Journal of Survey Statistics and Methodology (2020) 8, 903–931

WHO CAN YOU COUNT ON? UNDERSTANDING THE DETERMINANTS OF RELIABILITY

ROGER TOURANGEAU
TING YAN*
HANYU SUN

Using reinterview data from the PATH Reliability and Validity (PATH-RV) study, we examine the characteristics of questions and respondents that predict the reliability of the answers. In the PATH-RV study, 524 respondents completed an interview twice, five to twenty-four days apart. We coded a number of question characteristics and used them to predict the gross discrepancy rates (GDRs) and kappas for each question. We also investigated respondent characteristics associated with reliability. Finally, we fitted cross-classified models that simultaneously examined a range of respondent and question characteristics. Although the different models yielded somewhat different conclusions, in general factual questions (especially demographic questions), shorter questions, questions that did not use scales, those with fewer response options, and those that asked about a noncentral topic produced more reliable answers than attitudinal questions, longer questions, questions using ordinal scales, those with more response options, and those asking about a central topic. One surprising finding was that items raising potential social desirability concerns yielded more reliable answers than items that did not raise such concerns. The respondent-level models and cross-classified models indicated that five adult respondent characteristics were associated with giving the same answer in both interviews—education, the Big Five trait of conscientiousness, tobacco use, sex, and income. Hispanic youths and non-Hispanic black youths were less likely to give the same answer in both interviews. The cross-classified model

ROGER TOURANGEAU, TING YAN, and HANYU SUN are with Westat, MD, USA. The work reported here was funded by a grant from the National Institute on Drug Abuse, National Institutes of Health (5R01DA040736-02 to RT). The views and opinions expressed in this manuscript are those of the authors only and do not necessarily represent the views, official policy or position of the U.S. Department of Health and Human Services or any of its affiliated institutions or agencies.
*Address correspondence to Ting Yan, Westat, 1600 Research Boulevard, Rockville, MD, 20850, USA. E-mail: TingYan@Westat.com.

doi: 10.1093/jssam/smz034 Advance access publication 3 October 2019

also found that more words were associated with less reliable answers. The results are mostly consistent with earlier findings but are nonetheless important because they are much less model-dependent than the earlier work. In addition, this study is the first to incorporate such personality traits as needed for cognition and the Big Five personality factors and to examine the relationships among reliability, item nonresponse, and response latency.

KEYWORDS: reliability; gross discrepancy rate; Kappa, response times; item nonresponse.

1. INTRODUCTION

Although most survey researchers acknowledge the importance of eliciting reliable answers from the respondents, it is not routine to collect reliability data in surveys. A few surveys, such as the Current Population Survey (CPS), do regularly collect reinterview data to estimate the reliability of the answers (Forsman and Schreiner 1991; Sinclair and Gastwirth 1996), but in general, such surveys are rare. In a reinterview study, respondents are recontacted a short period of time after the initial interview (say, within two weeks) and asked some or all of the same questions they answered in the initial interview. A major obstacle to doing this routinely is the added expense of recontacting and reinterviewing respondents.

More common are special studies done once to assess the reliability of specific survey instruments. For example in 2006, the Substance Abuse and Mental Health Services Administration (SAMHSA) did a study to evaluate the reliability of answers to the National Survey on Drug Use and Health (NSDUH; SAMHSA 2010). That study also used the reinterview method to examine the reliability of answers to the NSDUH questions. Similar one-time studies have been done to estimate the reliability of answers to the Behavioral Risk Factor Surveillance System questions (Stein, Lederman, and Shea 1993), the Youth Behavioral Risk Factor Survey questions (Brener, Collins, Kann, Warren, and Williams 1994), diagnoses derived from the National Epidemiological Study of Alcohol and Related Conditions data (Grant, Dawson, Stinson, Chou, Kay, and Pickering 2003; Grant, Goldstein, Smith, Jung, Zhang, et al. 2013), selected questions on the 2002–2003 Tobacco Use Supplement to the CPS (Soulakova, Hartman, Liu, Willis, and Augustine 2012), and the General Social Survey (Smith 1980; Hout and Hastings 2016). The assessment of the CPS Tobacco Use Supplement took advantage of the CPS panel design in which some respondents completed the Tobacco Use Supplement twice. The PATH-RV study is another example of a one-time study to estimate the reliability of the answers to the questionnaires in a particular survey, in our case the Wave 4 Population Assessment of Tobacco and Health study (PATH).

