Statistical Journal of the IAOS 31 (2015) 305–315
DOI 10.3233/SJI-140844
IOS Press

# An empirical examination of the relationship between nonresponse rate and nonresponse bias

Graham Wright
*Cohen Center for Modern Jewish Studies, MS 014, Brandeis University, P.O. Box 549110, Waltham, MA 02454-9110, USA*
*Tel.: +1 781 736 2134; E-mail: gwwri@brandeis.edu*

**Abstract.** The dramatic decline in survey response rates over the past three decades raises significant concerns about the possibility of bias in survey results. Current theory emphasizes that it is the relationship between response propensity and variables of interest that determines the extent of the bias, and that a low response rate in itself does not necessarily imply a high level of bias. This assertion is supported by a number of studies which have shown that response rate alone is a fairly poor predictor of nonresponse bias. However, most of these studies suffer from methodological features that in some way compromise their attempts to isolate the relationship between response rate and bias. This paper describes the results of a pair of studies which allow for a near-ideal examination of this relationship. The results support the conclusions of prior research, showing that even achieved samples with response rates as low as 10 percent may produce highly accurate estimates in certain cases.

Keywords: Online survey, survey nonresponse, response rate, nonresponse bias

## 1. Introduction

The dramatic decline in survey response rates over the past three decades has been well documented [1–3]. Equally well documented is the accompanying shift in the way survey researchers think about nonresponse bias. The classical paradigm of probability sampling requires a 100 percent response rate to guarantee unbiased estimates [4,5]. Survey researchers working within this paradigm are thus willing to absorb significant costs in order to achieve a response rate sufficiently close to 100 percent that the traditional techniques of statistical inference can be confidently applied. Although survey researchers know that 100 percent response rates are generally impossible to achieve, and design their surveys accordingly, the classical paradigm still implies that results from surveys with particularly low response rates should be treated with relative skepticism, and that significant efforts should be made to minimize nonresponse as much as possible [6]. However, in an era of declining response rates and increasing per-unit costs, even achieving a response rate of over 50 percent has become a virtual impossibility for many researchers.

Recent work emphasizes a more nuanced understanding of the importance of response rates. Groves et al. [7] and Groves and Couper [8] note that nonresponse bias is the product of *both* the nonresponse rate *and* the distinctiveness of nonrespondents, relative to respondents, on a given variable of interest. Similarly, Bethlehem and Kersten [9] see nonresponse bias as being determined by both the relative size of the nonrespondent group and the "contrast", defined as the difference between the means for respondents and nonrespondents, on a variable of interest. Therefore, while, *ceteris paribus*, a low response rate may increase the *risk* of nonresponse bias, it is the relationship between response propensity and the variable of interest that determines the extent of the bias [7,10]. If there is no relationship between the probability of response and the variable of interest, then a higher response rate will not lead to any decrease in bias. Conversely, if response propensity is closely related to the variable of interest, then significant bias can easily coexist with a high response rate. This framework also implies that nonresponse bias is manifested on a variable by variable basis. A survey with a low response rate may simultaneously produce relatively unbiased estimates on items that are not correlated with response propensity, and highly biased estimates of items that are.

This theoretical framework is supported by a number of studies that have increasingly shown that response rate alone is a fairly poor predictor of nonresponse bias. An observational study by Merkle and Edelman [11] found no relationship between nonresponse bias and nonresponse rates when estimates from exit polls were compared to recorded vote totals. In 2006 Keeter et al. [12] replicated a 1996 methodological experiment that compared results of a "standard" survey to those of a "rigorous" survey (with a higher response rate) administered to a different sample of the same population. In both decades the estimates from the pairs of surveys were statistically indistinguishable for most items. Other studies have simulated the impact of lower response rate, by excluding respondents who either responded later in the field period [13] or who only responded after additional effort [14], and found similarly small relationships between response rate and bias. A 2008 meta-analysis of 59 separate studies for which data on nonrespondents exists found that the nonresponse rate of a survey was, by itself, a poor predictor nonresponse bias [4]. In an analysis of the accuracy of probability and nonprobability internet surveys. Yaeger et al. [15] found that higher response rates were actually correlated with *less* accurate estimates. Schouten, Cobben, and Bethlehem found similar results in their analysis of a 1998 Dutch survey of household living conditions [16].

This research has given powerful support to the argument that response rate, considered alone, is a poor predictor of nonresponse bias. In light of these findings Shouten et al. [16,17] have argued against the use of overall or subgroup response rates as measures of nonresponse bias and have develop alternative indicators (denoted "R-indicators"), based on the representativeness of the respondent population. However, most of the studies discussed above suffer from methodological features that in some way compromise their attempts to fully isolate the relationship between response rate and bias. For Keeter et al. [12] and Yaeger et al. [15], many of the population benchmarks used to asses bias due to nonresponse rate were derived from other national surveys, which were deemed to be unbiased precisely by virtue of their low nonresponse rates. Consequently, these studies, as well as that of Curtin and Presser [14], focus mainly on comparing estimates from subsamples with different (actual or simulated) response rates to each other as opposed to comparing these survey estimates to known population values. Merkle and Edelman [11] compared reported voting behavior from exit polls to actual vote totals, and did not randomize interviews at the polling locations, potentially conflating interviewer effects and nonresponse bias [7]. The variable response rates discussed by Schouten et al. [16] were achieved through different modes (in person vs. phone interviews) potentially conflating nonresponse bias with mode effects. Groves and Peytcheva's 2008 meta-analysis was disproportionately comprised of self-administered surveys (56 percent of surveys included), which were heavily drawn from the medical field, potentially complicating generalizations to other forms of survey research [4]. Heerwegh et al.'s 2007 study [13] looked only at bias with respect to a single variable, leaving open question about how bias might, in line with the theories described above, manifest itself differently for different variables in the same study.

