American Economic Review 2013, 103(7): 3001–3021
http://dx.doi.org/10.1257/aer.103.7.3001

Conclusions Regarding Cross-Group Differences in Happiness Depend on Difficulty of Reaching Respondents[†]

By Ori Heffetz and Matthew Rabin*

A growing literature explores differences in subjective well-being across demographic groups, often relying on surveys with high nonresponse rates. By using the reported number of call attempts made to participants in the University of Michigan's Surveys of Consumers, we show that comparisons among easy-to-reach respondents differ from comparisons among hard-to-reach ones. Notably, easy-to-reach women are happier than easy-to-reach men, but hard-to-reach men are happier than hard-to-reach women, and conclusions of a survey could reverse with more attempted calls. Better alternatives to comparing group sample averages might include putting greater weight on hard-to-reach respondents or even extrapolating trends in responses. (*JEL* C83, I31)

In the literature studying the determinants of subjective well-being, much of the evidence is based on surveys with significant and growing nonresponse rates.[1] Except for some checks, such as verifying that the demographics of respondents match the general population, happiness researchers have for the most part had to live with the potential selection bias. This paper takes advantage of a variable collected in a prominent telephone survey—the University of Michigan's "Surveys of Consumers" (SOC)—to shed some light on the existence and nature of such selection bias, as well as on a potential way of mitigating it.

In particular, the SOC records how many call attempts it took to reach each respondent who completed the survey; the median in our sample is five attempts, and 10 percent of the respondents were reached after more than 20 attempts. We show that happiness levels compare differently across commonly studied demographic groups among the easy-to-reach versus the difficult-to-reach. For instance, the more

* Heffetz: S.C. Johnson Graduate School of Management, Cornell University, 324 Sage Hall, Ithaca, NY 14853 (e-mail: oh33@cornell.edu); Rabin: Department of Economics, University of California, Berkeley, 508-1 Evans Hall No. 3880, Berkeley, CA 94720 (e-mail: rabin@econ.berkeley.edu). We thank Angus Deaton, Winston Lin, Edward Miguel, Nora Cate Schaeffer, Justin Wolfers, and seminar participants at RAND and LSE for useful discussions and comments; Miles Kimball and Noah Smith for their help with the data; and three anonymous referees for suggestions that improved the paper. This research was assisted by scientific collaborations supported by NIH/NIA grant R01-AG040787 to the University of Michigan (with subcontract to Cornell University). The authors have no financial or other material interests related to this research to disclose.

† Go to http://dx.doi.org/10.1257/aer.103.7.3001 to visit the article page for additional materials and author disclosure statement(s).

[1] Response rates in large surveys vary by survey methodology. Recent response rates in the General Social Survey—a personal visit survey—have been slightly above 70 percent. Response rates in the Gallup-Healthways Well-Being Index—a daily telephone survey—have been around 31 percent. Recent response rates in the data we use in this paper lie roughly halfway between these two figures. See Curtin, Presser, and Singer (2005) for a discussion of the increasing trends in nonresponse in recent years.

attempts it took to reach child-free people below the median age, the higher their reported happiness is. The same is true for neither those of the same age group who have children nor for those above median age. Such group-differentiated gradients suggest that conclusions based on surveys with less complete or (more importantly) more complete data would systematically differ from those based on the SOC. They also indicate a potentially better approach to estimating population group differences than the current practice of comparing sample averages.

Our most striking results are between men and women, a cross-group comparison that has received much attention recently. The answer to the yes/no happiness question we use varies relatively little across demographics or call attempts, with 85.5 percent of respondents reporting happy (among the top and bottom income halves, 88.8 percent and 81.8 percent report happy). Easy-to-reach men are about as happy as easy-to-reach women without controlling for covariates—among the easiest-to-reach 30 percent of our respondents, 86.2 percent of men and 85.7 percent of women are happy—and are 1.3 points *less* happy than women when controlling for covariates. But among the hardest-to-reach 30 percent of respondents, men are 4.1 and 2.3 points happier than women, respectively, without and with controls.[2] This implies that a hypothetical survey that replicated the SOC's first two call attempts without further attempts would conclude, with the usual controls, that women are 1.3 percentage points happier than men. The actual survey concludes that *men* are 0.6 points happier than *women*.[3] Finally, assuming that those who were not reached in the SOC are similar to the harder-to-reach 50 percent of those reached—an assumption we discuss below—a hypothetical survey that reached the entire respondent pool would conclude that men are roughly 1.0 points happier than women.

In Section I, we discuss our approach. Long-standing literatures in statistics and in survey methodology have explored how to use response-difficulty patterns for making inferences about nonrespondents. Statistical models have been developed using the basic presumption that motivates our interpretation of the data: with differences in answering propensities across groups, the average unreached member of a group may be more similar to the average difficult-to-reach than to the average easy-to-reach. While a related intuition underlies, for example, economists' data-collection strategies for reducing potential attrition bias in panel surveys, we know of no papers that use our approach either within economics or within happiness research. We also discuss empirical work by Curtin, Presser, and Singer (2000), who use the number-of-calls variable in the SOC to study potential effects on the Index of Consumer Sentiment of the increasing difficulty, over the years, of reaching respondents. They find that sentiments among difficult-to-reach respondents are more positive than among easy-to-reach respondents, but conclude that this is unlikely to be problematic in the context of typical uses of the Index. Although this and related papers never studied—and, hence, never ruled out the possibility of—*group-differentiated* effects such as those we demonstrate, they have been repeatedly

[2] These are differences in percentages responding "Yes" to the happiness question. Regressing the binary happiness variable on a female indicator, a binary easy- versus hard-to-reach indicator, and an interaction, we reject the hypothesis of no interaction ($p = 0.03$ without controls and $p = 0.01$ with them).

[3] The 95 percent confidence intervals around the 1.3 and 0.6 point estimates are, respectively, men 0.2 points happier to women 2.8 points happier, and women 0.2 points happier to men 1.4 points happier.

cited as comfort that selection bias may not be a big concern in surveys on happiness and other topics. We conclude Section I with a simple theoretical example illustrating how the standard practice of comparing group averages could yield the wrong conclusions—and how using difficulty-of-reaching paradata could help identify the right conclusions.

In Section II, we review our dataset and empirical strategy. SOC respondents in the time period we examine reported whether they were happy "much of the time during the past week." A clear advantage of using the SOC for our purposes—beyond the availability of the number-of-calls variable—is that it reaches many subjects who would remain out of sample in other datasets. In the SOC, trained and experienced interviewers often make 30 or more call attempts to hard-to-reach households. For comparison, the Gallup-Healthways Well-Being Index—a daily survey of 1,000 Americans that calls landlines as well as cell phones—stops after 3–5 call attempts.

We present our results in Section III. We examine four characteristics that have received attention recently in the happiness-by-demographic-group literature: income (e.g., Kahneman and Deaton 2010), sex (Stevenson and Wolfers 2009), age (Stone et al. 2010), and having children in the household (Herbst and Ifcher 2011). Although we use many other variables as controls in our regressions, these are the only four characteristics we set out to examine.[4] Whether looking at raw, unadjusted means, or adjusting for covariates, we find that selection bias affects the size of the happiness-sex, -age, and -children gaps. These effects are large enough to affect conclusions about the gaps, and are statistically strong in many (though not all) of our specifications. The happiness-income gap is mostly unaffected.

We show the potential importance of the patterns we find in two ways. First, we show several cases where the conclusions regarding differences in happiness across groups would have been different if—as is the case with many other surveys—fewer attempts had been made to reach "hard" respondents. Second, we show how even conclusions based on the entire SOC sample are still biased if nonresponders are similar to hard-to-reach respondents—rather than to the *average* respondent, as is often implicitly assumed in current research.

Our proposed "corrective" obviously does not solve many other selection problems in survey analysis. It does not help assessing, for instance, whether the individuals chosen as potential respondents are representative of the population of interest. Nor does it address the significant problem in some surveys of those who are reached but refuse to participate; we have no evidence that such refuseniks are similar to the eventually-cooperative-but-difficult-to-reach respondents we study.[5] Notwithstanding these limitations, it is worth stressing that our empirical findings yield some simple conclusions that are virtually theory-free and interpretation-free: with measures and tests routinely used by researchers, conclusions about group happiness differences from a survey that targeted the SOC population but stopped calling sooner would be different from those based on the full dataset. Reasonable conjectures by researchers that selection bias would not affect results turn out, in

[4] We provide a detailed description of our hypothesis formation and what data analysis we conducted when discussing our empirical strategy in Section II.

