Subjective Well-being and Social Desirability

James Reisinger*

November 9, 2021

Abstract

Survey measures of depression are increasingly used by economics researchers to provide a nuanced account of well-being. I show that levels of depression reported using such measures are significantly understated and levels of happiness significantly overstated in survey interviews conducted using a response mode that does not allow for anonymous reporting compared to a mode that does in three longitudinal surveys widely used in economics research. I exploit randomized assignment to survey mode, as well as panel methods, to show that this reflects the causal effect of survey mode, not selection. The difference in reported depression and happiness between modes is comparable to the difference between individuals in the 25th and 75th income percentiles. This finding suggests perceptions of social desirability may substantially bias measures of subjective well-being.

Key words: Subjective well-being, depression, social desirability bias, survey methods.

JEL codes: D9, C8, I3

1 Introduction

Depression is costly in human and economic terms. Globally, it is a leading cause of disability (James et al., 2018); mental distress is associated with increasing mortality due to suicide and substance use (Case and Deaton, 2020); and there is strong evidence of a two-way link between depression and economic outcomes (de Quidt and Haushofer, 2019; Ridley, Rao,

*Reisinger: Harvard Kennedy School, 79 John F. Kennedy St, Cambridge, MA 02138. Email: reisinger@g.harvard.edu. I am grateful to Michela Carlana and Matthew Rabin for detailed feedback and guidance. I am also thankful to the staff at NORC at the University of Chicago, especially Quentin Brummet, for answering my questions about survey procedures. All mistakes are mine.

Schilbach, and Patel, 2020). Yet, even among individuals with diagnosable mental illness, less than 1/3 receive treatment (Layard, Chisolm, Patel, and Saxena, 2013). While our ability to understand the causes and consequences of depression hinges on our ability to measure it, in many contexts, this measurement requires the use of survey-based depression questionnaires. Since clinical interviews are time-intensive and require substantial training to administer, survey measures of mental health are widely used by researchers to gauge mental health both in a developed (e.g., Case and Deaton, 2015, 2017; Finkelstein et al., 2012; Katz, Kling, and Liebman, 2001; Kling, Liebman, and Katz, 2007; Lindqvist, Östling, and Cesarini, 2020; Ludwig et al., 2012) and developing world context (e.g., Adhvaryu, Fenske, and Nyshadham, 2018; Baird, De Hoop, and Özler, 2013; Banerjee et al., 2015; Haushofer and Shapiro, 2016). The validity of these measures depends on whether survey respondents answer truthfully.

I present evidence that higher levels of depression and lower levels of happiness are reported by individuals interviewed in person compared to individuals interviewed by phone in three longitudinal surveys widely used across the social sciences: the National Longitudinal Survey of Youth, 1997 cohort (NLSY 97); the National Longitudinal Survey of Youth, 1979 main cohort (NLSY 79); and the National Longitudinal Survey of Youth, 1979 Children and Young Adult cohort (NLSY 79 CYA). The effect is equivalent to the difference in reported depression between respondents with college degrees and those who did not graduate high school; it is more than twice the difference between non-Hispanic White and Black respondents; and it is roughly equal to the difference between the 75th and 25th percentiles of income. I show that this difference reflects the causal effect of survey mode, not selection, using random assignment to phone or in-person survey mode in the 2015 wave of the NLSY 97. I then show the estimated effect is quantitatively similar when estimated in others years and surveys using a regression specification with individual fixed effects.

This effect is driven by anonymity provided for some sensitive questions during in-person

interviews but not phone interviews, suggesting social desirability bias is the primary driver. In most waves of the NLSY 97 and NLSY 79 CYA, depression measures are typically included in a self-administered questionnaire ("in-person SAQ") for in-person interviews but not for phone interviews. Questions included in the SAQ are completed by respondents anonymously, without the interviewer reading the question or observing the answer. For phone interviews, the respondent must report their answer to all questions directly to the interviewer. However, there are also questions related to depression included in the NLSY 97, NLSY 79, and NLSY 79 CYA but asked directly by the interviewer during in-person interviews ("in-person non-SAQ"). I find that the effect of interview mode on depression reporting is large and robust when comparing phone interviews to in-person SAQ, but the effects are small and not statistically significant for most depression measures when comparing phone interviews to in-person non-SAQ.

This paper extends an existing literature documenting differences in depression reporting between SAQ and non-SAQ surveys in limited contexts to three large, longitudinal surveys commonly used in social science research (Aquilino, 1998; Cernat, Couper, and Ofstedal, 2016; Chan et al., 2004; Epstein, Barker, and Kroutil, 2001). Additionally, while previous evidence on differences between phone and in-person non-SAQ interviews is mixed, I show that phone interviews induce lower levels of depression reporting but that the effects are small relative to the difference between phone interview and in-person SAQ (Aneshensel, Frerichs, Clark, and Yokopenic, 1982; Aquilino, 1998; Henson, Cannell, and Roth, 1978; Li et al., 2012). Surveys that rely entirely on phone interviews, like the Panel Study of Income Dynamics (PSID), and mixed mode surveys, like the British Household Panel Survey (BHPS) and recent waves of the National Longitudinal Study of Adolescent to Adult Health (ADD Health), may reflect differences between groups in the effect of survey mode, rather than true differences between the groups. This possibility is consistent with Heffetz and Rabin (2013) who find differences in the Michigan Survey of Consumers in measures of happiness between easy to reach and hard to reach respondents as measured by number of contact

attempts and show that these differences bias cross-group comparisons.

By isolating the role of anonymity in driving mode effects, I also extend the existing literature on social desirability bias to survey measures of depression. Previous research suggests anonymity (either provided through an SAQ or indirect elicitation techniques like the list method) may ameliorate the effects of social desirability bias and encourage honest reporting on sensitive topics (Bertrand and Duflo, 2017; Coffman, Coffman, and Ericson, 2016; Conti and Pudney, 2011; John, Loewenstein, and Prelec, 2012; Karlan and Zinman, 2012; Krumpal, 2013; Levitt and List, 2007; List et al., 2004; Tourangeau and Yan, 2007). Social desirability bias is a sub-genre of the experimenter demand effects illustrated by de Quidt, Haushofer, and Roth (2018) in which a respondent infers the interviewer's objective and is motivated to provide answers that further that objective. If a survey respondent is concerned with self-image or privacy, and they believe acknowledging depression or unhappiness is perceived negatively by the interviewer, the respondent may be motivated to minimize their symptoms. Such concerns may be due to a social stigma against mental illness (Bharadwaj, Pai, and Suziedelyte, 2017; Corrigan, 2004; Hatzenbuehler, Phelan, and Link, 2013). However, a respondent's attitudes about depression and their beliefs about others' attitudes may be more important than others' beliefs in biasing reports of mental illness (Corrigan, Watson, and Barr, 2006). Likewise, in the framework of de Quidt, Haushofer, and Roth (2018), it is a respondent's beliefs about the experimenter's goal that drive biased reporting, rather than the goal itself.

This paper also extends the existing literature on social desirability bias to measures of happiness and anxiety. The consistency of mode effects across domains suggests that these effects do not simply reflect hesitance to acknowledge one's depression to a stranger; rather, they suggest a tendency to exaggerate positive affect and a reticence to acknowledge negative affect. While Chan et al. (2004) reports differences by mode in response to the constituent questions of the CES-D, they do not find an effect on questions related to happiness or

anxiety, possibly due to a small sample size. Dolan and Kavestos (2016) find higher reported happiness in phone interviews than in-person interviews but use non-random assignment into survey mode. Happiness questions are widely used as measures of emotional or affective well-being (among many others, see Kahneman and Deaton, 2010; Kahneman and Krueger, 2006). The results of this study suggest the need to carefully consider the mode by which such happiness question are asked.

The remainder of this paper proceeds as follows. Section 2 describes the NLSY 97, the NLSY 79 main cohort, and the NLSY 79 Children and Young Adult cohort. Section 3 presents evidence that mode effects bias the reporting of depression and happiness in each survey. Section 4 concludes.

2 Data

I draw on three longitudinal surveys commonly used across the social sciences: the National Longitudinal Survey of Youth, 1997 cohort (NLSY 97); the National Longitudinal Survey of Youth, 1979 main cohort (NLSY 79); and the National Longitudinal Survey of Youth, 1979 Children and Young Adult cohort (NLSY 79 CYA). The NLSY 97 is a panel survey that began following a sample ages 12 to 18 in 1997. It was conducted yearly from 1997 to 2011, and then every other year. The NLSY 79 and NLSY 79 CYA are similar in structure and content to the NLSY 97. The NLSY 79 began following a cohort of individuals ages 14 to 22 in 1979 and was conducted annually until 1994 and every other year thereafter. The NLSY 79 CYA follows the children born to women in the NLSY 79 beginning in 1986 and is conducted every other year.