Additionally, efforts to "simulate" lower response rates by sequentially analyzing portions of the achieved sample which had responded prior to a given time point (as in Heerwegh et al. [13] and Curtin et al. [14]) may run the risk of conflating nonreponse bias with another outcome of lower response rate – a smaller achieved sample size. All else being equal a sample with a lower response rate will likewise contain fewer observations, and estimates derived from such a sample will thus have larger standard errors. Consequently, even if substantial bias exists in low response rate estimates, the larger standard errors associated with these estimates may obscure this relationship. Another, perhaps obvious, implication of the negative relationship between response rates and standard errors is that reducing the risk of nonresponse bias is not the only reason to desire a higher response rate.

A complication that many of the studies above have faced when exploring the relationship between response rate and bias is the absence of population data to use as a benchmark for assessing nonresponse bias. Of course the existence of population data for survey nonrespondents is not uncommon for surveys which make use of administrative data to define the survey frame, or which have access to census data for relevant population groups. Census records, or auxiliary variables which are extant in administrative data, are frequently used as weighting targets to perform nonresponse adjustments, with the explicit goal of remedying bias in the achieved sample on these variables due to survey nonresponse [8,18]. To the extent that there is correlation between the auxiliary variables and survey items of interest, then adjusting for bias with regards to the former will also reduce bias on the latter [10]. Directly correcting for nonresponse bias by adjusting on the survey items themselves would obviously be preferable, but in general, if population data is extant on the variables of interest then there is little need to conduct a survey in the first place.

However, in order to use auxiliary variables to explore the relationship between response rate and nonreponse bias, the level of bias with respect to these variables must be assessed at a number of different response rates, holding the sample frame and other features of survey methodology, including the standard errors of estimates, as near constant as possible. An ideal study for examining this relationship would thus utilize the methodological experiments described in Keeter et al. [12], instead of simulating lower response rates by excluding respondents. However, in contrast to previous experiments, such a study would also have access to observed, rather than estimated, population data for all sample members, as in Heerwegh et al. [13] and Schouten et al. [16].

This paper makes use of a data source that fulfills exactly these criteria. As in other studies, separate random samples of a population were subjected to different levels of "effort", which produced differing response rates. In this case, however, a large number of auxiliary variables are available via administrative records for all members of the population, including nonrespondents. Respondents in different samples can thus be compared to the underlying population, as well as to each other. All respondents were interviewed in the same fashion, regardless of which sample they were in, negating the possibility of mode effects. In addition, the samples were constructed in such a way that the sample with a lower achieved response rates actually contains *more* respondents than the sample with a higher response rate. This removes the possibility that bias in the lower response rate samples might be masked by larger standard errors. Finally, the study, including the methodological experiments described below, was administered in near identical fashion to two different populations of vastly different sizes in two different countries. This increases the reliability of the findings, and allows for an examination of the influence of sample size overall. In combination, these features allow for a truly explicit examination of the relationship between nonresponse rate and nonresponse bias.

## 2. Study design

In 2011 the Cohen Center for Modern Jewish Studies at Brandeis University conducted a pair of online surveys of applicants to the "Taglit-Birthright Israel" program, which provides free, 10 day trips to Israel for Jewish young adults. The program attracts tens of thousands of applicants each year, and a portion of applicants are selected, more or less randomly, to participate. Applications for trips are accepted in two "rounds" each year: in February for trips taking place over the summer and in September for trips taking place in the winter. Applicants must provide basic demographic data (including age, gender and self-identified Jewish denomination) as well as contact information when they register for the program. The Cohen Center has access to this auxiliary data for all applicants to the program, including survey nonrespondents. The two studies considered in this paper were of applicants who applied to the program between 2006 and 2010. An online survey of applicants to the program from the United States was administered between February and March of 2011. Another online survey of Canadian applicants was administered between July and September of 2011 and also includes applicants who applied to participate in winter 2010/2011 trips, which are not included in the US frame. For each study the eligible population (131,804 eligible applicants for the US study, 12,686 for the Canadian study) were stratified by year of application, participant status (whether or not they went on the trip), age (over or under 25 at time of application) and gender. For both studies two separate stratified random samples of dramatically different sizes were drawn, each corresponding to a specific incentive structure. For the US frame, a smaller "guaranteed incentive" (GI) sample was drawn containing 3,000 cases and a much larger "raffle" (R) sample was drawn containing 64,400 cases. In the Canadian frame the smaller GI sample contained 1,270 cases and the larger R sample contained the entire remaining 11,416 cases. Even though the raffle sample in the Canadian study contains all remaining population members it can still be treated as random sample with a particularly large sampling fraction, rather than a failed census, because it is the complement of another random sample (i.e. the guaranteed incentive Canadian sample). The survey instruments administered to the US and Canadian populations were very similar, with only slight differences on questions explicitly dealing with American or Canadian cultural or political matters.

