variance of the trait studied to different sources. The CTD decomposes variance into additive genetic, common, and unique environmental proportions of the variation (Medland and Hatemi, 2009). This is done through the comparison of how similar MZ co-twins are to each other and how similar DZ co-twins are to each other. We know that MZ twins share their genome; they are genetically identical. We also know that DZ twins share 50 percent of their genome just like all other siblings. Both MZ and DZ twins grow up in the same household and therefore share a sizable, but on average equal, portion of their environments. And every twin, as an individual, is also exposed to environmental stimuli that are unique to him or her.

Considering all these affects together we can draw a two-group structural equation model where three "latent" sources of influence impact the variation in the studied trait. These three sources of influence collect all additive genetic components (A), all common environmental components (C), and unique environmental components (E). As shown in Figure 1, the latent additive genetic influence is perfectly correlated (as denoted by the curved arrow) across MZ twins who are genetically identical but is only correlated by 0.5 for DZ twins

FIGURE 1

ACE Model



who share only half of their genome on average. This is the only difference between the two groups of the model. The common environment is common independent of zygosity and the unique environment is not correlated across the individuals.

If the traits are valid measures of a certain construct of interest beyond their intercorrelations, we can also expect that the same sources of variation will drive variance in the twin model. However, this is only a necessary but not a sufficient condition of validity. If different sources (i.e., genetic, common environment, and/or unique environment) are driving the variation for the different indicators, that would mean that they are manifestations of very different underlying traits. Hence, they are not valid measurements of a unidimensional underlying construct.

Data

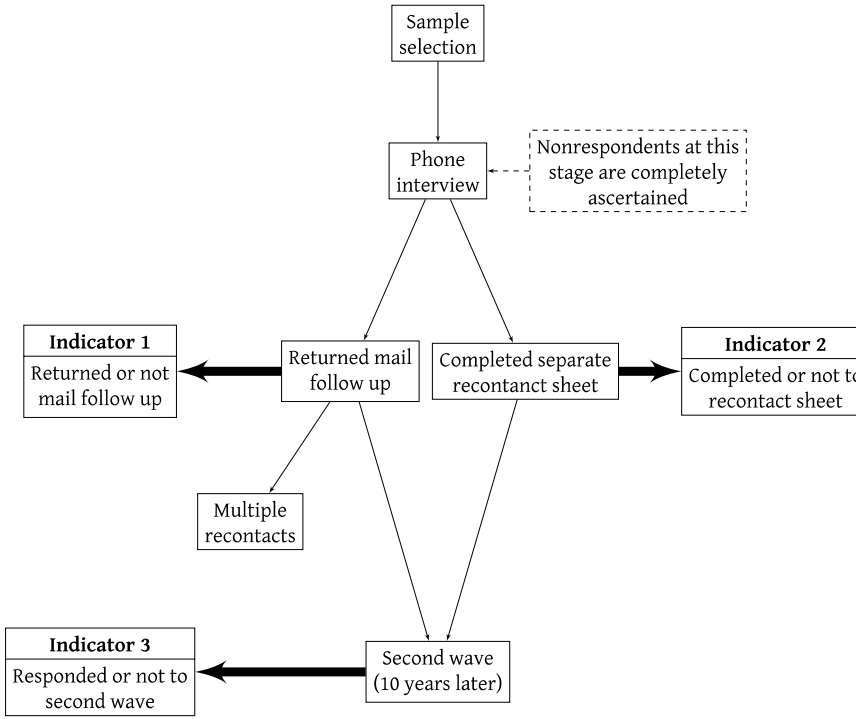
To test the validity of response propensity we use the the National Survey of MIDUS. MIDUS is a large, representative multiwave survey of people between the ages of 25 and 74 with a moderate oversample of the older cohort (MIDUS Methodology, 1995–1996). As it is clear from the age group included, this survey goes well beyond the understanding of midlife and has a somewhat misleading title. Due to the complexity of the survey design, we also present the various survey processes and response propensity indicators derived from it in Figure 2.

The first wave of MIDUS was fielded in 1995–1996 using random digit dialing (RDD). The contacted individual was informed that the study was about health and well-being in the middle years of life and was conducted by the Harvard Medical School. Once within-household selection of the respondent occurred, no other respondent was chosen if the person was not available for an interview. Rather, multiple recontacts were attempted. Response rate for the phone interview was 70 percent.