In addition to these efforts to evaluate specific survey questionnaires, there have also been several systematic attempts to investigate the characteristics of respondents and questions that produce reliable answers. The most comprehensive of these efforts are by Alwin and his colleagues (Alwin 2007; Alwin, Baumgartner, and Beattie 2018) and by Saris and Gallhofer (2007a, 2007b) and their colleagues (Saris, Revilla, Krosnick, and Schaeffer 2010; Revilla, Saris, and Krosnick 2014). Earlier efforts along these lines include Andrews (1984) and O'Muircheartaigh (1991). Alwin, Saris, and Gallhofer are each critical of the reinterview methods that have traditionally been used to estimate the reliability of survey items and advocate alternative methods instead.

Alwin (2007) argues that memory effects inflate the reliability estimates from the typical reinterview study. According to Alwin, respondents can remember their answers from the initial interview and repeat them in the second interview, although he does not provide much empirical backing for this claim. He advocates modeling data from panel surveys to estimate reliabilities. In a three-wave panel survey, it is possible to fit a model—the "quasi-simplex" model—that provides separate estimates of random error variance (that is, unreliability) and variance in true change over time. Additional assumptions are needed to make the parameters of the model identifiable. He advocates the use of panel data in which interviews are conducted at least two years apart to minimize any memory effects.

Saris and Gallhofer (2007a, 2007b) note another potential issue with the reliability estimates from reinterview studies. They argue that respondents may give the same answer to a survey question because of a "method" effect. For example, if the question uses an agree-disagree format, respondents may select "agree strongly" as their answer in both interviews partly because they are prone to select that response option regardless of the content of the question. The presence of such methods effects inflates the estimated reliability of the answers. They advocate deriving reliability estimates from multitrait, multimethod (MTMM) experiments. In an MTMM experiment, multiple "traits" (that is, constructs) are measured, each by several methods (say agree-disagree items and items with item-specific response options). It is possible to separate the methods variance from the valid variance in a specific question via structural equation modeling. In a few of Alwin's (2007) analyses, he adopts the MTMM approach, as well. It is not always clear how important the methods effects are; they often seem to contribute little to the reliable variation in answers.

Despite these methodological differences, there is at least some convergence in the conclusions reached by the investigators. Alwin (2007; see also Alwin et al. 2018 and Hout and Hastings 2016) finds that items eliciting facts produce more reliable answers than those measuring subjective constructs (say attitudes or beliefs); open-ended questions produce more reliable answers than closed-ended; unipolar questions produce more reliable answers than bipolar questions (although this difference seems small); two-category scales produce more

reliable answers than scales with more response categories; fully-labeled scales produce more reliable answers than scales in which only the endpoints are given verbal labels; and shorter questions produce more reliable answers than longer questions (but only for stand-alone questions as opposed to questions in batteries). Finally, questions with long introductions seem to produce *less* reliable answers than those with short or no introductions. This last finding illustrates a potential limitation of trying to identify question features associated with unreliability—the underlying causal relationships may be unclear. Questions with long introductions often involve complicated or unfamiliar tasks, which is why the investigators provide the lengthy introductions. The difficulty of the task rather than the length of the introduction may account for the low reliability of answers to such items.

Saris and Gallhofer (2007a, 2007b) also conclude that questions with more fully labeled scales and unipolar (or "asymmetric") questions are associated with greater reliability and that longer questions (those with more subordinate clauses) produce less reliable answers than shorter questions. Table 1 in Saris and Gallhofer (2007b) summarizes their conclusions, which are based on a meta-analysis of the results from eighty-seven MTMM experiments, mostly in the Netherlands, involving more than 1,000 survey items. In many cases, it is difficult to compare their results with Alwin's because they examine different question characteristics.

We summarize some of the noteworthy findings from the prior research in table 1. Two studies are worth highlighting because they examine the reliability of reports about tobacco use. Johnson and Mott (2001) examined age at reported first use of tobacco and other substances using longitudinal data from the National Longitudinal Survey of Youth. They found that female respondents, white respondents, and more educated respondents were more likely to report consistently across waves; for some substances, older respondents gave more consistent reports, but for others, younger respondents did. Soulakova et al. (2012) examined the CPS Tobacco Use Supplement and found few main effects; for one item, males were less reliable than females, and for another item, telephone respondents gave more consistent answers than those interviewed in-person.

