Difficulty to Reach Respondents and Nonresponse Bias: Evidence from Large Government Surveys[*]

Ori Heffetz   Daniel B. Reeves

December 20, 2017

Abstract

How high is unemployment? How low is labor force participation? Is obesity more prevalent among men? How large are household expenditures? We study the sources of the relevant official statistics—the Current Population Survey (CPS), the Behavioral Risk Factor Surveillance System (BRFSS), and the Consumer Expenditure Survey (CEX)—and find that the answers depend on whether we look at easy- or at difficult-to-reach respondents, measured by the number of call and visit attempts made by interviewers. A challenge to the (conditionally-)random-nonresponse assumption, these findings empirically substantiate the theoretical warning against making population-wide estimates from surveys with low response rates.

KEYWORDS: nonresponse bias, selection bias, survey data, difficulty of reaching, contact attempts, number of calls, paradata

JEL CLASSIFICATION: C18, C83, J60, I18.

[*] Heffetz: S.C. Johnson Graduate School of Management, Cornell University, 324 Sage Hall, Ithaca, NY 14853, and Department of Economics, Hebrew University of Jerusalem (e-mail: oh33@cornell.edu); Reeves: Department of Economics, Cornell University, 474-B Uris Hall, Ithaca, NY 14853 (e-mail: dbr88@cornell.edu). We thank Adrian Chadi, Tatiana Homonoff, Bruce Meyer, Ted O'Donoghue, Matthew Rabin, participants in seminars at Cornell, and two anonymous referees for useful early discussions and comments on previous drafts. We are grateful to John Dixon, Brian Meekins, and Polly Phipps for patiently responding to our questions regarding the CPS. The authors have no financial or other material interests related to this research to disclose.

An online appendix is available at `http://nber.org/~heffetz`.

I Introduction

To what extent do survey-based estimates depend on the difficulty of reaching respondents? The answer can hint at how cautious one should be when making population-wide inferences from surveys with low response rates. If within a survey sample, after controlling for other observables, outcomes are systematically different across easy-to-reach and hard-to-reach respondents, then one may question the routinely made assumption that nonrespondents—those out of sample who are, effectively, the hardest to reach—are similar to the average within-sample respondent. In other words, within-sample comparisons across difficulty-of-reaching groups can help assess the assumption of (conditional) random selection into the survey sample. This paper reports findings from such within-sample comparisons. Its goal is to help in assessing how sensitive population-wide estimates of important outcomes are to survey response rates and to assumptions regarding nonrespondents.

Our (purely empirical) investigation proceeded in three steps. First, we identified three large and widely used government surveys that report the number of phone or in-person visit attempts made in the course of reaching respondents: the Census/BLS's Current Population Survey (CPS), CDC's Behavioral Risk Factor Surveillance System (BRFSS), and BLS's Consumer Expenditure Survey (CEX). The three show a wide range of response rates, which have all been generally declining over the years, averaging in our data from just under 45% in the BRFSS to around 70% in the CEX and to just over 90% in the CPS. We note that these are the only three datasets that we investigated, and we chose in advance to report our findings from all three.

Second, within each dataset, we sought to identify one or two key outcomes to analyze. We were guided by our goal to focus on the outcomes of most interest to researchers, policymakers, and the public. Our search resulted in choosing four key outcomes: the labor force participation rate and the unemployment rate from the CPS, obesity prevalence from the BRFSS, and total household expenditures from the CEX. All four are national statistics that are closely watched, analyzed, and discussed in the academic literature, business world (or, in the case of obesity, health-policy world), and popular media. Importantly, unlike opinion-poll and social-survey outcomes such as voter intentions or consumer and investor confidence—outcomes that some researchers fundamentally mistrust—our four key outcomes are regarded by many as (more or less) objective indicators, based on survey questions designed to elicit factual reports rather than perceptions, feelings, or convictions.[1]

Third, based on the number of contact attempts made to each respondent, we divided each dataset into three or four difficulty-of-reaching groups, as similar in size as we could. We then compared, within each dataset, easy-to-reach respondent groups versus hard-to-reach groups with regard to outcome averages (both unadjusted and adjusted for observables) and to cross-demographic-group differences.

We find strong and robust differences between easy-to-reach and hard-to-reach respondents in all four primary outcome variables. Briefly, the labor force participation rate (65.1% overall in our 2012–2013 CPS sample) is 4.9 percentage points lower among the easiest-to-reach respondents (reached in a single visit attempt) than among the hardest-to-reach respondents (reached in 3 or more attempts), after controlling for demographic differences across difficulty-of-reaching groups. Similarly, the unemployment rate (7.6% in our sample) monotonically decreases 1.5 percentage points from easiest- to hardest-to-reach; the obesity rate (28.4% in our 2012 BRFSS sample) monotonically decreases 3.1 points from easiest (1 call attempt) to hardest (7+ attempts); and average log quarterly household expenditures ($9,459 in our 2008–2013 CEX sample, exponentiated back to dollars) increase by $465 from easiest- (1 contact attempt) to hardest-to-reach (5+ attempts). In addition, for labor force participation and for obesity—but not for the other two outcomes—the male-female gap and other cross-demographic-group gaps consistently shrink or increase with difficulty of reaching.

Overall, our analysis reveals a consistent picture: in our data, difficulty-of-reaching is strongly correlated with important outcomes of interest, even after controlling for the main observables that typical weighting schemes are based on. How important is this finding?

In principle, the finding of systematic differences between easy- and hard-to-reach respondents is not, by itself, necessarily worrisome. As long as survey nonrespondents are randomly selected (unconditionally, or on observables) from the population-representative

[1] While not a key outcome of our paper, we also analyzed a life satisfaction question asked in the BRFSS, and used it to probe the robustness of past research on subjective well-being data (see section III.1).

sample targeted by the survey, sample averages (unconditional, or conditional on observables) could be made generalizable to the population. But a difficult-to-reach respondent in one survey could be a nonrespondent in another survey that had a higher nonresponse rate due to time, budget, or other constraints. By the same token, nonrespondents in a given survey can be viewed simply as respondents so difficult that they remained out of reach. Indeed, the relatively high response rates in surveys such as the CPS suggest that additional effort and resources can bring many difficult (non)respondents into the sample. From this perspective, a finding of within-sample differences across difficulty-of-reaching groups challenges the random-selection assumption: if difficulty of reaching is correlated with outcomes, the likelihood of nonresponding may also be correlated with outcomes. Moreover, if in-sample trends in differences across difficulty-of-reaching groups extend to (out-of-sample) nonrespondents then not only are nonrespondents different from the average respondent; their average outcomes are not even within the range of average outcomes observed within sample.[2]

If we could somehow observe nonrespondents' outcomes—e.g., by matching respondents and nonrespondents with administrative measures of the survey outcomes we investigate—we could directly examine whether the difficulty-outcome trends we find extend to nonrespondents, and hence whether these trends are indeed evidence of nonresponse bias. (We could then also directly investigate another important question that our paper cannot examine: that of survey measurement error.) In practice, such direct tests are impractical for the same reason that the government surveys we study are so widely used in the first place: nationally representative administrative datasets containing "true" measures of the outcomes we study are not readily available.[3]

While we cannot use our data to directly assess whether the within-sample trends we

[2] Heffetz and Rabin (2013, p. 3007) provide a step-by-step numerical example that illustrates this point.

[3] In fact, it is not even clear that administrative data could exist for outcomes such as labor force participation and unemployment rates as defined and measured in the CPS, since they require reports of job search activity. On the other hand, in certain other special cases researchers *are* able to match administrative and survey data. We are aware of two papers—Lin and Schaeffer (1995) regarding child-support awards and payments, and Kreuter, Müller and Trappmann (2010) regarding welfare receipts and other outcomes—that investigate difficulty trends where true outcomes are known for respondents and nonrespondents. As we discuss in detail in section V, the former finds that the difficulty trends among respondents mostly do not extend to nonrespondents, but its survey context and difficulty measure are rather different from ours, while the latter, whose setup and difficulty measure are arguably more comparable to ours, finds that they do.

find extend to nonresponders, the wide range of response rates in the three datasets we study provides indirect, circumstantial evidence. In the CPS, with a nonresponse rate just under 10%, we do observe many who would be nonrespondents in another survey. The consistent in-sample trends we find in the CPS may hence tentatively provide some rationale for out-of-sample extrapolation of in-sample trends in higher-nonresponse-rate surveys—while acknowledging, of course, that such cautious rationale may apply to certain types of nonresponders more than to others. More generally, our finding of consistent trends in three surveys with, respectively, around 55%, 30%, and 10% nonresponse rate suggests that nonrespondents may on average look more like very difficult-to-reach respondents than like the average in-sample respondent. (Indeed, as hinted above, the very coexistence of such a wide range of nonresponse rates may in itself suggest it.)

Beyond these cautious arguments, the assumption that nonrespondents look like the in-sample average is simply hard to defend when said average is a moving target, changing systematically as increasingly difficult respondents are added to the sample. Moreover, when tracking outcomes over time, that assumption may *differentially* impact estimates of both long-run trends—as response rates slowly decline—and shorter-run fluctuations—as the outcome-difficulty gradient could itself vary with the outcome (we discuss such evidence in our concluding section). In the typical case—such as ours—where researchers do not know much about nonresponders, one may therefore view a finding of large within-sample difficulty-of-reaching differences as a reason for concern, the more so the higher is the nonresponse rate.

The rest of our paper proceeds as follows. In section II we study the CPS. We find large cross-difficulty-of-reaching differences in the raw (i.e., unadjusted) subsample means of our two key outcomes. The labor force participation rate climbs from 63.0% [SE 0.1] in the easiest-to-reach group to 72.3% [0.2] in the hardest-to-reach group—an increase of over 9 percentage points—while the unemployment rate declines from 8.1% [0.1] to 6.7% [0.2]. Of course, if these differences in outcomes were explained entirely by demographic differences between the easy- and hard-to-reach respondents then—assuming the demographic composition of the population is known from outside sources—applying the correct weights would yield unbiased population estimates. To show that this is not the case, at least not in a simple way, we compare adjusted means (controlling for observables including age, sex, race, education, and others) and still find that the adjusted labor force participation rate increases from 64.1% [0.1] to 69.0% [0.3], and the adjusted unemployment rate decreases from 8.0% [0.1] to 6.5% [0.2].

We also find that differences in labor force participation across age, sex, and other demographic groups change systematically with difficulty of reaching, albeit less dramatically. Finally, in robustness analysis we find that our CPS findings are even stronger when limiting the sample to self reports, for whom our difficulty-of-reaching measure is likely cleaner than for proxy reports; and are weaker when limiting the sample to telephone-completed interviews, for whom the measure is likely noisier.

In section III we study obesity prevalence in the BRFSS. Here the uncontrolled means decrease from 29.4% [SE 0.1] to 27.1% [0.1] from the easiest- to the hardest-to-reach groups, and the adjusted means decrease from 29.7% [0.1] to 26.6% [0.2]. We also find some evidence of cross-demographic-group differences in obesity that change with difficulty of reaching. For example, among BRFSS's easiest-to-reach respondents, males' adjusted obesity rate is indistinguishable from females' rate (male−female diff: 0.4% [SE 0.3]), however in the hardest-to-reach group males are a further 1.8 [0.4] percentage points more likely to be obese than females (diff: 2.2%). We illustrate the potential practical implications of these findings using a simple extrapolation. We also explore the relationship between difficulty of reaching and the self-reported height and weight variables that underlie the BRFSS obesity measure. Finally, we show that using the survey weights that the BRFSS recommends for making population-wide inferences does not change our conclusions.