[5] Similarly, those who *never* answer the phone may be different from the merely difficult to reach.

our data, to be false in important cases. Our analysis shows how taking advantage of difficulty-to-reach data when available can improve empirical evidence on happiness.

We conclude the paper in Section IV. We begin with discussing how our analysis might shed light on some past happiness findings. Clearly, nonresponse bias and the potential for using difficulty-of-reaching data are not per se connected to the study of happiness, and we also explore what our findings imply more broadly regarding the collection and analysis of survey data. Finally, moving beyond survey data, we comment on the interpretation of evidence from laboratory experiments, for example those that study cross-gender differences in behavior.

I. Difficulty-of-Reaching and Selection Bias

Within-sample indications of nonresponse bias are the subject of a large body of work. Widely cited contributions, both those focused on general formal statistical models and those focused on specific applications and survey-design implementations, are published in leading outlets such as *Journal of the American Statistical Association* and *Public Opinion Quarterly*. However, we are not familiar with any research or discussion along these lines within the happiness literature. In fact, with few exceptions the happiness literature has not engaged with nonresponse bias much at all.[6, 7]

Here we briefly review the papers on nonresponse bias that are most relevant to ours. We note that none of them studies subjective well-being, and that none of them focuses on empirically studying nonresponse bias that *differentially* affects different groups. Our focus on such differential effects is crucial: indeed, we find no *general* link between happiness and difficulty-of-reaching when our data are pooled together.

Focusing, as we do, on nonresponse that is caused by nonavailability, Potthoff, Manton, and Woodbury (1993) propose a model where an individual's probability of completing an interview is related to the outcome of interest. Participation probabilities are distributed in the population of potential respondents, and the authors use published number-of-call distributions (with mostly up to five calls) to estimate parameters of that distribution. Our paper can be viewed as a straightforward demonstration that the link between number of calls and an outcome of interest—in our case, self-reported happiness—not only exists but differs across groups. Our proposed correction is simpler and is based on more conservative assumptions than theirs.

[6] Of the four happiness papers cited in the introduction, for instance, only Stone et al. (2010) make an explicit reference to previous work on nonresponse bias, citing Curtin, Presser, and Singer (2005) as evidence suggesting that "nonresponse bias is not proportional to response rate"; for reasons argued in this article, we believe that neither Curtin, Presser, and Singer (2005) nor the papers they in turn cite provide much comfort on selection and nonresponse bias in comparing groups' happiness.

[7] While not explicitly engaging with the statistics and survey-design papers we review below, some economists' data collection efforts are designed to reduce potential nonresponse bias by implementing an approach that explicitly recognizes the potential link between outcomes and difficulty of reaching. Examples include Kling, Liebman, and Katz (2007) and Baird et al. (2011) who, to reduce potential bias due to attrition in panel data, concentrate resources on a special "second phase" tracking effort targeting a random subsample of panel participants who were not reached in a more standard "first phase" tracking effort (the resulting data are reweighted to account for this scheme). In addition to yielding relatively high "effective" (i.e., reweighted) response rate, this approach also yields a difficulty-of-reaching measure that these researchers can use to conduct analyses similar to ours. Recently, Behaghel et al. (2012) suggest that randomized tracking effort could, under certain conditions, be replaced with ex post information on difficulty of reaching (such as number of calls or visits) for dealing with survey attrition.

Lin and Schaeffer (1995) assess models such as Potthoff, Manton, and Woodbury (1993) in a specific application: they match external data from court records with data from a telephone survey regarding child-support awards and payments, and compare court amounts both between respondents with different numbers of calls and between respondents and nonrespondents. They find only a weak association between their variables of interest and number of calls. But they repeatedly emphasize the specific context of their exercise and "the somewhat specialized study design and the nature of this sample," to urge caution in drawing general lessons. Using a similar design in a similar context, Lin, Schaeffer, and Seltzer (1999) find differences between responding and nonresponding divorced fathers. For a recent meta-analysis of 59 studies that estimate nonresponse bias in specific applications, see Groves and Peytcheva (2008).

In a different strand of the literature, Keeter et al. (2000) run an experiment to assess the extent to which reducing nonresponse—by applying more "rigorous" methods—affects outcomes in a national survey. They conduct two telephone surveys, one over five days with response rate of 36 percent and one over eight weeks with response rate of 61 percent. The two surveys differ in several additional ways, but the authors conclude that, for the vast majority of the 91 questions they ask in both surveys, the difference in outcomes is small.[8] While they ask no happiness or other subjective well-being questions, their findings are generally in line with our finding of no whole-population link between difficulty of reaching respondents and happiness. Since they conduct no cross-group analysis, we do not know if links within specific subpopulations such as those we find would be found in their non-happiness questions.

Recently, Kreuter, Müller, and Trappmann (2010) link administrative data to the German Panel Study "Labor Market and Social Security" (PASS), and study a subsample of recent unemployment benefits recipients who were interviewed by telephone. They find that increasing the sample by including increasingly high number-of-call respondents reduces nonresponse bias in employment status and other outcomes. Like others, they point to their specific population as a possible weakness. While they note that their data do not allow for more than simple subgroup analysis, they report briefly that "[r]eplicating our analyses for subgroups defined by age and gender showed similar results to the overall analyses," but that some differences were found for one outcome variable (as their Appendix figures show).

Finally, two papers by Curtin, Presser, and Singer use SOC data predating the 2005–2010 data we use and shed light on important features of these data.[9] Curtin, Presser, and Singer (2000) study the impact of response rate on the widely monitored Index of Consumer Sentiment (ICS). Observing that "the mean number of calls to complete an interview more than doubled from 3.9 in 1979 to 7.9 in 1996, and interviews from refusal conversion likewise rose from 7.4 percent to 14.6 percent," they ask: "Would the survey results have been affected if

[8] Exceptions include demographics and frequency of volunteer work. Consistent with the latter, Abraham, Helms, and Presser (2009) find that Current Population Survey (CPS) respondents who also became respondents in the American Time Use Survey (ATUS, the sample for which is drawn from the CPS) volunteer much more than CPS respondents who became ATUS nonrespondents. They additionally find that this nonresponse bias in volunteering does not in general differentially affect different demographic groups.

[9] Their motivation to investigate nonresponse issues is also practical—Curtin was and is the director of the SOC.

this additional effort had not been made and lower response rates achieved?" Comparing easier-to-reach respondents with harder-to-reach respondents, they find that the latter have higher ICS scores (are more optimistic), but that the difference remains constant over time and, hence, has little effect on cross-time comparisons. However, they emphasize that the only association between variables they examine involves time, and, hence, that "further work on nonresponse's impact on associations between variables is needed." Such associations are at the core of the cross-group happiness comparisons we examine. To the best of our knowledge, we are the first to focus on such cross-group comparisons—not only in the happiness domain, but in any domain. In online Appendix A, we revisit Curtin, Presser, and Singer's (2000) conclusions regarding consumer sentiment. As an illustrative example, we repeat their analysis using the 2005–2010 data but do so separately for women and men, which had the most dramatic difficulty-of-reaching group-difference effects in our analysis of happiness. We replicate their findings of substantial aggregate difficulty-of-reaching effects but find little indication of systematic differential effects between women and men.

In a second paper, Curtin, Presser, and Singer (2005) examine nonresponse in the SOC from 1979 to 2003. They identify data errors that affected the conclusions of previous papers and report that from 1979 to 1996 the response rate declined from about 72 percent to 60 percent, and since then has been declining twice as fast, reaching 48 percent by 2003.[10] In the last paragraph of their paper, they conclude that "without better approaches to both contacting respondents and persuading them to be interviewed, the long-term future of telephone survey research does not appear promising." Such long-term trends of deterioration in response rates are likely to make the development of techniques to help deal with selection bias grow in importance.

Our study of group-differentiated nonresponse bias may be especially relevant in the happiness context both because much of the happiness literature is focused on group comparisons and because these may be especially subject to the effects we identify. Specifically, in stark contrast with the questions examined by the papers reviewed above—questions on demographics, political affiliation, concrete opinions, and behaviors such as volunteering, voting, child-support payment, and employment—a central feature of happiness and similar questions is how little stock researchers place on the cardinal value of the answers; ordinal rankings across groups or situations seem easier to interpret. But many of the same things that might lead to cross-group differences in happiness might also lead to differences in response rates; e.g., having children in households might affect happiness in either direction in ways that are directly related to how likely a person is to answer the phone. It is this relation between difficulty of reaching and the topic under investigation that underlies the bias.