Alwin's work is based on an analysis of the survey response process that draws on the model proposed by Tourangeau and his colleagues (Tourangeau 1984; Tourangeau, Rips, and Rasinski 2000; see also Cannell, Miller, and Oksenberg 1981). He distinguishes "six critical elements" of the response process (see table 2.1 in Alwin 2007):

- content validity (how well the question measures the phenomenon of interest)
- comprehension of the question (how well the respondent understands the question)

Table 1. Key Results from Prior Reliability Studies

Variable Question characteristics	Studies	Finding
Type of question (Factual vs. non-factual)	Smith (1980), Alwin (2007), Hout and Hastings (2016)	Factual questions produce more reliable answers
Position in questionnaire	Saris and Gallhofer (2007a, 2007b)	Later items in questionnaire produce more reliable answers
Question length	Alwin (2007)	Shorter questions produce more reliable answers
Syntactic complexity (Number of subordinate clauses)	Saris and Gallhofer (2007a, 2007b)	Questions with fewer subordinate clauses produce more reliable answers
Polarity	Alwin (2007), Saris and Gallhofer (2007a, 2007b), Alwin et al. (2018)	Unipolar questions produce more reliable answers than bipolar
Response format (Open vs. closed for factual Items)	Alwin (2007)	Factual questions with numeric open-ended responses produce more reliable answers than those that use closed categories with vague quantifiers
Number of response categories	Revilla, Saris, and Krosnick (2014), Alwin et al. (2018)	Fewer response options produce more reliable answers
Middle categories	Saris and Gallhofer (2007a, 2007b), Alwin et al. (2018)	Questions without a middle response option produce more reliable answers
Verbal labeling	Alwin and Krosnick (1991), Alwin (2007), Saris and Gallhofer (2007a, 2007b)	Questions in which every option is labeled verbally produce more reliable answers
Type of scale (Agree-disagree vs. item-specific)	Saris et al. (2010)	Item-specific response scales produce more reliable answers
Reference period	Saris and Gallhofer (2007a, 2007b)	Questions asking about the past produce more reliable answers than those asking about the present and future.

Continued

Table 1. *Continued*

Variable Question characteristics	Studies	Finding
Salience of Topic	Saris and Gallhofer (2007a, 2007b)	Questions about salient/central topics produce more reliable answers than questions about less salient topics
Instructions for Respondents	Saris and Gallhofer (2007a, 2007b)	Respondent instructions produce less reliable answers
Respondent Characteristics		
Age	Alwin (1989), Alwin and Krosnick (1991), Rodgers et al. (1992)	Younger respondents provide more reliable answers
Education	Smith (1980), Alwin (1989, 2007), Alwin and Krosnick (1991), Saris and Gallhofer (2007a, 2007b)	More educated respondents provide more reliable answers
Race/ethnicity	Stein et al. (1993)	Non-Hispanic Whites provide more reliable answers than Hispanic or non-Hispanic Blacks
Sex	Soulakova et al. (2012)	Females provide more reliable answers
Socioeconomic Status	Smith (1980)	Those in higher prestige occupations provide more reliable answers

- accessibility of the information (whether the respondent has the information sought by the question)
- retrieval (how well the respondent can remember the information sought)
- motivation (how willing the respondent is to report accurately)
- communication (how easily the respondent can translate his or her answer onto the response scale provided)

According to Tourangeau and his colleagues, comprehension, retrieval, judgment, and reporting are the major components of the response process. The reporting component encompasses two subprocesses: editing the answer (the respondent's altering the answer he or she reports to avoid losing face or appearing inconsistent) and mapping it onto the format required by the question. These two subcomponents show considerable conceptual overlap with the "motivation" and "communication" elements in Alwin's framework.