In section IV we analyze log quarterly expenditures of CEX households, again finding a notable trend across difficulty-to-reach groups. Transformed back into dollars, unadjusted means rise from $8,225 [SE $74] to $9,990 [64], and adjusted means rise from $9,120 [59] to $9,585 [43]—a 5% increase—from easiest to hardest. We again demonstrate the robustness of our estimates in several ways. We also analyze separately total food and total health expenditures and find that they increase and decrease, respectively, with difficulty of reaching.

In section V we draw connections with previous work that investigates the relationship between difficulty of reaching and outcomes. To our knowledge, such work is almost nonexistent in economics. We review work—mostly in statistics and survey methodology—that has been exploring the potential use of difficulty-of-reaching measures and other paradata (data about how the data were collected) for making inferences regarding nonrespondents. Importantly, we review evidence that suggests that higher response rates do not always reduce overall bias: while potentially reducing unit nonresponse bias, they may increase item nonresponse bias or measurement error if, for example, the harder to reach are less likely to accurately answer certain survey questions. We also review evidence on the quality of number-of-contact-attempts paradata as difficulty-of-reaching measures.

Our paper's closest predecessor is Heffetz and Rabin (2013), whose closest predecessor is Curtin, Presser and Singer (2000). These earlier papers study outcomes from the University of Michigan's Surveys of Consumers: self-reported happiness, and the Index of Consumer Sentiment (ICS), respectively. Both papers find that conclusions regarding outcome variables depend on the difficulty of reaching respondents. Indeed, the outcome measures they study—designed to elicit mostly unverifiable emotions, attitudes, and beliefs—may be directly affected by the momentary context in which they are asked and hence may be more prone to difficulty-of-reaching differences. (For example, busy people may only be reached late in the day, when they are more pessimistic; or over the weekend, when they are more optmistic.) In contrast, our main outcomes of interest—labor force participation, unemployment, obesity, and expenditures—are measures that are designed to reflect verifiable states that do not change moment to moment, and as such are supposedly unaffected by the momentary situation related to the availability (or busyness) of respondents. A main contribution of our paper is to document that conclusions regarding such outcomes *also* depend on the difficulty of reaching respondents, in a way that simple reweighting schemes cannot correct.

We conclude in section VI, where we discuss the practical implications of our findings—both for users of the specific outcomes we study and more broadly for designers and users of other surveys. As a concrete application, we plot unemployment data over two decades (1994–2013) and demonstrate that cyclical unemployment fluctuations have been more extreme among easy than among difficult respondents.

II CPS: Labor Force Participation and Unemployment

II.1 Data

The Current Population Survey (CPS) is a monthly survey conducted by the Bureau of Labor Statistics (BLS) and the U.S. Census Bureau. Among its many uses, it provides closely watched labor force statistics for the U.S., such as the labor force participation and unemployment rates. The survey consists of a rotating panel of households. A participating household provides data for four consecutive months, is not contacted for eight months, and then provides data for four more consecutive months. The eight interviews are referred to as month-in-sample (MIS) 1 through 8. The CPS samples addresses, so each household's first interview (MIS 1) must begin with a personal visit by a field interviewer, but it can be completed by phone if requested by the respondent. The majority (around 85%) of the remainder of the interviews are conducted via telephone, with the exception of MIS 5, the one following the eight-month break, which is also usually completed in person.

Within each household, the interviewer attempts to interview the "most knowledgeable" member, often the owner or renter of the unit (CPS, 2015). During each interview the respondent is asked to help create a roster of eligible household members and answer questions to determine, for each member, whether she is: employed, unemployed, or out of the labor force. Specifically, for each civilian member of the household who is 15 or older, a series of questions is asked to determine whether or not she was employed in the interview month's reference week (almost always the week containing the 12th of the month). If she was, the CPS codes her as employed. Otherwise, the interview determines, among other things, whether or not she was actively looking for work in the past four weeks and was available for work during the reference week. If she was, the CPS codes her as unemployed. Otherwise—i.e., she neither had a job (employed) nor was actively looking for one (unemployed)—she is coded as out of the labor force.

We analyze CPS data from January 2012 to December 2013. The CPS response rate in each of these 24 months is between 89% and 91%, and is 90.1% overall for the period.[4] Our

[4] These are calculated from the CPS monthly files: the number of households with a partial or complete interview is divided by the number of eligible households in the month. This calculation corresponds to the American Association of Public Opinion Research (AAPOR) Response Rate #2, #4, and #6 (AAPOR,

analysis sample consists of all 307,603 MIS 1 observations that the BLS would include when calculating the labor force participation and unemployment rates—i.e., civilians aged 16 and up with complete labor force records.[5] The labor force participation rate is defined as the share of these who are in the labor force (i.e., they are either employed or unemployed); the unemployment rate is defined as the share of those in the labor force who are unemployed.

Finally, our difficulty-of-reaching measure is a variable recording the "number of actual and attempted personal contacts." It is a noisy measure. Importantly, it counts only personal visit attempts, so it likely understates the difficulty of reaching respondents who were also contacted by the telephone.[6] In addition, since it is reported by the interviewer (and not by an automatic system, as in some telephone surveys), it may be affected by intentional or unintentional misreporting.[7] Indeed, 24,807 observations (8.1% of our analysis sample) have nil contact attempts reported, and we do not know how difficult to reach they were; in the tables below we classify their number of contact attempts as "None Reported," or "NR." The remainder of the observations have 1–9 (top coded) contact attempts recorded. We classify them into three categories: 1 attempt (64.3%), 2 attempts (17.0%), and 3 or more attempts (10.7%).

2016). (The three response rate definitions differ in how households of unknown eligibility are treated, but the CPS contains no such households because interviewers classify the eligibility of each address.) Krueger, Mas and Niu (2017, Figure 4) show that nonresponse rate in the CPS has been generally increasing, from slightly above 4% in 1990 to roughly 8% in 2010, and then steeply climbing to nearly 11% by 2014.

[5] We restrict our analysis to MIS 1 data for two reasons. First, the number-of-contact-attempts variable only counts *in-person* contact attempts, so it is mostly non-zero only for MIS 1 (where 91.9% of the observations in our analysis sample are non-zero) and, to a lesser extent, for MIS 5 (76.1% non-zero). Second, the CPS is subject to rotation group bias, i.e., some of its outcomes vary systematically by MIS (for recent evidence see and Dixon (2013) and Krueger, Mas and Niu (2017)); focusing on a single MIS avoids this confounding bias. (While the sources of this bias are still being investigated, Krueger, Mas and Niu (2017) find suggestive evidence that the unemployment rate calculated from early interviews—MIS 1, 2, 3, and 5—is a stronger predictor of other measures of economic slack than that calculated from later interviews.)

[6] 52,624 of the observations in our sample (17.1%) are known to have been conducted at least in part via telephone, and we also analyze them separately below. Another 6,398 observations (2.1%) have a missing value for interview mode.

[7] While not explicitly stated in the data dictionary, it was confirmed to us by researchers at the BLS that this variable, as well as the mode variable, are manually entered by the interviewer.

II.2 Analysis by Difficulty of Reaching

II.2.1 Sample composition

Table 1 reports basic demographics for our sample. Each of its first four columns is based on a single number-of-contact-attempts category (1, 2, 3+, and NR); the fifth column is based on the entire sample (All). The first three columns show that on some demographics the sample's composition changes systematically with number of attempts. Notably, the young are harder to reach than the old: those aged 20–39 comprise 30.7% of the 1-attempt sample compared with 37.0% of the 3+-attempts sample, while those aged 65 and up comprise 20.2% and 12.5% respectively. (In both cases, the 2-attempt percentages lie in between.) On the other hand, women and men are of overall similar difficulty of reaching: the share of women, 52.2% overall, does not vary much with number-of-attempts category. We also note that demographic-group shares in the NR column are in some cases inside, and in other cases outside (below or above), the range of shares in the three leftmost columns. This makes it difficult to hypothesize about the difficulty-of-reaching of NR respondents, and we have little to say about them in the rest of this section; we include them in the tables only for completeness.

Looking at our variables of interest at the bottom of the table, the labor-force-status composition of the sample changes dramatically with contact attempts: a sharp increase in the share employed, from 57.9% (1 attempt) to 67.5% (3+ attempts), is mirrored almost entirely by a sharp decrease in the share not in the labor force, from 37.0% to 27.7%, while the share unemployed is pretty stable and hovers around 5%. In other words, as contact attempts increase, household members are more likely to be recorded as employed and less likely to be recorded as not employed who are not looking for employment. As a result, and as discussed in the introduction, our two key outcomes—the labor force participation rate and the unemployment rate—increase, and decrease, respectively, with contact attempts. In the rest of this subsection we show that this increase and decrease are not eliminated by controlling for the (first order) changes in demographic composition noted above.

Table 1: CPS Demographics

Attempts		1	2	3+	NR	All
Age:	16–19 (%)	6.4	7.0	7.1	6.1	6.5
		(0.1)	(0.1)	(0.1)	(0.2)	(0.0)
	20–39	30.7	33.4	37.0	28.4	31.6
		(0.1)	(0.2)	(0.3)	(0.3)	(0.1)
	40–49	16.5	18.1	18.8	17.3	17.1
		(0.1)	(0.2)	(0.2)	(0.2)	(0.1)
	50–64	26.2	25.7	24.7	28.9	26.2
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
	65 and up	20.2	15.8	12.5	19.3	18.6
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
Children in household		25.5	28.4	28.5	23.4	26.2
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
Female		52.2	52.0	52.0	52.8	52.2
		(0.1)	(0.2)	(0.3)	(0.3)	(0.1)
Educ:	Less than high school	14.9	14.7	14.1	11.1	14.5
		(0.1)	(0.2)	(0.2)	(0.2)	(0.1)
	High school	30.2	29.1	28.0	27.1	29.6
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
	Some college or tech. school	27.3	28.0	28.3	27.0	27.5
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
	College graduate	27.6	28.2	29.6	34.8	28.5
		(0.1)	(0.2)	(0.3)	(0.3)	(0.1)
Race:	White	82.9	81.1	78.8	84.8	82.3
		(0.1)	(0.2)	(0.2)	(0.2)	(0.1)
	Black	9.7	10.1	11.5	8.6	9.8
		(0.1)	(0.1)	(0.2)	(0.2)	(0.1)
	Asian	4.4	5.6	6.4	4.4	4.8
		(0.0)	(0.1)	(0.1)	(0.1)	(0.0)
	Other	3.1	3.2	3.3	2.3	3.1
		(0.0)	(0.1)	(0.1)	(0.1)	(0.0)
L.F.P.:	Employed	57.9	62.5	67.5	63.5	60.2
		(0.1)	(0.2)	(0.3)	(0.3)	(0.1)
	Unemployed	5.1	5.2	4.8	4.0	5.0
		(0.0)	(0.1)	(0.1)	(0.1)	(0.0)
	Not in the labor force	37.0	32.3	27.7	32.5	34.9
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
Labor force participation		63.0	67.7	72.3	67.5	65.1
		(0.1)	(0.2)	(0.2)	(0.3)	(0.1)
Unemployment rate		8.1	7.6	6.7	5.9	7.6
		(0.1)	(0.1)	(0.2)	(0.2)	(0.1)
Median number of attempts (known)		1	2	3	n/a	1
Observations		197,751	52,275	32,770	24,807	307,603

Notes: Source: Current Population Survey, Jan. 2012–Dec. 2013. Sample: all MIS 1 observations who are qualified to be in the civilian labor force and gave enough employment information to be classified. All figures (and standard errors) reflect proportions within each column's difficulty-of-reaching category, except for those for unemployment rate (which are calculated as described in text) and number of attempts (which report medians). NR: No reported contact attempts.