Irrespective of the validity of these a priori arguments, we in fact find direct evidence of the effects of difficulty-of-reaching on conclusions regarding happiness.

[10] Of the 52 percent nonresponse cases in 2003, 27 percent were final refusals, around 16 percent were noncontacts, and most of the rest were considered "missed callbacks." Since 2003, response rate in the SOC kept decreasing, and was reported at 40 percent recently. In January 2009 the SOC started experimenting with adding a cell phone supplement to its landline sample; cell phone response rate was reported at 20 percent.

Free of further assumptions, the SOC and a survey that replicated it but stopped after fewer calls would reach different conclusions. We conclude this section with a stylized example of why it may be natural to expect misleading conclusions from sample averages. Potentially more importantly, the example shows how difficulty-of-reaching information can be used to bring either confidence or doubt to conclusions that rely on sample averages; and why using either difficult-to-reach respondents or even an extrapolation of within-sample trends may be more warranted than using sample averages when nonresponse is a problem.

Imagine a survey that made two call attempts to each potential respondent, where 1,000 Blue and 1,000 Red respondents were reached in one call attempt, and 800 additional respondents from each group were reached in exactly two calls. Imagine that among the Blue, 75 percent of those reached in one call, and 70 percent of those reached in two calls, are happy. Among the Red, the corresponding figures are 50 percent for both the first and second calls. A researcher conducting a typical analysis, with the implicit underlying no-nonresponse-bias model, would conclude that the Blue are (far) happier.

However, if within each group the happy and the unhappy each answer the phone in an i.i.d. manner and nobody refuses the survey once reached, then in fact the right conclusion would be that among the entire population the Blue are (far) *less* happy. Intuitively, unhappy Blue respondents must be dramatically less apt to answer than happy Blue in order for the proportions to drop that quickly, so we know that a higher percentage of unreached are unhappy. In fact, with the strong assumptions made, we know that only 32 percent of Blues are happy—rather than the sample average of 73 percent!

To see this, let N_B be the number of Blue potential respondents, h_B the true fraction of them who are happy, x_B the probability that a happy Blue answers the phone given a call attempt, and y_B the probability that an unhappy Blue answers given a call attempt. Then $N_B h_B x_B = 750\ (= 0.75 \times 1{,}000)$ report being happy on the first attempt and $N_B h_B (1 - x_B)\, x_B = 560\ (= 0.70 \times 800)$ report happy on the second attempt. The corresponding numbers of unhappy Blues are $N_B (1 - h_B)\, y_B = 250$ and $N_B (1 - h_B)\, (1 - y_B)\, y_B = 240$. Together these imply $x_B = 25\frac{1}{3}$ percent and $y_B = 4$ percent, and $h_B = 32$ percent. As another way to parse these numbers, we know that more than 44 percent of happy Blues have been reached after only two calls, but less than 8 percent of unhappy Blues have!

By contrast, we know that unhappy Reds answer at the same rate as happy Reds: there is no within-Red selection bias in happiness, and the lack of "trend" would rightly comfort researchers that it is really true that 50 percent of Reds are happy. Notice that the representativeness of the groups among the reached is not a test, and, hence, reweighting the sample using the survey weights is not a solution. In this example almost nothing is inferable about happiness from the average in-sample happiness. But everything can be inferred from a richer interpretation of the data that includes trends.

This, of course, is just illustrative, and more generally the inability to guess group difference from averages is not so dire, and the ability to do so from difficulty-of-reaching trends is not so powerful. In fact, our data are inconsistent with this and other natural simplified models, and we do not use this or any structural model in our empirical analysis. Instead, we show empirically that the

logical possibility of selection bias leading to misleading conclusions indeed manifests itself, and that using difficulty-of-reaching data seems to prove useful. And our extrapolation illustration at the end of Section III shows how conclusions can change even with the relatively conservative assumption that the levels among the unreached are more similar to the levels among the difficult to reach than to the levels among the *average* respondent.

II. Data and Empirical Strategy

We analyze data from the University of Michigan's Surveys of Consumers (SOC).[11] The SOC is a monthly telephone survey of about 500 individuals, conducted by the University's Survey Research Center from its Ann Arbor facility. The sample is designed to be representative of US households (that own a telephone) excluding Alaska and Hawaii. The survey incorporates a rotating panel, with each monthly sample including about 300 new respondents as well as 200 "reinterviews." Households are selected using list-assisted random digit dialing (landline only), and a single adult respondent (18 or older) is selected within each household using the Kish procedure. Reinterviews are respondents who were interviewed as new respondents six months earlier and are interviewed a second time.

Our data include the 64 full monthly samples from August 2005 through November 2010. Of the 32,250 interviews that were completed in this time period, the happiness question was asked in 31,227 and answered in 31,007 (99.3 percent of cases asked); of these, four have a missing number-of-calls variable.[12] Our main analysis is based on the remaining 31,003 observations, treated as independent (unweighted) cases. As alternative, robustness-check specifications—conducted well after we completed our main analysis—we also (a) allowed for clustering by respondent in our regressions, and (b) conducted our analysis on only new respondents (18,987 observations). Clustering has little effect on our findings—the increase in standard errors is often comparable to the size of our tables' rounding error. Excluding reinterviews—39 percent of the observations—has more impact on our estimates: while remaining generally similar, they become less precise; we report the relevant figure and tables in online Appendix B.

The happiness question is our main dependent variable. It was included in the SOC for the first time in August 2005 and was dropped in early 2011, when its funding ceased. It is based on the happiness measure in the Center for Epidemiologic Studies Depression Scale. Specifically, subjects were asked "Would you say yes or no?" to whether they agree with the statement "Much of the time during the past week, you were happy."

The main independent variable in our analysis—in addition to demographic variables—is the number of phone calls made in the course of completing each interview. For each monthly sample, call attempts start around the beginning of the month. Interviews are conducted seven days a week. Attempts to interview

[11] Our brief description in this section of the SOC design and methodology is based on the documentation available on the SOC website (www.sca.isr.umich.edu), as well as on Curtin, Presser, and Singer (2000, 2005), Kimball et al. (2006), and our conversations with Survey Research Center researchers.

[12] For budgetary reasons, the happiness question was not asked in reinterviews from September 2005 to January 2006.

unreached respondents continue with no limit on the number of calls but have to stop when a new cycle begins. Our respondent-level number-of-calls variable ranges from 1 to 82, with a mode at 1 (16 percent of respondents) that decays quickly and has a long tail. We provide further information—e.g., regarding the distribution of this variable in our data, and regarding demographic variables—in the next section.

Finally, we wish to provide information regarding the evolution of our project and analysis, to aid the reader in judging how compelling our evidence is and in interpreting the statistical significance of our results. Studying the issues discussed in this paper was not our original intent in looking at the SOC, and we began our analysis only after noticing data patterns indicating some sort of in-sample group-differentiated selection effects. We then started thinking about selection issues and subsequently looked for—and found—a difficulty-of-reaching measure: the number-of-calls variable. We then began to explore how this variable was related differently to the happiness of different groups. We looked at the four groups reported in this paper, since they have recently been the focus of demographic happiness comparisons. We started with neither a clear model nor a predetermined list of hypotheses. But we report comparisons for exactly the four groups we initially looked at. We considered exploring race and ethnicity but determined we did not have statistical power and have not done any statistical analysis on these categories. For the two continuous variables we look at, age and income, we originally compared those above and below the median, without looking at finer categorizations, and we report our results below in this original order. We originally examined most of the data without controls, and we report all such analyses. We then added the standard controls reported below, which we chose ahead of time. These controls mildly strengthened some of our results without much changing their qualitative nature. Having completed much of the analysis and after presenting the paper, we also briefly explored, graphically, two additional comparisons: married versus divorced, and those with versus without a college degree. We did not find more than nuanced differences across these groups. We did not pursue these explorations and assume that a more careful analysis would show little selection bias, much like one of our four included groups and the questions we report on in online Appendix A. Had we found dramatic results, we introspect that we would have likely added them to the paper. We would also like to think that we would have noted in that contingency that we would in *this* contingency have not included them. After completing a draft nearly identical to this one in which all regressions used OLS, we recently changed all regressions in our main analysis to use probit instead. The OLS regressions were common in this literature and allowed us to readily interpret regression coefficients as percentage-point differences in happiness across demographic cells, but we replaced them with probit at the convincing request of a referee. The original regressions are now reported in online Appendix D. This change had little quantitative or statistical effect and did not affect any of our interpretations.