The contact procedure was similar for both the US and Canadian surveys. As with Keeter et al. [12], the different samples in both surveys were subject to different levels of effort, which produced significantly different response rates. In both surveys the R samples were contacted by email only, with entry into a raffle as the only incentive. Members of the GI samples were contacted by email and offered a guaranteed incentive. Members of a small random sub-sample of the US GI sample were offered $25 Amazon.com gift cards, and all other members of the GI samples were offered $15 Amazon.com gift cards. Approximately two weeks after the initial email invitation all members of the GI samples who had not yet responded were called (using phone numbers provided during the registration process) and encouraged to complete the survey online. The callers did not actually administer the survey to the respondents, but simply encouraged the respondents to complete it on their own, and, in many cases, re-emailed the respondent's unique survey URL to an email address of the respondent's choosing. Because the survey was, in all cases, self-administered online, there are no mode effects across the different samples.

This methodology allows for the analysis of three different groups of respondents in each survey. The $\mathrm{R}_{\mathrm{TOTAL}}$ groups contain all respondents to the R samples in the US and Canadian frames. The $\mathrm{GI}_{\mathrm{PRECALL}}$ groups contain respondents to the US and Canadian GI samples who responded prior to the beginning of the calling. The $\mathrm{GI}_{\mathrm{TOTAL}}$ groups contain all respondents to the GI samples, including those who responded after the onset of calling. Note that, for both the US and Canadian surveys, $\mathrm{GI}_{\mathrm{PRECALL}}$ is a strict subset of $\mathrm{GI}_{\mathrm{TOTAL}}$, containing only those GI sample respondents who took the survey prior to the beginning of the calling operation. Analogous to earlier efforts [13,14] to simulate lower response rates by excluding certain respondents $\mathrm{GI}_{\mathrm{PRECALL}}$ thus represents an approximation of what the final response rate of the GI samples would have been in the absence of calling operations. In contrast, $\mathrm{R}_{\mathrm{TOTAL}}$ and $\mathrm{GI}_{\mathrm{TOTAL}}$ represent mutually exclusive samples drawn from the same population.

As would be expected, the varying levels of effort produced markedly different response rates (see Table 1). In both the US and Canadian studies the $\mathrm{R}_{\mathrm{TOTAL}}$ groups, which received email invitations and a raffle incentive, achieved response rates of around 10 percent.[1] Prior to the beginning of calling the GI samples achieved 18 percent response rate after the initial email invitation and reminders ($\mathrm{GI}_{\mathrm{PRECALL}}$). The intense calling effort then raised the response rates for the small samples to almost 50 percent ($\mathrm{GI}_{\mathrm{TOTAL}}$). It should be reiterated that respondents to all three groups were drawn from random, stratified samples of the same population, and as such the only differences between the groups are due to differing methodological protocols, leading to different response rates. As mentioned above the R samples contain a substantially larger number of respondents than the two GI groups despite their lower response rates, owing the much larger initial size of the R samples. Consequently, in the analyses discussed below estimates from the R sample will have smaller standard errors and confidence intervals compared to those from the GI groups. On the other hand, despite similarities in response rates between the two surveys, estimates derived from the Canadian survey will have slightly larger standard errors, due to the smaller size of both Canadian samples.

By examining auxiliary demographic variables available for all population members, it is possible to examine how the characteristics of the achieved sample change, with respect to these variables, as the response rate rises from 10 percent to 18 percent to 50 percent. This procedure is essentially identical to the technique utilized in Schouten et al. [16] where administrative data were "artificially treated as survey items" (p. 102) by deleting their values for non-respondents. The mean for each of these variables will be computed for each of the three groups described above, as well as for the entire population. This allows an empirical measure of the correlation between response rate and bias, with respect to these specific variables.

## 3. Comparison of demographic estimates to population values

Table 2 shows demographic estimates from the US survey derived from the three groups described above compared to the true population values, as well as 95 percent confidence intervals.[2] In regard to demo-

[1] All response rates reported are AAPOR RR2.

[2] There are no confidence intervals around the population means. In Tables 2 and 3 a single star (*) has been used to denote cases where the population mean is included within the 95 percent confidence interval of the sample mean for that item.

Table 1
Response rate (AAPOR RR2) by survey and subgroup

| | $\mathrm{R}_{\mathrm{TOTAL}}$<br>(Raffle incentive.<br>Email invitation/reminder) | $\mathrm{GI}_{\mathrm{PRECALL}}$<br>(Guaranteed incentive.<br>Email invitation/reminder only.) | $\mathrm{GI}_{\mathrm{TOTAL}}$<br>(Guaranteed incentive.<br>Email invitation/reminder + Calling.) |
|---|---|---|---|
| US Survey | RR2: 9.61% N: 6,194 | RR2: 17.69% N: 529 | 49.0% N: 1,468 |
| Canadian Survey | RR2: 11.42% N: 1,304 | RR2: 18.12% N: 230 | RR2: 48.11% N: 611 |