II.2.2 Labor force participation

Table 2 reports our main labor force participation results. Since the table's structure is shared by all other main-results tables in the rest of this paper (with unemployment, obesity, and expenditures as dependent variables), as well as with many appendix tables, we describe it below in some detail. We also note here that our main findings are not affected more than trivially by including a larger set of age indicators and interactions or, for our *binary* dependent variables (labor force participation, unemployment, and obesity), by replacing the OLS specification in our main-results tables with probit or logit.[8]

Table 2's four columns report results from a *single* OLS regression. The dependent variable is a 0/1 labor force participation indicator. The regressors are sets of demographic indicators (those reported in table 1, plus unreported indicators for marital status (6 categories), household size (5), state (51), urban/rural (3), interview month (12), interview year (2), and a constant); a set of difficulty-to-reach-category indicators; and a full set of interactions of the difficulty indicators × all demographic indicators (including those unreported). Panel A reports the estimated coefficients: the first column reports coefficients on the demographic indicators for the base (omitted) difficulty-to-reach category (1 contact attempt), and the other columns report the coefficients on the demographic indicators interacted with each of the three other difficulty categories (2, 3+, and NR). Notice the reported 4×13 coefficients are mechanically identical to those one would get from estimating a separate regression of the dependent variable on the set of demographics (and no interactions) within each of the four difficulty categories (that is, four separate regressions), and then subtracting the coefficients in the 1-attempt regression from those in each of the other three regressions.

Panel B reports adjusted means for the four difficulty-to-reach categories (calculated from the regression coefficients estimated in Panel A). Intuitively, the adjusted means are calculated as follows.[9] For each observation, one calculates the dependent variable's predicted

[8] Specifically, all the OLS-adjusted means discussed below remain within their reported standard errors (and are often within our tables' rounding error) when including a set of ten rather than five age categories, as well as all age-gender, age-education, and gender-education interactions; and all the probit/logit-adjusted means are within 0.13 percentage points of the OLS-adjusted means discussed below (indeed, most are exact matches, given rounding error), with the interaction coefficients showing the same patterns and statistical significance.

[9] In practice, we use the STATA 14.1 command "margins," which also computes standard errors using the delta method. Our intuitive exposition here draws in part on Williams (2011).

Table 2: Labor force participation

Attempts	1	2	3+	NR
		A: Regression with interactions		
	Base	Interactions		
Age: 16–19	-0.263***	-0.025**	-0.035**	-0.047***
	(0.006)	(0.012)	(0.015)	(0.018)
20–39	-0.011***	-0.009	-0.003	-0.024***
	(0.003)	(0.006)	(0.006)	(0.008)
50–64	-0.113***	0.018***	0.049***	0.028***
	(0.003)	(0.006)	(0.007)	(0.008)
65 and up	-0.574***	0.016*	0.069***	0.032***
	(0.004)	(0.008)	(0.011)	(0.011)
Children in household	0.039***	0.017***	0.035***	0.026***
	(0.003)	(0.007)	(0.007)	(0.009)
Female	-0.103***	-0.002	0.010**	0.007
	(0.002)	(0.004)	(0.005)	(0.005)
Educ: Less than high school	-0.136***	0.008	0.001	-0.020*
	(0.003)	(0.007)	(0.009)	(0.012)
Some college or tech. school	0.029***	-0.004	-0.010	-0.004
	(0.003)	(0.006)	(0.007)	(0.008)
College graduate	0.098***	-0.014***	-0.025***	-0.001
	(0.003)	(0.005)	(0.006)	(0.007)
Race: Black	-0.026***	-0.004	0.001	-0.016
	(0.004)	(0.008)	(0.009)	(0.011)
Asian	-0.055***	-0.000	0.013	0.011
	(0.005)	(0.010)	(0.012)	(0.015)
Other	-0.037***	-0.006	0.015	0.047**
	(0.006)	(0.013)	(0.016)	(0.020)
Constant	0.782***	0.016	0.105***	0.107***
	(0.011)	(0.024)	(0.027)	(0.038)
		B: Adjusted means		
Labor force participation	0.641***	0.662***	0.690***	0.665***
	(0.001)	(0.002)	(0.003)	(0.003)

Notes: Source: Current Population Survey, Jan. 2012–Dec. 2013. $N = 307,603$ (1 attempt: 197,751; 2: 52,275; 3+: 32,770; None Reported: 24,807). $R^2 = 0.29$. The table reports estimates from a single OLS regression. Dependent variable: 0/1 labor force participation indicator. See page 12 for a full explanation of table structure. Panel A: estimated coefficients from a fully interacted regression: each regressor is interacted with each difficulty-to-reach category (omitted category: 1 attempt). Regression also includes non-reported indicators (and their interactions) for marital status (6 categories), household size (5), state (51), urban/rural (3), interview month (12), and interview year (2); see appendix note A.1. Standard errors, clustered at the household level, in parentheses. Panel B: adjusted means, calculated from panel A regression. *** p<0.01, ** p<0.05, * p<0.1.

value four times, using that observation's true values for all the independent variables except for the set of difficulty-to-reach indicators, which are changed to indicate 1 attempt for the 1-attempt adjusted mean, are changed to indicate 2 attempts for the 2-attempts adjusted mean, and so on. The adjusted means row then reports these predicted values averaged across all the sample's observations. The four adjusted means are hence the average predicted value of the dependent variable in four hypothetical samples. Each of the four samples is identical to our actual (full) sample except that the number-of-contact-attempts category is hypothetically set to 1, 2, 3+, or NR, respectively, for all of that sample's observations.

We start with the bottom line: the adjusted means in Panel B are 64.1% (1), 66.2% (2), and 69.0% (3+)—a total increase of 4.9 percentage points, with each of the three point estimates statistically different from the others. In other words, the differences in composition across the difficulty-to-reach categories explain less than half of the 9.3-point raw increase from table 1.

Looking at panel A, the fully interacted regression shows that not only are respondents in different difficulty categories predicted to have different labor force participation rates even when they are otherwise identical (on observables) to the entire sample, but furthermore these predicted differences interact with demographics. For example, teenagers (aged 16–19) are 26.3 percentage points less likely to participate than those aged 40–49 among 1-attempt respondents but are 26.3 + 3.5 = 29.8 points less likely to participate among 3+-attempts respondents. In contrast, those 65 and up are 57.4 points less likely to participate with 1 attempt, but are 50.5 points less likely to participate with 3+ attempts, a (highly significant) 6.9 point reduction in the difference. As to other age-group coefficients, that on ages 20–39 seems relatively stable, while that on ages 50–64 shrinks from 11.3% (1 attempt) to 6.4% (3+). Other changes with difficulty of reaching include a 3.5, 2.5, and 1.0 point change of the difference between, respectively, those with and without children in the household, those with and without a college degree (baseline: only high school), and men versus women.

These differences in panels A and B between the difficult and the easy to reach—and hence potentially between nonrespondents and respondents—cannot be eliminated by standard reweighing schemes. The generalizability of in-sample estimates to population-wide estimates regarding labor force participation therefore depends crucially on maintaining low nonresponse rate in the CPS (while managing high response quality, including accurate responses to individual questions).

II.2.3 Unemployment rate

In table 3, which is otherwise identical to table 2, we switch to analyzing the unemployment rate by replacing the dependent variable with a 0/1 unemployment indicator and limiting the sample to labor force participants. Beginning again with the adjusted means in panel B, we see a distinctive trend as the unemployment rate drops from 8.0% (1 attempt) to 7.5% (2 attempts) to 6.5% (3+ attempts). This overall drop of 1.5 percentage points is similar to the drops in the annual unemployment rate from its Great Recession peak (9.6% in 2010) to its level in 2012 (8.1%) and from its 2001-recession peak (6.0% in 2003) to the next trough (4.6% in 2006–2007). It is again clear that unless nonrespondents are selected at random—a suspect assumption, given that even among respondents, difficulty of reaching is a strong predictor of the outcome—high response rates (with high-quality responses) are crucial for producing accurate population-wide estimates.

Looking at interactions in panel A, there appears to be less movement in table 3 than in table 2. (The only consistent trends with some statistical significance—up to the 5% level—are within some of the age rows.)

II.3 Robustness and Additional Results

II.3.1 Self reports versus proxies

Unlike the CEX—where a respondent provides expenditure information at the household level—and BRFSS—where a respondent provides individual information—in the CPS a respondent provides information about herself *and* about every other adult in the household. This means that in the CPS, number of contact attempts may be a noisier measure of the true, underlying difficulty of reaching household members who are not personally interviewed. Are the difficulty-of-reaching trends we report above indeed stronger among self reporters?[10]

[10] Notice that one should not expect the difficulty-of-reaching trends to entirely disappear among proxy reports, for reasons that could be both mechanical—e.g., all else equal, the household of a harder-to-reach

Table 3: Unemployment rate

Attempts	1	2	3+	NR
		A: Regression with interactions		
	Base	Interactions		
Age: 16–19	0.079***	0.012	0.022	-0.011
	(0.007)	(0.014)	(0.017)	(0.020)
20–39	0.014***	0.001	-0.011**	-0.004
	(0.002)	(0.004)	(0.005)	(0.006)
50–64	0.003	-0.005	-0.010**	0.001
	(0.002)	(0.004)	(0.005)	(0.006)
65 and up	0.002	-0.001	-0.008	0.014
	(0.003)	(0.007)	(0.008)	(0.009)
Children in household	-0.002	-0.009*	-0.006	-0.007
	(0.002)	(0.005)	(0.006)	(0.007)
Female	-0.004***	-0.005	-0.001	0.004
	(0.002)	(0.003)	(0.004)	(0.004)
Educ: Less than high school	0.043***	-0.001	-0.011	-0.002
	(0.004)	(0.008)	(0.009)	(0.013)
Some college or tech. school	-0.020***	0.003	0.000	0.003
	(0.002)	(0.004)	(0.005)	(0.006)
College graduate	-0.043***	0.003	0.008*	0.007
	(0.002)	(0.004)	(0.005)	(0.005)
Race: Black	0.064***	-0.018**	-0.010	-0.009
	(0.004)	(0.007)	(0.008)	(0.010)
Asian	-0.007*	0.006	0.010	-0.023***
	(0.004)	(0.007)	(0.008)	(0.008)
Other	0.047***	-0.012	0.003	-0.000
	(0.006)	(0.012)	(0.015)	(0.017)
Constant	0.061***	-0.032*	-0.025	-0.037
	(0.010)	(0.019)	(0.018)	(0.027)
		B: Adjusted means		
Unemployment rate	0.080***	0.075***	0.065***	0.066***
	(0.001)	(0.001)	(0.002)	(0.002)

Notes: Source: Current Population Survey, Jan. 2012–Dec. 2013. Sample: labor force participants, $N = 200,358$ (1 attempt: 124,530; 2: 35,376; 3+: 23,705; None Reported: 16,747). $R^2 = 0.05$. Table reports estimates from a single OLS regression (see page 12 for full explanation of table structure). Dependent variable: 0/1 unemployed indcator. Panel A: estimated coefficients from a fully interacted regression: each regressor is interacted with each difficulty-to-reach category (omitted category: 1 attempt). Regression also includes non-reported indicators (and their interactions) for marital status (6 categories), household size (5), state (51), urban/rural (3), interview month (12), and interview year (2); see appendix note A.1. Standard errors, clustered at the household level, in parentheses. Panel B: adjusted means, calculated from panel A regression. *** p<0.01, ** p<0.05, * p<0.1.