III. Results

To facilitate the presentation of our findings, as well as to follow the order in which we conducted our analysis, we begin by illustrating our main results graphically. We compare the mean response to the happiness question across pairs of

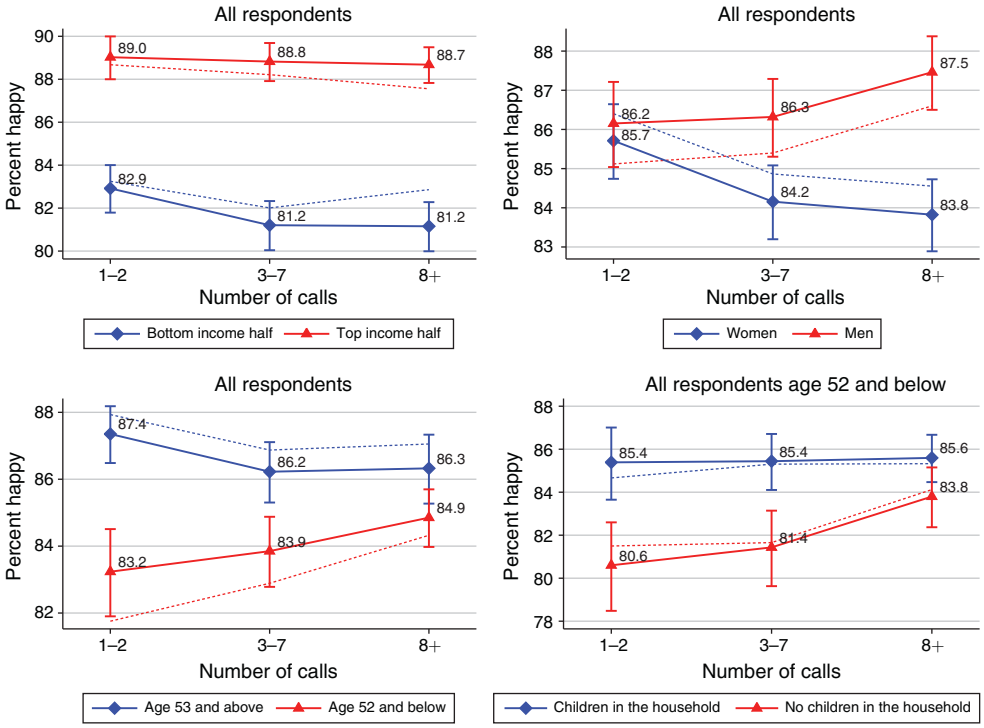

FIGURE 1. HAPPINESS BY RESPONDENT DEMOGRAPHICS AND NUMBER OF CALLS

Notes: Based on 31,003 observations divided into three groups by number of calls: 1–2 calls (9,263 observations), 3–7 (10,570), and 8 or more (11,170). Solid lines: unadjusted means. Capped ranges: 95 percent (binomial) confidence intervals. Dashed lines: adjusted means, using probit regressions with sets of indicator variables for race/ethnicity (5 categories), education (6), marital status (6), region (9), metropolitan status (5), sex (2; excluded in top right graph), income (5; excluded in top left graph), age (5; excluded in two bottom graphs), children in the household for respondents in bottom two age categories (2; excluded in two bottom graphs), and indicators for missing data.

Source: University of Michigan's Survey of Consumers, August 2005–November 2010.

complementary groups based on income, sex, age, and having children in the household. For simplicity, we aggregate the data into three large number-of-call categories, each consisting of roughly one-third of the data; and we report the above cross-group comparisons separately for each number-of-calls category. We then present more detailed analysis, with uncontrolled and controlled regressions comparing the same demographic groups but breaking down the variables age and income into quintiles rather than halves, and using nine rather than three number-of-call categories. We also report on tests that assess whether interactions of binary demographic variables with number-of-calls variables are statistically significant in happiness regressions that are based on the pooled data. We then replicate the uncontrolled and controlled regressions above using increasingly more inclusive subsets of our data, starting with the easiest-to-reach category and expanding the data to include increasingly harder-to-reach respondents. We close the section with a conservative form of out-of-sample extrapolations.

Figure 1 summarizes our main findings. Each of its four subplots reports—by two complementary demographic groups (diamonds versus triangles) and by three

categories of number-of-calls—the percent of respondents who responded "Yes" to the happiness question. The four pairs of demographic groups are: below- versus above-median income; women versus men; below- versus above-median age; and, among those below median age, whether or not they have children in the household.[13] The three number-of-call categories approximate, as much as the data allow, number-of-call tertiles for the entire sample: 30 percent of respondents were reached in one or two calls, 34 percent were reached in three to seven calls, and 36 percent were reached after eight or more calls.[14] Solid lines and capped ranges report uncontrolled means and 95 percent confidence intervals. Dashed lines report controlled means; the controls are listed in the figure notes and are discussed below.

The figure gives the simplest indication of how group differences in measured happiness depend on difficulty to reach respondents. We start with income: the figure's top left plot compares those with above-median income ($60,000 or more; triangles) with those below median (diamonds). The differences in happiness levels between the groups are much larger than the potential differences in difficulty-of-reaching trends; clearly, those with higher income are happier both in-sample and out-of-sample.

The figure's top right plot compares men (triangles) with women (diamonds). While in the easiest-to-reach category there is essentially no happiness-sex gap—0.4 not-statistically-significant percentage points ($p = 0.55$)—in the middle category men are 2.2 points happier than women ($p = 0.002$), and in the hardest-to-reach category men are 3.6 points happier ($p = 0.000$). The increasing gap reflects both an increase in men's happiness and a decrease in women's happiness as difficulty-to-reach increases: while men with 8-or-more calls are happier than men with 1–2 calls ($p = 0.07$), the opposite is true among women ($p = 0.005$). In sum, within our sample, the existence of a happiness-sex gap depends on the number-of-calls category one looks at. Below we show that once covariates are controlled for the *direction* of the gap depends on number-of-calls category.

The two bottom plots show that for the two other groups happiness gaps shrink with number of calls. The bottom left plot shows that while those above median age (diamonds) are happiest when easier to reach, those below the median (triangles) are happiest when harder to reach. The final plot shows that this increase among those below median age is driven almost entirely by those with children in their household: happiness among those without children appears virtually the same regardless of difficulty-to-reach. In both plots, a happiness gap exists regardless of

[13] Our income variable is based on the question: "To get a picture of people's financial situation we need to know the general range of income of all people we interview. Now, thinking about (your/your family's) total income from all sources (including your job), how much did (you/your family) receive in <previous year>?" Income is coded in (nominal) dollars. Our children-in-the-household variable is based on the question: "How many members of your household are 17 years of age or younger?" We limit our children-based comparisons to respondents below median age or, in the controlled regressions, to those in the bottom two age quintiles (while 58 percent of respondents below median age had children in household, only 7 percent of those above the median age did). For a discussion of the "appropriate" age group to look at when comparing those with and without children, see Herbst and Ifcher (2011).

[14] With the notable exception of the age-based groups, these three number-of-call categories also roughly approximate number-of-call tertiles within each demographic group, and, hence, the demographic composition of the number-of-call categories does not very dramatically. Specifically, the following groups account for the following percentages of respondents within the three number-of-call categories, respectively: top income: 41.8, 47.0, and 51.1 percent; women: 57.1, 55.5, and 56.2; young: 34.8, 45.8, and 60.6 (i.e., the young are systematically harder to reach than the old); children among the young: 54.6, 59.9, and 59.1.