Saris and Gallhofer's work is based on their analysis of the constituents of a survey item and the choices that the researchers make in crafting their items. According to Saris and Gallhofer, the key constituents of a survey item are (i) the item's introduction, including the motivation for the question, (ii) whatever information about the topic or definitions it provides, (iii) any instructions to the interviewer or respondent, (iv) the request for an answer (which does not always take the form of a question), and (v) the answer categories (cf. Figure 2.1 in Saris and Gallhofer 2007a; see also Saris and Gallhofer 2007b, p. 30). Based on this decomposition of an item's components, they derive a large number of item characteristics (75 in total) that might affect its reliability and validity. They coded all the items in their MTMM experiments for these characteristics and, using meta-analytic procedures, attempted to identify noteworthy predictors of reliability and validity.

Like Alwin's work (see also Revilla et al. 2014), our analysis of the PATH-RV data is guided by Tourangeau et al.'s (2000) discussion of the survey response process. We coded the 447 questions from the PATH Study Adult questionnaire, for which we were able obtain both initial and reinterview responses from at least 100 respondents; this article focuses on that set of items. In addition, we carried out similar coding and analyses of the 229 questions in the youth questionnaire, for which there were at least 100 responses in both interviews. With both questionnaires, we coded question characteristics that we thought might be related to difficulties in question comprehension, retrieval, judgment, or mapping. We also coded whether the answers were likely to be prone to "editing" due to social desirability concerns. Tourangeau and his colleagues argue that measurement error in surveys generally reflects difficulties with one or more components in the response process (for example, misunderstanding of the questions or retrieval failure). Here, we test the general hypothesis that such cognitive difficulties manifest themselves in reduced reliability.

In earlier work, Yan and Tourangeau (2008) used a similar approach to the one taken here to examine response times to survey questions, reasoning that difficulties with one or more components of the response process might lead to slower responses (see also Couper and Kreuter 2013; Olson and Smyth 2015). This suggests that the same question characteristics may be related both to slow *and* unreliable answers. It is also possible that similar problems in the response process may produce item nonresponse. For example, respondents who do not understand a question may give a "don't know" response rather than asking for clarification. Thus, some of our models of item characteristics related to the reliability of the answers include median response times to the question and the question's item nonresponse rate, and we present results showing how these three outcomes relate to one another.

In addition to examining question characteristics associated with reliability, we also examine relevant respondent characteristics. Two obvious candidates that have been explored in past work are the respondent's age and education. Because age is related to working memory capacity (Salthouse 1994), older

respondents may have a harder time understanding and processing questions than younger respondents. At least three prior studies indicate that older respondents give less reliable answers than younger ones (Alwin 1989; Alwin and Krosnick 1991; Rodgers, Andrews, and Herzog 1992). Similarly, education is thought to be a good summary measure of a broad range of cognitive skills and, therefore, related to the respondent's ability to give reliable answers to survey questions. Past work confirms that more educated respondents provide more reliable answers than their less educated counterparts (Alwin 1989, 2007; Alwin and Krosnick 1991).

Aside from these (and other) demographic variables, we examine the respondent's need for cognition—that is, how much someone enjoys carrying out challenging cognitive tasks (Cacioppo and Petty 1982). Finally, we also explore whether conscientiousness, one of the Big Five factors, is systematically related to the reliability of the answers. Conscientiousness is the tendency to be organized and to exercise self-discipline (for example, see Goldberg 1992). Conscientious respondents and respondents high in the need for cognition are less likely to engage in "survey satisficing" (Krosnick 1991, 1999), which refers to taking cognitive shortcuts to reduce the effort needed to answer survey questions. Satisficing may take various forms (say giving "don't know" responses, giving the same answer to every question in a battery of items, or selecting a random answer). Answers that are the product of satisficing are unlikely to be reliable. Finally, we examine the elapsed time between the interview and reinterview.

In summary, the PATH-RV study collected reinterviews with 524 respondents, using the PATH study wave 4 questionnaires. We used these data to assess the reliability of more than 400 items from the adult questionnaire and more than 200 from the youth questionnaire and coded selected characteristics of those items. The wordings of the items, reliability estimates, and coded characteristics are shown in the online supplementary materials. In choosing question characteristics to include in our models, we drew on past work by Alwin, Saris, Gallhofer and their collaborators and the model of the survey response process proposed by Tourangeau and his colleagues (Tourangeau 1984; Tourangeau et al. 2000; Tourangeau 2018). However, we used a different, less controversial method for estimating reliability than these previous efforts, reinterviewing respondents after a short period of time. We also investigated the characteristics of reliable respondents, including their age, education, and other demographic characteristics. Finally, we also explored the roles of need for cognition and conscientiousness.