To explore this hypothesis, appendix tables A.2 and A.3 (participation) and A.4 and A.5 (unemployment) reproduce tables 2 and 3 separately for self and proxy reports. As hypothesized, the difficulty-outcome trends are stronger—indeed, much stronger—among self reporters. Adjusted means for labor force participation show a monotonic 7.5-point increase (from 62.4 to 69.9 points) among self reporters, compared with an only 2.3-point increase (from 65.8 to 68.1) among the proxy reported. There are also generally more pronounced difficulty-of-reaching trends within the demographic-group estimates of self reporters. Similarly, when moving from 1 to 2 to 3+ attempts, adjusted means for unemployment display a monotonic 2.5-point decrease (8.4% to 7.2% to 5.9%) among self reporters, again roughly three times the 0.7-point decrease among the proxy reported (7.7% to 7.8% to 7.0%, no longer strictly monotonic).

While the dramatically steeper trends among self reporters in the adjusted means of both dependent variables support the idea that there is a link between the difficulty of contacting a respondent and her outcomes, and that the link cannot be accounted for with other observable demographic controls, the comparison of self versus proxy reports should be interpreted with caution, as it has its own set of limitations. Importantly, self reporters are likely to be different from those whose labor force status is reported by proxy. Indeed in our data self reporters are on average older, more highly educated, and more likely to be female (not reported); of course, they may also be different on unobservable characteristics. We therefore refrain from drawing strong conclusions based on self reporters alone. (Our main analysis above pools together proxy- and self-reports, more closely matching the sample used by the BLS when it estimates its statistics.)

II.3.2 Telephone-completed interviews only

As mentioned above, the CPS's number-of-contact-attempts variable only counts in-person attempts. In addition, the interview-mode variable only records the mode of the last interview (in person, telephone, or missing).[11] One may therefore wonder whether a respondent

individual is, on average, harder to reach—and circumstantial—e.g., demographic predictors of difficulty-of-reaching (such as age) are correlated across members within a household.

[11] We learned in conversations with BLS staff that the mode variable gets set every time the interviewer enters the case. As a result, for households whose data were collected over the course of multiple interviews,

whose last interview was by telephone may on average be harder to reach than a respondent whose last interview was in person, given the same number of reported in-person contact attempts, as the former required at least one (and possibly more than one) additional, unrecorded, telephone contact attempt.[12] Specifically, consider the hypothesis that at a given number of reported in-person attempts, telephone completes are more likely to have had, on average, an additional unobserved positive number of telephone attempts. This hypothesis yields two predictions: relative to in-person completes, the 17.1% telephone completes in our main sample should on average (a) have outcomes consistent with being more difficult than their difficulty measure suggests, and (b) show weaker outcome-difficulty trends (due to the difficulty variable being a noisier measure for them).

Appendix tables A.6 (participation) and A.7 (unemployment) repeat our main analysis restricting the sample to the 52,624 telephone completes in our data. (We do not report estimates restricted to in-person completes because, accounting for 80.8% of our data, they are rather close to the full-sample estimates above.) The adjusted means for labor force participation follow the same upward trend as in table 2, rising from 68.5% to 70.3% to 71.3% across the difficulty-to-reach categories, but are generally higher than the full-sample adjusted means (64.1%, 66.2% and 69.0%), consistent with (a) above, and their trend is less steep, consistent with (b) above. Likewise, the adjusted means for unemployment follow the same pattern as in table 3, declining from 6.4% to 6.0% to 5.1%, are lower than their full sample counterpart (8.0%, 7.5% and 6.5%, respectively), and show a slightly less steep trend although not statistically significantly so. The coefficient patterns in panel A of both appendix tables generally resemble those in the primary analysis, but at different levels (similar to the findings in the adjusted means). Overall, then, the evidence is consistent with the above hypothesis.

only the last interview's mode is recorded.

[12] Of course, as we do not observe the total number of telephone contact attempts, we can neither quantify the hypothesized difficulty difference nor can we even be certain that it exists, as unrecorded telephone attempts could have taken place also for individuals who completed the interview in person.

III BRFSS: Obesity

III.1 Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a large annual cross-sectional telephone survey designed to monitor health-related risk behaviors, chronic health conditions, and the use of preventative health services among the adult U.S. population. Most modules of the survey are designed and supported by the Centers for Disease Control and Prevention (CDC). However, data collection is decentralized and is individually administered by each of the 50 states, the District of Columbia, Guam and Puerto Rico (henceforth, the 53 locations). The survey is intended to be nationally representative of adults living in households in the United States and these territories. It uses random digit dialing for both landline and cellular telephone numbers; for landlines, one respondent is randomly chosen per household. The survey is conducted throughout the year, with differences in the exact timing between the 53 locations.[13] The 2012 BRFSS dataset—the latest publicly available at the time of conducting our analysis—includes 475,687 respondents. Across the 53 locations, its median survey response rate is 45.2%, ranging from 27.7% in California to 60.4% in South Dakota; weighted by location sample size, its average response rate is 44.6%.[14]

The BRFSS contains a variable with respondents' Body Mass Index (BMI, defined as $\frac{\text{mass in kg}}{(\text{height in meters})^2}$). It is calculated from respondents' reports of their height and weight (in either feet and pounds or meters and kilograms), and is missing for respondents who did not provide their height or weight ($n = 22,710$, 4.8% of the original sample); who reported being pregnant at the time of the interview ($n = 2,680$, 0.6%); or whose calculated BMI was considered erroneous by the CDC ($n = 85$). Our analysis is based on the remaining 450,212 respondents. Our main dependent variable is a 0/1 indicator of obesity, defined by the World Health Organization as BMI ≥ 30 (WHO, 1995, 2000).

[13] Appendix figures A.1 through A.3, each containing 54 mini-graphs representing the entire sample and the 53 locations, report the corresponding 54 distributions of contact attempts and of interview timing information (the month, day of the month, and day of the week in which the interview was conducted).

[14] The BRFSS calculates response rates using AAPOR Response Rate #4, separately for its landline and cellular phone samples within each location; it then reports location response rates as weighted averages, weighted by local landline/cellular full-sample sizes. See BRFSS (2013) for a detailed discussion of response rate calculation and variation across locations.

Our difficulty-of-reaching measure is a variable that records the number of calls made to each respondent. We use it to divide the data into four approximate difficulty quartiles: 1 call (25.5% of the data), 2–3 calls (30.5%), 4–6 calls (21.9%) and 7 or more calls (22.1%).[15]

For completeness and transparency, we note that chronologically, BRFSS data were the first we analyzed, and prior to analyzing obesity we analyzed an outcome that, while not a key outcome of our paper, was of great interest to us nonetheless: a life satisfaction question. We first attempted to replicate the main findings in Heffetz and Rabin (2013) and then, as BRFSS data have been used by economists to study the relationships between life satisfaction and properties of states (Oswald and Wu, 2010) and cities (Glaeser, Gottlieb and Ziv, 2016), we also explored the state relationships by difficulty of reaching and by cross-location differences in survey methodology. Our early analysis of the life satisfaction question used older (2005–2008) BRFSS data, matching the dataset in Oswald and Wu (2010); in general we replicated the published results that we explored.[16]

III.2 Analysis by Difficulty of Reaching

III.2.1 Sample composition

Table 4 reports the demographic composition of our BRFSS sample. As with the CPS, we find compositional differences in the makeup of the difficulty-to-reach subsamples. Notably, in both datasets the difficult to reach tend to be younger than the easy to reach. But in contrast with the CPS, in the BRFSS females are on average easier to reach: their share decreases monotonically from 61.2% (1-attempt respondents) to 56.3% (7+ attempts). The BRFSS sample also becomes less non-hispanic white (79.5% to 73.8%), more educated, and

[15] In the public-use data we analyze, the number-of-calls variable is globally top-coded at 35, and in some states it appears to be locally top-coded at 15 (top-coding does not affect our difficulty categories). Appendix figure A.4 reports the distribution of this variable in our sample and in each of the 53 locations.

[16] Specifically, Heffetz and Rabin (2013) found that cross-group differences in reported happiness depend on the difficulty of reaching respondents; our appendix tables A.23 and A.24, modeled after tables 1–4 from Heffetz and Rabin (2013), report that the original results replicate qualitatively (the data show the same directional patterns), but the estimated magnitudes are smaller. While this may be seen as a successful replication in a new dataset, we caution that the original outcome variable (happiness yesterday) is not directly comparable to the BRFSS variable (general satisfaction with one's life). Oswald and Wu (2010) found that average life satisfaction in a U.S. state is related to non-subjective quality-of-life measures for the state; we replicated their analysis and found that controlling for cross-state differences in survey methodology—including the faction of a state's population that was easy-to-reach, the faction interviewed in each third of the month, and the state's overall response rate—did not affect the original findings.

of higher reported income with difficulty of reaching. While the relevant variable definitions are not directly comparable across the CPS and BRFSS, these trends seem consistent across the two datasets. In particular, higher income is consistent with higher participation and lower unemployment (see also appendix table A.1, which replicates table 1 and adds CPS income data). Finally, there is a strong, monotonic and significant (both economically and statistically) difficulty-of-reaching trend in the fraction of the sample that is obese, which declines from 29.4% (1 attempt) to 27.1% (7+ attempts). Interestingly, this trend does not reflect a trend in the average weight of the difficulty-to-reach categories (as weight remains relatively constant), though it may in part reflect a 1 cm overall increase in average height from the 1- to 7+-attempt categories.

III.2.2 Obesity

Table 5 follows the same structure as the CPS outcome tables (see table 2's exposition on p. 12). It reports the results from a regression of a 0/1 obesity indicator on a set of demographic indicators (those in table 4 and additional unreported indicators for marital status (7 categories), location (53), urban/rural (6), and interview month (12)) and a constant, a set of difficulty-of-reaching indicators, and a full set of demographic-difficulty interactions. As panel B reports, adjusted obesity prevalence declines monotonically from 29.7% (1 attempt) to 26.6% (7+ attempts), an overall decrease of 3.1 percentage points that is in fact larger than the unadjusted mean decrease of 2.3 points. Not only do differences in demographic composition across the difficulty-to-reach groups not drive the differences in obesity prevalence; the demographic-composition differences in fact *mask* some of the adjusted differences in obesity across difficulty categories. Since our entire sample could be viewed as the 44.6% easiest to reach among eligible households, the sample average of 28.4% obesity may be an overestimate of the population average. For example, a simple out-of-sample extrapolation of the within-sample trend in table 5 suggests an overestimate of around 2 additional percentage points for the population obesity prevalence.[17] Of course, such a simple extrapolation leaves

[17] This back-of-the-envelope estimate makes the simplifying assumptions that our four observed difficulty-to-reach categories are equal-sized and represent the easiest-to-reach half of the total attempted population. The remaining, unobserved, half of the population is then split into 4 equal-sized difficulty-to-reach groups; a linear trend that approximates the trend observed in panel B's adjusted means is projected onto these groups; and then the overall average is taken.

