1-2013

Do non-response follow-ups improve or reduce data quality?: A review of the existing literature

Kristen Olson
University of Nebraska - Lincoln, kolson5@unl.edu

Published in *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176:1 (January 2013), pp. 129–145 ; doi: 10.1111/j.1467-985X.2012.01042.x

Submitted March 2011; revised January 2012; published online June 28, 2012.

Do non-response follow-ups improve or reduce data quality?: A review of the existing literature

Kristen Olson

University of Nebraska–Lincoln, 703 Oldfather Hall, Lincoln, NE 68588-0324, USA; email kolson5@unl.edu

Abstract

The paper systematically reviews existing literature on the relationship between the level of effort to recruit a sampled person and the measurement quality of survey data. Hypotheses proposed for this relationship are reviewed. Empirical findings for the relationship between level of effort as measured by paradata (the number of follow-up attempts, refusal conversion and time in the field) and question-specific item non-response rates, aggregate measures of item non-response rates, response accuracy and various measurement errors on attitudinal questions are examined through a qualitative review.

Keywords: item non-response, measurement error, non-response propensity, paradata, total survey error

1. Introduction

Under a total survey error framework (Biemer and Lyberg, 2003; Groves, 1989), decisions to reduce one source of survey error may have unintentional effects on other sources of error. Although no direct relationship exists between response rates and non-response bias (Groves and Peytcheva, 2008), it is not known whether efforts to increase response rates unintentionally lower the quality of data by increasing item non-response rates or measurement error. Respondents who require the most effort to recruit into the survey are hypothesized to provide lower quality responses (Cannell and Fowler, 1963). If this is so, field decisions to raise response rates such as additional follow-ups, converting refusals or extending the survey field period may also result in higher levels of measurement error in survey data.

In this paper, I review the existing empirical evidence on the relationship between level of effort as measured by paradata (Couper, 1998) about the recruitment process and quality of data as measured through item non-response rates, response accuracy, signed deviations, scale reliability, non-differentiation or variability of responses and attitudinal measurement error indicators including acquiescence, extreme and middle responses, and primacy or recency effects. First, four level-of-effort measures that divide the sample into "early" and "late" or "easy" and "difficult" respondents derived from paradata are described. Second, hypotheses that have been proposed in the existing literature to explain why a relationship may exist between level of effort derived from paradata and item non-response and/or measurement error are examined. Next, expected and empirical relationships between level of effort and quality of data are reviewed. Finally, the implications of these findings for practice are discussed.

2. Paradata about recruitment effort to reduce non-response rates

Information that is reported in paradata about the survey recruitment process is in the form of call records, including call attempts, refusal conversion indicators derived from call outcomes, the time spent in the field derived from the date and time of call attempts or a combination of these measures (see Kreuter *et al.* (2010) for a similar definition). These measures of levels of effort act as proxies for an individual's probability of participating in a survey, or their "response propensity," under a stochastic model for non-response (Bethlehem, 2002; Lessler and Kalsbeek, 1992; Oh and Scheuren, 1983). In general, higher levels of effort (e.g. more calls, converted refusal or a later interview date) are indirect measures of lower probabilities for survey participation. Both non-contact and refusal non-response lead to exertion of additional efforts (for example, both result in additional follow-up attempts to a sampled individual), but the types of efforts may differ (for example, refusal conversion occurs only for refusal non-response). Although these measures are discussed as mutually exclusive, there is clear overlap (for example, those who require additional follow-up attempts are interviewed later in the field period).

Survey organizations may engage in many other types of field effort to recruit sampled individuals into the sample pool, such as increasing incentives (e.g. Singer, 2002). Unfortunately, these various types of field operations are not consistently recorded in paradata (Couper, 1998; Chearo and Van Haitsma, 2010). This review focuses on the recruitment effort that is most consistently recorded in call records across organizations and across modes—that which can be constructed from the number of contact attempts, contact attempt outcomes and timing of the contact attempts.

3. Why might recruitment effort affect measurement quality?

Seven hypotheses, which are summarized in Table 1, exist for why efforts to recruit sampled individuals may affect item non-response and/or measurement error. These hypotheses include motivation, reaction against an attempt at persuasion, interest in the survey topic or sponsor, compositional differences, impressions of importance of the research, self-perceptual differences and changes in the survey design.

The most frequently posited hypothesis for why an association between level of effort and item non-response or measurement error may be observed is that the level of effort itself proxies for a sampled person's motivation to participate in the study (Cannell and Fowler, 1963; Dahlhamer *et al.*, 2006; Friedman *et al.*, 2004; Kaminska *et al.*, 2006, 2010; McDermott and Tan, 2008; Miller and Wedeking, 2006; Stoop, 2005; Tripplett *et al.*, 1996), which is sometimes referred to as a "latent co-operation continuum" (Bollinger and David, 2001). Under the motivation hypothesis, "general co-operativeness" causes participation and high quality survey answers (Bollinger and David, 2001). Individuals who lack motivation are simultaneously less likely to participate in a survey and, when eventually convinced to participate, have lower commitment to the respondent task (e.g. Cannell and Fowler, 1963), leading to short cuts when answering questions or "satisficing" (Krosnick, 2002; Krosnick and Alwin, 1987; see also Kaminska *et al.*, 2010; Sakshaug *et al.*, 2010; Tourangeau *et al.*, 2009). The motivational model applies most directly to refusal to participate with the survey request (refusal non-response) rather than difficulty in making contact with a sampled person (non-contact non-response) (Stoop, 2005). Additionally, the motivational model applies most directly to data quality issues that have a motivational component, such as item non-response (Beatty and Herrmann, 2002; Krosnick, 2002), especially on burdensome or difficult questions (Tripplett *et al.*, 1996).

Table 1. Summary of proposed hypotheses in empirical literature for the relationship between level of effort and measurement error

Hypothesis	*Type of survey participation*	*Items most affected*	*How measured*	*Limitations*
General motivation	Co-operation	Burdensome, difficult	Level of effort paradata used as motivation	Motivation is rarely parameterized separately from level of effort
Reactance	Co-operation	Uninteresting, private questions	Observed differences between low and high effort taken as evidence of reactance model	Difficulty in measuring "reactance"
Survey topic or sponsor interest	Co-operation	Topic-related or sponsor-related questions	Interviewer observation paradata of "I'm not interested" statement; observed responses to survey topic-related or sponsor-related questions	Survey topic is only one of many possible design features; topic and sponsor are often related, so an observed difference may be related to the sponsor rather than the topic
Personal characteristic or composition	Contact or co-operation	Any associated with the personal characteristic	Respondent demographic or other characteristics, typically age, education, sex and race	It is difficult to enumerate all possible personal characteristics that jointly affect survey participation and measurement errors
Research importance	Co-operation	All questions	Level of effort paradata taken as an indicator of research importance communicating greater importance	Whether respondents view higher numbers of contact attempts as is largely untested
Self-perception	Co-operation	All questions	Agreement for participation obtained before questionnaire administration	Requires time lapse between the act of agreeing to participate and completing the questionnaire
Changes in survey protocol	Contact or co-operation	Depends on the change in protocol	Paradata documenting the change in protocol	Change in protocol rarely randomly administered, so differences based solely on the protocol or on other characteristics cannot be disentangled