Table 2
US data – Demographic variables for achieved samples and population with 95 percent confidence intervals

| | | $\mathrm{R}_{\mathrm{TOTAL}}$<br>(10 percent RR) | $\mathrm{GI}_{\mathrm{PRECALL}}$<br>(18 percent RR) | $\mathrm{GI}_{\mathrm{TOTAL}}$<br>(50 percent RR) | Population |
|---|---|---|---|---|---|
| Achieved N | | 6,194 | 529 | 1,468 | |
| Demographic variables | % went on trip | 68% (66–69%) | 68% (64–72%) | 71%* (68–73%) | 66% |
| | % female | 60%* (58–61%) | 60%* (55–64%) | 57% (54–59%) | 54% |
| | % over 25 | 45%* (44–47%) | 45% (41–49%) | 41% (39–43%) | 41% |
| | % identify as Reform | 42% (41–43%) | 44% (40–49%) | 43% (40–45%) | 41% |
| | % identify as Conservative | 23% (22–24%) | 21% (18–25%) | 24% (21–26%) | 23% |
| | % Identify as Orthodox | 4%* (4–5%) | 5% (3–7%) | 5% (3–6%) | 6% |
| | % Identify as "Just Jewish" | 25% (24–26%) | 25% (22–29%) | 24% (22–26%) | 25% |
| Round applied | Winter 2006/2007 | 6% (5–7%) | 6% (4–8%) | 7% (5–8%) | 7% |
| | Summer 2007 | 11%* (10–12%) | 13% (10–16%) | 14% (13–16%) | 15% |
| | Winter 2007/2008 | 12% (11–12%) | 12% (9–14%) | 12% (10–13%) | 12% |
| | Summer 2008 | 15%* (15–16%) | 16%* (13–20%) | 19%* (17–21%) | 22% |
| | Winter 2008/2009 | 14%* (13–15%) | 13% (10–16%) | 12% (10–14%) | 11% |
| | Summer 2009 | 13%* (13–15%) | 13% (10–16%) | 15% (13–17%) | 15% |
| | Winter 2009/2010 | 14%* (13–15%) | 14%* (11–17%) | 12% (10–14%) | 10% |
| | Summer 2010 | 15%* (14–16%) | 13% (10–16%) | 9% (8–11%) | 9% |
| | Winter 2010/2011 | N/A | N/A | N/A | N/A |

graphic variables there is relatively little difference between estimates produced by different response rates, and little difference between these estimates and the population value. Even estimates from the $\mathrm{R}_{\mathrm{TOTAL}}$ group, with a 10 percent response rate, come within six percentage points of the population values for each of these variables. For the estimates of Jewish denominational identification there is no systematic change in the accuracy of the estimates as response rates rise. For the estimates of gender and age, estimates do generally become less biased as response rate rises. In contrast, the $\mathrm{GI}_{\mathrm{TOTAL}}$ estimates (representing a 50 percent response rate) are actually less accurate than the estimates derived from $\mathrm{R}_{\mathrm{TOTAL}}$, and $\mathrm{GI}_{\mathrm{PRECALL}}$ (representing response rates under 20 percent). This might be due to the script used by callers in the GI sample, which varied by participant status. Trip participants were asked to take part in a survey of "Taglit-Brithright Israel participants," evoking memories of the trip itself, while nonparticipants were merely asked to participate in a survey of "Jewish young adults." This feature might lead to participants being disproportionately influenced to participate by the calling protocol. However, as mentioned, even for age, gender and trip participation, the absolute difference between the $\mathrm{R}_{\mathrm{TOTAL}}$, $\mathrm{GI}_{\mathrm{PRECALL}}$ and $\mathrm{GI}_{\mathrm{TOTAL}}$ estimates is fairly small.

Table 2 also shows estimates and population values for the percentage of applicants who applied to the program in a specific round. On this measure a much more consistent, and substantial relationship between nonresponse rate and nonresponse bias is evident. The lower response rate estimates consistently overestimate the proportion of applicants who applied in more recent rounds and underestimate the proportion who applied in earlier rounds. As response rate rises this bias is reduced monotonically. This seems to imply the existence of a correlation between response propensity and round of application. One possible explanation for this correlation is that the contact information provided by applicants to earlier rounds was more likely to be out of date by the time the survey was fielded, leading to a greater probability of non-contact.

Table 3 shows the same comparisons for the Canadian survey. In regards to the demographic measures the 10 percent response rate $\mathrm{R}_{\mathrm{TOTAL}}$ estimates differ by no more than 5 percentage points from the population values, roughly similar to the trend in the US data. In addition, the overall relationship between bias and nonresponse remains the same as in the US data. Estimates of denominational identification do not differ systematically by response rate. For age and gen-

Table 3
Canadian data – Demographic variables for achieved samples and population with 95 percent confidence intervals

| | | $\mathrm{R}_{\mathrm{TOTAL}}$<br>(10 percent RR) | $\mathrm{GI}_{\mathrm{PRECALL}}$<br>(18 percent RR) | $\mathrm{GI}_{\mathrm{TOTAL}}$<br>(50 percent RR) | Population |
|---|---|---|---|---|---|
| Achieved N | | 1,304 | 230 | 611 | |
| demographic variables | % went on trip | 67% (65–70%) | 71% (62–78%) | 74%* (70–77%) | 67% |
| | % female | 58%* (55–61%) | 52% (46–59%) | 53% (49–57%) | 53% |
| | % over 25 | 71%* (69–74%) | 69% (63–76%) | 73% (69–77%) | 74% |
| | % identify as Reform | 22% (20–24%) | 20% (14–25%) | 22% (19–26%) | 23% |
| | % identify as Conservative | 23% (21–25%) | 28% (22–33%) | 29%* (26–33%) | 25% |
| | % Identify as Orthodox | 4% (3–5%) | 4% (2–7%) | 4% (2–6%) | 4% |
| | % Identify as "Just Jewish" | 40% (38–43%) | 41% (35–48%) | 36% (32–40%) | 38% |
| Round applied | Winter 2006/2007 | 5% (4–6%) | 6% (3–9%) | 6% (4–7%) | 6% |
| | Summer 2007 | 17% (15–19%) | 13%* (9–18%) | 18%* (15–22%) | 22% |
| | Winter 2007/2008 | 6% (5–7%) | 3% (1–6%) | 4% (3–6%) | 6% |
| | Summer 2008 | 15% (13–17%) | 20% (15–26%) | 22% (18–25%) | 23% |
| | Winter 2008/2009 | 6% (5–7%) | 7% (3–10%) | 5% (3–6%) | 4% |
| | Summer 2009 | 12% (10–14%) | 12% (8–16%) | 11% (9–14%) | 12% |
| | Winter 2009/2010 | 6% (4–7%) | 5% (2–8%) | 5% (3–6%) | 4% |
| | Summer 2010 | 26% (23–28%) | 27%* (22–33%) | 22% (19–26%) | 19% |
| | Winter 2010/2011 | 8% (6–9%) | 6% (3–9%) | 6% (4–8%) | 4% |