Following the first wave telephone interview, a paper and pencil self-administered questionnaire (SAQ) was sent to the respondents. Respondents were reminded several times both through the mail and over the phone if the questionnaire was not returned within a set time period (MIDUS Technical Report, 1995–1996). In the end, 86.8 percent of the people who responded to the phone questionnaire filled out and returned the SAQ. Failure to respond to the SAQ by those who responded to the phone survey is one of the measures of response propensity used in our analysis. This measure is widely used in nonresponse studies (Gmel, 2000; Groves, Singer, and Corning, 2000; Hill et al., 1997; Lahaut et al., 2002). Note that anyone who failed to respond to the paper and pencil questionnaire lost all chance of being in the second-wave follow-up. For this reason, we cannot intercorrelate this measure with subsequent measures for construct validation purposes. We refer to this process through which people are excluded as ascertainment throughout the article.

FIGURE 2

Survey Processes and Response Propensity Indicators for MIDUS



A second wave of data collection for MIDUS followed 10 years after the first wave (MIDUS Sample Descriptions 2004–2006). The second wave was a panel design where the respondents of the first wave were recontacted. A total of $60 of monetary incentive was provided for people who completed all steps of the second-wave procedure. Longitudinal retention rate was 65 percent for the main sample and 78 percent for the twin sample. Web Appendix, Table 1 contains a detailed breakdown of nonresponse at various stages of the survey. The MIDUS help desk provided the data to classify all individuals who did not participate in wave two into a refusal category, or other categories (such as deceased, unconfirmed deceased, unable to interview, and nonworking numbers) we coded as missing data. Hence, the indicator derived is those who completed both waves versus those who explicitly refused to complete the second wave. This indicator is in line with how nonresponse is analyzed throughout the literature (Currivan and Carley-Baxter, 2006; Lynn, 1998).

Building on the literature, we use additional variables as measures of response propensity. In addition to the mail questionnaire and in anticipation of a second wave of data collection 10 years in the future, the first wave of the

MIDUS mail questionnaire contained a recontact information sheet designed to ease the burden of tracking down the respondent by the research team. This sheet asked for information that could be perceived as highly personal and sensitive, including contact information for multiple close friends and/or relatives, and disclosure of the respondent's Social Security number. A variable was included in the MIDUS data file pertaining to the return of (or the failure to return) this recontact information. The refusal to provide this recontact information indicates the unwillingness to participate in a future survey and therefore is a good indicator of response propensity (Groves and Peytcheva, 2008). Additionally, not completing this distinct sheet of information, clearly independent of the questionnaire, becomes a direct measure of nonresponse. All in all, we consider that the failure to provide recontact information is at least as good an indicator of reluctant response or response propensity as the refusal to complete the mail follow up or the second wave.

In addition to the main (representative) MIDUS sample, multiple oversamples were also collected. The interesting oversample for the purposes of this study is the twin sample. To recruit a twin sample, 50,000 RDD calls were made inquiring if the respondent had a relative who was a twin. These twins were then contacted. Unfortunately, the research team appears to have recruited only twins where both twins signaled willingness to respond to the main questionnaire during a twin-only screening questionnaire. This, and the lack of information on zygosity of nonrespondents, eliminates the possibility of directly assessing the heritability of nonresponse or correcting for volunteer bias (Neale and Eaves, 1993). MIDUS reports a 60 percent response rate for the twin sample. This might appear low, as anecdotal evidence from twin studies suggests that twins understand their uniqueness and are often excited to participate in research. This figure is compounded both by noncompliance in the screening calls through the failure to acquire contact information and the research procedure where noncompliance by one of the twins in a pair led to the exclusion of both twins.

After the exclusion of the twin pairs with no available data on any of the indicators of interest, the exclusion of co-twins in families with multiple twin pairs, and the exclusion of unknown zygosity and different sex DZ twins, the procedure yielded 359 families with MZ and 337 families with DZ twin pairs. The validation that did not require a genetically informative sample utilized the general sample in MIDUS to take advantage of the larger sample ($n = 3{,}487$) and to avoid the violation of the independence assumption of the modeling techniques used.[3] Finally, the twin sample is used only to supplement the substantive findings using the main sample. This is important, as the generalizability of the twin study to nontwins cannot be taken for granted in light of the results.