Alternatively, increased levels of effort may directly affect a sampled person's motivation. The "reactance hypothesis" states that measurement quality will suffer as a result of a reaction of the sampled person against "prodding" (Eckland, 1965) or "harassment" (Diaz de Rada, 2005). In particular, the question answering process is viewed in terms of a product of a respondent's cognitive ability and motivation to complete the task (Krosnick, 2002; Beatty and Hermann, 2002). Reactance is one possible cause of lower motivation to engage in the respondent task. This hypothesis is derived from reactance theory (Brehm and Brehm, 1981) in which increased pressures to comply with a request lead to a reaction by an individual to reassert their freedom (by putting little effort into their task and either failing to answer survey questions or providing lower quality answers) when they feel that a freedom is eliminated or violated (through follow-ups or attempts at persuasion) (McKee, 1992; Miller and Wedeking, 2006). The reactance model requires respondents to be aware of multiple follow-up or persuasion attempts, permits multiple freedoms to be violated (e.g. taking away the time to do the study, or answering uninteresting or private questions), and the reaction is proportionate to the perceived violation of freedom.

Interest in the topic of the survey or the sponsor is the third explanation for an association between level of effort and poor quality of data (Couper, 1991; Currivan, 2005; Donald, 1960; Fricker, 2007; Martin, 1994; Stinchcombe *et al.*, 1981). Lack of interest in the topic is a potential cause of lower motivation to engage in the survey task, although lack of interest may also affect knowledge about a topic. Unlike general motivation and reactance, interest in the topic can be inferred from what is communicated on the doorstep to an interviewer (Couper, 1997; Dahlhamer *et al.* 2006, 2008) or through the distribution of responses to survey questions among those who are reluctant to participate in a survey compared with those who were not reluctant (Donald, 1960), although responses are not known for the non-respondents. Interest in the topic should play a stronger role in surveys where the topic is made salient, perhaps through the title and sponsor of the survey, and play more of a role when the items are directly related to the topic or sponsor of the survey. As with general motivation, interest in the topic is an influence on co-operation, not contactability (Groves *et al.*, 2000, 2004; Groves and Couper, 1998).

The fourth hypothesis is a "personal characteristic" or "composition" hypothesis. Here, people who vary on a particular characteristic, which may not be causal (i.e. manipulable), systematically differ in their likelihood to participate and the quality of reports that they provide. For example, students with high grade point averages are more likely to participate in surveys and are better reporters of their grade point average than low grade point average students, even in studies where the topic is not academic achievement (Olson, 2007; Olson and Kennedy, 2006). Demographic characteristics of the sampled person such as education, age or sex are other "personal characteristics" that may be related to both a person's response propensity and quality of data (Armenakis and Lett, 1982; Cannell and Fowler, 1963; Kaminska *et al.*, 2006; Mason *et al.*, 2002; Olson, 2006; Olson and Kennedy, 2006; Safir and Tan, 2009). The personal characteristic or composition hypothesis is often posed as an alternative hypothesis for motivation, harassment or topic interest, via including these characteristics as statistical control variables in multivariate analyses. The personal characteristic hypothesis permits the characteristic to be related to either contactability or co-operation, and it may apply to any type of measurement error. It also does not necessarily predict that individuals who are more likely to participate provide higher quality reports than those who are less likely to participate.

Two hypotheses predict that people who respond after much recruitment effort will be no different from or better reporters than those who participate more readily, with no clear expectation for differences across items. The "research importance" hypothesis is that higher levels of effort convey the importance of the research to the sampled person (Schmidt *et al.*,

2005). This importance encourages the respondent to work harder, potentially increasing the quality of their reports relative to those who had received fewer attempts. The second hypothesis comes from self-perception theory in which an individual observes his or her own behavior and then infers his or her attitudes from that behavior (Bem, 1967; Fazio, 1987). When the "self-perception" hypothesis is applied to the survey task, the respondent infers that they are a good survey respondent because they agreed to participate in a survey and thus provides high quality data (Jobber *et al.*, 1985). This should be especially true when the decision to participate in the survey is separated from the measurement task, such as when a phone survey is used to recruit sample participants who are then sent a mail survey to complete (e.g. Jobber *et al.*, 1985).

The final hypothesis is not related to respondent characteristics, but instead to protocol characteristics. In some studies, later respondents receive design features that are different from those received by earlier respondents; this change in design features results in different quality answers for later respondents. This model is invoked by those who use mode switches to recruit later respondents (Voogt and Saris, 2005), increase levels of incentive (Currivan, 2005) or permit higher rates of proxy responses (Tancreto and Bentley, 2005). This explanation has also been used when the reference period for key survey variables is constant, so later interviews have longer and potentially more error prone recall periods (Bilodeau, 2006). Here, the change in protocol may differentially affect certain items (e.g. questions with specific reference periods). The distinction between this hypothesis and other hypotheses is that the relationship between level of effort and measurement error is thought to be driven primarily by decisions that are made by the survey organization rather than respondent characteristics.

Any single hypothesis is unlikely to explain the relationship between the level of effort and quality of data completely. First, paradata can be used to separate survey participation into contactability and co-operation. The probability of contacting a sampled person depends on their at-home patterns and impediments to access, each influenced by sampled person, household and ecological characteristics, and the number and timing of contact attempts, which are influenced by interviewer characteristics and/or field management decisions (see the model of Groves and Couper (1998), page 28). One householder characteristic affecting co-operation may be general willingness to participate in a survey. Yet empirical research has shown that the probability of obtaining co-operation from a sampled person, conditional on making contact, is affected by characteristics of the sampled person and household, their neighborhood, the interviewer, the general "survey taking environment" and design features including mode, incentive, topic, sponsor and length of the interview (e.g. Groves and Couper, 1998; Groves *et al.,* 2002). Sample subjects may weigh these various design features against each other or use perceptions of the cumulative effects of all of the design features when making a participation decision as suggested by leverage saliency theory (Groves *et al.*, 2000) and social exchange theory (e.g. Dillman *et al.,* 2009). Similarly, item non-response or measurement error in reports depends on the type of question, question wording, length of the recall period, mode, topic, social desirability concerns, difficulty of retrieving the information, willingness to engage in all steps of the cognitive response process, respondent characteristics and interviewer characteristics (e.g. Tourangeau *et al.,* 2000; Sudman *et al.,* 1996; Krosnick, 2002; Beatty and Hermann, 2002).

Despite the wide assortment of causes for both survey participation and measurement error, the explanations that are used by the studies examined below tend to focus on only one or two "causes," usually general traits, with a relationship that is constant over items. Often, these causes are not parameterized but assumed to exist if a relationship is observed between paradata measuring levels of recruitment effort and the data quality measure.