Table 4: BRFSS Demographics

Attempts:		1	2–3	4–6	7+	All
Age:	18–39 (%)	18.5	20.6	23.6	20.9	20.8
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	40–49	12.1	13.6	15.5	17.6	14.5
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	50–59	18.3	19.8	21.1	23.7	20.6
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	60–69	22.1	21.4	20.0	20.5	21.1
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	70 and up	28.5	24.1	19.2	16.6	22.5
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	Missing	0.5	0.6	0.6	0.7	0.6
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Children in household		22.4	25.5	29.3	31.8	26.9
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Female		61.2	58.2	56.5	56.3	58.2
		(0.1)	(0.1)	(0.2)	(0.2)	(0.1)
Educ:	Less than high school	9.0	8.7	8.5	8.6	8.7
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
	High school	30.7	29.6	28.7	28.2	29.4
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	Some college or tech. school	27.8	27.5	27.1	25.6	27.1
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	College graduate	32.5	33.9	35.5	37.4	34.7
		(0.1)	(0.1)	(0.2)	(0.2)	(0.1)
	Missing	0.1	0.2	0.2	0.2	0.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Race:	White, non-hispanic	79.5	78.0	75.2	73.8	76.8
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	Black, non-hispanic	7.5	7.8	8.5	10.1	8.4
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
	Other, non-hispanic	6.0	6.4	6.7	6.4	6.4
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
	Hispanic	5.9	6.7	8.3	8.4	7.3
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
	Missing	1.0	1.1	1.2	1.3	1.2
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Inc:	Below $25,000	30.2	27.3	25.6	22.8	26.7
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	$25,000–49,999	23.9	23.4	22.7	21.3	22.9
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	$50,000–74,999	12.9	13.7	14.0	14.2	13.7
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	$75,000 and up	20.0	22.7	25.2	29.2	24.0
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
	Missing	13.0	12.8	12.5	12.5	12.7
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
Obese ($BMI \geq 30$)		29.4	28.5	28.1	27.1	28.4
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Avg. height (cm)		168.7	169.3	169.6	169.7	169.3
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Avg. weight (kg)		79.6	79.8	79.9	79.7	79.8
		(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
Median number of attempts		1	2	5	10	3
Observations		114,694	137,418	98,813	99,287	450,212

Notes: Source: Behavioral Risk Factor Surveillance System, 2012. Sample: Non-pregnant individuals who provided height and weight (and were not excluded due to erroneously high/low BMI). All figures (and standard errors) reflect proportions within each column's difficulty-of-reaching category, except for those for height and weight (which report averages) and number of attempts (which report medians).

out many important details (including, for example, the possibility of measurement error that is correlated with difficulty), and is hence only given as an illustration. Our point is that our data strongly question the practice of assuming that nonresponders are on average as obese as responders (unconditionally or conditionally on observables), and accordingly regarding the sample average (raw or reweighted) as the best population-wide estimate.

In panel A we note two cross-demographic-group obesity differences that change with difficulty of reaching and that we find of particular interest. First, women change from being essentially as obese as men (difference = -0.4 percentage points, insignificant) among 1-attempt respondents to being 2.2 points less obese than men among 7+-attempts respondents, a large and significant overall change of 1.8 points. Second, those with children in the household change from being a significant 2.3 points more obese among 1-attempt respondents to being essentially as obese as those in children-less households (difference = 0.5 points, insignificant), another large and significant overall change of 1.8 points. While neither of these changes is strictly monotonic (see table), in both cases conclusions regarding cross-demographic-group differences in obesity rates—comparing women versus men, and those with children versus without—qualitatively change. In both cases, whether one estimates a large and statistically significant difference or no difference depends on which difficulty category one looks at, and—given the BRFSS's relatively low response rate—population-wide inferences strongly depend on what one assumes about nonresponders.[18]

III.3 Robustness and Additional Results

III.3.1 Height and weight

The variables that underlie the obesity variable—self-reported weight (kg) and height (m)—may themselves be interesting to researchers as main outcomes. We repeat our analysis with each of the two as a dependent variable (appendix tables A.9 and A.10). We find a strong

[18] Appendix table A.8 reproduces table 5 with additional controls for reported employment status. The table shows some interesting patterns: relative to those employed for wages—other things equal—homemakers, the self-employed, and especially students, are less obese; while those out of work for more than a year, and especially those unable to work, are more obese. Also, as expected, some of the education and income coefficients change. At the same time, both the adjusted means in panel B and the patterns noted above in panel A change little, and always remain within their original standard errors.

Table 5: Obesity

Attempts	1	2–3	4–6	7+
		A: Regression with interactions		
	Base		Interactions	
Age: 18–39	-0.093***	0.007	0.009	0.031***
	(0.005)	(0.007)	(0.007)	(0.007)
40–49	-0.019***	0.013*	0.001	0.011
	(0.005)	(0.007)	(0.007)	(0.007)
60–69	-0.007*	0.004	0.005	0.004
	(0.004)	(0.006)	(0.006)	(0.006)
70 and up	-0.122***	0.012**	0.007	0.028***
	(0.004)	(0.006)	(0.007)	(0.007)
Children in household	0.023***	-0.007	-0.004	-0.018***
	(0.004)	(0.005)	(0.006)	(0.006)
Female	-0.004	-0.006*	-0.021***	-0.018***
	(0.003)	(0.004)	(0.004)	(0.004)
Educ: Less than high school	0.025***	-0.000	0.001	0.006
	(0.005)	(0.007)	(0.008)	(0.008)
Some college or tech. school	-0.007**	0.005	-0.002	0.005
	(0.003)	(0.005)	(0.005)	(0.005)
College graduate	-0.064***	0.002	-0.006	-0.007
	(0.004)	(0.005)	(0.005)	(0.005)
Inc: Below $25,000	0.035***	-0.003	-0.009	-0.016**
	(0.005)	(0.006)	(0.007)	(0.007)
$25,000-49,999	0.009*	-0.005	-0.006	-0.004
	(0.005)	(0.006)	(0.007)	(0.007)
$75,000 and up	-0.045***	0.003	0.001	0.008
	(0.005)	(0.006)	(0.007)	(0.007)
Race: Black, non-hispanic	0.123***	0.000	-0.010	0.013*
	(0.005)	(0.007)	(0.008)	(0.007)
Other, non-hispanic	-0.005	0.010	0.021**	-0.001
	(0.006)	(0.008)	(0.009)	(0.009)
Hispanic	0.033***	0.001	0.003	0.003
	(0.007)	(0.009)	(0.009)	(0.009)
Constant	0.370***	0.013	0.038**	-0.009
	(0.012)	(0.016)	(0.017)	(0.018)
		B: Adjusted means		
Obesity	0.297***	0.287***	0.282***	0.266***
	(0.001)	(0.001)	(0.001)	(0.002)

Notes: Source: Behavioral Risk Factor Surveillance System, 2012. $N = 450,212$ (1 attempt: 114,694; 2–3: 137,418; 4–6: 98,813; 7+: 99,287). $R^2 = 0.04$. Table reports estimates from a single OLS regression. Dependent variable: 0/1 obesity indicator. See page 12 for full explanation of table structure. Panel A: estimated coefficients from a fully interacted regression: each regressor is interacted with each difficulty-to-reach category (omitted category: 1 attempt). Regression also includes non-reported indicators (and their interactions) for missing data, marital status (7 categories), location (53), urban/rural (6), and interview month (12); see appendix note A.1. Standard errors in parentheses. Panel B: adjusted means, calculated from panel A regression. *** p<0.01, ** p<0.05, * p<0.1.

decreasing trend in adjusted mean weight, which drops from 80.4 [SE 0.06] to 79.1 [0.06] kg from the easiest to hardest category—a 1.3 kg ($\approx$ 2.9 lbs) difference. We also find evidence of cross-group differences—including across males and females and across some age and income groups—with overall changes in the range 0.8–1.6 [0.2–0.3] kg from easiest to hardest.

In contrast, adjusted mean height remains remarkably stable, at 169.3 [0.02–0.03] cm, across the four difficulty-to-reach categories, with small ($<$ 1 [0.1] cm) changes from easy to difficult in cross-sex and cross-race differences (table A.10).

III.3.2 Weighted regressions

Whereas the CPS and CEX are both multiple-interview panels from which we only analyze a single interview per household—MIS 1 in the CPS and second interview in the CEX (see next section)—the BRFSS is a cross section of individuals. As a result, unlike with the other two datasets, with the BRFSS we can use the provided (full sample) survey weights to probe the robustness of our unweighted analysis.

The BRFSS includes analysis weights designed to adjust for nonresponse and noncoverage in the survey and to make the number of cases sum to the population for each geographic region (usually state) (CDC, 2012). The weighting process takes into account the probability that someone was likely to be sampled. The weight is also raked on up to 12 demographic margins.[19] Appendix tables A.11, A.12, and A.13 recreate tables 5 (obesity), A.9 (weight), and A.10 (height), respectively, using the provided individual weights. In the adjusted means for obesity, the weighted prevalence trends from 29.4% to 28.4% to 27.2% to 26.1% [all SEs 0.3] across increasing difficulty-to-reach categories (compared with 29.7% to 28.7% to 28.2% to 26.6% in the unweighted table 5). We also find similar patterns in the adjusted means for weight and height across the weighted and unweighted tables, and for all three outcomes we additionally find the same general patterns, or lack thereof, in panel A's coefficients.

In summary, using the BRFSS weights does not change our in-sample conclusions.[20] This

[19] These are: age group by gender, race/ethnicity, education, marital status, home ownership status, gender by race and ethnicity, age group by race and ethnicity, phone ownership, region, region by age group, region by gender, region by race and ethnicity. See CDC (2012) for additional details.

[20] As an additional robustness check, appendix table A.14 recreates table 5 based on the 303,034 respondents (67.3% of our sample) who were interviewed by one of the 1,651 interviewers who conducted at least

finding questions the assumption that BRFSS nonrespondents are similar to the (properly weighted) average respondent.

IV CEX: Quarterly Household Expenditures

IV.1 Data

The Consumer Expenditure Survey (CEX), administered by the Bureau of Labor Statistics, measures the purchasing habits of U.S. consumers. The CEX's random sample of households is designed to be representative of the non-institutionalized, civilian population of the U.S.[21]

The CEX consists of two separate components—a diary survey and an interview survey—with two independent samples. We focus on the interview survey, which collects data on up to 95% of household expenditures. Its goal is to collect detailed data on large purchases that respondents can be expected to remember for three or more months (accounting for 60–70% of household purchases), and estimates for other major categories of expenses (accounting for 20–25% of purchases). The survey is a rotating panel: each quarter, approximately 20% of the sample is new; each household participates for five consecutive quarters.

The first interview collects demographic and other household details. Each of the second through fifth interviews collects expenditure data for the three preceding months; these data are used by the BLS to revise the weights in the Consumer Price Index, as well as to create national expenditure estimates. The survey has a quarterly target of approximately 7,000 participating households. Response rates among eligible households for 2008 to 2013 were, respectively, 73.8%, 74.5%, 73.4%, 70.4%, 69.5% and 66.7%—a fast decline of roughly 7 percentage points in five years.[22]

Beginning in 2008, the BLS releases detailed paradata for the CEX interview component.

10 interviews in each difficulty category; and appendix table A.15 does the same and adds a full set of interviewer controls and interviewer-difficulty interactions. As the tables show, our findings do not change, suggesting that using weights that accounted for such information would not change our conclusions. (We conduct this test only for BRFSS data because we do not observe interviewer ID in the CPS and CEX.)

[21] This section is based on the documentation available at `http://www.bls.gov/cex/` (accessed on May 26, 2016). This includes the percentages of household expenditures that the CEX is estimated to cover (reported in the next paragraph).

[22] As with the CPS, these figures correspond with AAPOR's Response Rate #2, #4, and #6. (There are no households of unknown eligibility, as CEX interviewers classify the eligibility of each address.)

The paradata contain information about each contact attempt, including when it occurred, its mode (in person or telephone), and its result. We create a number-of-contact-attempts variable by counting the contact-attempt entries in the paradata file. For consistency with our analyses in the previous sections, in our main analysis we use this variable for the first interview. However, since no relevant expenditure data are collected in the first interview, we use the second interview's expenditure data. (When we also use the second interview's number of contacts, our results below become stronger; see footnote 23.)

Our primary estimation sample is 2008:Q2–2013:Q4; we omit 2008:Q1 because we do not have paradata for its first interview, which occurred in 2007. We start with 39,277 observations, from which we drop 2,553 (6.5%) that could not be matched with first-interview paradata. We further exclude 2 observations from the main analysis of log total expenditures and 156 observations from the robustness analysis of log health expenditures due to negative expenditure values. We divide the resulting sample into four difficulty-of-reaching groups, as equal-sized as the contact-attempts distribution allows: 1 attempt (18.0% of the sample), 2 (19.7%), 3–4 (27.2%), and 5 or more (35.2%).