The wealth of indicators provided by MIDUS leaves us with three different measures of response propensity that are in line with both the theories and

[3]Each individual of a twin pair is not independent of the other.

practices of response propensity assessment. The availability of such information is extremely rare in studies not specifically designed to assess survey nonresponse. The multiple indicators allow for validation of these measures, which generally are accepted at face value. We are not aware of any studies that have compared multiple response propensity indicators in an attempt to validate them. Also, to date, no efforts have been made to compile response propensity measures for twin analysis. The availability of twin data allows for the introduction of twin studies as an approach to supplement traditional methods of validation for competing measures. Before turning our attention to our analysis, we briefly discuss the problem of volunteer bias for the twin studies.

Volunteer Bias in Twin Research

Independent of the efforts put forth by survey researchers to understand nonresponse bias, twin researchers also noticed that the volunteer samples used in twin studies suffer from a similar phenomenon. The problem has been dubbed selection bias or volunteer bias in the behavior genetic literature, as the researcher is only able to observe people who self-select (volunteer) to be part of the study. Lykken, Tellegen, and De Rubeis (1978) established the two-thirds rule that a typical twin sample will consist of two-thirds female and two-thirds identical (MZ) twins (vs. a third fraternal or DZ twins), possibly introducing differential nonresponse bias for the different groups included in the analysis. This is of particular concern as it threatens to bias heritability estimates, which use volunteer twin samples.

Twin researchers were also predominantly concerned about nonresponse when it is correlated with the dependent variable. This would occur if nonrespondents would be systematically different on the dependent variable from respondents. Martin and Wilson (1982) distinguish between two models of selection bias. Hard selection is where participants above a certain fixed threshold on the dependent variable do not participate, whereas soft selection is probabilistic where the probability of participation diminishes on a range (as opposed to the fixed threshold). Building on this study, Neale et al. (1989) provide a model where the probability of selection is a function of the cumulative normal distribution of the dependent variable and show that the "softer" the selection criteria, the lower the bias.

The jury is still out on the effects of volunteer bias in twin studies. Martin and Wilson suggest that co-twin correlations are biased downwards (1982). As explained earlier in this article in detail, twin models capitalize on the comparison of co-twin correlations for identical (MZ) and fraternal (DZ) twins to assess heritability. The greater the difference between MZ and DZ co-twin correlations, the higher the heritability. Theoretically speaking, if DZ twins suffer from higher self-selection bias, their co-twin correlations are biased downward. This increases the gap between the MZ and DZ co-twin

correlations; correlation heritability will be overestimated. On an empirical level, Tambs et al. (1989) point out that in the context of studying IQ, volunteer bias is not present. In a more extensive assessment of reluctant respondents Vink et al. (2004) found that "even for studies with moderate response rates, data collected on health, personality and lifestyle are relatively unbiased." Unfortunately, to successfully utilize this correction, zygosity must be known for all sampled individuals, and information has to be collected on twins whose co-twins did not respond to the survey. Interestingly, while volunteer bias and reluctant respondents have been considered by twin researchers, no study to date has assessed survey nonresponse directly as a phenotype. We will also not assess nonresponse directly, but if response propensity measures can effectively proxy nonresponse, the behavior genetic step in our validation will serve as heritability assessment of survey nonresponse. We note, however, that due to the cited issues with volunteer bias in twin samples, none of our results can be interpreted directly as the heritability of nonresponse. At best it is suggestive of this. Finally, for the purposes of validation, note that heritability estimates could be biased upward so only vast differences (which we do find) in the proportion of variance driven by A, C, and E can be interpreted as evidence against validity of the response propensity indicators driving a unidimensional construct.

Descriptive Statistics

Comparing the twin sample to the nontwin sample, it becomes clear that twins score significantly and substantively higher on all response propensity measures (see Web Appendix, Table 2 for more details). Unfortunately, these differences mean that the generalizability of the twin study results to the nontwin population may be problematic. On the other hand, the goal of the twin study is not broad generalization of the results to the general population, but the better understanding of the underlying causes of variation for nonresponse. While the difference between twins and nontwins is theoretically understandable, it could be argued that the sources of variation for the nonresponse indicators might be somewhat different for nontwins. We still argue that, as a supplement to the other approaches to validation, the results could be suggestive of the causes underlying the indicators, especially in light of the clear results presented below. In the twin research segment we also need to be concerned about the comparability of MZ and DZ twins on the studied trait. If there is a proportion difference between MZ and DZ twins in a heritability study, the results could be biased. This is of particular importance in light of the two-thirds rule described above (Lykken, Tellegen, and De Rubeis, 1978). Comparing self-reported MZ and DZ twins, response propensity score differences are in the expected direction, but in this sample none of these appear substantively different in magnitude or statistical significance.