4. Empirical relationships between level of effort and item non-response and measurement error

The question remains whether the relationships hypothesized above have been observed empirically. For each of the paradata-derived level-of-effort measures and measurement error, the relevant hypotheses that were discussed above are identified, as is whether the level-of-effort measure reflects non-contact non-response, refusal non-response or a mix of both types of non-response. The empirical literature on the relationship between level of effort and item non-response and/or measurement error is then examined. Finally, whether the observed empirical relationships support, contradict or provide insufficient evidence about the relevant hypotheses is reviewed.

4.1. Hypothesized relationships between follow-up attempts and item non-response and/or measurement error

The level-of-effort measure that is obtained from paradata about the recruitment process most often used is the number of mailings or calls made to a sampled case before completing the interview, which I shall refer to as "follow-up attempts." Follow-up attempts, as described here, are distinct from a strategic and deliberate change in protocol that may be used to increase response rates or to evaluate non-response bias, which is sometimes called a non-response follow-up (Groves, 2006). In this measure, respondents are sorted into at least two groups indicating "high" and "low" effort. In mail surveys, a follow-up attempt is a mailing. In interviewer-administered surveys, follow-up attempts are calls. Follow-up attempts can be made either to contact a household or to obtain co-operation, thus confounding these two sources of non-response.

Follow-up attempts may act as a proxy for *motivation* (under a general motivation hypothesis), may create feelings of *harassment* (under a reactance hypothesis) and may recruit people who have less *interest in the survey topic* (under a survey topic interest hypothesis). Under each of these hypotheses, additional follow-up attempts bring in sampled cases, but at the risk of a trade-off in increased measurement error and/or item non-response. Importantly, these hypotheses suggest that the types of items that are most sensitive to the trade-off are those that are burdensome, difficult, sensitive, private and/or topic related. In contrast, the *research importance* hypothesis suggests that additional follow-up attempts increase the sampled person's perception of the importance of the research being conducted and should increase a respondent's willingness to complete the survey questions.

4.2. Empirical relationships between follow-up attempts and item non-response and/or measurement error

Most of the studies that look at the association between follow-up attempts and measurement quality are mail surveys (Armenakis and Lett, 1982; Cannell and Fowler, 1963; De Leeuw and Hox, 1988; Diaz de Rada, 2005; Donald, 1960; Eckland, 1965; Gilbert *et al.*, 1992; Green, 1991; Helasoja *et al.*, 2002; Jobber *et al.*, 1985; Kaminska *et al.*, 2006; Korkeila *et al.*, 2001; Newman, 1962; Schmidt *et al.*, 2005; Treat and Stackhouse, 2002), followed by telephone (Kreuter *et al.*, 2010; Schoenman *et al.*, 2003; Stinchcombe *et al.*, 1981; Voogt and Saris, 2005; Yan *et al.*, 2004) and face-to-face studies (Kennickell, 1999; Miller and Wedeking, 2006; Robins, 1963; Safir and Tan, 2009; Stoop, 2005; Tancreto and Bentley, 2005). The studies' response rates range from 26.7% (Kreuter *et al.*, 2010) to 78.7% (Newman, 1962). Most of the studies examined a general population sample, although some focused on organization members (e.g. Donald, 1960; Newman, 1962), farmers (e.g. Stinchcombe *et al.*, 1981), benefit recipients (e.g. Kreuter

et al., 2010) or physicians (e.g. Schoenman *et al.,* 2003). Because of the need for having records available, all of the response accuracy studies are conducted on special populations.

When individual items are examined, respondents who required more follow-up attempts have higher item non-response rates on roughly 75% of the 64 items that were reported in these studies (Donald, 1960; Korkeila *et al.*, 2001; Newman, 1962; Stoop, 2005; Schoenman *et al.*, 2003; Stinchcombe *et al.*, 1981; Treat and Stackhouse, 2002). For example, in a telephone survey, Schoenman *et al.* (2003) found that 14% of physicians who responded with five or fewer calls failed to report their income, compared with 17% who responded after 6–10 calls (Schoenman *et al.,* 2003, page 35). The studies that examine individual items tend to select questions that the above hypotheses suggest should differ over successive levels of effort, such as items that are particularly relevant to the topic (e.g. Donald, 1960; Stinchcombe *et al.,* 1981), difficult (e.g. Newman, 1962) or sensitive (e.g. Stoop, 2005; Korkeila *et al.,* 2001; Schoenman *et al.,* 2003). Other non-topic-related, easy or non-sensitive items generally are not included in these analyses.

In contrast, when aggregate measures of item non-response created across multiple or all items in a survey are examined, only seven of 32 comparisons (about 22%) between those who participated after few *versus* many follow-ups were statistically different from 0 (De Leeuw and Hox, 1988; Donald, 1960; Diaz de Rada, 2005; Gilbert *et al.*, 1992; Green, 1991; Helasoja *et al.*, 2002; Jobber *et al.*, 1985; Kaminska *et al.*, 2006; Kennickell, 1999; Korkeila *et al.*, 2001; Miller and Wedeking, 2006; Safir and Tan, 2009; Schmidt *et al.*, 2005; Tancreto and Bentley, 2005; Wellman *et al.*, 1980). The aggregate comparisons that are statistically different across number of follow-up attempts tend to be those that are more focused, i.e. examine either fewer items or a specific type of item non-response (e.g. Kaminska *et al.,* 2006; Korkeila *et al.,* 2001; Safir and Tan, 2009; Miller and Wedeking, 2006; Tancreto and Bentley, 2005).

This contrast between the associations that were found between follow-up attempts and individual items and those with aggregate cross-survey measures is important, suggesting that reasonable evidence exists that item non-response rates on some, but not all, items increase with additional follow-ups. Evidence from the item-specific item non-response analyses suggest that the hypotheses about sensitive, difficult or topic-related questions being associated with this level-of-effort measure (e.g. general motivation, reactance and topic interest) have some support. However, additional analyses of the non-sensitive, easy and non-topic-related items are needed to lend further evidence to the hypothesis. These findings also provide strong evidence against the research importance hypothesis.

The relationship between accuracy of response and follow-up attempts has been evaluated in eight studies on topics such as hospital visits, academic performance, eligibility for dental insurance, unemployment benefits and voting behavior. The response accuracy studies have topics and/or sponsors that are related to the records themselves; the items that are contained in the records are of interest because measurement error is expected due to the item's difficulty (hospital visits), sensitivity (arrests) or social desirability (grades; voting). Across 41 comparisons in these studies, 20 were significantly different, with tendencies for higher rates of inaccurate reports among those requiring additional follow-ups (Cannell and Fowler, 1963; Eckland, 1965; Gilbert *et al.*, 1992; Kreuter *et al.*, 2010; Olson and Kennedy, 2006; Robins, 1963; Sakshaug *et al.*, 2010; Voogt and Saris, 2005). For example, in a study of former child counseling patients, Robins (1963) found that those who required multiple contacts to interview were significantly less likely to admit to having been arrested; the accuracy of reporting other characteristics (i.e. divorce, having problem spouses, attending high school and truancy) was not statistically different across levels of effort. The classic Cannell and Fowler (1963) study also falls into this category, in which hospital stays and visits are reported less accurately among those who require more recruitment effort. Thus, these findings lend weak support to the general motivation, reactance and topic or sponsor interest

hypotheses, but items that would not support these hypotheses tended not to be explicitly selected in these studies.