der higher response rate estimates are generally more accurate than lower response rate estimates, but again the overall bias is small even for low response rate estimates. As in the US survey, estimates of program participation become less accurate as response rate rises, although again the overall effect is small.

The tendency of lower response rate estimates to overestimate the proportion of applicants from more recent rounds is less dramatic in the Canadian data than in the US, but is still apparent. The absolute size of the bias in the lower response rate estimates is less than in the US data, and the accuracy of the estimates do not always increase monotonically as response rate rises. This implies that the correlation between time of application and quality of contact information is lower for Canadian applicants than Americans.

Overall, the most striking feature of these analyses is the relatively high accuracy of the $\mathrm{R}_{\mathrm{TOTAL}}$ estimates across virtually all measures, despite the low response rate. While bias does exist on some measures, a survey with a 10 percent response rate would still provide a highly accurate demographic profile of these populations. For the majority of demographic items examined estimates based on higher response rates do only marginally better, and sometimes worse, than estimates based on higher response rates. The exception to this trend, especially for the US survey, is the round of application, where a consistent bias towards more recent applicants is apparent for lower response rate estimates.

These findings suggest that, especially for the US survey, the low response rate to the initial email invitations was largely due to noncontact, and not refusal to take the survey by respondents. This hypothesis is supported by relatively low refusal rates for members of the GI sample. Only 11.1 percent of US applicants and 11.74 percent of Canadian applicants who were in the GI samples explicitly refused to take the survey when contacted by callers. Thus, even in the $\mathrm{GI}_{\mathrm{TOTAL}}$ group, which achieved a 50 percent response rate, almost 80 percent of the nonresponse (77 percent in the Canadian survey and 78 percent in the US survey) was due to noncontact, rather than refusal. Although it is not possible to determine the extent to which nonresponse to the email-only R samples was due to noncontact or refusal this result suggests that noncontact may be a primary driver of nonresponse to both surveys.

This hypothesis is further supported by the relationship between response rates and round of application. Since applicants to earlier rounds were more likely to have provided email addresses that they no longer check, or no longer have access to (in the case of applicants who were students at the time of application but who have since graduated), these individuals were disproportionately less likely to see, and therefore complete, the survey. When attempts were made to contact respondents by telephone, as opposed to by email, the response rate increased dramatically and the bias associated with round of application diminished, but there was little movement on other demographic estimates. This suggests that the email invitation to the survey (with or without a guaranteed incentive) was similarly appealing to individuals of different genders, ages and Jewish backgrounds. The relative lack of bias in regards to Jewish denomination is especially encouraging, since this variable is highly correlated with a number of variables of interest in the survey itself, such as attitudes towards Israel and Jewish ritual practice. For example, Orthodox Jews generally have extremely high levels of ritual practice whereas "Just Jews" are more likely to be unobservant. Since the low response rate estimates produced reasonably good estimates of the proportion of applicants that belong to specific denominations, it is less likely that they will produce estimates that are highly biased on other measures correlated with Jewish denomination.

This hypothesis can be tested by following the techniques used by Keeter et al. [12] and comparing the results of high and low response rate estimates on survey variables of interest. Population values for these variables are naturally not available (otherwise there would be no reason to do the survey in the first place), so the absolute bias due to nonresponse cannot be precisely calculated as it was above. However, the relationship between nonresponse rate and nonresponse bias for those who did respond can be calculated, and the analyses above can help contextualize the results presented above.

## 4. Comparison of survey items across groups

The analysis is based on 30 variables of interest available in both the US and Canadian surveys. The items selected can broadly be characterized into four broad categories: Israel attitudes (6 items), Jewish engagement during high school (9 items), current Jewish engagement (11 items) and non-Jewish demographic items (4 items). To aid comparability all non-dichotomous items were recorded into binary variables. Estimates for each binary variable were computed for the mutually exclusive $\mathrm{R_{TOTAL}}$, and $\mathrm{GI_{TOTAL}}$ groups (representing response rates of 10 and 50 percent respectively), and for each comparison a chi-square test was computed to determine whether differences due to differing response rates were statistically significant. Because all variables analyzed were binary, all chi square tests reported have a single degree of freedom

In the US survey the observed differences between responses for the two groups were fairly small, and for only 6 of the 30 items tested were the differences statistically significant at the 95 percent confidence level.