2. METHODS

The PATH Study is a major national longitudinal study of tobacco use and health. It follows more than 40,000 members of the US household population

ages twelve and older and includes both tobacco users and nonusers. The fourth wave of interviewing was completed in 2017. The study uses audio computer-administered self-interviews (ACASI) to collect information on a wide range of topics, including use of tobacco products; attitudes and perceptions toward different tobacco products; knowledge of the contents of tobacco products and of their health consequences; tobacco-use cessation attempts, their outcomes, and rates of relapse; uptake of new products, switching of products or brands, and use of two or more tobacco products; and health conditions, including ones potentially related to tobacco use.

To the extent possible, the PATH-RV study replicated the systems and procedures of the main PATH Study. It used the same instruments and software to administer the questions, with trained PATH Study interviewers carrying out the field work (though independently of the main PATH Study). More details on the PATH-RV study can be found in Tourangeau, Yan, Sun, Hyland, and Stanton (2018); basic descriptive results from the study are presented there.

2.1 Sample Design and Selection

The target population for the PATH-RV study was the civilian household population twelve years of age or older in the United States (the 50 states and the District of Columbia). Active-duty members of the military and persons living in group quarters were excluded.

The sample was selected in four stages, beginning with thirty-nine primary sampling units (PSUs); the PSUs consisted of individual counties or groups of adjoining counties. The PATH-RV PSUs were a one-in-four subsample of the 156 PATH Study PSUs. The subsampling of PSUs for the PATH-RV study preserved the stratification in the original sample, ensuring a representative mix of areas across the four census regions and across different levels of urbanization. At the next stage, second-stage sampling units (SSUs), consisting of individual census blocks or groups of adjoining blocks, were selected. A systematic sample of 746 PATH-RV study SSUs was drawn. At the third stage of sampling, individual addresses were selected within the 746 sample SSUs. We selected a total of 9,782 addresses or about thirteen per SSU. The addresses came from the US Postal Service's Computerized Delivery Sequence File (CDSF). In addition, in a subsample of SSUs, the CDSF was supplemented, with field staff canvassing those areas and adding any addresses omitted from the CDSF to the final list for sampling. The addresses selected for the PATH-RV sample were addresses that had been held in reserve for the main PATH Study, assuring that no addresses were selected for both samples.

In the final stage of sample selection, persons living at the sample addresses were selected. Sample addresses were mailed a short screening questionnaire to identify members of three key population groups—adult (eighteen years and older) tobacco users, adult nonusers, and youth (twelve to seventeen years

old). We received screening questionnaires from 2,296 households. We selected both tobacco users and nonusers to ensure that we would have observations for every section of the adult questionnaire. We selected a total of 865 adults and 266 youth for the PATH-RV sample. In households where a youth was selected, we also randomly sampled one of the adults. In households with more than one youth, we selected one of them at random.

2.2 Data Collection

Data collection for the study took place in two phases. The first phase was the mail screening effort. Screening questionnaires were sent to the 9,782 sample addresses, and 2,296 of them returned completed questionnaires; at another 643 addresses, the mailings were returned as undeliverable. The overall response rate to the screening component of the study was 25.1 percent (AAPOR RR3). Sample addresses received up to six mailings: (i) an advance letter; (ii) an initial survey package (with a cover letter, screening questionnaire, and cash incentive—initially $5, later reduced to $2); (iii) a thank you/reminder postcard; (iv) a nonresponse mailing with a replacement questionnaire; and (v) and (vi) a final thank you/reminder postcard (later increased to two reminders).

The rest of the data collection was carried out by Westat field interviewers, all of whom were also conducting interviews for the main PATH Study. The PATH-RV study training covered the components of the study that differed from the main PATH Study, say the collection of saliva samples. The training consisted of self-paced home study lasting approximately four hours and an hour-long group session held via WebEx. In total, sixty-eight interviewers conducted the field data collection.