Our key outcome variable is log total quarterly expenditures (specifically, ln[1+expenditures]). When reporting and discussing results, we often transform the estimates back to dollar values (by exponentiating and subtracting 1); we use the delta method to transform SEs.

IV.2 Analysis by Difficulty of Reaching

IV.2.1 Sample composition

Table 6 reports substantial demographic differences across difficulty-to-reach groups in the CEX—a finding similar to those from the other two datasets. As in the CPS and BRFSS, the CEX's hard-to-reach are younger: the youngest two age categories, together covering ages 16–39, increase their share from 27.0% of the 1-attempt subsample to 33.5% of the 5+ subsample, while the 65+ age category falls from 30.9% to 15.6%. Also consistent with the other two datasets, the 5+ subsample is more educated and earns more than the 1-attempt subsample, with the proportions of college degree holders and of household income above $70,000 respectively increasing by 5.4 and 10.4 percentage points.

Table 6: CEX Demographics

Attempts		1	2	3–4	5+	All
Age:	16–29 (%)	13.2	10.0	11.0	13.5	12.1
		(0.4)	(0.4)	(0.3)	(0.3)	(0.2)
	30–39	13.8	15.2	17.1	20.0	17.2
		(0.4)	(0.4)	(0.4)	(0.4)	(0.2)
	40–49	14.3	18.7	19.8	22.1	19.4
		(0.4)	(0.5)	(0.4)	(0.4)	(0.2)
	50–64	27.8	28.9	30.2	28.8	29.0
		(0.6)	(0.5)	(0.5)	(0.4)	(0.2)
	65 and up	30.9	27.2	22.0	15.6	22.4
		(0.6)	(0.5)	(0.4)	(0.3)	(0.2)
Children in household		28.4	32.0	33.4	36.3	33.3
		(0.6)	(0.5)	(0.5)	(0.4)	(0.2)
Female		53.6	53.3	52.8	53.2	53.2
		(0.6)	(0.6)	(0.5)	(0.4)	(0.3)
Educ:	Less than high school	15.9	14.7	13.2	12.2	13.7
		(0.5)	(0.4)	(0.3)	(0.3)	(0.2)
	High school or GED	26.5	25.8	24.6	24.2	25.1
		(0.5)	(0.5)	(0.4)	(0.4)	(0.2)
	Some college or tech. school	30.4	28.8	29.9	30.9	30.1
		(0.6)	(0.5)	(0.5)	(0.4)	(0.2)
	College graduate	27.2	30.7	32.3	32.6	31.2
		(0.5)	(0.5)	(0.5)	(0.4)	(0.2)
Race:	White	84.3	83.0	82.3	77.8	81.2
		(0.4)	(0.4)	(0.4)	(0.4)	(0.2)
	Black	10.2	10.3	11.2	14.2	11.9
		(0.4)	(0.4)	(0.3)	(0.3)	(0.2)
	Asian	3.9	4.7	4.7	5.5	4.8
		(0.2)	(0.2)	(0.2)	(0.2)	(0.1)
	Other	1.7	2.0	1.9	2.5	2.1
		(0.2)	(0.2)	(0.1)	(0.1)	(0.1)
Inc:	Below $20,000	26.4	21.4	19.1	18.3	20.6
		(0.5)	(0.5)	(0.4)	(0.3)	(0.2)
	$20,000-39,999	24.8	23.0	21.3	21.1	22.2
		(0.5)	(0.5)	(0.4)	(0.4)	(0.2)
	$40,000-69,999	22.4	24.0	24.1	23.7	23.6
		(0.5)	(0.5)	(0.4)	(0.4)	(0.2)
	$70,000 and up	26.4	31.7	35.5	36.8	33.6
		(0.5)	(0.5)	(0.5)	(0.4)	(0.2)
Exp:	Total (log)	9.02	9.13	9.20	9.21	9.15
		(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
	Total ($)	8,225	9,200	9,865	9,990	9,459
		(74)	(79)	(72)	(64)	(36)
	Health (log)	5.12	5.20	5.12	4.83	5.03
		(0.03)	(0.03)	(0.03)	(0.03)	(0.01)
	Food (log)	7.15	7.24	7.31	7.33	7.27
		(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
Median number of attempts		1	2	3	7	3
Observations		6,603	7,230	9,973	12,918	36,724

Notes: Source: Consumer Expenditure Survey, 2008–2013. Sample: all second interview households that could be matched with a first-interview difficulty measure, excluding 2 and 156 observations with negative entries, respectively, in the total expenditure rows and in the health expenditure row. All figures (and standard errors) reflect proportions within each column's difficulty-of-reaching category, except for those for expenditures (which report average log expenditures and average log expenditures exponentiated into dollars), and number of attempts (which report medians).

There is also a qualitatively large and statistically significant trend in our key outcome, log total expenditures: exponentiated, they increase from $8,225 (1 attempt) to $9,200 (2), to $9,865 (3–4), to $9,990 (5+). Finally, the table also reports averages for two arguably important expenditure categories that have received much attention from economists and that we chose ahead of time (prior to looking at the data for specific categories): food and health. Average log food expenditures mirror the total expenditures pattern of monotonic (and highly statistically significant) increase, while health expenditures appear to generally (though nonmonotonically) decrease.

IV.2.2 Total expenditures

Table 7 has the same format as previous main-results tables. Panel A presents regression results for households' log quarterly total expenditures. Panel B's adjusted means are calculated from the panel A regression, and are reported both directly (in logs) and transformed back into dollars. The latter increase from $9,120 (amongst the 1-attempt subsample) to $9,331 (2), to $9,581 (3–4), and then stay flat at $9,585 (5+). Thus, accounting for the changing demographic composition of the subsamples, the difference between the easiest- and hardest-to-reach respondents shrinks from $1,765 in the unadjusted means (table 6) to $465—a still very significant 5% increase—in the adjusted means (table 7).[23] We find no significant trends in the coefficient estimates in Panel A.

[23] We view these estimates as somewhat conservative. Recall that we pair a household's difficulty-to-reach measure (from the first interview) with its outcome variable (expenditures from the second interview) across two different quarters. Therefore, if a household's difficulty of reaching in a specific interview is more strongly related to its outcome measures in that same interview, our estimates might be attenuated. To explore this possibility, appendix table A.16 replicates table 7, keeping the sample constant but using the difficulty-to-reach measure from each household's second interview. We indeed find similar patterns overall, with a larger expenditure gap between easiest- and hardest-to-reach respondents: the difference increases from $465 (table 7) to $769 (table A.16), reflecting both a drop from $9,120 to $8,942 among the easiest and a slightly smaller increase, from $9,585 to $9,711, among the hardest to reach. Furthermore, including additional controls for interview hour and duration—which are not available in our CPS and BRFSS data, and hence are not included in our main specification, but could in principle be incorporated into the survey weights—only decreases the (now nonmonotonic) expenditure gap in table 7 to $379 (table A.17)—still a very significant 4% increase—and the expenditure gap in table A.16 to $427 (table A.18).

Table 7: Total expenditures

Attempts		1	2	3–4	5+
		A: Regression with interactions			
		Base	Interactions		
Age:	16–29	-0.091***	-0.010	-0.004	0.012
		(0.025)	(0.035)	(0.032)	(0.030)
	30–39	-0.025	0.002	0.000	-0.022
		(0.023)	(0.030)	(0.028)	(0.027)
	50–64	0.000	-0.030	-0.002	-0.021
		(0.021)	(0.028)	(0.026)	(0.024)
	65 and up	-0.026	-0.029	-0.020	-0.046
		(0.023)	(0.031)	(0.029)	(0.028)
Children in household		0.063***	-0.048	-0.041	-0.022
		(0.023)	(0.031)	(0.029)	(0.027)
Female		-0.022*	0.019	0.006	0.025*
		(0.012)	(0.017)	(0.016)	(0.015)
Educ:	Less than high school	-0.152***	0.032	0.031	0.011
		(0.019)	(0.027)	(0.026)	(0.025)
	Some college or tech. school	0.104***	-0.018	-0.017	-0.031
		(0.016)	(0.023)	(0.021)	(0.020)
	College graduate	0.256***	-0.004	-0.013	-0.021
		(0.017)	(0.024)	(0.022)	(0.021)
Race:	Black	-0.076***	-0.019	-0.042	-0.016
		(0.020)	(0.028)	(0.026)	(0.024)
	Asian	-0.076**	-0.008	-0.030	0.024
		(0.032)	(0.042)	(0.039)	(0.037)
	Other	0.058	-0.125**	-0.115*	-0.037
		(0.047)	(0.063)	(0.059)	(0.055)
Inc:	Below $20,000	-0.614***	-0.020	0.008	0.034
		(0.019)	(0.027)	(0.025)	(0.024)
	$20,000-39,999	-0.257***	-0.011	0.014	0.025
		(0.018)	(0.025)	(0.023)	(0.022)
	$70,000 and up	0.396***	0.017	0.023	0.035
		(0.018)	(0.024)	(0.022)	(0.022)
Constant		9.172***	0.074	0.114**	0.022
		(0.037)	(0.051)	(0.048)	(0.046)
		B: Adjusted means			
Total expenditures (log)		9.118***	9.141***	9.168***	9.168***
		(0.006)	(0.006)	(0.005)	(0.004)
Total expenditures ($)		9,120***	9,331***	9,581***	9,585***
		(59)	(54)	(47)	(43)

Notes: Source: Consumer Expenditure Survey, 2008–2013. Sample: All households from table 6, excluding 2 households with negative net total quarterly expenditures. $N = 36,722$ (1 attempt: 6,603; 2: 7,229; 3–4: 9,972; 5+: 12,918). $R^2 = 0.56$. Table reports estimates from a single OLS regression. Dependent variable: ln(total quarterly household expenditures +1). See page 12 for full explanation of table structure. Panel A: estimated coefficients from a fully interacted regression: each regressor is interacted with each difficulty-to-reach category (omitted category: 1 attempt). Regression also includes non-reported indicators (and their interactions) for marital status (5 categories), size of consumer unit (3), urban/rural (2), interview month (12), and interview year (6); see appendix note A.1. Standard errors in parentheses. Panel B: adjusted means, calculated from panel A regression. *** p<0.01, ** p<0.05, * p<0.1.

IV.3 Robustness and Additional Results

IV.3.1 Health and food expenditures

Appendix tables A.19 and A.20 replicate table 7, with log health and log food expenditures as dependent variables. As with the unadjusted means, the trend in food expenditures is in line with the trend in total expenditures. It shows a 6% increase from easiest- to hardest-to-reach respondents—close to the 5% increase in total expenditures. In contrast, health expenditures show a highly significant (both statistically and economically) 13% *decrease* from easiest to hardest—opposite in sign and over double the percent increase in total (and in food) expenditures. As in table 7, there are no clear trends (and few significant interaction estimates) in panel A in either table A.19 or A.20.

This finding of significant and opposite trends in the adjusted means for food and health—the only two consumption categories we looked at—suggests that while some expenditure categories may be underestimated in the entire population, others may be *over*estimated. It highlights that without empirical evidence, it is difficult to know a priori the direction of potential bias.[24]

IV.3.2 Telephone and in-person interviews

As with the CPS, CEX interviews can occur either in person or over the telephone. In contrast with the CPS, in which only in-person attempts are recorded, in the CEX *all* contact attempt are supposed to be recorded, and we have no a priori reason to expect the number-of-contact-attempts variable to be a better or worse measure of difficulty across one mode or the other. Appendix tables A.21 and A.22 replicate table 7 separately for telephone and in-person interviews.[25] We observe the same general pattern in the in-person sample (24,305 interviews) and telephone sample (11,719 interviews) as we observe in the combined

[24] A posteriori, this and the finding of no clear trends in cross-demographic-group differences in either food or health may be consistent with some speculative interpretations of our results. For example, these findings are consistent with the untested idea that within the demographic groups we linearly control for (age, income, having children, etc.), households with healthier lives and lifestyles (spending more on food and less on health) are on average more difficult to reach. However, see footnote 26 for an alternative interpretation.