In each survey, the proportion of interviews conducted by phone has shifted over time (see Appendix Table A.1 for the exact proportions by year). From 1997 to 2013, the primary survey mode of the NLSY 97 was in-person interview, including both SAQ and non-SAQ

sections. However, in these years, around 10% of the sample was interviewed by phone. From 1998 to 2001, phone interviews were reserved for individuals who could not be reached any other way; e.g., respondents living overseas, respondents who were incarcerated, and respondents who refused an in-person interview. From 2002 to 2013, respondents could opt into a phone interview, but survey administrators tried to set up an in-person interview whenever possible. In 2015, the survey administrators began to prepare for a shift to the use of phone interviews as the primary mode. In this wave, the survey administrators randomly assigned respondents to phone interviews or in-person interviews to estimate the effect of phone interviews on respondent cooperativeness, but many respondents assigned to a phone interview chose to be interviewed in-person. In 2017, phone became the primary mode of interview, with in-person interviews conducted in about 10% of cases.[1]

Similarly, the NLSY 79 has also used a mix of phone and in-person reporting. Until 2000, most individuals were interviewed in person with around 5-15% of interviews in each wave conducted by phone. After 2000, most individuals were interviewed by phone with around 10-30% of interviews in each wave conducted in person. In the NLSY 79 Children and Young Adult cohort, from 1994 to 1998, the primary survey mode was in-person interview with 5-10% of the sample interviewed by phone. In 2000, the primary survey mode shifted to phone interview with 3-20% interviewed in-person.

All three surveys include measures of psychological well-being in various waves. Eight waves of the NLSY 97 include the five question Mental Health Inventory (MHI-5). This inventory includes questions asking how much of the time in the last 30 days the respondent felt nervous, calm, downhearted/blue, happy, or down in the dumps, answered on a 4 point scale. The usefulness of the MHI-5 for detecting depression has been confirmed by previous studies in a range of contexts and compares favorably with other commonly used questionnaires (Berwick et al., 1991; Cuijpers et al., 2009; Rumpf, Meyer, Hapke, and John, 2001). However,

[1]This information comes from private correspondence with survey administrators at NORC at the University of Chicago

there is no single established cut point to designate depression, so I interpret the total score as a gauge of overall mental health, converting it to z-score units for easier interpretation.

Several waves of the NLSY 79 and the NLSY 79 CYA include subsets of seven questions from the Center for Epidemiological Studies Depression (CES-D) questionnaire (Radloff, 1977). Respondents are asked to report how often they experienced certain feeling in the past week (e.g., "I felt depressed") using a four point scale. In the NLSY 79, these were administered at four points in time: the 1992 wave, the 1994 wave (around age 30), and in special health questionnaires administered in the waves immediately after a respondent turned 40 (1998-2006) and immediately after a respondent turned 50 (2008-2016). For the NLSY 79 CYA, respondents were asked the same seven questions in each wave from 1994 to 2016.

The NLSY 97 also intermittently includes three other questions related to mental health. In the 2009-11 waves, respondents were asked how many times in the past 12 months they had been treated by a mental health professional for an emotional, mental or psychiatric problem. Likewise, the NLSY 79 CYA also includes three related yes/no questions on whether the respondent received help for an emotional, behavioral, or family problem in the last 12 months (asked in 1994 to 2016); whether the respondent takes medication to control their activity level or behavior (asked in 1994 to 2016); and whether they ever considered attempting suicide (asked in 2012 to 2016).

3 The Effect of Survey Mode on Depression Reporting

3.1 Interview Timing, Interview Mode, and Depression

In each wave of the NLSY 97, depression exhibits a clear seasonal pattern. The mean is highest in October and decreases each month, reaching its lowest levels in late spring or early summer. This parallels the timing of interviews. Interviews begin in October of the

survey year and conclude in the spring or summer of the following year. Figure 1 panel a plots the average of depression by interview year and month. Vertical dotted lines demarcate the first month of interviews of each survey wave that included the depression questionnaire. Through the 2015 survey wave, average depression decreases each month after the start of the survey wave.

While this could reflect a seasonal pattern in underlying depression, I find that it reflects the causal effect of interview mode on reporting. In Figure 1 panel b, I show that the proportion of interviews conducted by phone follows the opposite trend. Interviews in the early months of a wave are mostly conducted in-person. Interviews in the later months are more likely to be conducted by phone. A difference in depression reporting between modes clearly explains the seasonal trend. In Figure 2 panel a, I plot the average of reported depression separately for each survey mode for each survey month, pooling across all years. The difference between modes is around 0.5 points in each month, and there is little difference in average depression across months.[2]

The difference between modes also persists across survey years with the exception of the 2017 wave where there is little difference between modes (depicted in Figure 2 panel b). The change in 2017 illustrates the mechanisms behind the effect. From 2000 to 2015, for in-person interviews, the depression inventory was included in a self-administered questionnaire (SAQ), along with sensitive questions on drug use, sexual activity, and criminal activity, but for phone interviews it was asked directly by the interviewer. In 2017, the depression questionnaire was administered directly by the interviewer in both phone and in-person interviews. Questions included in the SAQ are completed by respondents anonymously, without the interviewer reading the question or observing the answer. The interviewer turns their laptop to face the respondent who then completes the questionnaire by entering their

[2]Very few interviews were ever given in June, July, or August, so the appearance of convergence in these months is simply due to imprecision because of very small sample sizes. Nonetheless, I include these for completeness.

response directly, either reading the text on the screen or by listening to a recording of the questions using headphones. The administrators of the NLSY 97 implemented this approach to avoid interaction between the respondent and the interviewer during sensitive questions.[3] In this paper, I refer to the three methods of asking and responding to questions as 1) *in-person SAQ* for questions included in in-person interviews but answered anonymously via SAQ; 2) *in-person non-SAQ* for questions included in in-person interviews and asked directly by the interviewer; and 3) *phone interview* for all questions included in a phone interview and thus asked directly by the interviewer. In a survey wave, respondents are only asked a particular question via one of these three modes, but a respondent may be asked the same question by different modes in different waves.

The finding that respondents report lower levels of depression when they must respond directly to the interviewer, whether in person or by phone, suggests that perceptions of social desirability bias reporting. Of course, individuals interviewed by phone may be different from individuals interviewed in person. Thus, a naive comparison of depression between individuals interviewed by phone and in person by SAQ would suffer from omitted variable bias. In the next section, I leverage random assignment to survey mode in the 2015 wave of the NLSY 97 to show that differences in reporting reflects the causal effect of survey mode.

3.2 Estimates of the Causal Effect of Survey Mode on Depression Reporting Using the NLSY 97 2015 Phone Experiment

In 2015, the NLSY 97 survey administrators conducted an experiment to assess the difference between survey modes in the difficulty of completing interviews and in respondent cooperativeness in preparation for a planned shift to all phone interviews in 2017. To do this, they randomly assigned some respondents to be surveyed by phone and some to be surveyed

[3] https://www.nlsinfo.org/content/cohorts/nlsy97/intro-to-the-sample/interview-methods

in-person.[4] I leverage this random assignment to identify the causal effect of interview mode on reported depression. I show that being interviewed by phone has a clear, negative effect on total reported depression and happiness, measured using the MHI-5 inventory, when compared to reports given by in-person SAQ. The effect is qualitatively large and robust to numerous specifications. In contrast, there is no significant difference in reported depression, gauged by a report of ever missing a day of work due to poor mental health, between phone interviews and in-person non-SAQ interviews. This suggests the anonymity provided by the in-person SAQ drives differences in depression reporting between survey modes.

Eligibility for the experiment was restricted to respondents who had completed at least one interview in the 2010, 2011, and 2013 waves, and for whom no more than one of those three interviews was conducted by phone. From a population of 6905 deemed eligible, in a first randomization, the administrators assigned roughly 1/7 (966) to be interviewed by phone. This selection occurred before interviews began for the 2015 wave. No special incentives were given, but respondents were told at the beginning of the wave that they would be interviewed by phone. However, 287 of these individuals nonetheless chose to complete an interview in-person. To increase the number of individuals interviewed by phone, in December of 2015, two months after the start of interviews, another roughly 1/7 (983) of the original eligible sample was randomly selected for a phone interview. Since interviews had already begun, many of these individuals had already been interviewed. Thus, only 192 individuals from the second randomization were interviewed by phone.

I use assignment to the phone interview condition in the first randomization as an instrumental variable for survey mode in two-stage least squares (2sls) to estimate the effect of a phone interview on reported depression (the effect of treatment on the treated). Under the monotonicity assumption of Imbens and Angrist (1994) (I.e., if there are no defiers: individuals who would have been interviewed by phone but chose not to be solely because of

[4]The source of this information is personal correspondence with NLSY 97 survey administrators at NORC at the University of Chicago

the treatment assignment), this identifies the local average treatment effect of being interviewed by phone for compliers (individuals who were only interviewed by phone because of the treatment assignment). To avoid complications that may arise due to the difference in timing of the two randomizations, I do not use assignment in the second randomization as an instrument.