TABLE 1—DEMOGRAPHIC HAPPINESS GAPS BY NUMBER OF CALLS (*Separate*)

Number of calls:	1	2	3	4	5–6	7–9	10–13	14–20	21+
Inc. below $30,000	−0.096*** (0.016)	−0.087*** (0.017)	−0.094*** (0.022)	−0.070*** (0.026)	−0.116*** (0.021)	−0.090*** (0.022)	−0.150*** (0.024)	−0.111*** (0.024)	−0.093*** (0.022)
Inc. $30,000–49,999	−0.041** (0.016)	−0.013 (0.017)	−0.014 (0.021)	0.011 (0.023)	−0.050** (0.020)	−0.042** (0.021)	−0.026 (0.022)	−0.030 (0.022)	−0.041* (0.021)
Inc. $75,000–109,999	−0.007 (0.016)	0.019 (0.016)	0.042** (0.019)	0.029 (0.023)	−0.004 (0.018)	0.030 (0.018)	0.019 (0.020)	0.025 (0.019)	0.008 (0.018)
Inc. $110,000 and up	0.006 (0.016)	0.019 (0.016)	0.038** (0.019)	0.042* (0.022)	0.008 (0.017)	0.022 (0.018)	0.012 (0.019)	0.052*** (0.018)	0.035** (0.017)
Female	−0.014 (0.010)	0.008 (0.010)	−0.022* (0.013)	−0.032** (0.014)	−0.020* (0.012)	−0.013 (0.012)	−0.045*** (0.013)	−0.039*** (0.013)	−0.039*** (0.012)
Age 18–39	0.030 (0.019)	0.002 (0.018)	0.033 (0.021)	−0.004 (0.023)	0.033* (0.018)	0.024 (0.018)	0.022 (0.020)	0.019 (0.019)	0.010 (0.017)
Age 40–49	0.005 (0.019)	−0.032* (0.018)	−0.001 (0.022)	−0.002 (0.024)	0.000 (0.019)	0.009 (0.019)	0.027 (0.021)	0.009 (0.019)	−0.011 (0.019)
Age 60–69	0.049*** (0.016)	0.021 (0.016)	0.055*** (0.019)	0.033 (0.022)	0.056*** (0.018)	0.037* (0.019)	0.045** (0.023)	0.066*** (0.021)	0.040* (0.020)
Age 70 and up	0.081*** (0.015)	0.039*** (0.015)	0.057*** (0.019)	0.052** (0.021)	0.031* (0.019)	0.017 (0.021)	0.045* (0.024)	0.052** (0.025)	0.018 (0.026)
Age $\leq$ 49 and children	0.046** (0.021)	0.070*** (0.021)	0.046** (0.023)	0.032 (0.026)	0.030 (0.020)	0.013 (0.019)	0.024 (0.020)	−0.004 (0.018)	0.015 (0.017)
Age $>$ 50	0.058*** (0.018)	0.080*** (0.019)	0.049** (0.020)	0.051** (0.023)	0.029 (0.019)	0.008 (0.017)	0.015 (0.019)	0.012 (0.017)	0.024 (0.016)
Observations	4,974	4,289	3,140	2,473	3,612	3,457	2,978	2,916	3,164

Notes: Dependent variable: "Much of the time during the past week, you were happy. Would you say yes or no?" Each column reports average marginal effects from four different probit regressions, separated by horizontal lines. Regressions also include indicators for missing data (not reported). Regressions within a column are conducted on the same subsample, based on number-of-calls category indicated in top row; size of each subsample is indicated in bottom row (up to three observations are dropped in children regressions due to missing-data indicator perfectly predicting dep. variable). Standard errors in parentheses.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: University of Michigan's Survey of Consumers, August 2005–November 2010.

number-of-calls category (all six pairwise comparisons' p-values < 0.05); but its size among those with eight or more calls is less than half its size among those with 1–2 calls.

The dashed lines in all four plots tell a qualitatively similar story: while adding controls shifts the levels, the trends are less affected. Interestingly, in the sex plot these shifts generate a crossing of the curves: with controls, easy-to-reach women are happier than easy-to-reach men, but difficult-to-reach men are happier than difficult-to-reach women. We discuss this reversal of the controlled happiness-sex gap, as well as its statistical significance, below.

Table 1 reports average marginal effects from 4×9 separate probit regressions of happiness on sets of demographic indicators. Each column summarizes four regressions, separated by horizontal lines; each of the four regressions is based on the same number of observations, reported at the bottom row of the table. The columns correspond to nine number-of-call categories—1, 2, 3, 4, 5–6, 7–9, 10–13, 14–20, and 21-or-more—chosen to approximate number-of-call deciles; however, since the first two categories together account for roughly 30 percent of the data, there are only nine categories in total.

We start again with income. Each of the nine income regressions has four (rounded) income quintiles, and a constant, as regressors; the middle quintile—$50,000–74,999—is omitted (and is, hence, the base category for the marginal effects). We highlight two findings. First, consistent with past findings using a measure of positive affect (Kahneman and Deaton 2010), happiness increases steeply from the bottom quintile to the second quintile, but then increases less and less steeply, and—with the possible exception of the two rightmost columns—we find little evidence of further increases beyond $75,000.[15] Second, consistent with Figure 1 above, the happiness-income gap generally increases as the number of calls increases. The increase is far from monotonic, however, and is too modest to dramatically affect any happiness-income conclusions.

The nine sex regressions—with a Female indicator and a constant as regressors—also confirm what Figure 1 suggested. In the two easiest-to-reach categories the difference between men's and women's happiness is 1.4 and −0.8 percentage points and is statistically insignificant in spite of the fact that these two categories include on average 50 percent more observations than other categories. But the difference grows to 3.9–4.5 percentage points in the three hardest-to-reach categories, statistically significant at the 1 percent level in each of the three. The 95 percent confidence intervals around the estimates in the two leftmost categories are $[-3.4, 0.5]$ and $[-1.3, 2.8]$. The point estimates in each of the three rightmost categories lie outside of either interval.

The age regressions are again consistent both with past findings and with Figure 1 above. Indeed, the overall age-happiness pattern that emerges resembles that reported in Stone et al. (2010, Figure 2): like these authors, we find a U-shaped pattern with its nadir located in the 50s (the omitted quintile). Yet, the happiness lead of those at higher age generally shrinks with number-of-calls category. Moreover, our point estimates suggest that while in the four leftmost categories those at age 70 and up are 0.2–3.2 points happier than those at age 60–69—in line with Stone et al. (2010), whose data resulted from up to five call attempts—in the five rightmost categories those 70 and up are 0–2.5 points *less* happy.[16]

Finally, the fourth set of regressions reveals that among those below age 50, respondents with children are dramatically happier than those without children (the omitted category)—by 4.6–7.0 percentage points, significant at the 5 percent level—in the three easiest-to-reach categories; are somewhat happier—by 1.3–3.2 points, not statistically significant—in the three middle categories; and are only slightly happier on average—by −0.4–2.4 points, not significant—in the three hardest-to-reach categories. Hence, difficulty-to-reach strongly affects the size of the estimated children-happiness gap.

To directly test the statistical significance of these effects, we pool all our data and—separately for each of the four demographic groups—use probit to regress

[15] The relevant happiness question in Kahneman and Deaton (2010) is a yes/no "happiness" item on a list of feelings following the preamble: "Did you experience the following feelings during a lot of the day yesterday?" Summarizing the relevant findings from their positive affect measure—which averages the fractions reporting happiness, smiling, and enjoyment—and two other measures, they write: "We infer that beyond about $75,000/y, there is no improvement whatever in any of the three measures of emotional well-being."

[16] This latter result is not a central one in Stone et al. (2010); and, depending on how the data are aggregated, its apparent reversal in our data may not be statistically strong.

TABLE 2—DEMOGRAPHIC HAPPINESS GAPS BY NUMBER OF CALLS (*Separate*)