Four studies looked at differences in the variability in responses across people or non-differentiation across items for respondents who received few *versus* many follow-up attempts. Under a general motivation hypothesis, people who require more follow-up attempts should have more variable responses due to satisficing. The standard deviation of responses (variability across people) on seven items significantly changed, both increasing and decreasing, over repeated follow-up attempts (Donald, 1960; Green, 1991), and non-differentiation (variability across items within people) also increased and decreased across call attempts (Miller and Wedeking, 2006; Yan *et al.*, 2004). Thus, there is reasonable evidence that the variability of answers changes with additional numbers of follow-ups, although not always in the direction of more measurement error.

There is no clear prediction from the hypotheses on the types of outcomes for attitudinal items what might be particularly sensitive to additional follow-up attempts, unless there is a general reactance to the survey questions. Across the analyses of attitudinal questions, the findings are similarly equivocal—additional follow-ups do not systematically bring in respondents whose answers to scale questions that are less reliable or internally consistent or are more likely to have other forms of measurement error on attitudinal items than those recruited with fewer follow-ups. Two studies examined scale reliability, finding no meaningful difference in reliability for those who required more follow-up attempts (De Leeuw and Hox, 1988; Green, 1991). For example, Green (1991) examined the reliability of four scales in a mail survey of teachers and found no difference in the reliability of three scales across three mailings. In the fourth scale, although a statistically significant difference between the mailings was observed, reliability did not monotonically decrease. Six additional measures examined are acquiescence, extreme responses, middle responses, inconsistent responses over logically related questions, multiple responses to single-response questions or a combination thereof (Armenakis and Lett, 1982; Diaz de Rada, 2005; Kaminska *et al.*, 2006; Yan *et al.*, 2004). Across 10 comparisons, only three are statistically different from 0, and one of the three yields higher quality data among the high recruitment effort respondents.

Overall, respondents who were recruited with more calls or mailings differed on some data quality measures from those recruited with fewer calls or mailings. They have higher item non-response rates, less accurate answers and answers that differ in variability, but they have no systematic difference in quality of responses to attitudinal questions.

4.3. Hypothesized relationships between refusal conversion and item non-response and/or measurement error

Paradata also contain information about call outcomes, including whether a sampled unit refused to be interviewed. *Refusal conversion* studies use these paradata to compare respondents who refused to participate at some point during the field period (but eventually participated) with all other respondents. Because converted refusals for mail surveys are rare, refusal conversion studies are conducted for interviewer-administered surveys. Converted refusals are a subset of those who receive additional follow-up attempts since at least one additional attempt is needed to convert a refusal. Other paradata are not collected "for free" but instead are recorded by interviewers. In interviewer-administered surveys, concerns that are voiced by a household member about survey participation to the interviewer "on the doorstep" during the recruitment request such as "I'm not interested" may be recorded by interviewers on standardized forms (Campanelli *et al.*, 1997; Morton-Williams, 1993). These doorstep statements are strongly associated with refusing to participate in a survey (e.g. Bates *et al.*,

2008; Groves and Couper, 1998) and are used to provide an alternative measure of respondent reluctance. In general, refusal conversion measures may not clearly identify whether the selected respondent was the one who voiced reluctance (especially if the respondent was not yet selected) or another household member.

In general, hypotheses about the association between refusal conversion and item non-response or measurement error are the same as those for follow-up studies, with a particular relevance for the reactance and topic interest hypotheses. Higher levels of item non-response or measurement error among converted refusals are expected because of a "reaction" against complying with the recruitment request that they had previously declined or because refusal occurs due to a general lack of interest in the topic. In fact, stronger associations between refusal conversion and item non-response or measurement errors may be expected because these hypotheses are specifically about the likelihood of co-operating with a request, and converted refusals clearly measure non-co-operation non-response. These associations should be strongest if the respondent selected is the person who refused, not another household member.

4.4. Empirical relationships between refusal conversion and item non-response and/or measurement error

The refusal conversion studies are roughly equally split between telephone surveys (Blair and Chun, 1992; Currivan, 2005; Keeter *et al.*, 2000; Kreuter *et al.*, 2010; Mason *et al.*, 2002; Olson and Kennedy, 2006; Retzer *et al.*, 2004; Schoenman *et al.*, 2003; Stinchcombe *et al.*, 1981; Tripplett *et al.*, 1996; Yan *et al.*, 2004) and face-to-face surveys (Campanelli *et al.*, 1997; Couper, 1997; Dahlhamer *et al.*, 2006, 2008; McDermott and Tan, 2008; Miller and Wedeking, 2006; Robins, 1963; Smith, 1983; Stoop, 2005). They are also primarily general population studies. Response rates range from 42% to 71% in the six studies that used refusal conversion indicators from call records. Response rates range from 61% to 72% in the four studies that recorded "not interested" doorstep statements. Between 10% and 20% of the respondent pool in these studies are converted refusals.

Question level item non-response rates for converted refusals *versus* respondents who had never refused were examined in nine studies. In these studies, item non-response rates are much higher for converted refusals than for other respondents on virtually all the items examined. For example, seven studies examine item non-response on income-related questions by using bivariate (Campanelli *et al.*, 1997 (two studies); Retzer *et al.*, 2004; Schoenman *et al.*, 2003; Stoop, 2005) and multivariate analyses (Carroll and Chong, 2006; Dahlhamer *et al.*, 2008). In each instance, the item non-response rate on the income question was higher among those who had previously refused or stated that they were not interested than among other respondents. Differences between converted refusals and never refusals on income item non-response remained even after multivariate controls (Carroll and Chong, 2006), although this varies for other items, by whether item non-response was a "don't know" or "refusal," and by the type of concerns voiced on the doorstep (Dahlhamer *et al.*, 2008).

10 studies examined an aggregate measure of item non-response for converted refusals. In the four politics-related studies, converted refusals consistently had statistically higher item non-response rates than never refusers (Campanelli *et al.*, 1997; Couper, 1997; Keeter *et al.*, 2000; Miller and Wedeking, 2006). Among the other studies, the direction is similar, although not quite as consistent.

Four studies had validation information available for examining response accuracy. Only three of the 19 comparisons in two refusal conversion studies were statistically different from 0, and converted refusers were equally likely to be more or less accurate reporters (Olson and

Kennedy, 2006; Robins, 1963; Kreuter *et al.*, 2010). In the final study, the "not-interested" respondents had strikingly lower political knowledge compared with the "interested" respondents (Couper, 1997). In sum, unlike the number of follow-ups, no consistent evidence exists that converted refusers provide less accurate answers than those who participated without refusing.