[25] We use a variable indicating interview mode as reported by the interviewer. Interviews reported as occurring through mixed modes (telephone and in-person) and interviews with missing mode data are excluded.

sample, but the in-person exponentiated adjusted means seem a few hundred dollars lower on average: \$8,923 [SE \$66] to \$9,387 [54] (in person, a 5% increase) versus \$9,344 [124] to \$9,920 [71] (telephone, a 6% increase) from easiest to hardest.

V Previous Related Work

V.1 Research Utilizing Difficulty-of-reaching Measures

Our paper contributes to an empirical literature that seeks to shed light on nonresponse bias by investigating the links between difficulty measures and survey-based outcomes. Within economics, we are aware of only two such investigations, both recent: Heffetz and Rabin (2013), discussed above; and Behaghel et al. (2014), discussed below. Outside of economics the literature is longer-established, larger, and growing. Heffetz and Rabin (2013) review it in some detail. This section focuses on new work that postdates their review, but we present it in the context of the earlier literature, organized by three strands: theoretically focused, empirically focused, and experimental.

First, on the theory-focused front, models have been developed where the probability of survey participation is related to the outcome of interest; see Potthoff, Manton and Woodbury (1993), who also analyze published number-of-calls data to fit parameters in their model. Our findings in the present paper appear consistent with such models. A more recent literature builds on such earlier work and aims at further utilizing paradata, including number of contact attempts, to directly improve analysis. One idea is to incorporate paradata in reweighting methods. Biemer, Chen and Wang (2013) present one such technique, a call-back model, in which the response propensity is modeled using information on the number of contact attempts. Krueger and West (2014) examine a similar but richer model that adds additional sources of auxiliary data, such as interviewer observations and characteristics of the interview location. Of particular interest to our analysis, the corrected weights sometimes move different subgroups' estimated outcome prevalence in opposite directions, suggesting that nonresponse bias may be differentially affecting different groups—consistent with our findings that the difficulty-outcome gradient may differ across demographic groups.

Coming from economics rather than from statistics and survey methodology, Behaghel et al. (2014) (mentioned above) propose using difficulty-of-reaching paradata for dealing with survey attrition, and apply their proposed corrective to a job-search experiment.

Second, on the empirically focused front, we noted above two studies that link administrative data to survey data (see footnote 3). Such studies can shed light on both nonresponse bias—by comparing administrative-data outcomes across survey nonrespondents and respondents (of different difficulty)—and measurement error—by comparing respondent outcomes (by difficulty) across administrative and survey data. Lin and Schaeffer (1995) match child-support awards and payments with a telephone survey and examine nonresponse bias. They provide some evidence that hard-to-reach respondents may not actually be more like nonrespondents than the easy-to-reach. However, their survey is conducted in a very different context than the large government surveys we study, as most of their nonrespondents were never located from the initial court records; and their resulting difficulty measure is also rather different, as they exclude contact attempts made prior to locating the respondent. Kreuter, Müller and Trappmann (2010) link administrative and survey data for a subsample of German unemployment-benefits recipients who were interviewed by telephone, and provide insight on both nonresponse and measurement: in their data, adding hard-to-reach respondents reduces nonresponse bias (the true averages of survey participants become closer to the true target-population averages), but increases measurement error (survey reports of the hard to reach are farther from their true values). For three of the four outcomes they study—employment status, age, and foreign citizenship—the net effect is to decrease the overall bias in survey outcomes (the reduction in nonresponse bias outweighs the increase in measurement error). But for the receipt of unemployment benefits the addition of hard-to-reach respondents actually increases the overall bias in the survey measure. These findings highlight the concern that unwilling—and thus hard-to-reach—respondents may be more prone to misreport and, in particular, misreport sensitive behaviors.

Related to the potential unwillingness of the hard to reach, two recent papers propose respondents' motivation as a potential explanation of the link between difficulty and outcomes. In the first, a working paper, Chadi (2014) documents a connection between a respondent's motivation to participate in the survey, number of contact attempts, and subjective happiness. In the second, Meyer, Mok and Sullivan (2015) examine the declining accuracy of survey data by combining survey and administrative measures of government transfers. They find that measurement error contributes as much to the observed bias in survey estimates as item nonresponse and unit nonresponse combined; they speculate that an increase in "over-surveyed" and unmotivated participants could increase measurement error. While the idea that the hard-to-reach are less motivated and more prone to measurement error cannot be directly assessed in our data, it could not alone easily explain our findings. For example, to explain the cross-demographic-group trends we find for some outcomes, such difficulty-motivation links would have to differ across demographics. For another example, they would have to *increase* food expenditures and *decrease* health expenditures from easy to difficult respondents.[26] At the same time, the idea that for some variables, item nonresponse increases with difficulty of reaching finds support in our data. For example, based on reported imputation flags in the CPS, we calculate imputation rates for twelve-month family income to be 19.2 percent (1 attempt), 20.8 (2), 23.9 (3+), and 27.8 percent (NR). (There is no clear way to directly link imputation flags to any of our main-outcome variables.)

Third, on the experimental front, in an influential study Keeter et al. (2000) conduct two telephone surveys—one over five days with a response rate of 36%, the other over eight weeks with a response rate of 61%—and find mostly small differences in outcomes, with demographics a notable exception. While our evidence in the present paper is consistent with the finding of significant demographic-composition changes as response rates increase, we also find significant changes in all four (non-demographic) key outcome variables we examine. The latter suggests that the earlier findings should not be misinterpreted as providing a blanket justification for drawing population-wide conclusions regarding non-demographic variables from surveys with low response rates.

While we are not aware of recent work investigating the generalizability of the above ex-

[26] Of course, the less motivated may be more reluctant, differentially across demographics and expenditures, to accurately report certain kinds of information. We cannot rule out, for example, that due to social-image considerations, the harder to reach over-report behaviors perceived as "positive"—labor force participation, total expenditures, and food expenditures—while under-reporting behaviors perceived as "negative"—unemployment, health expenditures, and (especially among women, all else equal) obesity. However, our finding (in table 2) that the increase in labor force participation with difficulty of reaching is *less* pronounced (all else equal) among men, the childless, and college graduates does not easily fit within such a social-image explanation.

perimental findings by running more experiments, the experimental finding of demographic differences across easy- and hard-to-reach respondents has been highlighted in recent empirical work. That work acknowledges the possibility of a link between difficulty and outcome variables, but leaves open the possibility that demographic controls may alleviate the problem—something that we show in this paper is not generally possible. Legleye et al. (2013), for example, examine the effect of increasing the number of contact attempts in a French telephone survey designed to measure sexual and reproductive health (SRH) issues. The inclusion of harder-to-reach respondents, who are found to differ on SRH behaviors from the rest of the sample, is shown to make the sample closer to the demographic composition of the target population. Similarly, a working paper by Pudney and Watson (2013) simulates the impact of reducing the number of call attempts in two health and employment longitudinal surveys—the British Household Panel Survey (BHPS) and the Household, Income and Labour Dynamics in Australia (HILDA)—and find that it would change the samples' demographic composition and outcomes of interest such as disability, ill health and employment. Other recent papers also find significant differences in demographics and outcome variables between the easy and difficult to reach, including Cohen, Rohde and Yu (2013), and Hetschko and Chadi (2017).

Recall that we find a consistent demographic and socioeconomic gradient across our three datasets: hard-to-reach respondents are younger and more educated, and they have higher household income, than easy-to-reach respondents. In addition, our outcome variables, which are associated with socioeconomic status, exhibit similar patterns: the difficult-to-reach are more likely to be employed (in the CPS) and are generally healthier (having lower obesity rates in BRFSS and lower medical expenditures in CEX). These findings are broadly consistent with those in the works reviewed above, although differences in methodology prevent direct comparisons.[27] This broad consistency highlights the fact that researchers interested in outcomes related to socioeconomic status may need to pay particular attention to the potential for nonresponse bias in their analyses.

[27] For example, Legleye et al. (2013) and Pudney and Watson (2013) both find that the employed are harder to reach, while Curtin, Presser and Singer (2000) find that those with higher income are harder to reach. Similarly, the existing evidence is consistent with the notion that the healthy are generally harder to reach than the unhealthy across a variety of measures such as annual medical expenditures, self-reported health, and disability status (Cohen, Rohde and Yu, 2013, Pudney and Watson, 2013).

V.2 Evidence on the Quality of Difficulty-of-reaching Measures

A limitation of paradata is that they are rarely collected for direct use by analysts. Often a mere by-product of the data collection process, their quality may be lower than that of other survey data. Bates et al. (2010) examine the quality of paradata across three federal surveys: the National Health Interview Survey (NHIS), the CEX, and the CPS. The three collect the data through a common Contact History Instrument (CHI).[28] The authors find that while most CHI entries in the CPS and NHIS were recorded immediately after the contact attempt, in the CEX almost 20% of the attempts were recorded with some delay. In all three datasets, attempts that did not result in a contact were more likely to be recorded later, and hence presumably had a higher chance of being forgotten, than those resulting in a contact. The authors also discuss an internal report according to which CEX interviewers estimated that they ever complete a record for only around 85% of their attempts. Similarly, Biemer, Chen and Wang (2011) conduct an informal survey of field interviewers and supervisors for the National Survey on Drug Use and Health (NSDUH), and find greater incentives to underreport than to overreport the number of visit attempts.[29] To the extent that the uneven underreporting found in these studies adds measurement error to our difficulty-of-reaching independent variables, our estimates may be attenuated. Our estimated difficulty trends may hence be viewed as lower bounds.

VI Discussion and Conclusion

Investigating three of the most commonly used government surveys, we find significant, systematic differences in key outcomes across number-of-contacts groups. These differences persist within demographic cells—indeed, some of them systematically differ across demographic groups—and cannot be eliminated by standard reweighting. They qualitatively replicate within subsamples of the data, and they generally grow when the number-of-contacts

[28] The CHI information is publicly available—and is the source of our contact attempts data—for the CEX but not for the CPS.

[29] According to the authors, overreporting could be caught through timesheet reviews while underreporting might help keep a case considered alive, thus helping the interviewer avoid being perceived as using time inefficiently.

measure seems cleaner.

That selection and nonresponse may bias survey outcomes is well-known theoretically but—judging by common practices—under-appreciated empirically. By demonstrating that even after adjusting for demographics, key outcome variables are strongly related, empirically, to respondents' difficulty of being reached—and hence, potentially, to their likelihood of survey participation—we hope to have convinced readers that a routine assumption of random nonresponse is hard to justify in these data.

In practice, how concerned should users of these and similar data be?