In Appendix Table A.2, I show that randomization was successful in balancing observable characteristics between the phone interview condition and in-person interview condition. To do this, I regress an indicator variable for treatment assignment in the first randomization on a series of lagged values of covariates. Since values of variables measured in 2015 may reflect the effect of assignment to the phone interview treatment, I use the most recent lagged value of each covariate. I show that survey mode assignment in either randomization is not associated with whether the respondent's 2013 wave interview was conducted by phone, total depression in 2010, or a set of lagged individual-level covariates, including self-reported health in 2013, household income in 2013, weeks worked in 2013, number of children in 2013, marital status in 2013, Census region of residence in 2013, a rural household indicator, number of address changes in 2013, age, sex, race, ethnicity, or a college graduate indicator.

While treatment compliance was not perfect, being assigned to the phone interview condition had a strong effect on whether an individual was interviewed by phone. Column 1 of Table 1 reports results from a regression of an indicator variable for being interviewed by phone on an indicator for assignment to the phone treatment in the first randomization:

$$P_{i,2015} = \alpha_0 + \alpha_1 T_{i,2015} + u_{i,2015} \tag{1}$$

where $P_{i,2015}$ is an indicator for whether individual i was interviewed by phone in the 2015

wave; and $T_{i,2015}$ is an indicator for assignment to the phone interview condition in the first randomization. For all estimates in this table, I report heteroskedasticity robust standard errors in parenthesis.

Assignment to the phone interview condition in the first randomization increased the probability of being interviewed by phone by 52.9 percentage points, significant at the 1% level. The heteroskedasticity-robust first-stage F-statistic is 1042.9.

In Columns 2 through 4, I report two-stage least squares estimates of the effect of a phone interview on total depression, happiness, and days of work missed due to poor mental health using treatment assignment as an instrumental variable. Importantly, in Columns 2 and 3, when using the MHI-5 depression total and happiness question as outcomes, the comparison is between individuals asked the question by phone and those asked by in-person SAQ, as the MHI-5 questionnaire included in the SAQ section of all in-person interviews. In contrast, in Column 4, the comparison is between phone and in-person non-SAQ, as the question about days of missed work was part of the non-SAQ section of all in-person interviews. The second stage specification is

$$Y_{i,2015} = \beta_0 + \beta_1 P_{i,2015} + u_{i,2015} \tag{2}$$

where $Y_{i,2015}$ is the outcome of interest for individual i in the 2015 wave.

In Column 2, I show that individuals interviewed by phone report a 0.211 SD lower value of total depression than individuals interviewed by in-person SAQ, significant at the 1% level. For comparison, in 2015, men report depression levels that are 0.252 SD lower than women. Black respondents report depression levels that are 0.111 SD lower than White respondents. Respondents with college degrees report depression levels that are 0.213 SD

lower than respondents who did not graduate high school. Respondents at or above the 75th percentile of household income ($97,000) report depression that is 0.215 SD lower than respondents at or below the 25th percentile of household income ($29,400).

In Column 3, I show that individuals interviewed by phone report a 0.185 SD higher value on the MHI-5 happiness question than individuals interviewed by in-person SAQ, significant at the 1% level. For comparison, in 2015, men report happiness levels that are 0.0807 SD higher than women. Black respondents report happiness levels that are 0.0979 SD higher than White respondents. Respondents with college degrees report happiness levels that are 0.0558 SD higher than respondents who did not graduate high school. Respondents at or above the 75th percentile of household income ($97,000) report happiness levels that are 0.138 SD higher than respondents at or below the 25th percentile of household income ($29,400).

In Appendix Table A.3, I also report the estimates for each of the other constituent variable of the MHI-5 depression index, including nervousness, calmness/peacefulness, downheartedness/blueness, and feelings of being down in the dumps. I find that mode effects are consistent across these variables, with phone interviews resulting in more "positive" responses.

In Column 4, I show that individuals interviewed by phone are 2.79 percentage points less likely than individuals interviewed by in-person non-SAQ to report they had missed at least one day of work, school, or other activities due to mental health in the last 12 months, but the estimate is not statistically significant. The fact that this comparison is not statistically significant suggests that the SAQ drives the effect, rather than some other aspect of in-person or phone interviews.

3.3 Panel Estimates of Mode Effects

I this Section, I extend the analysis of mode effects to other waves of the NLSY 97, as well as the NLSY 79 and NLSY 79 CYA. I show that there are large and robust differences in depression reporting when comparing phone interviews to in-person SAQ, estimated using a regression specification that includes individual and survey wave fixed effects, as well as time-varying controls. In comparisons of phone interviews to in-person non-SAQ, the effects are smaller and often not statistically significant. This suggests the primary mechanism is the anonymity of the in-person SAQ, but that other differences between phone and in-person interviews may still play a role. One reason to expect a difference between phone interviews and in-person non-SAQ is that there is greater "social distance" in phone interviews, so that it is more difficult for interviewers to build rapport with respondents or assuage concerns about privacy (Aquilino, 1998).

In Table 2, I report a series of regressions using various years and depression measures from the NLSY 97, NLSY 79, and NLSY 79 CYA. Panel A reports comparisons between phone interview and in-person SAQ. Panel B reports comparisons between phone interview and in-person non-SAQ. Each regression takes the following general form:

$$Y_{it} = \alpha_i + \tau_t + \beta_1 P_{it} + X'_{it}\delta + u_{it} \tag{3}$$

where Y_{it} is a measure of depression for individual i in survey wave t; P_{it} is an indicator for whether individual i was interviewed by phone in survey wave t; α_i is an individual fixed effect; τ_t represents a survey wave or year fixed effect; X_{it} is a vector of time-varying individual controls; u_{it} is an idiosyncratic error term assumed to be independent between individuals but not within individuals. Standard errors in all specifications are clustered at

the individual level.

In each specification, controls include a quadratic in age, number of children, marital status, log of household income, a rural household indicator, Census region of residence (including residences outside the US), whether the individual is incarcerated, and the interview calendar month. I code missing values of control variables as zero and include missing value indicator variables for each.

I report coefficient estimates on P_{it} for each comparison both without (model 1) and with individual fixed effect (model 2). Implicitly, the estimate using individual fixed effects is based on a comparisons of each individual's reported depression in waves where they were interviewed by phone to waves where they were interviewed in-person. Thus, it controls for fixed individual characteristics that may determine depression or selection into survey mode. However, there are still plausible sources of omitted variable bias in this comparison. For instance, if someone experiences the death of a family member, they may be more reluctant to participate and thus more likely to be interviewed by phone in that year. If such time-varying unobserved factors are systematically correlated with both interview mode and reported depression, these estimates may be biased.

Panel A of Table 2 establishes that reported depression is consistently lower in phone interviews than in-person SAQ. Each estimate in this panel is significant at the 5% or 1% level. The effect of a phone interview estimated with individual fixed effects on the standardized total of the MHI-5 questions in the 2000 to 2015 waves of the NLSY 97 is -0.244 SD, significant at the 1% level. Reassuringly, this estimate is similar in magnitude to the IV estimates reported in Table 1. The effect on the standardized value of reported happiness is 0.134 SD, significant at the 1% level.

Using the NSLY 79 CYA, I find that respondents interviewed by phone are 3.71 percentage points less likely to report receiving help for emotional, family or personal problems in the

last 12 months than individuals interviewed by in-person SAQ, significant at the 1% level. They are 3.46 percentage points less likely to report seriously considering suicide in the last 12 months, significant at the 5% level. And they are 1.08 percentage points less likely to report taking medication to control their behavior or activity level, significant at the 1% level.

Panel B of Table 2 establishes that the difference in reported depression between phone interviews and in-person non-SAQ is smaller and less robust. However, the point estimates consistently indicate that lower levels of depression are reported in phone interviews. In the NLSY 97, there is not a statistically significant difference between phone and in-person non-SAQ interviews for reports of whether the respondent had missed work, school, or other activities due to mental health problems or whether the respondent had sought professional help for a mental health problem in the last 12 months. However, for both outcomes and both specification, the point estimate is negative.

Pooling the 1992 and 1994 waves of the NLSY 79, the difference in reported depression measured using seven questions from the CESD between phone and in-person non-SAQ interviews is -0.150 SD when excluding individual fixed effects, significant at the 1% level. The effect is only -0.0647 SD when including individual fixed effects, significant at the 10% level. When pooling surveys administered at age 40 and age 50, the estimated difference without individual fixed effects is -0.0792 SD, significant at the 1% level. With individual fixed effects, the estimate is 0.00208 SD but is not statistically significant.

Finally, in the NLSY 79 CYA, the difference in reported depression measured using seven questions from the CESD between phone and in-person non-SAQ interviews is -0.0825 SD when excluding individual fixed effects, significant at the 1% level. With individual fixed effects, the estimate is -0.0142 SD and is not statistically significant.