Number of calls:	1	2	3	4	5–6	7–9	10–13	14–20	21+
Inc. below $30,000	−0.096***	−0.068***	−0.066***	−0.054**	−0.091***	−0.092***	−0.105***	−0.069***	−0.057**
	(0.018)	(0.019)	(0.024)	(0.027)	(0.022)	(0.024)	(0.026)	(0.026)	(0.023)
Inc. $30,000–49,999	−0.045***	−0.005	−0.008	0.013	−0.044**	−0.047**	−0.016	−0.019	−0.040*
	(0.016)	(0.017)	(0.022)	(0.023)	(0.020)	(0.021)	(0.022)	(0.023)	(0.021)
Inc. $75,000–109,999	−0.003	0.021	0.048**	0.023	−0.008	0.027	0.009	0.026	0.003
	(0.016)	(0.016)	(0.020)	(0.023)	(0.018)	(0.018)	(0.020)	(0.020)	(0.019)
Inc. $110,000 and up	0.013	0.019	0.048**	0.034	0.002	0.017	−0.008	0.048**	0.029
	(0.016)	(0.017)	(0.020)	(0.023)	(0.018)	(0.019)	(0.021)	(0.019)	(0.018)
Female	0.002	0.026**	−0.004	−0.017	−0.008	0.001	−0.025*	−0.017	−0.029**
	(0.010)	(0.011)	(0.013)	(0.015)	(0.012)	(0.012)	(0.014)	(0.013)	(0.013)
Age 18–39	0.006	−0.044*	0.021	−0.003	0.026	0.037	0.011	0.033	−0.003
	(0.025)	(0.026)	(0.027)	(0.031)	(0.025)	(0.024)	(0.026)	(0.024)	(0.021)
Age 40–49	−0.019	−0.074***	−0.025	−0.023	−0.018	0.013	0.006	0.011	−0.029
	(0.023)	(0.024)	(0.027)	(0.031)	(0.025)	(0.024)	(0.025)	(0.024)	(0.021)
Age 60–69	0.066***	0.037**	0.065***	0.047**	0.070***	0.057***	0.057***	0.083***	0.049**
	(0.016)	(0.015)	(0.019)	(0.022)	(0.018)	(0.019)	(0.022)	(0.021)	(0.019)
Age 70 and up	0.121***	0.072***	0.093***	0.097***	0.064***	0.062***	0.080***	0.089***	0.046*
	(0.016)	(0.015)	(0.019)	(0.022)	(0.020)	(0.021)	(0.023)	(0.024)	(0.025)
Age $\leq$ 49 and children	0.024	0.043***	0.028	0.015	0.016	−0.001	0.024	−0.015	0.009
	(0.017)	(0.016)	(0.020)	(0.023)	(0.019)	(0.019)	(0.020)	(0.019)	(0.017)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,965	4,280	3,135	2,471	3,603	3,449	2,978	2,916	3,151

Notes: Dependent variable: "Much of the time during the past week, you were happy. Would you say yes or no?" Average marginal effects from probit regressions, conducted on subsamples based on number-of-calls category. All regressions include the following additional control variables (not reported): sets of indicators for race/ethnicity (five categories), education (6), marital status (6), region (9), metropolitan status (5), and indicators for missing data (if indicator perfectly predicts dependent variable, observations are dropped). Standard errors in parentheses.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: University of Michigan's Survey of Consumers, August 2005–November 2010.

happiness on a binary demographic indicator (coded as in Figure 1), number-of-call variables, and interactions. Using different number-of-call measures, we find as follows (regressions are not reported): with a binary measure (one to four calls versus five or more), the interaction term is statistically significant ($p < 0.05$) in the sex and in the age regressions, and $p < 0.1$ in the income and in the children regressions; with a continuous measure—log(number of calls)—the interaction is significant ($p < 0.05$) in the sex and in the children regressions, and $p = 0.06$ in the age regression; with three number-of-call categories (as in Figure 1), the interaction terms are jointly significant ($p < 0.05$) in the sex and in the age regressions; with nine categories (as in Table 1), the interactions are (just) jointly significant ($p = 0.05$) only in the sex regression. Adding controls (as in Figure 1) to the regressions does not change the general picture.[17]

Table 2 is similar to Table 1, with the exception that each column reflects one probit regression rather than four. Its regressors include all the regressors from Table 1,

[17] Specifically, in the controlled regressions the income interactions are not statistically significant in any of the four specifications; the female interactions are significant ($p < 0.02$) in all four specifications; the age interactions are significant ($p < 0.01$) in all but the nine-categories specification, where $p = 0.051$; and the children interactions have $p = 0.08$ and $p = 0.09$, respectively, in the binary and log specifications.

as well as an additional set of indicators (not reported) for race/ethnicity (five categories), education (6), marital status (6), region (9), metropolitan status (5), and additional indicators for missing data; see online Appendix C for definition of these control variables.

Comparing the income coefficients in Table 2 with those in Table 1 confirms that the additional controls do not qualitatively change the main happiness-income findings discussed above. Similarly, comparing the female coefficients confirms that the added controls do not change the finding that the happiness-sex gap varies dramatically with number of calls. In fact, as discussed in the context of Figure 1, in the controlled regressions not only the size, but also the direction of the happiness-sex gap depends on difficulty to reach respondents.

Finally, although both the age and the children coefficients from Table 1 change once the indicators for age, children, and other variables are all included in one regression, the finding above that difficulty of reaching respondents could strongly affect cross-group conclusions is seen to hold. For example, while the difference in happiness between age 70-and-up and age 60–69 in the four leftmost categories is a sizable 2.8–5.5 percentage points, it shrinks in the five rightmost categories to an insignificant difference, including two sign-changes, ranging -0.6–2.3 points. Similarly, within the two bottom age quintiles, those with children are 2.4–4.3 points happier than those without them in the three leftmost categories, but the gap shrinks to -0.1–1.6 points in the three middle categories, and to -1.5–2.4 in the three rightmost categories.

Tables 3 and 4 are similar to Tables 1 and 2, but with regressions based on increasingly large subsamples of number-of-calls categories. The leftmost column in each of the tables is identical to the leftmost column in Tables 1 and 2, respectively, and is based on only respondents with one call. However, the next column is based on those with *either* one or two calls; the third column is based on those with 1–3 calls; and so on, with the rightmost column based on the entire sample.

Consider the sex regressions. Starting with the Female row in Table 3, it shows that if one stopped calling respondents after one, two, or even three calls (reaching the 40 percent easiest-to-reach of our respondents) one would fail to reject the hypothesis that women are as happy as men. But if one stopped collecting data after four, six, or nine calls, one would reject equality across men and women and estimate women around 1.4 percentage points less happy than men. If one further continued making calls, one's estimate of the happiness gap would keep increasing and would eventually lie outside the 95 percent confidence intervals around the two- and the three-or-fewer-calls estimates ($[-1.9, 1.0]$ and $[-2.1, 0.4]$ respectively). Table 4 once more illustrates how not only the magnitude, but also the sign of the estimated happiness-sex gap could change as survey response rate increases. Someone who stopped calling respondents after two calls would estimate in the controlled regression that women are 1.3 percentage points *happier* than men ($p = 0.09$), with a 95 percent confidence interval $[-0.2, 2.8]$. As the sample includes increasingly harder-to-reach respondents, however, this estimate monotonically shrinks. Although the -0.6 percentage point estimate based on the entire sample is not statistically different from zero, it lies outside the above two-calls-or-less confidence interval.

Trivially, the 2.2 percentage-point estimate in the rightmost column of Table 3 and the corresponding 0.6 estimate in Table 4 are the population estimates of the

TABLE 3—DEMOGRAPHIC HAPPINESS GAPS BY NUMBER OF CALLS (*Cumulative*)

Number of calls:	1	≤ 2	≤ 3	≤ 4	≤ 6	≤ 9	≤ 13	≤ 20	All
Inc. below $30,000	−0.096*** (0.016)	−0.092*** (0.012)	−0.092*** (0.010)	−0.089*** (0.010)	−0.094*** (0.009)	−0.093*** (0.008)	−0.099*** (0.008)	−0.100*** (0.007)	−0.099*** (0.007)
Inc. $30,000–49,999	−0.041** (0.016)	−0.028** (0.011)	−0.024** (0.010)	−0.018** (0.009)	−0.024*** (0.008)	−0.027*** (0.008)	−0.026*** (0.007)	−0.027*** (0.007)	−0.028*** (0.007)
Inc. $75,000–109,999	−0.007 (0.016)	0.005 (0.011)	0.015 (0.010)	0.017* (0.009)	0.013 (0.008)	0.015** (0.007)	0.016** (0.007)	0.017*** (0.006)	0.016*** (0.006)
Inc. $110,000 and up	0.006 (0.016)	0.013 (0.011)	0.019** (0.010)	0.023*** (0.009)	0.020** (0.008)	0.020*** (0.007)	0.019*** (0.007)	0.023*** (0.006)	0.024*** (0.006)
Female	−0.014 (0.010)	−0.004 (0.007)	−0.009 (0.006)	−0.013** (0.006)	−0.014*** (0.005)	−0.014*** (0.005)	−0.018*** (0.004)	−0.020*** (0.004)	−0.022*** (0.004)
Age 18–39	0.030 (0.019)	0.017 (0.013)	0.021* (0.011)	0.016* (0.010)	0.020** (0.009)	0.021*** (0.008)	0.021*** (0.007)	0.021*** (0.007)	0.020*** (0.006)
Age 40–49	0.005 (0.019)	−0.013 (0.013)	−0.009 (0.011)	−0.008 (0.010)	−0.006 (0.009)	−0.003 (0.008)	0.001 (0.008)	0.002 (0.007)	0.001 (0.007)
Age 60–69	0.049*** (0.016)	0.035*** (0.011)	0.040*** (0.010)	0.039*** (0.009)	0.042*** (0.008)	0.041*** (0.007)	0.042*** (0.007)	0.044*** (0.007)	0.043*** (0.006)
Age 70 and up	0.081*** (0.015)	0.061*** (0.010)	0.061*** (0.009)	0.059*** (0.008)	0.055*** (0.008)	0.051*** (0.007)	0.051*** (0.007)	0.051*** (0.006)	0.048*** (0.006)
Age ≤ 49 and children	0.046** (0.021)	0.058*** (0.015)	0.054*** (0.012)	0.050*** (0.011)	0.046*** (0.010)	0.040*** (0.009)	0.037*** (0.008)	0.031*** (0.007)	0.029*** (0.007)
Age > 50	0.058*** (0.018)	0.068*** (0.013)	0.063*** (0.011)	0.061*** (0.010)	0.055*** (0.009)	0.046*** (0.008)	0.043*** (0.007)	0.038*** (0.007)	0.035*** (0.006)
Observations	4,974	9,263	12,403	14,876	18,488	21,945	24,923	27,839	31,003