TABLE 1

Age and Sex Corrected Bivariate Correlation (Probit Link Function Used for Dichotomous Variables)

Correlations	Did Not Return Recontact Info	Did Not Complete Wave Two
Did not return recontact info	1	
Did not complete wave two	**0.285**	1
Controls		
Age (standardized)	**−0.101**	**−0.104**
Sex (female)	0.048	−0.082

$p < 0.001$ are bolded. Correlations are presented after regressing out the effects of age and sex.

Results

Construct Validity

Table 1 displays the correlation between the return of the recontact information sheet and wave two refusal. The presented correlation is a partial correlation with the effect of age and sex removed. We also present the impact of age and sex on the response propensity measures as derived from the partial correlations.[4] Response to the SAQ after the phone interview is not included in this analysis since for all cases of wave one SAQ nonresponse the rest of the response propensity measures are missing. All steps of validation were conducted using Mplus 5 statistical software (Muthén and Muthén, 2008). Dichotomous measures were defined as such within Mplus, using a probit link function.

More interestingly, while the correlation is statistically significant ($p < 0.001$), it is not strong. One contributor to the lack of strength is the dichotomous nature of the indicators that generally bias correlations downward, but even with this property in mind the correlations are modest at best. If the same underlying latent response propensity construct is the primary driver of both these measures, we would expect the correlation to be much higher. These results question the validity of these nonresponse propensity indicators on the grounds of construct validity. Nevertheless, how weak or strong is a correlation of roughly 0.3 is not too clear. Also, what drives this correlation is not something a construct validity test could answer, and thus we turn to behavior genetic analysis.

[4]The impact of age and sex on the response propensity measures can be interpreted as regression coefficients.

Validation Through Genetic Analysis

As discussed, twin studies capitalize on knowledge about MZ and DZ twin differences and the variance of traits (or phenotypes) is decomposed into three sources of influence that collect all additive genetic components (A), all common environmental components (C), and unique environmental components (E).

Note that for this stage of validation we can add a third measure of response propensity: return of SAQ, as those who did not return this questionnaire were not included in the second wave (hence could not be correlated). Considering that the three outcomes studied in this article are dichotomous, and that the continuous one is not measured for twins, the above-described model needs to be adjusted to handle a dichotomous outcome. This is done through the same methodology that extends a linear regression into a probit regression. We assume that the trait has an underlying normal distribution (with a mean of 0 and a standard deviation of 1) and a threshold that separates the distribution into the presence and the absence of the condition. The location of this threshold and where the individual falls on the underlying normal distribution jointly will determine the presence or the absence of the outcome for each individual. Calculations were done on the raw data with the software default estimator, and thresholds are age and sex corrected (McGue and Bouchard, 1984). Summary tables are also available in the Web Appendix, Table 3.

To reiterate, if we find with the twin model relatively uniform sources of variation, we can say that a necessary condition of validity is fulfilled. However, if the results suggest that different sources are driving the variation for the different indicators, this would mean that these are invalid indicators. And looking at the results, the latter scenario appears to be the case. Full ACE model results with bootstrapped confidence intervals are displayed in Table 2. In fact, follow-up contact refusals appear to be influenced by additive genetic factors (67.5 percent for SAQ nonresponse and 61.7 percent for wave two nonresponse, $p < 0.05$ for both), with no considerable common environmental influences. Nonresponse on the recontact information sheet is predominantly influenced by common environmental effects (48.1 percent, $p < 0.05$) and no additive genetic effects.