These estimates indicate that mode effects in depression reporting are driven mainly by the

anonymity of the in-person SAQ. Nonetheless, the point estimates in Panel B consistently suggest that there are lower reports of depression and related outcomes in phone interviews than in-person non-SAQ. Thus, phone interviews may induce more socially desirable reporting than in-person non-SAQ interviews.

3.4 Robustness

In Appendix Section A.2, I present analyses showing the robustness of these estimates to item non-response to depression questions and survey attrition. In Appendix Table A.4, I show that non-response to depression questions is higher in in-person SAQ than phone interviews, but in Appendix Section A.2.1, I present worst case treatment-effect bounds that show that non-response does not substantially bias the estimated effect of interview mode. In Appendix Section A.2.2, I show that attrition likely does not bias the result, as attriters respond similarly to non-attriters in waves in which they are interviewed.

4 Conclusion

This paper presents evidence that socially desirable reporting biases measures of depression and happiness in three widely used longitudinal surveys: the National Longitudinal Survey of Youth, 1997 cohort; the National Longitudinal Survey of Youth, 1979 cohort; and the the National Longitudinal Survey of Youth, Children and Young Adult cohort. Respondents consistently report lower levels of depression and higher levels of happiness in phone interviews than in-person self-administered questionnaires. Using a 2015 experiment in which NLSY 97 respondents were randomly assigned to be interviewed by phone or in-person SAQ, I show that these differences are due to the causal effect of mode on reporting, not selection. Panel regressions with individual fixed effects produce similar estimates across other years of the NLSY 97 and the NLSY 79 CYA. I show that these differences are likely attributable

to the anonymity provided for some questions during in-person but not phone interviews, highlighting social desirability bias as the mechanism.

References

Adhvaryu, A., J. Fenske, and A. Nyshadham (2018, November). Early Life Circumstance and Adult Mental Health. *Journal of Political Economy 127*(4), 1516–1549.

Aneshensel, C. S., R. R. Frerichs, V. A. Clark, and P. A. Yokopenic (1982). Measuring Depression in the Community: a Comparison of Telephone and Personal Interviews. *Public Opinion Quarterly 46*(1), 110–121.

Aquilino, W. (1998). Effects of Interview Mode on Measuring Depression in Younger Adults. *Journal of Official Statistics 14*(1), 15–29.

Baird, S., J. De Hoop, and B. Özler (2013). Income Shocks and Adolescent Mental Health. *Journal of Human Resources 48*(2), 370–403.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015, May). A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries. *Science 348*(6236), 1260799.

Bertrand, M. and E. Duflo (2017, January). Chapter 8 - Field Experiments on Discrimination. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 1 of *Handbook of Field Experiments*, pp. 309–393. North-Holland.

Berwick, D. M., J. M. Murphy, P. A. Goldman, J. E. Ware, A. J. Barsky, and M. C. Weinstein (1991, February). Performance of a Five-Item Mental Health Screening Test. *Medical Care 29*(2), 169–176.

Bharadwaj, P., M. M. Pai, and A. Suziedelyte (2017, October). Mental Health Stigma. *Economics Letters 159*, 57–60.

Case, A. and A. Deaton (2015, December). Rising Morbidity and Mortality in Midlife among White non-Hispanic Americans in the 21st Century. *Proceedings of the National Academy of Sciences 112*(49), 15078–15083.

Case, A. and A. Deaton (2017). Mortality and Morbidity in the 21st Century. *Brookings Papers on Economic Activity*, 397–476.

Case, A. and A. Deaton (2020, March). *Deaths of Despair and the Future of Capitalism*. Princeton: Princeton University Press.

Cernat, A., M. P. Couper, and M. B. Ofstedal (2016, December). Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models. *Journal of Survey Statistics and Methodology 4*(4), 501–524.

Chan, K. S., M. Orlando, B. Ghosh-Dastidar, N. Duan, and C. D. Sherbourne (2004). The Interview Mode Effect on the Center for Epidemiological Studies Depression (CES-D) Scale: An Item Response Theory Analysis. *Medical Care 42*(3), 281–289.

Coffman, K. B., L. C. Coffman, and K. M. M. Ericson (2016, August). The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated. *Management Science 63*(10), 3168–3186.

Conti, G. and S. Pudney (2011, April). Survey Design and the Analysis of Satisfaction. *The Review of Economics and Statistics 93*(3), 1087–1093.

Corrigan, P. (2004, October). How Stigma Interferes with Mental Health Care. *The American Psychologist 59*(7), 614–625.

Corrigan, P. W., A. C. Watson, and L. Barr (2006). The Self-Stigma of Mental Illness: Implications for Self-Esteem and Self-Efficacy. *Journal of Social and Clinical Psychology 25*(8), 875–884.

Cuijpers, P., N. Smits, T. Donker, M. ten Have, and R. de Graaf (2009, August). Screening for Mood and Anxiety Disorders with the Five-Item, the Three-Item, and the Two-Item Mental Health Inventory. *Psychiatry Research 168*(3), 250–255.

de Quidt, J. and J. Haushofer (2019). Depression through the Lens of Economics: A Research Agenda. In *The Economics of Poverty Traps*, pp. 127–152. Cambridge, MA: National Bureau of Economic Research.

de Quidt, J., J. Haushofer, and C. Roth (2018, November). Measuring and Bounding Experimenter Demand. *American Economic Review 108*(11), 3266–3302.

Dolan, P. and G. Kavetsos (2016, June). Happy Talk: Mode of Administration Effects on Subjective Well-Being. *Journal of Happiness Studies 17*(3), 1273–1291.

Epstein, J. F., P. R. Barker, and L. A. Kroutil (2001, February). Mode Effects in Self-Reported Mental Health Data. *Public Opinion Quarterly 65*(4), 529–549.

Finkelstein, A., S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group (2012, August). The Oregon Health Insurance Experiment: Evidence from the First Year. *The Quarterly Journal of Economics 127*(3), 1057–1106.

Hatzenbuehler, M. L., J. C. Phelan, and B. G. Link (2013, May). Stigma as a Fundamental Cause of Population Health Inequalities. *American Journal of Public Health 103*(5), 813–821.

Haushofer, J. and J. Shapiro (2016, November). The Short-term Impact of Unconditional Cash Transfers to the Poor: Experimental Evidence from Kenya. *The Quarterly Journal of Economics 131*(4), 1973–2042.

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

Henson, R., C. F. Cannell, and A. Roth (1978, June). Effects of Interview Mode on Reporting of Moods, Symptoms, and Need for Social Approval. *The Journal of Social Psychology 105*(1), 123–129.

Horowitz, J. L. and C. F. Manski (2000, March). Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. *Journal of the American Statistical Association 95*(449), 77–84.

Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

James, S., D. Abate, K. H. Abate, and Others (2018, November). Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet 392*(10159), 1789–1858.

John, L. K., G. Loewenstein, and D. Prelec (2012, May). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science 23*(5), 524–532.

Kahneman, D. and A. Deaton (2010). High Income Improves Evaluation of Life but Not Emotional Well-Being. *Proceedings of the National Academy of Sciences 107*(38), 16489–16493.

Kahneman, D. and A. B. Krueger (2006). Developments in the Measurement of Subjective Well-Being. *The Journal of Economic Perspectives 20*(1), 3–24.

Karlan, D. S. and J. Zinman (2012, May). List Randomization for Sensitive Behavior: An Application for Measuring Use of Loan Proceeds. *Journal of Development Economics 98*(1), 71–75.

Katz, L., J. Kling, and J. Liebman (2001). Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment. *Quarterly Journal of Economics cxvi*(2), 607–654.

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

Krumpal, I. (2013, June). Determinants of Social Desirability Bias in Sensitive Surveys: a Literature Review. *Quality & Quantity 47*(4), 2025–2047.

Layard, R., D. Chisolm, V. Patel, and S. Saxena (2013). Happiness and Mental Health. In J. Helliwell, R. Layard, and J. Sachs (Eds.), *World Happiness Report 2013*. UN Sustainable Development Solutions Network.

Levitt, S. D. and J. A. List (2007, June). What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World? *Journal of Economic Perspectives 21*(2), 153–174.

Li, C., E. S. Ford, G. Zhao, J. Tsai, and L. S. Balluz (2012, February). A Comparison of Depression Prevalence Estimates Measured by the Patient Health Questionnaire with Two Administration Modes: Computer-Assisted Telephone Interviewing versus Computer-Assisted Personal Interviewing. *International Journal of Public Health 57*(1), 225–233.

Lindqvist, E., R. Östling, and D. Cesarini (2020). Long-Run Effects of Lottery Wealth on Psychological Well-Being. *The Review of Economic Studies*.