Surprisingly little evidence exists about differences in measurement error on attitudinal questions for converted *versus* never refusers. Non-differentiation was examined by two studies—one analyzing the American National Election Studies and the other an attitude survey—on three sets of items, with only one comparison showing marginally statistically significant differences between the two groups (Miller and Wedeking, 2006; Yan *et al.*, 2004). No study compared scale reliability for converted and never refusers. No clear differences were observed for primacy or recency effects for converted refusals (Blair and Chun, 1992; Jans, 2007) or in acquiescence, middle or extreme responses (Yan *et al.*, 2004).

Overall, converted refusals have higher item missing data rates than respondents who had not previously refused, but they do not necessarily provide less accurate or variable answers. Few studies have examined the relationship between measurement error on attitudinal items and refusal conversion. Although the item non-response findings support the reactance hypothesis, the other findings clearly do not. This discrepancy is particularly interesting as we might expect the empirical evidence for refusal conversion and these data quality outcomes to be stronger than for follow-up attempts given that converted refusals are prior non-co-operators.

4.5. Hypothesized relationships between time in the field and item non-response and/or measurement error

Paradata about the recruitment process also contain information about the day and time that call attempts are made. From these types of paradata is derived the third effort metric, the *time in the field period*. Studies that use the time in the field period divide respondents into groups that are defined by the number of weeks or months that have elapsed in the study period before an interview was completed. "Early" respondents participate at the beginning of the study period; "late" respondents participate at the end. As with the number of follow-up attempts, this measure confounds non-contact and refusal non-response. Furthermore, the time in the field period will not directly reflect effort to a case if first-contact attempts are made on different days to different cases (e.g. a late sample release).

There is no clear reason why a sampled case that has been in the field longer should have lower motivation, independent of effort exerted on the case. Plausible explanations here include the *personal characteristic* hypothesis or the *change in the recruitment protocol* hypothesis. Neither of these hypotheses is definitive on predictions for what will occur late in the field period—the personal characteristic hypothesis requires those who are recruited late to differ on observable characteristics that are also related to item non-response or measurement error; the change in the recruitment protocol hypothesis depends on what changes occurred or whether the item itself is related to date (e.g. memory of an event occurring or date-related questions).

4.6. Empirical relationships between time in the field and item non-response and/or measurement error

With no clear directional hypothesis for time in the field and item non-response rates, it is not surprising to find a lack of empirical studies using this level-of-effort measure. Only two studies report question-specific item non-response rates by the time in the field period

(Schoenman *et al.*, 2003; Voigt *et al.*, 2003). The item non-response rates for the early and late respondents are quite similar in these studies. The same story is found when examining the three studies that looked at aggregate measures of item non-response and response accuracy over the dates of the field period. Two studies found that item missing rates significantly increased over the course of the field period (Bilodeau, 2006; Friedman *et al.*, 2004), although these relationships depended on the type of item non-response examined. In the third study, no significant relationship was found (Wellman *et al.*, 1980).

For other measurement error outcomes, there is virtually no empirical evidence. Only one record check study looked at the time in the field in a mail survey as the level of effort measure (Voigt *et al.*, 2005), finding less accurate reports for late respondents for one group of respondents ("cases") than another group ("controls"). One study looked at measurement errors to attitudinal questions or variability of responses by the date of interview (Yan *et al.*, 2004). This telephone study found no differences between early and late respondents on acquiescence, middle or extreme responses or non-differentiation.

In sum, few studies have examined the relationship between time in the field and quality of data. For the few existing empirical examinations, as expected, no clear statistical difference is found in item non-response on individual items or in aggregate, for reporting accuracy or for measurement error in attitudinal items for early and late respondents.

4.7. Hypothesized and empirical relationships between combination measures and item non-response and/or measurement error

Combinations of the measures are also used. The combination measures generally indicate "any extraordinary effort" *versus* "no extraordinary effort" through high numbers of calls, extended time in the field and/or refusal conversion. As with the above measures, measures that combine levels of effort are indicators of the respondent's contactability and willingness to be interviewed along with survey organizational procedures about following up sampled householders. To the extent that these combination measures better identify the reluctant, unmotivated or otherwise uninterested respondents, we would expect that the same theories as discussed in the follow-up attempts and refusal conversion apply here.

Five studies looked at a combination level-of-effort measure and item non-response; no other data quality outcomes were examined in these studies. Two studies—an examination of the National Health Interview Survey and a telephone survey of physicians—looked at question level item non-response rates (Chiu *et al.*, 2001; Thran *et al.*, 1987). In both studies, the respondents who required extra recruitment effort had higher item non-response rates than those who participated without extra effort. Differences in "late or difficult" respondents on income item non-response held even in multivariate models (Bates and Creighton, 2000). In contrast, neither of the studies that examined aggregate item non-response rates by a combination measure of level of effort found differences in the mean or median number of item non-response answers (Keeter *et al.*, 2000; Kennickell, 1999).

5. Summarizing remarks and discussion

The findings for all the types of levels of effort and data quality outcomes are summarized in Table 2.

This review has five main findings.

(a) Many hypotheses have been posited to explain the relationship between recruitment effort and item non-response and/or measurement error.

(b) Respondents who are recruited with higher numbers of follow-up attempts and converted refusals tend to have higher item non-response rates on specific items than their easier-to-recruit counterparts. Limited and mixed evidence is available for respondents who are recruited later in the field period. This relationship is attenuated when looking at aggregate measures of item non-response.

(c) Respondents who required higher numbers of follow-up attempts provide less accurate answers on some items and more variable answers than those who required fewer follow-up attempts. No evidence of a consistent difference in accuracy or variability of responses is found for those who are converted refusals or are interviewed later in the field period.

(d) Little research has examined whether respondents who are recruited with more effort provide less reliable answers, are more acquiescent, are more prone to providing extreme or middle answers or provide less internally consistent answers to attitudinal questions than those recruited with less effort.

(e) Items that are predicted to be the most likely to show a relationship between effort and the data quality indicators that were examined here tend to do so. However, items that are predicted not to have a relationship are not examined in these studies.