A traditional extension of this model would be a multivariate Cholesky decomposition of the correlation between the indicators, but in this case there is no point in estimating this model. The SAQ completion ascertains the other two variables so intercorrelation is not feasible. Additionally, since wave two completion is driven by AE and response to the recontact information inquiry is driven by CE effects, the source of covariation can only be through (E), the unique environmental effects.[5]

[5] Despite the obvious source of covariation, an ACE Cholesky decomposition was attempted but the model did not converge. Given the inclusion of zero variance effects, the dichotomous outcome and the small sample size make this model failure unsurprising.

TABLE 2

Genetic Analysis Decomposing Variance into (A) Additive Genetic, (C) Common Environmental, and (E) Unique Environmental Effects with Corresponding Bootstrapped 95 Percent Confidence Intervals

	A	2.50%	97.50%	C	2.50%	97.50%	E	2.50%	97.50%
Did not complete SAQ	0.675	0.192	0.882	0	0	0.524	0.235	0.147	0.641
Did not return recontact information	0	0	0.644	0.481	0.182	0.7	0.529	0.34	0.798
Did not complete wave two	0.617	0.083	0.838	0.062	0	0.577	0.321	0.161	0.534

Limitations

Survey researchers have long been forced to come to terms with the reality of nonresponse. In the absence of sufficient auxiliary data, we simply cannot observe any information from people who do not respond to our surveys. While this might not introduce bias in all instances of survey data analysis, it is clearly a problem when the exact information we are interested in is the characteristics of the nonrespondents. We can place nonresponse on a continuum and find indicators that best describe this continuum, but there will always be that person who simply does not respond and therefore does not provide us with any information. For this reason, all such assessments of response propensity are ascertained at the point where the observed units become the missing units, and this study is no exception.

This issue is aggravated in the genetic analysis, where the thresholds of response propensity ascertainment could be different for MZ and DZ twins. This is a problem as response propensity, by definition, is correlated with nonresponse so our dependent variable is ascertained by survey nonresponse. We know that MZ twins respond more to survey requests than DZ twins (Lykken, Tellegen, and De Rubeis, 1978). This is suggested to decrease DZ co-twin correlations, inflating the heritability estimates (Martin and Wilson, 1982). While we have provided a wealth of evidence that traditional nonresponse indicators are not necessarily valid, all we can do is turn to these exact indicators to assess the possibility of MZ and DZ differences in response propensity. Based on the indicators at hand, the difference appears to be minimal and not significantly different from zero in this sample. And even if a differential nonresponse bias of MZ and DZ respondents produces systematic inflation of heritability, this clearly could not account for the results at hand, where two of the nonresponse indicators exhibited high heritability while the other produced no heritability with high shared environmental effects.

A bigger problem for assessing the heritability of the response propensity indicators is the vast and significant difference between twins and nontwins. Twins, independent of zygosity, have a larger propensity to respond than nontwins. This raises concerns about generalizability to the nontwin population (Medland and Hatemi, 2009). If twinness and the recognition of scientific importance are the sole source that drives the differences between twins and nontwins, it is, at least theoretically, likely that the sources of variance in nonresponse propensity are not different for nontwins—just the magnitude. Additionally, the difference is small and even the introduction of additional sources of variance could only add modest effects unaccounted for. Yet, the results should be considered with caution and only as a supplement to the previous two methods of validation.

In addition to the possible problems discussed, the results of twin studies rest on a broader set of assumptions. This analysis is not immune to these, either. (For an extensive discussion, see Alford, Funk, and Hibbing, 2008a, 2008b; Hannagan and Hatemi, 2008; Medland and Hatemi, 2009; Littvay 2012.)

Finally, it is important to note that response propensity is a highly context-dependent measure. Someone's willingness to respond to a survey conducted by a prestigious university on general health and life style issues, even if it includes several political questions, might be completely different from someone's willingness to respond to a political survey right before or right after an election. Unfortunately, we did not have multiple surveys with readily available response propensity indicators and oversamples of twins at our disposal to assess how the variation of the specific survey context influences people's response propensity or nonresponse levels. Still, we argue that the results presented are a good first step in understanding the validity of response propensity measures. Political psychologists, on the other hand, have ignored the impact of psychological individual differences on the survey process. We cannot expect survey researchers to invest in the understanding of these processes in the specific context of political surveys.