List, J. A., R. P. Berrens, A. K. Bohara, and J. Kerkvliet (2004, June). Examining the Role of Social Isolation on Stated Preferences. *American Economic Review 94*(3), 741–752.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling, and L. Sanbonmatsu (2012, September). Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science 337*(6101), 1505–1510.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement 1*, 385–401.

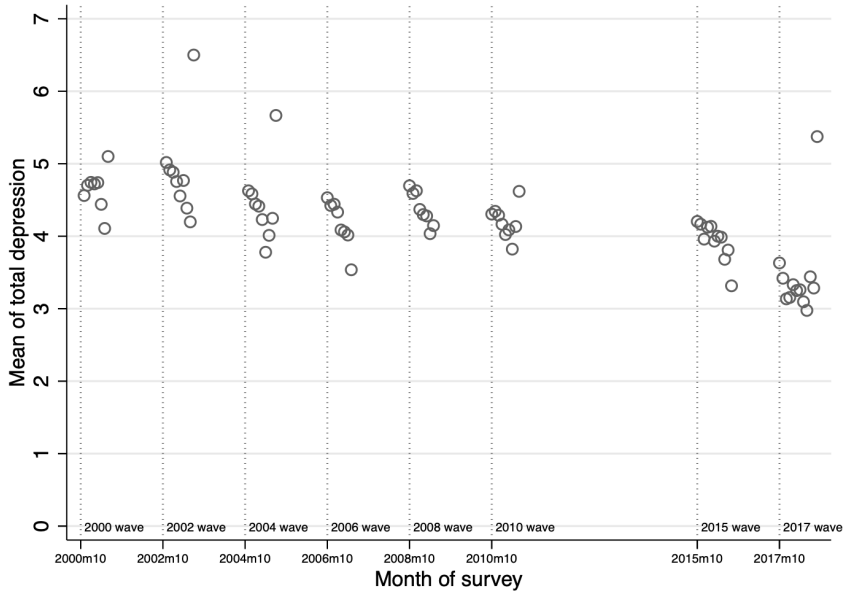
Ridley, M. W., G. Rao, F. Schilbach, and V. H. Patel (2020, May). Poverty, Depression, and Anxiety: Causal Evidence and Mechanisms. Technical Report 27157, National Bureau of Economic Research, Inc.

Rumpf, H. J., C. Meyer, U. Hapke, and U. John (2001, December). Screening for Mental Health: validity of the MHI-5 using DSM-IV Axis I Psychiatric Disorders as Gold Standard. *Psychiatry Research 105*(3), 243–253.

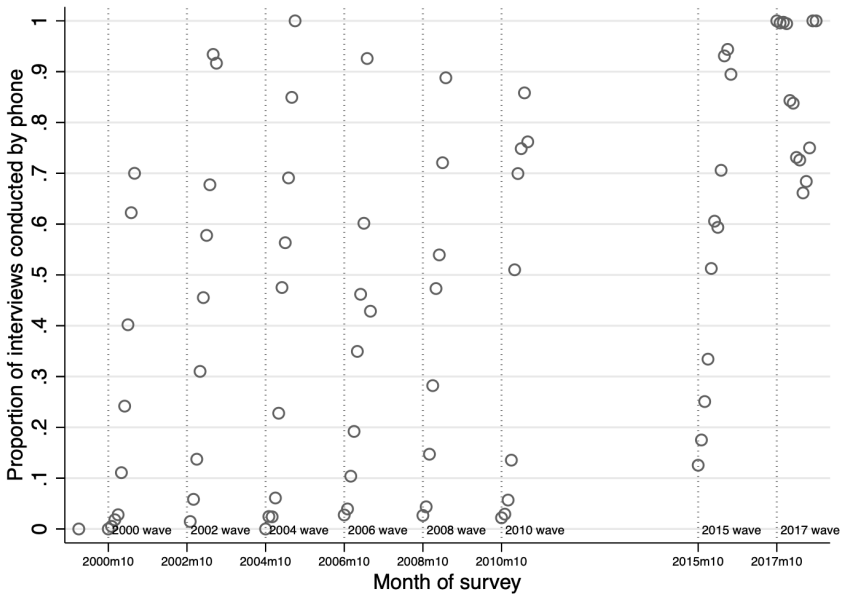
Tourangeau, R. and T. Yan (2007). Sensitive Questions in Surveys. *Psychological Bulletin 133*(5), 859–883.

Figure 1: Depression and Phone Interviews by Survey Month, NLSY 97

(a) Depression by Year and Month of Survey, NLSY 97



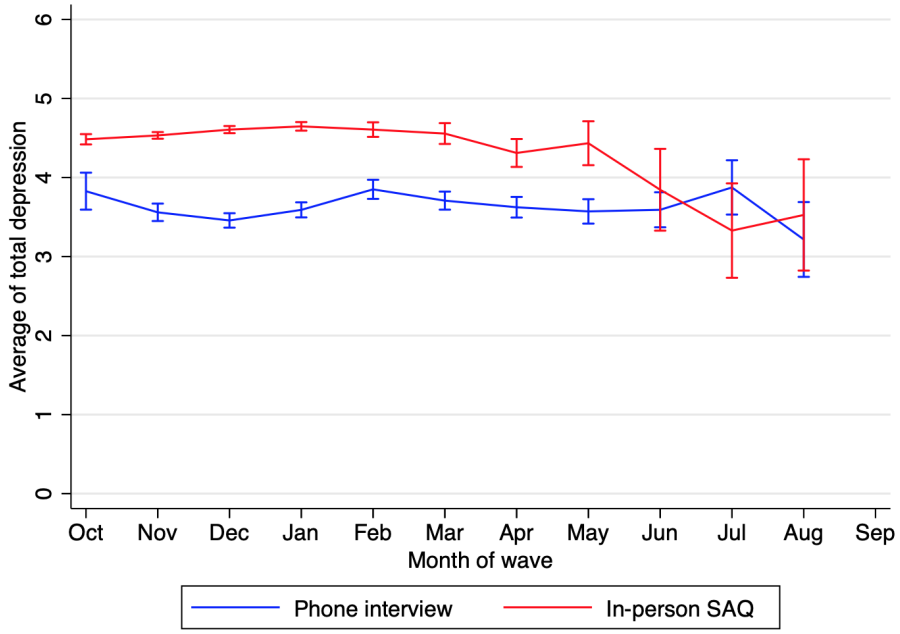
(b) Phone Interviews by Year and Month of Survey, NLSY 97



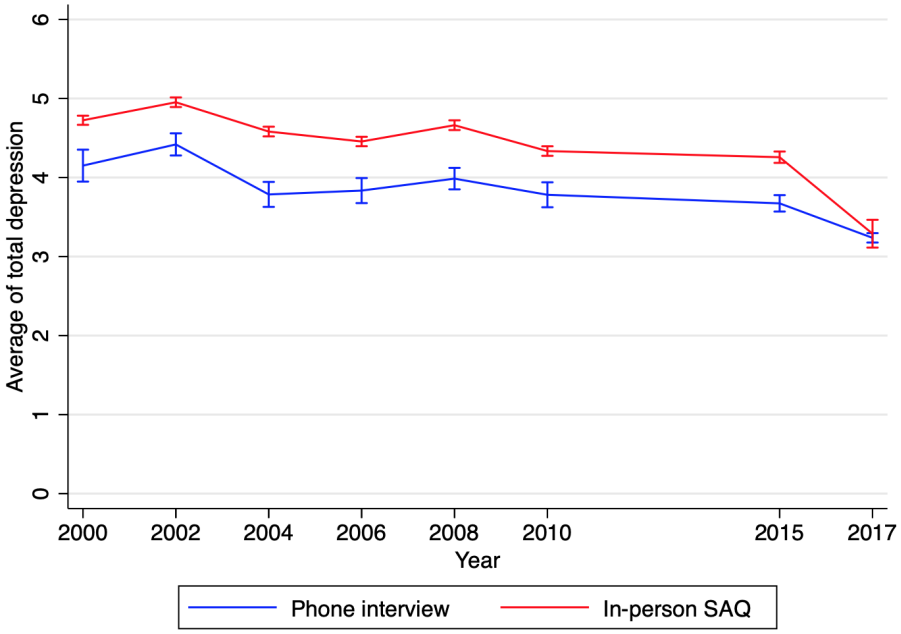
Note: The top panel depicts the average total of the MHI-5 questionnaire by the month a survey was conducted for the 2000, 2002, 2004, 2006, 2008, 2010, 2015, and 2017 waves of the NSLY 97. Total are calculated using the method described in Appendix Section A.1.1. The bottom panel depicts the proportion of surveys conducted by phone by survey month.

Figure 2: Average of MHI-5 Total, NLSY 97

(a) Average by survey month (all waves)



(b) Average by survey year



Note: The top figure depicts the average of the total MHI-5 questionnaire score for each mode of administration by month of survey and survey mode, pooling responses in the the 2000, 2002, 2004, 2006, 2008, 2010, 2015, and 2017 waves of the NSLY 97. The bottom figure reports the average for each mode of administration by survey year. Total are calculated using the method described in Appendix Section A.1.1.