Notes: Dependent variable: "Much of the time during the past week, you were happy. Would you say yes or no?" Each column reports average marginal effects from four different probit regressions, separated by horizontal lines. Regressions also include indicators for missing data (not reported). Regressions within a column are conducted on the same subsample, based on number-of-calls categories of increasing size, indicated in top row; size of each subsample is indicated in bottom row (up to nine observations are dropped in children regressions due to missing-data indicator perfectly predicting dependent variable). Standard errors in parentheses.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Source: University of Michigan's Survey of Consumers, August 2005–November 2010.

happiness-sex gap *under the assumption that unreached respondents look like the average reached respondent*. But the evidence we presented so far—that even among reached respondents, the "easy" may not look like the "difficult"—casts doubt on this assumption. Moreover, our analysis so far is virtually assumption-free: we report what estimates would look like if one stopped calling potential respondents after fewer call attempts.[18]

We conclude our analysis with a more "daring" question regarding counterfactuals: what would estimates look like if the sample included respondents who are harder to reach than even the hardest-to-reach respondents in the data? Although answering this question requires making an assumption that is not directly testable, we contrast it with the typical daring assumption that almost all researchers implicitly make—that there is random selection into their dataset. What if we replaced this implicit assumption with the alternative assumption that the average nonresponder is similar to the average *harder-to-reach* respondent? Of course, even if seemingly conservative

[18] Of course, excluding respondents with number-of-calls above a certain cutoff from the analysis is not necessarily the same as conducting a survey that is designed to stop calling after that cutoff.

TABLE 4—DEMOGRAPHIC HAPPINESS GAPS BY NUMBER OF CALLS (*Cumulative*)

Number of calls:	1	≤ 2	≤ 3	≤ 4	≤ 6	≤ 9	≤ 13	≤ 20	All
Inc. below $30,000	−0.096***	−0.084***	−0.081***	−0.077***	−0.080***	−0.082***	−0.084***	−0.083***	−0.081***
	(0.018)	(0.013)	(0.011)	(0.011)	(0.010)	(0.009)	(0.008)	(0.008)	(0.008)
Inc. $30,000–49,999	−0.045***	−0.027**	−0.023**	−0.017*	−0.023***	−0.025***	−0.024***	−0.023***	−0.025***
	(0.016)	(0.012)	(0.010)	(0.009)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)
Inc. $75,000–109,999	−0.003	0.008	0.018*	0.019**	0.013*	0.015**	0.014**	0.016**	0.014**
	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)
Inc. $110,000 and up	0.013	0.016	0.023**	0.025***	0.021**	0.020***	0.016**	0.020***	0.021***
	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)
Female	0.002	0.013*	0.009	0.005	0.002	0.002	−0.001	−0.003	−0.006
	(0.010)	(0.008)	(0.007)	(0.006)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)
Age 18–39	0.006	−0.017	−0.006	−0.008	0.000	0.007	0.007	0.012	0.011
	(0.025)	(0.018)	(0.015)	(0.014)	(0.012)	(0.011)	(0.010)	(0.009)	(0.008)
Age 40–49	−0.019	−0.046***	−0.040***	−0.039***	−0.034***	−0.026**	−0.022**	−0.017*	−0.017**
	(0.023)	(0.017)	(0.014)	(0.013)	(0.011)	(0.010)	(0.009)	(0.009)	(0.008)
Age 60–69	0.066***	0.051***	0.055***	0.053***	0.056***	0.056***	0.057***	0.058***	0.056***
	(0.016)	(0.011)	(0.010)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)
Age 70 and up	0.121***	0.097***	0.097***	0.097***	0.092***	0.089***	0.089***	0.088***	0.084***
	(0.016)	(0.011)	(0.009)	(0.009)	(0.008)	(0.007)	(0.007)	(0.007)	(0.006)
Age ≤ 49 and children	0.024	0.034***	0.032***	0.029***	0.027***	0.022***	0.022***	0.017***	0.016**
	(0.017)	(0.012)	(0.010)	(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.006)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,965	9,245	12,380	14,851	18,454	21,903	24,923	27,839	31,003

Notes: Dependent variable: "Much of the time during the past week, you were happy. Would you say yes or no?" Average marginal effects from probit regressions, conducted on subsamples based on number-of-calls categories of increasing size. All regressions include the following additional control variables (not reported): sets of indicators for race/ethnicity (five categories), education (6), marital status (6), region (9), metropolitan status (5), and indicators for missing data (if indicator perfectly predicts dependent variable, observations are dropped). Standard errors in parentheses.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Source: University of Michigan's Survey of Consumers, August 2005–November 2010.

(in contrast, say, to the assumption that the happiness/difficulty-of-reaching trends we observe in sample continue out of sample), our alternative assumption regarding the unreached might be wrong. The unreached may be inherently unreachable and thus inherently different from any number-of-calls-based subpopulation of current respondents. Moreover, estimates based on only harder-to-reach respondents come with wider confidence intervals. Nonetheless, when difficulty-of-reaching data are available, projections based on the no-selection-bias assumption may be less attractive than the extrapolations we now turn to.

As a concrete simple example, we again focus on the happiness-sex gap and explore the implications of the assumption that unreached (non)respondents are similar to the hardest-to-reach half of our respondents. We start with the no-controls regressions. As can be calculated from Table 1, the average happiness-sex gap among the 52 percent of respondents with five or more calls is 3.0 percentage points in favor of men. Since among our entire sample the average gap is 2.2 points, if response rate is 50 percent then the average gap among the entire population under our assumption is roughly $\frac{2.2 + 3.0}{2} = 2.6$ points.[19] In the context of our discussion

[19] Assuming that response rate in our data is 50 percent is, again, quite conservative. See Curtin, Presser, and Singer (2005) for response rate estimates under different definitions and assumptions.

of the estimates in Table 3, this estimate of the gap is larger than the estimate based on the entire sample, and it lies well outside the 95 percent confidence intervals of most estimates based on up to the 70 percent of respondents who completed an interview within nine calls. This suggests that if SOC interviewers stopped calling potential respondents after nine calls, the standard assumption of no selection bias would lead to an estimate of the gap that is roughly half the size of our 2.6-points estimate, with a confidence interval that rejects our estimate. Conducting a similar exercise with the controlled regressions in Tables 2 and 4 yields an estimate that among the hardest-to-reach 52 percent of respondents men are 1.5 percentage points happier than women. Using our extrapolation method, we would estimate that among the entire population men are 1.0 points happier. This estimate is larger than the 0.6 points estimate based on our entire sample, and is opposite in sign to—and outside the 95 percent confidence intervals around—the point estimates based on respondents reached in nine or fewer calls.

IV. Discussion and Conclusion

Declining response rates over time suggest that selection bias in cross-group happiness comparisons might become ever more severe, and analysis such as ours might prove ever more useful.[20] Our findings suggest the usefulness of reporting difficulty-of-reaching data as the SOC does, and more generally of making information about calling algorithms more easily accessible to researchers. They also suggest the value of simply making more call attempts; when big enough differences across groups in happiness/number-of-calls correlations are found, extra caution should be taken.