Is increased level of recruitment effort detrimental to survey data quality? The strongest evidence is for item non-response—increased follow-up attempts and refusal conversion tend to be associated with increased item non-response rates. However, this does not hold for all items, with sensitive, difficult or burdensome items showing clearer trends than other types of

Table 2. Summary of findings for changes in six categories of measurement error by four level-of-effort measures†

	Results for the following levels of recruitment effort paradata:			
	Follow-up attempts	Refusal conversion	Date of interview	Combination
Question-specific item non-response	Strong evidence for higher item non-response rates (16 studies)	Strong evidence for higher item non-response rates (7 studies)	No clear difference (2 studies)	Weak evidence for higher item non-response rates (2 studies)
Aggregate item non-response	Weak evidence for higher item non-response rates (15 studies)	Strong evidence for higher item non-response rates (11 studies)	No clear difference (2 studies)	No clear difference (2 studies)
Accuracy	Weak evidence for less accurate responses (6 studies)	No clear difference (3 studies)	No clear difference (1 study)	No available evidence (0 studies)
Scale reliability	No clear difference (2 studies)	No available evidence (0 studies)	No available evidence (0 studies)	No available evidence (0 studies)
Variability in answers	Clear evidence of difference, but not on direction (4 studies)	No clear difference (2 studies)	No clear difference (1 study)	No available evidence (0 studies)
Measurement error—attitudinal questions	No clear difference (5 studies)	No clear difference (3 studies)	No clear difference (1 study)	No available evidence (0 studies)

†"No available evidence" indicates that, at the time of writing, no empirical studies have e xamined this combination of level of effort and the measurement error indicator. "No clear difference" indicates that there are empirical studies, but there is no clear evidence of a difference between high and low effort respondents.

item. Surprisingly little empirical evidence exists for measures other than item non-response rates. In addition, few studies measure motivation, harassment or the other potential causes that were enumerated in Table 1, other than asserting that level of effort *is* a measure of motivation (e.g. Cannell and Fowler, 1963).

Parameterization of causes for the two sources of error separate from level of effort is important for understanding and being able to anticipate when a relationship will or will not occur. This is difficult and will require careful parsing of non-contact from refusal non-response as well as measurement of concepts such as motivation and topic interest.

Across these studies, level of effort is used to identify "easy" *versus* "difficult" cases. However, an easy case for a survey with an 80% response rate may be different from an easy case for a survey with a 25% response rate. Many studies compared demographic and socio-economic characteristics of the easy and difficult cases; however, few took these comparisons to multivariate analyses, examining the association between level of effort and measurement error or item non-response, controlling for these characteristics. This type of multivariate modeling is necessary to identify whether the general tendency to be difficult is associated with quality of data independently of who the difficult respondents may be.

Selection or publication bias in which items are reported in the item non-response evaluations is plausible. Whereas large differences between high and low recruitment effort respondents exist when examining item non-response rates for a single question or for a block of highly related questions, aggregate item non-response rates across the entire questionnaire show much smaller differences. This may be because study researchers identify questions that they think are particularly burdensome, sensitive or difficult. Thus, questions where the theory would predict a significant difference in item non-response rates between the high and low effort respondents are over-reported in these studies relatively to questions where the theory would not predict such a difference. If item non-response rates for all questions had been reported individually, a more definitive conclusion about the relationship between recruitment effort and item non-response on specific types of questions could be drawn.

Given the relationship between item non-response and level of effort, should high effort sampled individuals not be recruited into the respondent pool? Item non-response has a frequently used set of adjustment methods—imputation—that can "fill in" the missing values for these cases (see Mason *et al.,* 2002). Even those respondents who participate with little recruitment effort have some level of item non-response; thus, imputation methods are needed with or without these extra respondents. Concern arises if those respondents who require higher effort are more likely to have non-ignorable item non-response (Little and Rubin, 2002), but few studies directly examined differences between item respondents and item non-respondents on the survey variables of interest, with or without conditioning on covariates (see Sakshaug *et al.,* 2010). Only one study examined key survey estimates with and without imputation and showed that imputation yielded estimates with lower mean-square error (Mason *et al.*, 2002). Whether these differences in item non-response rates across levels of effort also translate to non-ignorable item non-response deserves future investigation.

Most of the hypotheses in Table 1 focus on non-co-operation, suggesting that converted refusals, which are the most direct measure of refusal non-response, should be most susceptible to item non-response and measurement errors than never refusers. In particular, lack of motivation, reaction against an attempt at persuasion or lack of interest in the survey topic all predict that converted refusals are less motivated and thus more likely to "satisfice" (Krosnick, 2002; Krosnick and Alwin, 1987) than those who did not previously refuse. Striking differences in item non-response rates between converted refusals and non-refusers were reported in these studies. However, other types of measurement error, including

those who are suspected to be most sensitive to satisficing (e.g. acquiescence or non-differentiation), showed no clear differences or had no available evidence. Since motivation is difficult to measure directly, motivation alone cannot be disentangled from other factors to explain these results.

Higher levels of inaccurate reports among those who require additional follow-ups are more troubling and more dangerous for the quality of data. It is imperative that additional empirical research examines this measurement error by using record check studies. To advance knowledge about the relationship between effort of recruitment and accuracy of response, studies should be designed specifically for items where inaccurate answers are expected. If responses to a question tend to be accurate, then there will be little variation in measurement error to explain with level of effort. Recent investigations of the relationship between survey participation and measurement error (e.g. Tourangeau *et al.*, 2009, 2010) have started this investigation by experimentally varying individual survey protocol characteristics (e.g. sponsor and sensitive questions) or by attempting to separate non-contact non-response from refusal non-response through level-of-effort paradata (e.g. Kreuter *et al.*, 2010).

Recent research also has expanded our understanding of how to measure "response propensity" in this type of analysis (e.g. Fricker and Tourangeau, 2010; Olson, 2006, 2007; Peytchev *et al.*, 2010). These studies use predicted probabilities from a logistic model using a response indicator derived from paradata as the outcome variable and available auxiliary characteristics for respondents and non-respondents. These studies implicitly use a personal characteristic model to discriminate between those who have higher and lower probabilities to participate, given the propensity model. Although early work assumed a general motivation model (e.g. Olson, 2006, 2007; Olson and Kennedy, 2006; Yan *et al.*, 2007), more recent work using this approach has attempted to pre-identify or sort questions into categories that are likely to be more or less sensitive to changes in response propensity (e.g. Fricker and Tourangeau, 2010; Peytchev *et al*,. 2010). This work is likely to be fruitful for further exploiting paradata to understand when there will or will not be a relationship between participation and measurement errors.

The four measures of recruitment effort that were examined here are clearly only a first set of measures. Other measures could include mode switches, changes in interviewers or interviewer workload or the use of persuasion letters, among other design features. The paradata-derived level-of-effort measures that were examined here are commonly used in existing published and unpublished literature to proxy for response propensity. Future research should examine other protocol features as well.

Acknowledgments — Thanks go to Lindsey Witt, Chun Feng and Lauren Walton for research assistance and to Sonja Ziniel, Frauke Kreuter, Paul Biemer, Gabi Durrant and three reviewers for comments on earlier drafts. An earlier version of this paper was presented at the International Total Survey Error Workshop in June 2008.

References

Armenakis, A. A. and Lett, W. L. (1982) Sponsorship and follow-up effects on response quality of mail surveys. *J. Bus. Res.*, **10**, 251–262.

Bates, N. and Creighton, K. (2000) The last five percent: What can we learn from difficult/late interviews? *Proc. A. Meet. Am. Statist. Ass.*

Bates, N., Dahlhamer, J. and Singer, E. (2008) Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. *J. Off. Statist.*, **24**, 591–612.