Table 1: IV Estimates of the Effect of Phone Interviews on Depression Reporting, NLSY 97 2015 Wave

	(1) Phone interview	(2) Depression Total z-score	(3) Happy z-score	(4) Missed work
	ols	2sls	2sls	2sls
Assigned to phone, 1st randomization	0.529***			
	(0.0164)			
Phone interview		−0.211***	0.185***	−0.0242
		(0.0660)	(0.0682)	(0.0227)
HR first stage F-stat	1042.90			
N	6325	6325	6325	6314

Note: The sample is restricted to 2015 observations of individuals designated as eligible for the 2015 phone experiment. Eligible individuals were randomly selected to be surveyed by phone in 2015. Not all of the individuals selected for a phone survey were ultimately surveyed by phone, so I use assignment as an instrumental variable and estimate effects with 2sls. Column 1 reports the first-stage coefficient using ols. The outcome variable is an indicator for whether the interview was conducted by phone in 2015. Columns 2 through 4 report the 2sls estimate of the effect of a phone interview. The outcome variable in Column 2 is standardized total depression measured using the MHI-5 in 2015. The outcome variable in Column 3 is the standardized response to happiness question on the MHI-5 in 2015. The outcome variable in Column 4 is an indicator for reporting ever missing work due to poor mental health. Heteroskedasticity robust standard errors are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Panel Estimates of the Effect of Interview Mode on Depression Reporting in the NLSY 97, NLSY 79, and NLSY 79 CYA

	Survey	(1)	(2)	N
Panel A: Phone vs. in-person SAQ				
Depression total (z-score)	NLSY 97 (2000-15)	−0.251844***	−0.242797***	53099
		(0.015149)	(0.013590)	
Happy (z-score)	NLSY 97 (2000-15)	0.143079***	0.135104***	53099
		(0.015888)	(0.015489)	
Received help for emotional problem	NLSY 79 CYA (1994-2016)	−0.040677***	−0.037128***	49163
		(0.005263)	(0.005061)	
Ever seriously considered suicide	NLSY 79 CYA (2012-16)	−0.036753***	−0.034618**	16672
		(0.011877)	(0.014491)	
Uses medication to control behavior	NLSY 79 CYA (1994-2016)	−0.013235***	−0.010853***	49260
		(0.003938)	(0.003629)	
Panel B: Phone vs. in-person non-SAQ				
Missed any days due to mental health	NLSY 97 (2009-17)	−0.004230	−0.002047	43323
		(0.005197)	(0.005363)	
Treated by mental health professional	NLSY 97 (2009-11)	−0.002517	−0.002796	22413
		(0.006252)	(0.006872)	
Depression total (z-score)	NLSY 79 (1992-1994)	−0.150072***	−0.064732*	17511
		(0.025222)	(0.035108)	
Depression total (z-score)	NLSY 79 (Age 40 and 50)	−0.079218***	−0.002084	15858
		(0.025643)	(0.032124)	
Depression total (z-score)	NLSY 79 CYA (1994-2016)	−0.082542***	−0.014162	38484
		(0.019477)	(0.018310)	
Individual controls		Yes	Yes	
Survey wave FEs		Yes	Yes	
Individual FEs			Yes	

Note: Each row reports the estimated effect of a phone interview versus in-person interview from two separate regression models. Panel A reports the effect of phone interviews versus in-person SAQ. Panel B reports the effect of phone interviews versus in-person non-SAQ. The outcome variable is listed in the first column of each row. The regression is restricted to the indicated survey and years. The unit of observation is individual by survey wave. Model (1) includes both time-varying and invariant individual controls, and year fixed effects. Model (2) includes time-varying individual controls, year fixed effects, and individual fixed effects. Time invariant controls include an indicator for Black respondents, an indicator for Hispanic ethnicity, an indicator for sex, an indicator for ever graduating high school during the years of the survey, and an indicator for ever graduating college. Time-varying controls include a quadratic in age, number of children, marital status, log of household income, a rural household indicator, Census region of residence, whether the individual is incarcerated, and calendar-month of the interview. I code missing values of control variables as zero and include missing value indicator variables for each. Standard errors clustered at the individual level are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Online Appendix

A.1 Details on Depression Measurement

A.1.1 Details of Depression Questions in the NLSY 97

The MHI-5 inventory included in the NLSY 97 includes includes the following 5 questions:

- How much of the time during the last month have you been a very nervous person?
- How much of the time during the last month have you felt calm and peaceful?
- How much of the time during the last month have you felt downhearted and blue?
- How much of the time during the last month have you been a happy person?
- How much of the time during the last month have you felt so down in the dumps that nothing could cheer you up?

Importantly, unlike most studies using the MHI-5 in which responses are measured on a 6 point Likert scale, the version included in the NLSY-97 uses a 4-point scale:

- All of the time
- Most of the time
- Some of the time
- None of the time

I calculate a total depression score as follows. For negative questions (nervous, downhearted, down in the dumps), 3 points are assigned for "all of the time," 2 points for "most of the

time," 1 point for "some of the time," and 0 points for "none of the time." The coding is reversed for positive questions (calm, happy). I use a standardized version of this measure to aid in interpreting units. I first replace any non-responses to one of the five questions with the mean taken by year and survey mode. I standardize the measure by subtracting of the sample mean calculated in the NLSY-97, pooling years, and divide by the sample standard deviation.

I also consider three related questions asked only in in-person non-SAQ and phone interviews. In the 2009 to 2011, respondents were asked:

- During the past 12 months, how many times were you treated by a mental health professional for emotional, mental or psychiatric problems?
  - None
  - 1 time
  - 2 times
  - 3 times
  - 4 or more times

To use this question as an outcome variable, I construct a binary indicator variable for any response of 1 or more times. Second, in the 2009 to 2017 waves, respondents were asked the following two questions:

- Some conditions are not treated by a professional. During the past 12 months, how many times did you have an emotional, mental or psychiatric problem so that you missed at least one full day of usual activities such as work or school, but were not treated by a professional?

- How many times did you miss work because you were just not feeling right – for example, you were 'too blue' to get up in the morning, or feeling too anxious to conduct your usual activities? Please do not include times that you missed work that you've already told me about.

Each question includes the following response scale:

- None
- 1 time
- 2 times
- 3 times
- 4 or more times

To use these two questions as outcome variables, I construct a binary indicator variable for any respondent who indicates an answer of 1 or more times to either questions.

A.1.2 Details on Depression Questions in the NLSY 79

A subset of questions from the Center for Epidemiological Studies Depression (CES-D) questionnaire (Radloff, 1977) was administered to all respondents at four points in time: the 1992 wave, the 1994 wave (around age 30), and in special health questionnaires administered in the waves immediately after a respondent turned 40 (1998-2006) and immediately after a respondent turned 50 (2008-2016). The original scale has twenty questions, but to cut down on the length of the survey, only seven question were used consistently in the NLSY 79. The questions are asked as follows: Now I am going to read a list of the ways that you might

have felt or behaved recently. After each statement, please tell me how often you felt this way during the past week. During the past week...

- I did not feel like eating; my appetite was poor.
- I had trouble keeping my mind on what I was doing.
- I felt depressed.
- I felt that everything I did was an effort.
- My sleep was restless.
- I felt sad.
- I could not get "going."

Each question is answered on a four point Likert scale:

- Rarely/None of the time/1 Day
- Some/A little of the time/1-2 Days
- Occasionally/Moderate amount of the time/3-4 Days
- Most/All of the time/5-7 Days

I calculate a total score on this scale by assigning 0 points for each response of "Rarely/None of the time/1 Day"; 1 point for "Some/A little of the time/1-2 Days"; 2 points for "Occasionally/Moderate amount of the time/3-4 Days"; and 3 points for "Most/All of the time/5-7 Days." I calculate a total score on this scale by assigning 0 points for each response of "Rarely/None of the time/1 Day"; 1 point for "Some/A little of the time/1-2 Days"; 2 points for "Occasionally/Moderate amount of the time/3-4 Days"; and 3 points for "Most/All of

the time/5-7 Days." I standardize the measure by subtracting of the sample mean calculated in the NLSY-79, pooling years, and divide by the sample standard deviation.

A.1.3 Details on Depression Question in the NLSY 79 Children and Young Adult Cohort

In each wave from 1994 to 2016, respondents to NLSY 79 Children and Young Adults cohort are asked the same seven questions from the CES-D as the main NLSY 79 cohort. I create a total depression score using the same method as the main NLSY 79 cohort.

I also consider three related yes/no questions:

- During the last 12 months, have you received any help for an emotional, behavioral, or family problem?
- Do you regularly take any medicine or prescription drugs to help control your activity level or behavior?
- Have you ever seriously considered attempting suicide?