Figure 1 shows the distribution of observed percentage point differences between the two subsamples for both the entire set of thirty items and the six where the differences were statistically significant.

The results here echo those of Keeter et al. (compare to Fig. 1 in Keeter et al.'s 2006 paper [12]) who found similarly small differences between high and low response rate samples. However, the "low" response rate for this analysis (10 percent RR2) is far lower than the 25 percent RR3 for the "standard" survey utilized by Keeter et al., while the high response rate examined here (50 percent RR2) is comparable to the 50 percent RR3 achieved by Keeter er al.'s "rigorous" survey. This begs the question of whether Keeter et al. would have observed similarly small differences if the "standard" response rate had been even lower.

The six items for which significantly different estimates were obtains across the two samples are shown in Table 4. The only item related to Israel attitudes in this group is a question about respondents' confidence in discussing the topic of minority populations in Israel. The higher response rate samples produced higher estimates of the proportion of applicants with more Jewish friends in high school, and more Jewish friends now, and who had attended a Jewish wedding. The higher response rate sample also produced significantly higher estimates for the proportion of applicants who were employed and who had volunteered in the past 12 months. On four of the measures, the higher response rate survey estimated a slightly higher proportion of individuals with greater levels of Jewish engagement. This is contrary to the naive hypothesis that those with higher levels of Jewish engagement would be more inclined to take a survey on Jewish matters with little prompting, which would imply that estimates of Jewish engagement would decline as response rate rises. However, the relatively small size of the differences reported here, and the lack of statistically significant differences for most other measures of Jewish engagement frustrate further speculation.

Comparing the same set of 30 items in the Canadian survey produces results remarkably similar to those from the US survey. As seen in Fig. 2 there was a slightly larger differences between the $\mathrm{R_{TOTAL}}$, and $\mathrm{GI_{TOTAL}}$ estimates in the Canadian survey relative to the US, but in only 4 of those cases (see Table 5) were the differences statistically significant at the 95% level of confidence.[3] This is to be expected considering the

[3] There were, however an additional seven items where the differences in estimates (generally 4 or 5 percentage points) were significant at the 90% confidence level. These included the "Jewish wedding" and "volunteer" items that were significant in the US survey, as well as confidence in discussing Israeli art, confidence in following along in Jewish religious services, being a member of a synagogue, being married and having attended a bar or bat mitzvah in the past 12 months.

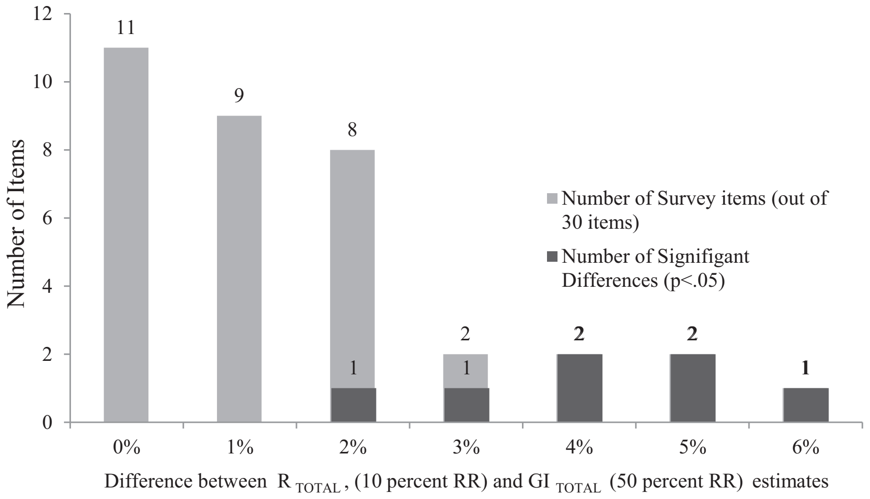

Fig. 1. Histogram of the difference between $R_{TOTAL}$, (10 percent RR) and $GI_{TOTAL}$ (50 percent RR)-US survey.

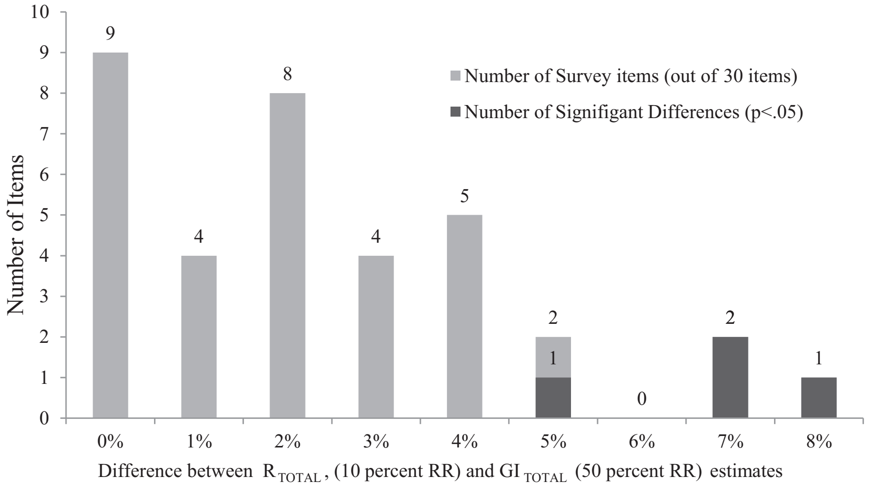

Fig. 2. Histogram of the difference between $R_{TOTAL}$, (10 percent RR) and $GI_{TOTAL}$ (50 percent RR) – Canadian survey.

smaller number of observations in the Canadian samples.