Beatty, P. and Herrmann, D. (2002) To answer or not to answer: Decision process related to survey item nonresponse. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 71–85. New York: Wiley.

Bem, D. J. (1967) Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychol. Rev.*, **74**, 183–200.

Bethlehem, J. (2002)Weighting nonresponse adjustments based on auxiliary information. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), chapter 18. New York: Wiley.

Biemer, P. and Lyberg, L. (2003) *Introduction to Survey Quality*. New York: Wiley.

Bilodeau, A. (2006) Non-response error versus measurement error: A dilemma when using mail questionnaires for election studies. *Aust. J. Polit. Sci.*, **41**, 107–117.

Blair, J. and Chun, Y. I. (1992) Quality of data from converted refusers in telephone surveys. *A. Meet. American Association for Public Opinion Research, St Petersburg*.

Bollinger, C. R. and David, M. H. (2001) Estimation with response error and nonresponse: Food stamp participation in the SIPP. *J. Bus. Econ. Statist.*, **19**, 129–141.

Brehm, S. S. and Brehm, J. W. (1981) *Psychological Reactance: A Theory of Freedom and Control*. New York: Academic Press.

Campanelli, P., Sturgis, P. and Purdon, S. (1997) *Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates*. London: Social and Community Planning Research.

Cannell, C. F. and Fowler, F. J. (1963) Comparison of a self-enumerative procedure and a personal interview: A validity study. *Publ. Opin. Q.*, **27**, 250–264.

Carroll, M. and Chong, Y. (2006) Are refusal conversions different from willing respondents with respect to item non-response, demographics and selected health characteristics?: The National Health and Nutrition Examination Survey, 1999-2002. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 2795–2802.

Chearo, D. and Van Haitsma, M. (2010) Standardizing paradata. *A. Meet. American Association for Public Opinion Research, Chicago*.

Chiu, P.-L., Riddick, H. and Hardy, A. M. (2001) A comparison of characteristics between late/difficult and non-late/difficult interviews in the National Health Interview Survey. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*

Couper, M. P. (1991) Modeling survey participation at the interviewer level. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 98–107.

Couper, M. P. (1997) Survey introductions and data quality. *Publ. Opin. Q.*, **61**, 317–338.

Couper, M. P. (1998) Measuring survey quality in a CASIC environment. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 41–49.

Currivan, D. (2005) The impact of providing incentives to initial telephone survey refusers on sample composition and data quality. *A. Meet. American Association for Public Opinion Research, Miami*.

Dahlhamer, J. M., Simile, C. M. and Taylor, B. (2006) Exploring the impact of participant reluctance on data quality in the National Health Interview Survey (NHIS). In *Proc. Methodological Issues in Measuring Population Health*. Ottawa: Statistics Canada.

Dahlhamer, J. M., Simile, C. M. and Taylor, B. (2008) Do you really mean what you say?: Doorstep concerns and data quality in the National Health Interview Survey (NHIS). *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 1484–1491.

De Leeuw, E. and Hox, J. (1988) Artifacts in mail surveys: The influence of Dillman's total design method on the quality of responses. In *Sociometric Research*, vol. 2, *Data Analysis* (eds. W. E. Saris and I. N. Galhofer), pp. 61–73. New York: St Martin's.

Diaz deRada, V. (2005) The effect of follow-up mailings on the response rate and response quality in mail surveys. *Qual. Quan.*, **39**, 1–18.

Dillman, D. A., Smyth, J. D. and Christian, L. M. (2009) *Internet, Mail and Mixed-mode Surveys: The Tailored Design Method*, 3rd edition. New York: Wiley.

Donald, M. N. (1960) Implications of nonresponse for the interpretation of mail questionnaire data. *Publ. Opin. Q.*, **24**, 99–114.

Eckland, B. K. (1965) Effects of prodding to increase mail-back returns. *J. Appl. Psychol.*, **49**, 165–169.

Fazio, R. H. (1987) Self-perception theory: A current perspective. In *Social Influence: The Ontario Symposium* (eds. M. P. Zanna, J. M. Olson and C. P. Herman), pp. 129–150. Hillsdale: Erlbaum.

Fricker, S. (2007) The relationship between response propensity and data quality in the Current Population Survey and the American Time Use Survey. *Dissertation*. University of Maryland, College Park. Unpublished.

Fricker, S. and Tourangeau, R. (2010) Examining the relationship between nonresponse propensity and data quality in two national household surveys. *Publ. Opin. Q.*, **74**, 934–955.

Friedman, E. M., Clusen, N. A. and Hartzell, M. (2004) Better late?: Characteristics of late respondents to a health care survey. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 992–998.

Gilbert, G. H., Longmate, J. and Branch, L. G. (1992) Factors influencing the effectiveness of mailed health surveys. *Publ. Health Rep.*, **107**, 576–584.

Green, K. E. (1991) Reluctant respondents: Differences between early, late and nonresponders to a mail survey. *J. Experimental Educ.*, **59**, 268–276.

Groves, R. M. (1989) *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Publ. Opin. Q.*, **70**, 646–675.

Groves, R. M. and Couper, M. P. (1998) *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R. M., Dillman, D. A., Eltinge, J. L. and Little, R. J. A., eds. (2002) *Survey Nonresponse*. New York: Wiley.

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

Groves, R. M., Presser, S. and Dipko, S. (2004) The role of topic interest in survey participation decisions. *Publ. Opin. Q.*, **68**, 2–31.

Groves, R. M., Singer, E. and Corning, A. (2000) Leverage-saliency theory of survey participation: Description and an illustration. *Publ. Opin. Q.*, **64**, 299–308.

Helasoja, V., Prattala, R., Dregval, L., Pudule, I. and Kasmel, A. (2002) Late response and item nonresponse in the Finbalt Health Monitor Survey. *Eur. J. Publ. Health*, **12**, 117–123.

Jans, M. (2007) Agree to disagree or vice versa: Response order effects in a phone survey of attitudes toward science and technology. *A. Meet. Midwest Association of Public Opinion Research, Chicago*.

Jobber, D., Allen, N. and Oakland, J. (1985) The impact of telephone notification strategies on response to an industrial mail survey. *Int. J. Res. Marketing*, **4**, 291–296.

Kaminska, O., Goeminne, B. and Swyngedouw, M. (2006) Satisficing in early versus late responses to a mail survey. *A. Meet. Midwest Association of Public Opinion Research, Chicago*.

Kaminska, O., McCutcheon, A. and Billiet, J. (2010) Satisficing among reluctant respondents in a cross-national context. *Publ. Opin. Q.*, **74**, 956–984.

Keeter, S., Miller, C., Kohut, A., Groves, R. M. and Presser, S. (2000) Consequences of reducing nonresponse in a national telephone survey. *Publ. Opin. Q.*, **64**, 125–148.

Kennickell, A. B. (1999)What do the "late" cases tell us?: Evidence from the 1998 Survey of Consumer Finances. *Int. Conf. Survey Nonresponse, Portland*.