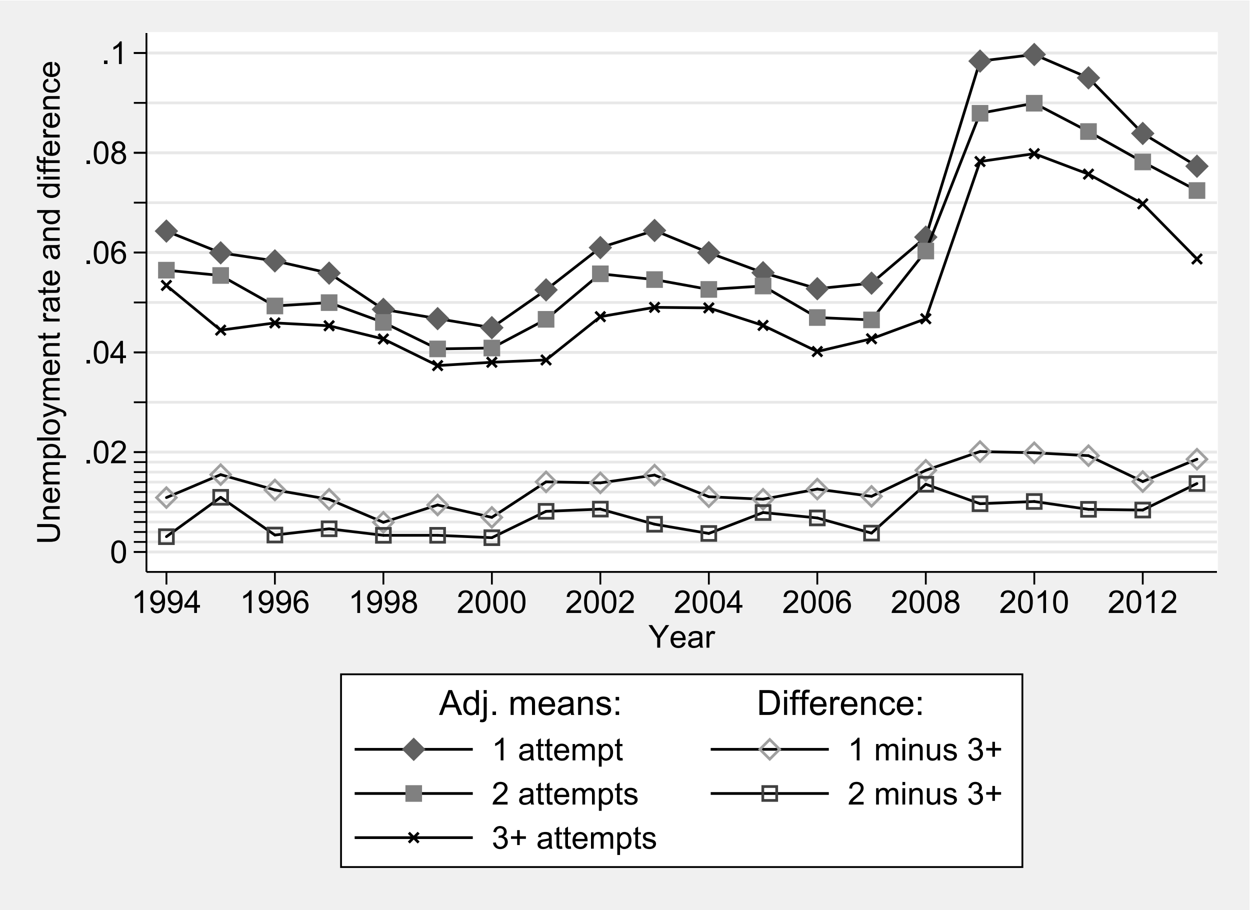
The answer depends on application and, importantly, on survey response rates: the lower these are, the more population-wide inferences from sample estimates rely on the assumption of random nonresponse (or random on observables). As we discussed when analyzing the BRFSS, our findings are consistent with the possibility that the population-wide prevalence of obesity could be 1 or 2 percentage points below the BRFSS sample mean, and, furthermore, cross-group comparisons of obesity (women versus men, for example) could change from an estimate of no difference to an estimate of a significant difference, depending on assumptions regarding nonresponders. Like the BRFSS, many telephone surveys nowadays have response rates below 50%—sometimes, well below that figure. In-person surveys still enjoy generally higher response rates, but they too show worrying trends: for example, as mentioned earlier, response rate in the CEX lost 7 percentage points from 2008 to 2013; this recent drop accelerated a previous, slower decline of 4 points from 2001 to 2008 (NRC, 2013). Brick and Williams (2012), who examine the causes of increasing nonresponse, note decreases in the response rate of several large, telephone and in-person, U.S. cross-sectional surveys from 1997–2007, including in the National Health Interview Survey (NHIS), National Household Education Survey (NHES), General Social Survey (GSS), and National Immunization Survey (NIS). Even among the most prominent in-person surveys, such as the CPS, response rates have been declining, dropping most recently from around 92% in 2010 to just above 89% in 2014 (Krueger, Mas and Niu, 2017). While the CPS's presently high response rates mean that from a practical point of view, our CPS findings are mostly academic, and (measurement error aside) the official CPS-based population-wide estimates are likely rather accurate, our findings underline the importance of keeping response rates high—something that may or may not be achievable in the future.

We conclude with a concrete demonstration of the value of high (and accurate) response rates, in the context of one important application. Figure 1 tracks the unemployment rate by difficulty of reaching over time. The solid shapes depict the panel-B adjusted-mean estimates from twenty replications of table 3, for the twenty years starting in 1994—the earliest year for which we have number-of-attempts data—with one exception: the "Children in household" control, which is not available for all years, is excluded from the regressions. The hollow shapes depict differences between these adjusted means. The figure generalizes our main finding from table 3: in each of the twenty years, the unemployment rate is significantly higher among the easy than among the difficult to reach. However, the easy-minus-difficult difference (hollow squares) is both highly variable and, importantly, highly correlated with the unemployment rate (solid shapes).[30] In other words, unemployment fluctuations are more extreme among easy than among difficult respondents. Qualitatively, the figure therefore suggests that the common practice of treating nonrespondents as average respondents yields larger cyclical fluctuations in unemployment estimates relative to treating them as difficult respondents. Quantitatively, this striking finding is of negligible significance *at present* (given current CPS response rates)—but it demonstrates the importance of keeping response rates and accuracy high.

While we know little about the increasingly many nonrespondents in the datasets we examine, we interpret our findings as, at the very least, suggesting that the burden of proof should lie with researchers whenever they make the random-nonresponse assumption in surveys with low response rates. Moreover, in specific applications this burden may extend across time and space. In particular, to the extent that the outcomes we examined are used for tracking the economy over time (as in the above example) or for making cross-country comparisons, our findings suggest that users should be concerned about, and try to control for, possible differences across periods and locations in the relation between outcomes and difficulty of reaching. Finally, this burden of proof may extend beyond outcomes' *means*—

[30] The correlation between the 1-attempt unemployment rate (solid squares) and the 1-attempt-minus-3+-attempts difference (hollow squares) is 0.83 ($N = 20$). This high correlation is accompanied by substantial variation in the 1-attempt-minus-3+-attempts difference: for example, in 2000, when the U.S. annual unemployment rate was the lowest since the late 1960s, the difference was 0.7%; in 2010, when unemployment was the highest since the early 1980s, the difference was 1.9%.

Figure 1: Unemployment rate, by number of contact attempts and year



Notes: Source: Current Population Survey, 1994–2013. See details in text.

the sole focus of the present paper. What is the relation, for example, between outcomes' *variance* and difficulty of reaching? If nonresponders' outcomes vary more than responders' outcomes, statistics that measure inequality may be affected too. We leave this question to future research.

Once the burden of proof lies with researchers, we trust that the ensuing demand for higher quality paradata, including difficulty-of-reaching measures, will hasten the process of these measures getting better, cleaner, and more widely available.

References

AAPOR. 2016. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys." 9th ed.

Bates, Nancy, James Dahlhamer, Polly Phipps, Adam Safir, and Lucilla Tan. 2010. "Assessing Contact History Paradata Quality Across Several Federal Surveys." In *JSM Proceedings of the Survey Research Methods Section of the American Statistical Association.* 91–105.

Behaghel, Luc, Bruno Crépon, Marc Gurgand, and Thomas Le Barbanchon. 2014. "Please Call Again: Correcting Non-Response Bias in Treatment Effect Models." *Review of Economics and Statistics*, 97(5): 1070–1080.

Biemer, Paul, Patrick Chen, and Kevin Wang. 2011. "Errors in the Recorded Number of Call Attempts and their Effect on Nonresponse Adjustments Using Callback Models." In *58th World Statistical Congress (Session IPS033).*

Biemer, Paul P., Patrick Chen, and Kevin Wang. 2013. "Using Level-of-Effort Paradata in Non-response Adjustments with Application to Field Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1): 147–168.

BRFSS. 2013. "2012 Summary Data Quality Report." http://www.cdc.gov/brfss/annual_data/2012/pdf/summarydataqualityreport2012_20130712.pdf (accessed May 26, 2016).

Brick, J. Michael, and Douglas Williams. 2012. "Explaining Rising Nonresponse Rates in Cross-sectional Surveys." *The ANNALS of the American Academy of Political and Social Science*, 645(1): 36–59.

CDC. 2012. "Weighting the Data." `http://www.cdc.gov/brfss/annual_data/2012/pdf/Weighting-the-data_webpage-content-20130709.pdf` (accessed May 26, 2016).

Chadi, Adrian. 2014. "Dissatisfied With Life or With Being Interviewed? Happiness and Motivation to Participate in a Survey." *SOEPpaper*, No. 639.

Cohen, Steven B., Fred Rohde, and William Yu. 2013. "Building Wave Response Rates in a Longitudinal Survey: Essential for Nonsampling Error Reduction or Last In-First Out?" *Field Methods*, 25(4): 361–387.

CPS. 2015. "Current Population Survey Interviewing Manual." `http://www2.census.gov/programs-surveys/cps/methodology/intman/CPS_Manual_April2015.pdf` (accessed May 26, 2016).

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): 413–428.

Dixon, John. 2013. "Using Paradata to Understand Panel Effects in the Current Population Survey Unemployment Rate." In *Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference*. 2–9.

Glaeser, Edward L., Joshua D. Gottlieb, and Oren Ziv. 2016. "Unhappy Cities." *Journal of Labor Economics*, 34(S2): S129–S182.

Heffetz, Ori, and Matthew Rabin. 2013. "Conclusions Regarding Cross-Group Differences in Happiness Depend on Difficulty of Reaching Respondents." *American Economic Review*, 103(7): 3001–3021.

Hetschko, Clemens, and Adrian Chadi. 2017. "The Magic of the New: How Job Changes Affect Job Satisfaction." *Journal of Economics and Management Strategy*, forthcoming.

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): 125–148.

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.

Krueger, Alan B., Alexandre Mas, and Xiaotong Niu. 2017. "The Evolution of Rotation Group Bias: Will the Real Unemployment Rate Please Stand Up?" *Review of Economics and Statistics*, 2(99): 258–264.

Krueger, Brian S., and Brady T. West. 2014. "Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments." *Public Opinion Quarterly*, 78(4): 795–831.

Legleye, Stéphane, Géraldine Charrance, Nicolas Razafindratsima, Aline Bohet, Nathalie Bajos, and Caroline Moreau. 2013. "Improving Survey Participation: Cost Effectiveness of Callbacks to Refusals and Increased Call Attempts in a National Telephone Survey in France." *Public Opinion Quarterly*, 77(3): 666–695.

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

Meyer, Bruce D., Wallace Mok, and James X. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives*, 29(4): 199–226.

NRC. 2013. "Nonresponse in Social Science Surveys: A Research Agenda." Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education, Washington, DC.

Oswald, Andrew J., and Stephen Wu. 2010. "Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A." *Science*, 327(5965): 576–579.

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): 1197–1207.

Pudney, Stephen, and Nicole Watson. 2013. "If at First You Don't Succeed? Fieldwork, Panel Attrition, and Health-Employment Inferences in BHPS." *ISER Working Paper Series*, 2013-27.

WHO. 1995. "The Use and Interpretation of Anthropometry Physical Status: Report of a WHO Expert Committee." WHO Techincal Report Series, No. 854, Geneva.

WHO. 2000. "Obesity: Preventing and Managing the Global Epidemic: Report of a WHO Consultation." WHO Techincal Report Series, No. 894, Geneva.

Williams, Richard. 2011. "Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects." In *Stata Conference, Chicago.*