The first two are asked in each wave from 1994 to 2016. The last was asked in 2012, 2014, and 2016.

A.2 Robustness of Mode Effects to Non-Reporting and Attrition

There are two types of non-response in the NLSY 97 on these estimates. First, there is attrition from wave to wave in the NLSY 97. Not every participant is ultimately surveyed in a given wave. Despite substantial efforts to complete interviews with all members of the cohort, some refuse to be interviewed; some cannot be located; and others die or are too

sick to be interviewed. Second, there is also non-response to individual depression questions conditional on participation. These are individuals who refuse to answer a question or respond "I don't know." Panel A of Table A.4 details the proportion of respondents who were not interviewed (attrition), and Panel B details the proportion of individuals surveyed with any non-response to one of the five depression questions, broken out by phone interview and in-person SAQ. I only report these statistics for survey waves in which the depression questionnaire was included.

Attrition is around 10.1% in 2000 and increases to around 25.0% in 2017. The primary reason for attrition is refusal to participate, including both direct refusals, refusals by gate keepers (e.g., individuals who control access to an apartment building), as well as any individuals who could not schedule an interview before the end of the wave and individuals who were too sick to participate. Such refusals increase from around 9.77% in 2000 to 19.3% in 2017. Deaths increase from around 0.17% of the cohort in 2000 to 2.30% in 2017. Individuals who cannot be located increase from 1.93% in 2000 to 6.71% in 2017. For some individuals, no reason for non-interview is provided.

Individuals are more likely to refuse to answer a depression question or answer "I don't know" in an in-person SAQ than a phone interview. Across years, Around 0.98% of respondents refuse or answer answer "I don't know" to at least one of the give depression questions in phone interviews. Aside from relatively high levels of 5.81% in 2000 and 1.56% in 2002, non-responses in phone interviews are typically less than 1%. Individuals offering at least one non-responses are only 0.24% of the sample interviewed by in-person SAQ in 2000 and 0.15% in 2002 but more than 4% in 2004 to 2015. In 2017, they only account for 0.43% of the sample.

One possible driver of this difference is that the visual display used in in-person SAQ explicitly offers answer choices corresponding to a refusal and "I don't know," but these options are not read by the interviewer during phone interviews or in-person non-SAQ. Rather, they are

left implicit.[5] This likely makes the option not to respond more salient in the in-person SAQ. Respondents may also find it more difficult to refuse to answer a question when responding directly to the interviewer than when entering their response anonymously in a computer.

A.2.1 Bounding Treatment Effects Given Item Non-Response

Refusal to answer may bias estimation of the effect of survey mode on the reporting of depression. In the previous analyses, I impute missing values for any of the five depression questions using the year and interview mode mean. To assess whether this may impact my findings, in this Section, I present worst-case bounds for each estimate of the survey mode effects. To do this, I re-estimate the effect after imputing a value for refusals that is most opposed to the effect I identify previously. Even under this worst-case scenario, estimates of the mode effect are still similar to the estimate with no imputation.

These bounds are conceptually similar to the worst case bounds proposed by Horowitz and Manski (2000). To estimate these bounds, I use the fact that responses to each of the five depression questions are bounded between 0 and 3. Importantly, among individuals with any non-responses to a depression question, most only fail to respond to one of the five questions. So, I can still use the responses to the other questions in calculating total depression. I adjust the scale so that 3 corresponds to the answer choice indicating the highest level of depression, and 0 corresponds to the answer choice indicating the lowest level of depression. For the "lower bound," I impute a value of 0 for any refusal or response of "I don't know" for individuals interviewed by phone. I impute a value of 3 for any refusal or response of "I don't know" for individuals interviewed by in-person SAQ. I then recalculate a total score for each respondent. Since I find a negative difference in reported depression between phone interviews and in-person SAQ in the last two Sections, this should result in an even "more negative" estimate of the difference. Likewise, for the "upper bound," I impute a value of 3

[5]Source: personal correspondence with survey administrators at NORC at the University of Chicago

for any refusal or response of "I don't know" for individuals interviewed by phone. I impute a value of 0 for any refusal or response of "I don't know" for individuals interviewed by in-person SAQ. This should result in an estimate of the difference in reported depression between phone interviews and in-person that is closer to zero or even positive.

Table A.5 reports the upper and lower bounds from this procedure using two specifications: the IV specification described in Section 3.2; and the panel specification with individual fixed effects described in Section 3.3. Using the panel specification with individual fixed effects, the estimated effect of a phone interview on reported depression is -0.290 SD for the lower bound and -0.151 SD for the upper bound, both significant at the 1% level. Using the IV specification, the estimated effect of a phone interview on reported depression is -0.367 SD for the lower bound and -0.182 SD for the upper bound, both significant at the 1% level. These results suggest that non-response minimally biases my estimates of the difference in depression reporting between phone interviews and in-person SAQ.

A.2.2 Accounting for Attrition

In the main analysis, I ignore attrition. Thus, the estimates reported in the main paper are the the causal effect of survey mode on depression reporting conditional on participation in the survey. Of course, any estimates based on survey data implicitly condition on participation in the survey. So, this estimate has direct implications for other surveys measuring depression. All surveys have non-responders, and if attriters in the NLSY 97 are similar to attriters in other surveys, their exclusion is unlikely to bias the external validity of the result.

However, we may still be interested in the effect of a phone interview among attriters, but worst case bounds that account for attrition would be so wide as to be meaningless. To see this, consider a hypothetical estimate of the average treatment effect of a phone interview

on depression reporting calculated using both attriters and non-attriters Let α_0 denote the average treatment effect among individuals who participate in the survey, and let p denote the proportion of the sample that does not attrit. Let α_1 denote the average treatment effect among attriters. So, $1-p$ denote the proportion of the sample that attrits. The hypothetical estimate of the average treatment effect of a phone interview would be:

$$\beta = p\alpha_0 + (1-p)\alpha_1$$

The variables p (around 0.8326) and α_0 (around -0.5 points) are known, assuming the estimates above are correct. However, the value of α_1 is unknown. Since the total depression score can range from 0 to 15, the average effect of a phone interview among attriters could be anywhere from -15, if all attriters have a total depression score of 0 when interviewed by phone and 15 when interviewed by in-person SAQ, to 15 given the reverse. Plugging these values in, the worst case bounds on the overall average treatment effect β would range from -2.93 points to 2.09 points.

Instead, I introduce an analysis to gauge whether the effect of survey mode is the same among individuals who attrit from the NLSY 97. Since some individuals who are not interviewed in a wave are interviewed in later or earlier waves, I estimate the difference in the effect of a phone interview for individuals who have never missed a survey wave and individuals who have missed any survey waves. I also estimate the difference in effect based on the number of survey waves an individual has missed. To the extent that these individuals are similar to individuals who never respond to a survey, this analysis should offer reassurance that the estimate of the causal effects of interview mode on depression reporting would not be substantially difference among attriters as among non-attriters.

The analysis is of the following general form:

$$Y_{it} = \alpha_i + \tau_t + \beta_1 P_{it} + \beta_2 P_{it} \times A_i + X'_{it}\delta + u_{it} \tag{4}$$

where A_{it} is a count of the number of survey waves in which individual i did not participate between 1998 and 2017. Thus, β_2 estimates the difference in the effect of phone interview based on the number of waves in which an individual did not participate. I also include a variant in which I include interactions between P_{it} and separate indicators for the number of waves an individual did not participate. That is, an indicator taking the value of 1 for every respondent who missed exactly one wave; every respondent who missed exactly two waves; etc. I include one indicator for all individuals who missed 6 or more waves.

Table A.6 reports the results. Columns 1 and 3 omit individual fixed effects, and Columns 2 and 4 include them. Columns 1 and 2 include an interaction between the phone interview indicator and a count of the number of waves an individual did not participate between 1998 and 2017. Columns 3 and 4 include interactions between the phone interview indicators and indicators for each number of waves an individual did not participate. The coefficients on interactions terms are not statistically significant in any of the four specifications. Moreover, in Column 3 and 4, tests of the joint significance of the six interaction give p-values of 0.887 and 0.793 respectively. Thus, there does not appear to be a difference in the effect of a phone interview between individuals who participate in all waves and those who do not or based on the number of waves an individual does not participate. The point estimates on interaction terms are also quite small in comparison to the estimate of the coefficient on the phone interview indicator.

While it is, of course, impossible to determine the effect among individuals who never participate in a survey, this result suggests individuals who have a higher propensity to attrit are not substantially different in their response to survey mode.