Korkeila, K., Suominen, S., Ahvenainen, J., Ojanlatva, A., Rautava, P., Helenius, H. and Koskenvuo, M. (2001) Non-response and related factors in a nation-wide health survey. *Eur. J. Epidem.*, **17**, 991–999.

Kreuter, F., Müller, G., and Trappmann, M. (2010) Nonresponse and measurement error in employment research: Making use of administrative data. *Publ. Opin. Q.*, **74**, 880–906.

Krosnick, J. A. (2002) The causes of no-opinion responses to attitude measures in surveys: They are rarely what they appear to be. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 87–100. New York: Wiley.

Krosnick, J. A. and Alwin, D. F. (1987) An evaluation of a cognitive theory of response-order effects in survey measurement. *Publ. Opin. Q.*, **51**, 201–219.

Lessler, J. T. and Kalsbeek, W. D. (1992) *Nonsampling Error in Surveys*. New York: Wiley.

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

Martin, C. L. (1994) The impact of topic interest on mail survey response behavior. *J. Market Res. Soc.*, **36**, 327–338.

Mason, R., Lesser, V. and Traugott, M. W. (2002) Effect of item nonresponse on nonresponse error and inference. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 135–148. New York: Wiley.

McDermott, N. and Tan L. (2008) The effect of refusal conversion on data quality in the Consumer Expenditure Interview Survey. *Consumer Expend. Surv. Anthol.*, 23–32.

McKee, D. O. (1992) The effect of using a questionnaire identification code and message about non-response follow-up plans on mail survey response characteristics. *J. Market Res. Soc.*, **34**, 179–192.

Miller, J. M. and Wedeking, J. (2006) Examining the impact of refusal conversions and high callback attempts on measurement error in surveys. *A. Meet. American Association for Public Opinion Research, Chicago*.

Morton-Williams, J. (1993) *Interviewer Approaches*. Cambridge: Cambridge University Press.

Newman, S. W. (1962) Differences between early and late respondents to a mailed survey. *J. Advertising Res.*, **2**, 37–39.

Oh, H. and Scheuren, F. (1983) Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*, vol. 2, *Theory and Bibliographies* (eds. W. Madow, I. Olkin, and D. B. Rubin), pp. 143–184. New York: Academic Press.

Olson, K. (2006) Survey participation, nonresponse bias, measurement error bias, and total bias. *Publ. Opin. Q.*, **70**, 737–758.

Olson, K. (2007) An investigation of the nonresponse/measurement error nexus. *PhD Dissertation*. University of Michigan, Ann Arbor. Unpublished.

Olson, K. and Kennedy, C. (2006) Examination of the relationship between nonresponse and measurement error in a validation study of alumni. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 4181–4188.

Peytchev, A., Peytcheva, E. and Groves, R. M. (2010) Measurement error, unit nonresponse and self-reports of abortion experiences. *Publ. Opin. Q.*, **74**, 319–327.

Retzer, K. F., Schipani, D., and Cho, Y. I. (2004)Refusal conversion: Monitoring the trends. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 4984–4990.

Robins, L. N. (1963) The reluctant respondent. *Publ. Opin. Q.*, **27**, 276–286.

Safir, A. and Tan, L. (2009) Using contact attempt history data to determine the optimal number of contact attempts. *A. Meet. American Association for Public Opinion Research, Hollywood*.

Sakshaug, J., Yan, T. and Tourangeau, R. (2010) Nonresponse error, measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Publ. Opin. Q.*, **74**, 907–933.

Schmidt, J. B., Calantone, R. J., Griffin, A. and Montoya-Weiss, M. M. (2005) Do certified mail third-wave follow-ups really boost response rates and quality? *Marketing Lett.*, **16**, 129–141.

Schoenman, J. A., Berk, M. L., Feldman, J. J. and Singer, A. (2003) Impact of differential response rates on the quality of data collected in the CTS Physician Survey. *Evaln Health Profess.*, **26**, 23–42.

Singer, E. (2002) The use of incentives to reduce nonresponse in households. In *Survey Nonresponse* (eds. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little), pp. 163–177. New York: Wiley.

Smith, T. W. (1983) The hidden 25 percent: An analysis of nonresponse on the 1980 General Social Survey. *Publ. Opin. Q.*, **47**, 386–404.

Stinchcombe, A. L., Jones, C. and Sheatsley, P. B. (1981) Nonresponse bias for attitude questions. *Publ. Opin. Q.*, **45**, 359–375.

Stoop, I. A. L. (2005) *The Hunt for the Last Respondent*. The Hague: Social and Cultural Planning Office of the Netherlands.

Sudman, S., Bradburn, N. M. and Schwarz, N. (1996) *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Tancreto, J. G. and Bentley, M. (2005) Determining the effectiveness of multiple nonresponse follow-up contact attempts on response and data quality. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 3626–3632.

Thran, S., Olson, L. and Strouse, R. (1987) The effectiveness and costs of special data collection efforts in a telephone survey of physicians. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 761–766.

Tourangeau, R., Groves, R. M., Kennedy, C. and Yan, T. (2009) The presentation of a web survey, nonresponse and measurement error among members of web panel. *J. Off. Statist.*, **25**, 299–321.

Tourangeau, R., Groves, R. M. and Redline, C. D. (2010) Sensitive topics and reluctant respondents: Demonstrating a link between nonresponse bias and measurement error. *Publ. Opin. Q.*, **74**, 413–432.

Tourangeau, R., Rips, L. J. and Rasinski, K. A. (2000) *The Psychology of Survey Response*. New York: Cambridge University Press.

Treat, J. B. and Stackhouse, H. F. (2002) Demographic comparison between self-response and personal visit in Census 2000. *Popln Res. Poly Rev.*, **21**, 39–51.

Tripplett, T., Blair, J., Hamilton, T. and Kang, Y. C. (1996) Initial cooperators vs. converted refusers: Are there response behavior differences? *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 1038–1041.

Voigt, L. F., Boudreau, D. M., Weiss, N. S., Malone, K. E., Li, C. I. and Daling, J. R. (2005) RE: "Studies with low response proportions may be less biased than studies with high response proportions." *Am. J. Epidem.*, **161**, 401–402.

Voigt, L. F., Koepsell, T. D. and Daling, J. R. (2003) Characteristics of telephone survey respondents according to willingness to participate. *Am. J. Epidem.*, **157**, 66–73.

Voogt, R. J. J. and Saris, W. E. (2005) Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *J. Off. Statist.*, **21**, 367–387.

Wellman, J. D., Hawk, E. G., Roggenbuck, J. W. and Buhyoff, G. J. (1980) Mailed questionnaire surveys and the reluctant respondent: An empirical examination of differences between early and late respondents. *J. Leis. Res.*, **12**, 164–173.

Yan, T., Tourangeau, R. and Arens, Z. (2004) When less is more: Are reluctant respondents poor reporters? *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 4632–4651.