As with the US survey the two samples gave significantly different estimates on the "minority populations in Israel" question, as well as current and past proportion of Jewish friends (Table 5). The two Canadian samples also give different estimates of the proportion of respondents who feel confident discussing Israeli religion. Once again, however, the absolute difference between estimates is rather small, especially considering the larger standard errors and smaller sample size. As with the US data speculation about the cause of these differences is frustrated by their small magnitude and the small proportion of variables where they manifest.

It should also be noted that, for both the US and Canadian surveys, the large number of statistical tests performed dramatically increases the likelihood of observing significant differences between the GI and R samples which are due to chance alone [19]. Given 30 independent comparisons tested at an error rate of 0.05, there is a 78.54% likelihood of observing at least one significant difference between two identical populations. Applying a Bonferroni adjustment to correct for this increased likelihood of type 1 error would effectively involve lowering the alpha of each test from 0.05 to 0.0016. None of the differences observed in the Canadian survey are significant at this level, and only two items in the US survey are: the proportion of current and high school friends who are Jewish. Although

Table 4
Items with statistically significant differences across samples – US survey

| | $GI_{TOTAL}$<br>(50 percent RR)<br>estimate | $R_{TOTAL}$<br>(10 percent RR)<br>estimate | Percent difference | Chi-square (1 df) |
|---|---|---|---|---|
| More than half of high school friends were Jewish | 54% | 48% | 6% | 15.2796*** |
| More than half of current friends are Jewish | 56% | 51% | 5% | 10.3271*** |
| Currently employed | 50% | 45% | 5% | 9.6369** |
| Volunteered in the past 12 months | 66% | 70% | 4% | 8.6509** |
| "Somewhat" or "very" confident in discussing minority populations in Israel | 25% | 28% | 4% | 7.7548** |
| Attended Jewish wedding in the past 12 months | 38% | 35% | 3% | 4.454* |
| Achieved N | 1,468 | 6,194 | | |

Table 5
Items with statistically significant differences across samples – Canadian survey

| | 50%RR estimate | 10%RR estimate | Percent difference | Chi-square (1 df) |
|---|---|---|---|---|
| More than half of high school friends were Jewish | 58% | 50% | 8% | 9.4251** |
| "Somewhat" or "very" confident in discussing minority populations in Israel | 26% | 33% | 7% | 8.9738** |
| More than half of current friends are Jewish | 58% | 51% | 7% | 7.9984** |
| "Somewhat" or "very" confident in discussing Israeli religion | 52% | 57% | 5% | 4.0307* |
| Achieved N | 611 | 1,304 | | |

the Bonferroni test is conservative, due to its assumption that all items are perfectly uncorrelated with each other, it is still likely that the GI and R samples are even more similar to one another, with respect to the 30 items tested, than the analyses above suggest.

## 5. Discussion

The results presented here concur with the previous research cited above that show a low correlation between nonresponse rate and nonresponse bias in many situations. In addition, these results are less subject to criticisms which might be leveled against previous research. The auxiliary variables utilized here are not estimates but known parameters, and were measured in exactly the same way for respondents and nonrespondents. Differences in response rates were not conflated with other methodological complications, such as mode effects, differences in sampling scheme, achieved sample size, or other "house effects" that might intrude when comparing surveys conducted by different organizations. The replication of the experiments on populations of dramatically different sizes allows for an exploration of the impact, or lack thereof, of overall sample size. The wide variance between the high and low response rates examined allows for much stronger conclusions about the lack of an *a priori* relationship between response rate and bias.

In particular, the results here suggest that for the population in question (applicants to Taglit-Birthright Israel) there is little correlation between response propensity to complete an internet survey on Jewish matters and the variables of interest (viz. Jewish engagement, Jewish background, and attitudes towards Israel). There are a number of possible explanations for this phenomenon. First, the population is relatively homogeneous from a demographic perspective. By virtue of the program's eligibility criteria and the rounds chosen, the entire eligible population being surveyed was between 18 and 36 years old at the time of the survey, and only around 10 percent were married. This lessens the possibility that lifecycle factors or differing patterns of email usage might contribute to differential probabilities of response for different applicants. In addition, American Jews as a whole comprise a relatively homogeneous group in terms of socioeconomic status, with relatively high levels of education and income [20], further lowering the potential for differential nonresponse due to cultural or economic factors. Furthermore, by definition, all members of the population were Jewish and had applied to participate in a program offering a free, Jewish-focused trip to Israel. One could therefore assume that there was a high baseline level of interest in the topics addressed in the survey. It is likely that for a population that was more heterogeneous, with respect to either demographic or attitudinal factors, a 10 percent response rate might have produced more substantial bias.

The sort of results presented here have increasing relevance in regard to the current debate about the validity of various forms of "nonprobability sampling" which includes such diverse techniques as opt-in online panels, snowball samples, and river sampling. These forms of sampling have generally been viewed with extreme skepticism by the survey research community, since the selection probability for any given case is generally not known. However, nonprobability sampling proponents sometimes argue that the distinction between probability and nonprobability based sampling isn't as clear cut as it seems since, it is claimed, "Nonresponse is a form of self-selection [21]." Ergo the true probability of selection for respondents to a probability based sample is unknown, just as with nonprobability samples.