Discussion

A quick review of the social science reveals that most individual-level data we use come from surveys. For this reason, we should not only care about the specific responses people put down to survey questions, but the transactions that overarch the survey process as well. These transactions manifest behaviors that can be better understood through the lens of psychology and should be studied to minimize bias in the data and the cost of data-collection efforts. Emerging research in political science exploring biological processes of behavior gives us a unique opportunity to view the survey process in a different light (Hatemi et al., 2011; Hatemi and McDermott, 2012; Littvay et al., 2011). There are a large number of prescriptions in the literature that increase response rates, but none of them produce flawless results. Increasing response rates often comes with increasing other sources of error (Olson, 2006). Nonresponse is surely a concern for all survey researchers, but it is not necessarily a problem as long as it is not correlated with the variables of interest. Since we can never test this, scholars have attempted to produce theories that view response propensity as a continuum and invent corresponding measures of such a continuum. The traditional and less orthodox tests of validity presented in this article question the value of such measures.

The data available allowed for the exploration of three indicators of response propensity. Some of these measures were direct ascertainments of the others, but the ones that were observed simultaneously intercorrelated with each other only modestly. If these indicators do indeed measure some underlying latent response propensity continuum, they do not do a very good job of it. We introduced a new test of validity using twin studies. Using this test we were able to show that the different indicators of nonresponse are driven by vastly different effects. While nonresponse in a follow-up or panel setting appears to be highly heritable (also found in Thompson et al., 2010), nonresponse to recontact information inquiry seeking contact information of close friends

and/or relatives and the respondent's Social Security number is predominantly driven by shared environmental influences.

This result is not overly surprising. Even though response propensity indicators are generally used based on ad-hoc availability, the results suggest that these indicators are heavily influenced by context. The privacy considerations concerning giving even a reliable organization such as the Harvard School of Medicine very sensitive information are also very different from giving someone the time of day to discuss your health and lifestyle. Such measures of response propensity cannot be used without close considerations of these properties unique to the specific operationalization.

Conclusion

Beyond the substantive findings, we demonstrated how genetic variance decomposition with a twin design could contribute to our understanding of validity. The approach is a natural extension of convergent validity that assesses if various predictors of the construct indicators have consistent predictive power. To the best of the authors' knowledge nobody, to date, has proposed this type of analysis to test validity. Twin studies could become a valuable source of validity assessment, as they can point to lack of validity without strong theoretical expectations needed to test convergent validity the conventional way with theoretically sound predictors. We caution anyone who wishes to use twin studies as a method of validation, as the method can only provide evidence against validity. A finding such as all indicators of a construct are driven by common environmental sources is insufficient evidence to claim that the same predictors are responsible for the variation.

Turning to our specific behavior genetic findings, we can only speculate about the specific sources of variation as derived by the genetic analysis. Heritability, common, and unique environment are admittedly extremely vague predictors. Theoretically sound speculation about specific genetic mechanisms is definitely beyond the scope of this article, but the place to look for these would be the known genetic determinants of the psychological correlates that significantly covary with the nonresponse indicators. Multivariate extension of the ACE model could provide information about the source of this covariation. If it is genetic, specific genes could be considered as theoretically sound predictors. More interestingly, it could be argued that the privacy concern associated with giving out sensitive information is a new age problem. In most of human history identity theft was not an issue, nor did people feel the need to protect close friends and relatives from the unwanted inquiries of strangers or organizations that intruded in their lives (such as pollsters and researchers). For this reason, it is not surprising to find that this trait is heavily influenced by familial socialization and not by (genetic) factors that evolved over a large number of generations. On the other hand, decisions to talk about one's self or to give somebody the time of the day are age-old considerations

people faced even in prehistoric times. A survey request might be a 20th- and 21st-century phenomena, but the nature of the transaction might not be an unfamiliar event from the distant perspective of human history.

REFERENCES

Abraham, K. G., A. Maitland, and S. M. Bianchi. 2006. "Nonresponse in the American Time Survey. Who Is Missing from the Data and How Much Does it Matter?" *Public Opinion Quarterly* 70:676–703.

Adcock, R., and D. Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95:529–46.

Alford, J. R., C. L. Funk, and J. R. Hibbing. 2008a. "Beyond Liberals and Conservatives to Political Genotypes and Phenotypes." *Perspectives on Politics* 6:321–28.

———. 2008b. "Twin Studies, Molecular Genetics, Politics, and Tolerance: A Response to Beckwith and Morris." *Perspectives on Politics* 6:793–97.