A.3 Appendix Tables and Figures

Table A.1: Proportion of Interviews Conducted by Phone

	Proportion of interviews conducted by phone		
	NLSY 97	NLSY 79	NLSY 79 CYA
All years	0.23	0.41	0.83
1992		0.13	
1994		0.10	0.05
1996			0.06
1998		0.26	0.10
2000	0.08	0.36	0.86
2002	0.16	0.68	0.80
2004	0.13	0.80	0.88
2006	0.12	0.78	0.82
2008	0.15	0.90	0.87
2010	0.11	0.89	0.89
2012		0.90	0.91
2014		0.94	0.96
2015	0.28		
2016		0.79	0.97
2017	0.90		

Note: This table only lists years in which depression questionnaires were administered. The unit of observation is individual by year, restricted to individuals who responded to the depression questionnaire.

Table A.2: NLSY 97: Regression of 2015 Treatment Assignment Indicator on Lagged Covariates

	Assigned phone 1st randomization	Assigned phone 2nd randomization
	(1)	(2)
Phone interview, 2013	0.0117	−0.0221
	(0.0166)	(0.0149)
Total depression, 2010	0.00437	−0.00257
	(0.00471)	(0.00476)
Health self-report, excellent or great, 2013	−0.00548	−0.00592
	(0.00907)	(0.00916)
Log HH income, 2013	0.00136	−0.00367
	(0.00205)	(0.00246)
Missing: HH income, 2013	0.0101	−0.0439
	(0.0274)	(0.0312)
Weeks worked, past 12 months, 2013	0.0000600	0.000395*
	(0.000235)	(0.000233)
No. of children, 2013	−0.00599*	−0.00700*
	(0.00343)	(0.00366)
Married, 2013	0.00948	0.0144
	(0.00994)	(0.00993)
Northeast region, 2013	−0.00858	−0.00573
	(0.0132)	(0.0130)
North central region, 2013	0.00585	0.00398
	(0.0122)	(0.0120)
West region, 2013	−0.0175	0.00359
	(0.0119)	(0.0123)
Lives outside US, 2013	0.109	−0.0324
	(0.115)	(0.0653)
Rural HH, 2013	−0.00810	−0.00339
	(0.0125)	(0.0127)
No. of address changes, 2013	−0.00962**	0.00138
	(0.00460)	(0.00549)
Age	−0.00423	−0.0000879
	(0.00315)	(0.00319)
Male	−0.0103	−0.0130
	(0.00906)	(0.00909)
Black	0.00821	−0.00345
	(0.0112)	(0.0110)
Hispanic	0.00917	−0.00264
	(0.0119)	(0.0117)
HS graduate	0.00612	−0.0224
	(0.0156)	(0.0173)
College graduate	0.00528	−0.0420**
	(0.0172)	(0.0187)
Joint F-test, p-value	0.446	0.569
N	6310	6310

Note: Each column reports a regression of an indicator assignment to the treatment group in the 2015 phone experiment in the first or second randomization respectively on the indicated covariates. The unit of observation is individual by survey wave. The sample is restricted to 2015 observations of individuals designated as eligible for the 2015 phone experiment. Each regression also includes a control for assignment to the treatment group in the other randomization and missing value indicators for each covariate. Heteroskedasticity robust standard errors are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.3: Effect of Phone Interview on Constituent Variables of MHI-5 Depression Scale, NLSY 97

	(1) Panel (2000-15)	(2) IV (2015)
Nervous	−0.105206***	−0.141464***
	(0.008631)	(0.048520)
Calm and peaceful	0.065311***	0.080536*
	(0.009422)	(0.048749)
Downhearted and blue	−0.159908***	−0.103680**
	(0.008747)	(0.044032)
Happy	0.064954***	0.129344***
	(0.009033)	(0.044450)
Down in the dumps	−0.122921***	−0.122293***
	(0.007329)	(0.035369)
N	59620	6325

Note: Each row reports the coefficients from two separate regressions using the indicated variable as an outcome, restricting the sample to years these questions were asked by in-person SAQ (2000, 2002, 2004, 2006, 2008, 2010, 2015). The unit of observation is individual by survey wave. Each outcome variable is scaled so that 3 correspond to "All of the time"; 2 corresponds to "Most of the time"; 1 corresponds "Some of the time"; and 0 corresponds to "None of the time." Columns 1 uses all waves with the individual fixed effect specification and control variables used in Column 2 of Table 2. Columns 2 uses the 2sls specification in Column 4 of Table 1 and is restricted to the 2015 wave. Standard errors clustered at the individual level are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4: Survey Mode and Attrition, NLSY 97 2000-2017 Waves

	overall	2000	2002	2004	2006	2008	2010	2015	2017
Panel A: Attrition									
Total not interviewed	0.1674	0.1006	0.1211	0.1650	0.1586	0.1663	0.1675	0.2094	0.2504
Deceased	0.0107	0.0017	0.0033	0.0050	0.0086	0.0115	0.0131	0.0193	0.0230
Not located	0.0332	0.0193	0.0282	0.0308	0.0213	0.0305	0.0302	0.0383	0.0671
Refusal	0.1397	0.0977	0.1112	0.1546	0.1413	0.1313	0.1256	0.1636	0.1925
Panel B: Non-response to depression questions									
Any refusal or don't know, phone interview	0.0098	0.0581	0.0156	0.0097	0.0045	0.0103	0.0133	0.0074	0.0038
Any refusal or don't know, in-person SAQ	0.0394	0.0024	0.0015	0.0473	0.0587	0.0547	0.0516	0.0757	0.0043

Note: Only years in which depression questions were included in the survey are shown. Panel A shows the overall proportion of the original sample that is not interviewed in the indicated year; as well as the proportion of the original sample that is not interviewed for each indicated reason. For some individuals who were not interviewed, no specific reason is given, so the total of deceased, not located, and refusals may not add up to the overall total. Refusals include direct refusals, individuals unable to be interviewed due to health or incarceration, and gatekeeper refusals. Panel B reports the proportion of individuals surveyed in a given wave that refuse to answer or answer "I don't know" to one or more of the five depression questions.

Table A.5: Worst Case Bounds on Mode Effects, Conditional on Being Surveyed, NLSY 97

	Panel (2000-15)		IV (2015)	
	(1)	(2)	(3)	(4)
	lower	upper	lower	upper
Phone interview	−0.337***	−0.177***	−0.369***	−0.183***
	(0.0153)	(0.0160)	(0.0684)	(0.0665)
Time-varying controls	Yes	Yes		
Survey wave FEs	Yes	Yes		
Individual FEs	Yes	Yes		
N	52853	52853	6325	6325

Note: Each column presents coefficient estimates from a regression of standardized total depression from the MHI-5 on an indicator for participating in a phone interview. The outcome variable in each column uses the imputations for refusals and responses of "I don't know" described in Section A.2.1, restricting the sample to years these questions were asked by in-person SAQ (2000, 2002, 2004, 2006, 2008, 2010, 2015). Columns 1 and 2 use the same specification and control variables as Column 2 of Table 2. Columns 3 and 4 use the same specification as Column 4 of Table 1 and are restricted to the 2015 wave. The unit of observation is the individual by survey wave. Standard errors clustered at the individual level are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6: NLSY 97: Comparison of Mode Effects between Never and Ever Attriters

	Depression total, standardized			
	(1)	(2)	(3)	(4)
Phone × count of waves not surveyed	−0.00104	0.00168		
	(0.00488)	(0.00517)		
Phone × waves not surveyed = 1			0.0354	0.0110
			(0.0419)	(0.0356)
Phone × waves not surveyed = 2			0.0118	0.0272
			(0.0502)	(0.0415)
Phone × waves not surveyed = 3			−0.00446	0.0189
			(0.0596)	(0.0545)
Phone × waves not surveyed = 4			−0.0359	−0.0627
			(0.0604)	(0.0596)
Phone × waves not surveyed = 5			−0.0607	0.0717
			(0.0708)	(0.0644)
Phone × waves not surveyed ≥ 6			0.0325	0.00436
			(0.0438)	(0.0445)
Phone interview	−0.248***	−0.250***	−0.251***	−0.252***
	(0.0180)	(0.0160)	(0.0206)	(0.0175)
Count of waves not surveyed	−0.0170***			
	(0.00319)			
Waves not surveyed = 1			0.00515	
			(0.0256)	
Waves not surveyed = 2			0.00872	
			(0.0330)	
Waves not surveyed = 3			−0.0818*	
			(0.0421)	
Waves not surveyed = 4			−0.0474	
			(0.0458)	
Waves not surveyed = 5			−0.0989*	
			(0.0508)	
Waves not surveyed ≥ 6			−0.116**	
			(0.0518)	
p-value, joint significance of interactions			0.8435	0.7997
Controls	Yes	Yes	Yes	Yes
Survey wave FEs	Yes	Yes	Yes	Yes
Individual FEs		Yes		Yes
N	51166	50915	51166	50915

Note: Each column presents coefficient estimates from a regression of standardized total depression from the MHI-5 on the indicated RHS variables, restricting the sample to years these questions were asked by in-person SAQ (2000, 2002, 2004, 2006, 2008, 2010, 2015). The unit of observation is the individual by survey wave. Standard errors clustered at the individual level are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.