Although a systematic examination of its implications for happiness research is beyond the scope of this paper, our analysis might shed light on some of the past findings. For instance, some of the most recent findings in the cross-group happiness literature use data from the Gallup-Healthways Well-Being Index (GHWBI), a daily telephone survey of US residents with rather low response rate. Describing the GHWBI data, Kahneman and Deaton (2010) write: "Up to five callbacks were made in the case of no answer. Of all calls that resulted in contacts with an eligible candidate, 31 percent of the candidates agreed to be interviewed; of these, 90 percent completed the entire interview. Despite the sampling limitations, available evidence suggests that the estimates of population parameters were not compromised; for example, the survey predicted recent election results within an acceptable margin of error." Stone et al. (2010), who also use the GHWBI, write: "Up to three to five callbacks were made in the case of no answer. … Although a higher response rate would be preferred, we note that studies suggest that nonresponse bias is not proportional to response rate." They cite Curtin, Presser, and Singer (2005) as evidence supporting the last sentence. Our analysis suggests caution in using findings like Curtin, Presser, and Singer (2005) to alleviate worries regarding nonresponse bias in *cross-group* outcomes of the type that these happiness papers explore and report on.

[20] Moreover, we demonstrate the possibility that even when response rate does not seem to affect mean outcomes—as has been found in some of the above-cited empirical work examining nonresponse—it could still affect differences in outcomes across groups.

Moreover, our analysis may be helpful in suggesting *which* group comparisons might be less or more sensitive to nonresponse bias. For example, we find no evidence that income comparisons such as those reported by Kahneman and Deaton (2010) are affected by nonresponse. At the same time, and although very far from being central results in their paper, Kahneman and Deaton (2010) also report (in their Table 1) that, e.g., those age 60 or above report higher positive affect than those below 60, and, to a much lesser extent, that women report higher positive affect relative to men. These results may be more consistent with early respondents in our data than with later respondents or, indeed, than with our entire sample. Although we do not find old-versus-young happiness reversals in our data, we find enough of a trend to worry that cross-age differences could be reversed with more complete samples. Our finding of a female-male reversal obviously gives even more cause for worry.

This female-male reversal may provide some insight beyond the way it identifies nonresponse bias. It suggests that busier women are less happy than less busy women, while the opposite is true for men, among whom the busier are happier. If this reflects a causal relationship, and if it happened that the aspects of busyness that differentially affect women and men's happiness increased among the population over time, then this would predict a trend in both real and measured cross-sex differences in happiness: women would be decreasingly happy, over time, both absolutely and relative to men. Stevenson and Wolfers (2009) find such trends in the GSS and other surveys—trends that our findings may help explain.[21, 22]

More broadly, our findings suggest that users of survey data—not just about happiness—should look at information such as number of calls as part of their regular data analysis routine. When trends are found, we would be hard pressed to think of reasons to use average respondents as a better proxy for missing respondents than the hard-to-reach. Going back to the data from existing surveys and reanalyzing them by difficulty-of-reaching might be one easy robustness check on existing results, especially when response rate is low.

Such recommendations should be neither hard nor expensive to implement. In telephone surveys, paradata such as number of calls either are already recorded automatically by computerized systems or could be recorded almost costlessly. Such data should be made readily available to researchers, and researchers should demand them. For other modes of interview, other data might prove useful. In the case of web surveys, researchers could perhaps look at information such as number of invitations and reminders sent, or time between invitation and participation—however, introspection and empirical evidence would need to be employed to determine how to interpret any trends found. More generally, our findings may suggest considering

[21] Indeed, what Stevenson and Wolfers refer to as "the paradox of declining female happiness" would then be intertwined with "the paradox of cross-sex differences in the relationship between happiness and busyness" generated by our data.

[22] Calibrationwise, consider the following very rough calculation. The SOC's mean number of calls in 1979 was 3.9—less than half our sample mean of 8.3. In Table 4, the column that gets closest to the 1979 mean is the "≤ 9" column, with mean 3.6. Under assumptions in the spirit of those mentioned above, the busyness channel alone would, hence, imply a change in the Female coefficient from 0.2 to -0.6 over roughly the 30-year period studied by Stevenson and Wolfers (2009). Of course, this is only illustrative. Neither of the coefficients is statistically significant, so nonresponse bias may explain none of the trend or much more of it. And, clearly, comparing coefficient sizes across surveys and happiness questions is tenuous.

a broader adoption of standards of reporting the details of surveys' sample design, mode of interviewing, response rate, number of calls/reminders, etc.[23]

Finally, our findings could have implications that go beyond survey data. Economists' increasingly rich set of evidence from lab experiments most often relies on small samples of recruited subjects—often volunteer students—with little attention to potential selection issues. Are the differences, e.g., in risk attitudes or social preferences between volunteer wealthy and volunteer poor subjects a reliable predictor of the differences between rich and poor among the rest of the population? Wealthy subjects who volunteer to participate in an experiment may be among the most intellectually curious, while poor participants may be those most in need of money. Similarly, do cross-gender differences in lab behavior carry over to nonparticipants? If gender differences in lab behavior are interpreted as evidence of different underlying preferences or experience across men and women, how justified is it to assume at the same time that men and women do not differ regarding how they select into participation? Analysis of speed-of-signup, for example, may provide a lens into laboratory selection bias.[24]

REFERENCES

Abraham, Katharine G., Sara E. Helms, and Stanley Presser. 2009. "How Social Processes Distort Measurement: The Impact of Survey Nonresponse on Estimates of Volunteer Work." *American Journal of Sociology* 114 (4): 1129–65.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel. 2011. "Worms at Work: Long-run Impacts of Child Health Gains." Unpublished.

Behaghel, Luc, Bruno Crepon, Marc Gurgand, and Thomas Le Barbanchon. 2012. "Please Call Again: Correcting Non-response Bias in Treatment Effect Models." Institute for the Study of Labor Discussion Paper 6751.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64 (4): 413–28.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2005. "Changes in Telephone Survey Nonresponse Over the Past Quarter Century." *Public Opinion Quarterly* 69 (1): 87–98.

Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72 (2): 167–89.

Heffetz, Ori, and Matthew Rabin. 2013. "Conclusions Regarding Cross-Group Differences in Happiness Depend on Difficulty of Reaching Respondents: Dataset." *American Economic Review*. http://dx.doi.org/10.1257/aer.103.7.3001.

Herbst, Chris M., and John Ifcher. 2011. "A Bundle of Joy: Does Parenting Really Make Us Miserable?" Unpublished.

Kahneman, Daniel, and Angus Deaton. 2010. "High Income Improves Evaluation of Life but not Emotional Well-Being." *Proceedings of the National Academy of Sciences of the United States of America* 107 (38): 16489–93.

Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M. Groves, and Stanley Presser. 2000. "Consequences of Reducing Nonresponse in a National Telephone Survey." *Public Opinion Quarterly* 64 (2): 125–48.

Kimball, Miles, Helen Levy, Fumio Ohtake, and Yoshiro Tsutsui. 2006. "Unhappiness after Hurricane Katrina." National Bureau of Economic Research Working Paper 12062.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica* 75 (1): 83–119.

[23] Such standards are set, for example, by the American Association for Public Opinion Research (AAPOR).

[24] As with web surveys, however, one would worry that delays in signing up for participation may be less indicative of how nonresponders would behave than are unanswered calls in telephone surveys.

Kreuter, Frauke, Gerrit Müller, and Mark Trappmann. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74 (5): 880–906.

Lin, I-Fen, and Nora Cate Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59 (2): 236–58.

Lin, I-Fen, Nora Cate Schaeffer, and Judith A. Seltzer. 1999. "Causes and Effects of Nonparticipation in a Child Support Survey." *Journal of Official Statistics* 15 (2): 143–66.

Potthoff, Richard F., Kenneth G. Manton, and Max A. Woodbury. 1993. "Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks." *Journal of the American Statistical Association* 88 (424): 1197–207.

Stevenson, Betsey, and Justin Wolfers. 2009. "The Paradox of Declining Female Happiness." *American Economic Journal: Economic Policy* 1 (2): 190–225.

Stone, Arthur A., Joseph E. Schwartz, Joan E. Broderick, and Angus Deaton. 2010. "A Snapshot of the Age Distribution of Psychological Well-being in the United States." *Proceedings of the National Academy of Sciences of the United States of America* 107 (22): 9985–90.