Allison P. D. 2001. *Missing Data*. *Sage University Press Series: Quantitative Applications in the Social Sciences* (Vol. 136). Thousand Oaks: Sage.

Bergman, L., R. Hanve, and J. Rapp. 1978. "Why Do Some People Refuse to Participate in Interview Surveys." *Statistik Tidskrift* 4:341–56.

Brehm, J. 1993. *The Phantom Respondents; Opinion Surveys and Political Representation*. Ann Arbor: University of Michigan Press.

Campanelli, P., P. Sturgis, and S. Purdon. 1997. *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*. London: Survey Methods Centre at SCPR.

Carmines, E. G., and Zeller, R. A. 1979. *Reliability and Validity Assessment*. *Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–017*. Newbury Park, CA: Sage.

Currivan, D., and L. Carley-Baxter. 2006. "Nonresponse Bias in a Longitudinal Follow-up to Random-Digit Dial Survey." Paper presented at the Method of Longitudinal Surveys Conference. Colchester, UK: University of Essex. Available at <https://www.iser.essex.ac.uk/files/survey/ulsc/methodological-research/mols-2006/scientific-social-programme/papers/Currivan.pdf.>

Dillman, D. A., J. L. Eltige, R. M. Groves, and R. J. A. Little. 2002. "Survey Nonresponse in Design, Data Collection and Analysis." Pp. 3–26 in D. A. Dillman, J. L. Eltige, R. M. Groves, and R. J. A. Little, eds., *Survey Nonresponse*. New York: John Wiley & Sons, Inc.

Etter, J. F., and T. V. Perneger. 1997. "Analysis of Nonresponse Bias in a Mailed Health Survey." *Journal of Clinical Epidemiology* 50:1123–28.

Gmel, G. 2000. "The Effect of Mode of Data Collection and of Nonresponse on Reported Alcohol Consumption: A Split-Sample Study in Switzerland." *Addiction* 95:123–34.

Goyder, J. C. 1986. "Surveys on Surveys: Limitations and Potentialities." *Public Opinion Quarterly* 50:27–41.

Groves, R. M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Survey." *Public Opinion Quarterly* 70:646–75.

Groves, R. M., and M. P. Couper. 1996. "Contact-Level Influences on Cooperation in Face-to-Face Surveys." *Journal of Official Statistics* 12(1):63–83.

———. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., M. P. Couper, S. Presser, E. Singer, R. Tourangeau, and G. Piani Acosta. 2006. "Experiments in Producing Nonresponse Error." *Public Opinion Quarterly* 70:720–36.

Groves, R. M., and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72:167–89.

Groves, R. M., E. Singer, and A. D. Corning. 2000. "Leverage-Saliency Theory of Survey Participation: Description and an Illustration." *Public Opinion Quarterly* 64:229–308.

Hatemi, P. K., C. T. Dawes, A. Frost-Keller, J. E. Settle, and B. Verhulst. 2011. "Integrating Social Science and Genetics: News from the Political Front." *Biodemography and Social Biology* 57:67–87.

Hatemi, P. K., and R. McDermott. 2012. "The Genetics of Politics: Discovery, Challenges, and Progress." *Trends in Genetics* 28:525–33.

Heiskanen, M., and S. Laaksonen. 1995. "Non-Response at the SLC and the Depression Trap." Paper presented at the Sixth International Workshop on Household Survey Nonresponse. Helsinki, Finland: Statistic Finland.

Hill, A., J. Roberts, P. Ewings, and D. Gunnell. 1997. "Nonresponse Bias in a Lifestyle Survey." *Journal of Public Health Medicine* 19:203–07.

Johnson, T. P., Y. I. Cho, R. T. Campbell, and A. L. Holbrook. 2006. "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey." *Public Opinion Quarterly* 70:759–79.

Lahaut, V. M. H., H. A. M. Jansen, D. van de Mheen, and H. F. L. Garretsen. 2002. "Nonresponse in a Sample Survey on Alcohol Consumption." *Alcohol and Alcoholism* 37:256–60.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: John Wiley.

Littvay, L. 2012. "Do Heritability Estimates of Political Phenotypes Suffer from an Equal Environment Assumption Violation? Evidence from an Empirical Study." *Twin Research and Human Genetics* 15:6–14.