However, a more nuanced understanding of nonresponse bias, which is supported by the findings presented here, somewhat undermines this argument. These results show that probability based samples can, in certain circumstances, achieve a very high level of accuracy with a surprisingly low response rate. However, this is so precisely because of a low correlation between response propensity and the variables of interest. The data here suggest that the very low 10 percent response rate achieved by the initial email invitation was largely due to the quality of contact information associated with the respondent. If an individual fails to respond to a survey because he or she never receives an invitation then this form of nonresponse ought not to be considered "self-selection", since the respondent never even has the opportunity to make a conscious choice about whether to respond to the survey. Clearly noncontact can be correlated with other relevant factors (such as age or gender in an RDD survey) but these demographic factors seem less likely to be correlated with variables of interest relative to the psychological and sociological factors that contribute to an individual's personal decision about whether to participate or refuse a survey. Insofar as the decision to "opt-in" to a nonprobability survey is correlated with these same factors, the possibility for bias may be far more severe than for a probability based survey with even a 10% response rate.

## 6. Conclusion

These findings lend considerable support to the idea that response rate itself is a poor predictor of nonresponse bias. Of far greater importance are the causes and correlates of the nonresponse that does exist, regardless of its magnitude. This research provides additional clarity as to the ways in which survey nonresponse does or does not contribute to bias. In the studies examined here the demographic and attitudinal homogeneity of the population, and the apparently low rates of refusal, appeared to be the key factors in reducing nonresponse bias. The comparison between US and Canadian surveys also showed that given these two factors, significant differences in sample size have only a negligible impact on results. In combination with previous research, these findings provide powerful support for a more nuanced approach to analyzing response rates. In addition, the consistency of these results may be seen as continued justification of probability based sampling, even in an era of low response rates.

## References

[1] E.D. Leeuw and W.D. Heer, Trends in Household Survey Nonresponse: A Longitudinal and International Perspective, in *Survey Nonresponse*, R.M. Groves et al., Wiley: New York, 2002, pp. 41–54.

[2] R. Curtin, S. Presser and E. Singer, Changes in Telephone Survey Nonresponse over the Past Quarter Century, *Public Opinion Quarterly* **69**(1) (2005), 87–98.

[3] U.S. Department of Education – National Center for Education Statistics, An Overview of Response Rates in the National Household Education Survey: 1991, 1993, 1995 and 1996, in *NCES 97-948*, J.M. Brick, M. Collins and K. Chandler, eds, U.S. Department of Education: Washington D.C, 1997.

[4] R.M. Groves and E. Peytcheva, The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis, *Public Opinion Quarterly* **72**(2) (2008), 167–189.

[5] K. Olson, Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias, *Public Opinion Quarterly* **70**(5) (2006), 737–758.

[6] Y. Baruch, Response Rate in Academic Studies – A Comparative Analysis, *Human Relations* **52**(4) (1999), 421–438.

[7] R.M. Groves et al., Survey Methodology. 2004, New Kersey: Wiley.

[8] R.M. Groves, Theoretical Motivations for Post-Survey Nonresponse Adjustment in Household Surveys, *Journal of Official Statistics* **11**(1) (1995), 1995.

[9] J.G. Bethlehem and H.M.P. Kersten, On the Treatment of Nonresponse in Sample Surveys, *Journal of Official Statistics* **1**(3) (1985), 287–300.

[10] C.-E. Sarndal, The 2010 Morris Hansen Lecture Dealing with Survey Nonresponse in Data Collection, in Estimation, *Journal of Official Statistics* **27**(1) (2011), 1–21.

[11] D.M. Merkle and M. Edelman, Nonresponse in Exit Polls: A Comprehensive Analysis, in *Survey Nonresponse*, R.M. Groves et al., eds, Wiley: New York, 2002.

[12] S. Keeter et al., Guaging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey, *Public Opinion Quarterly* **70**(5) (2006), 759–779.

[13] D. Heerwegh, K. Abts and G. Lossveldt, Minimizing survey refusal and noncontact rates: do our efforts pay off? *Survey Research Methods* **1**(1) (2007), 3–10.

[14] R. Curtin, S. Presser and E. Singer, The Effects of Response Rate Changes on the Index of Consumer Sentiment, *Public Opinion Quarterly* **64**(4) (2000), 413–428.

[15] D.S. Yaeger et al., Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted With Probability and Non-Probability Samples, *Public Opinion Quarterly* **75**(4) (2011), 709–747.

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

[17] B. Schouten, N. Schlomo and C. Skinner, Indicators for Monitoring and Improving Representativeness of Response, *Journal of Official Statistics* **27**(2) (2011), 231–253.

[18] S. Lundstrom and C.-E. Sarndal, Calibration as a Standard Method for Treatment of Nonresponse, *Journal of Official Statistics* **15**(2) (1999), 305–327.

[19] D. Curran-Everett, Multiple comparisons: philosophies and illustrations, *American Journal of Physiology* **279**(1) (2000), R1–R8.

[20] B.R. Chiswick and C.U. Chiswick, The Economic Status of American Jews in the Twentieth Century, in *Encyclopedia of American Jewish History*, S.H. Norwood and E.G. Pollack, eds, ABC-CLIO: Santa Barbara, CA, 2007, pp. 62–66.

[21] D. Rivers, Second Thoughts About Internet Surveys. 2009; Available from: http://www.pollster.com/blogs/doug_rivers.php?nr=1.