Littvay, L., P. T. Weith, and C. T. Dawes. 2011. "Sense of Control and Voting: A Genetically-Driven Relationship." *Social Science Quarterly* 92:1236–52.

Lykken, D. T., A. Tellegen, and R. De Rubeis. 1978. "Volunteer Bias in Twin Research: The Rule of Two-Thirds." *Social Biology* 25:1–9.

Lynn, P. 1998. "Data Collection Mode Effects on Responses to Attitudinal Questions." *Journal of Official Statistics* 14:1–14.

Martin, N. G., and S. R. Wilson. 1982. "Bias in Estimation of Heritability from Truncated Samples of Twins." *Behavior Genetics* 12:467–73.

McGue, M., and T. J. Bouchard. 1984. "Adjustment of Twin Data for the Effects of Age and Sex." *Behavior Genetics* 14:325–43.

Medland, S. E., and P. K. Hatemi. 2009. "Political Science, Biometric Theory, and Twin Studies: A Methodological Introduction." *Political Analysis* 17:191–214.

Muthén, L. K., and B. O. Muthén. 1998–2007. *Mplus User's Guide*, 5th ed. Los Angeles, CA: Muthén & Muthén.

National Survey of Midlife Development in the United States (MIDUS). Methodology. (1995–1996).

———. Technical Report. (1995–1996).

———. Sample Descriptions: MIDUS1 and MIDUS2, All Parts. (2004–2006).

Neale, M. C., and L. J. Eaves. 1993. "Estimating and Controlling for the Effects of Volunteer Bias with Pairs of Relatives." *Behavior Genetics* 23:271–77.

Neale, M. C., L. J. Eaves, K. S. Kendler, and J. K. Hewitt. 1989. "Bias in Correlations from Selected Samples of Relatives: The Effects of Soft Selection." *Behavior Genetics* 19:163–9.

Olson, K. M. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70:737–58.

Rao, P. S. R. S. 1983. "Nonresponse and Double Sampling: Randomization Approach." Pp. 97–105 in W. Madow, I. Olkin, and D. B. Rubin, eds., *Incomplete Data in Sample Surveys* (Vol. 2). New York: Academic Press.

Sarndal, C. E., and B. Swensson. 1987. "A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse." *International Statistical Review* 55:279–94.

Shadish, W. R., M. H. C. Clark, and P. M. Steiner. 2008. "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment." *Journal of the American Statistical Association* 103:1334–44.

Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In R. M. Groves, D. A. Dillman, J. L. Eltige, and S. Laaksonen, eds., *Survey Nonresponse*. New York: Wiley.

———. 2006. "Introduction. Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70:637–45.

Singh, B. 1983. "Nonresponse and Double Sampling: Bayesian Approach." Pp. 107–19 in W. Madow, I. Olkin, and D. B. Rubin, eds., *Incomplete Data in Sample Surveys* (Vol. 2). New York: Academic Press.

Smith, T. W. 1984. "Estimating Nonresponse Bias with Temporary Refusals." *Sociological Perspectives* 27:473–89.

Tambs, K., J. M. Sundet, P. Magnus, and K. Berg. 1989. "No Recruitment Bias for Questionnaire Data Related to IQ in Classical Twin Studies." *Personality and Individual Differences* 10:269–71.

Thompson, L. F., Z. Zhang, and R. D. Arvey. 2010. "Genetic Underpinnings of Survey Response." *Journal of Organizational Behavior* 32:395–412.

Tourangeau, R., R. M. Groves, and C. D. Redline. 2010. "Sensitive Topics and Reluctant Respondents Demonstrating a Link Between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74:1–22.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*, 1st ed. Cambridge: Cambridge University Press.

Vink, J. M., G. Willemsen, J. H. Stubbe, C. M. Middeldorp, R. S. Lighthart, K. D. Baas, H. J. Dirkzwager, E. J. Geus, and D. I. Boomsma. 2004. "Estimating Non-Response Bias in Family Studies: Application to Mental Health and Lifestyle." *European Journal of Epidemiology* 19:623–30.

Voigt, L. F., T. D. Koepsell, and J. R. Daling. 2003. "Characteristics of Telephone Survey Respondents According to Willingness to Participate." *American Journal of Epidemiology* 157:63–74.

Wagner, J. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74:1–21.