Respondents were interviewed twice, with the reinterview done five to twenty-four days after the initial interview. Both interviews were done using ACASI, as in the main PATH Study. With ACASI, the computer displays the questions directly to the respondents on screen and also plays the question aloud to them via earphones. In the PATH-RV study (and the main PATH Study), a text-to-speech synthesized voice was used to generate the audio version of the questions. Both the initial interview and the reinterview used the PATH Study wave four questionnaires. On average, both interviews took about an hour for adults to complete and about 40 minutes for the youths. Adults were offered $35 for completing each interview, and youths were offered $25. Both adults and youths were offered $10 to provide a saliva sample after the reinterview. Before contacting the members of the youth sample, interviewers first obtained parental consent for the interview.

With a few exceptions, the protocol was the same in the reinterview as in the initial interview. The reinterview questionnaire included some additional items that all came after all of the regular PATH Study questions had been

administered. These consisted of alternative measures of some of the variables, a ten-item battery designed to assess the Big Five personality traits (Goldberg 1992), a multi-item scale to assess the need for cognition (Cacioppo and Petty 1982; Cacioppo, Petty, and Kao 1984) that was administered only to adult respondents, and questions asking about the reasons for any discrepant answers to selected items about tobacco use. The possible reasons for discrepancies included true change, misunderstanding the question, memory problems, inattention, and reluctance to answer truthfully (Cottler, Compton, Brown, Shell, Keating, et al. 1994). The reinterview program had a record of the respondent's answers in the initial interview, and the probes (which were administered after the respondent completed the main PATH questionnaire) were triggered when discrepant answers on selected questions were detected. Respondents were asked to explain why their answers changed but were not given the chance to change their answers; in addition, they were not forewarned they might be asked to explain some of their answers. The results from discrepancy results are described in more detail by Tourangeau et al. (2018).

After the reinterview, respondents were asked to provide a saliva sample. They were not told beforehand they would be asked to do this. Respondents who agreed were then administered a saliva test done by the interviewer, using the Alere iScreen screening device. This test detects cotinine (a metabolite of nicotine) for up to four days after tobacco use. Interviewers recorded the test results using the data collection software. Finally, if an adult respondent reported using any tobacco products in the second interview, the interviewer asked to photograph the products. For the respondents who agreed (110 of them), the interviewer photographed the product(s) using the camera built in to the computer used to administer the questionnaire.

Table 2 shows the outcomes of the field work—that is, the number of sampled adults who completed the PATH-RV interviews and the number who provided saliva samples. Overall, 46.3 percent of the sample members completed the two interviews; 89.5 percent of those also provided a saliva sample.[1] Appendix table 1 shows the composition of the PATH-RV sample and compares it with the main PATH study sample and with population figures from the CPS.

1. The main PATH Study response rates were considerably higher than those of the PATH-RV study. The fourth wave of the main PATH Study, conducted at roughly the same time as the PATH-RV study (December 2016 through January 2018 for wave four of the main PATH Study versus March 2017 through February 2018 for the PATH-RV study), included a replenishment sample. The screener response rate for the newly sampled cases was 53.0 percent (versus 25.1 percent in our study); 68.3 percent of the newly selected adults and 70.4 percent of the newly added youths completed wave four interviews (as compared with 51.6 percent of the adults and 48.4 percent of the youths who completed initial interviews in our study). The main PATH Study collected screener data in person.

Table 2. Data Collection Results for the PATH-RV Study, by Sample Subgroup

	Selected	Completed initial interview	Completed reinterview	Provided saliva sample
Adults	865	446	407	366
Users	329	176	161	142
Nonusers	536	270	246	224
Youths	266	129	117	102

3. RESULTS

Most of our analyses look at two reliability measures: the gross discrepancy rate (GDR) and kappa. The gross discrepancy rate is the proportion of respondents who give different answers to the question in both interviews. Kappa is the chance-corrected agreement rate. Lower values for the GDR and higher values for kappa indicate greater reliability.

We present four sets of results. First, we analyzed the results for each item. We fit models that examined the reliability of the item (as measured by kappa or the GDR) as a function of the item's characteristics. The item-level models took the following form:

$$y_i = \beta_0 + \sum_j \beta_{1j} x_{ij} + \varepsilon_i,$$

where y_i is the item-level GDR or kappa for item i, and x_{ij} represents characteristic j of item i (say the number of sentences in the item). The GDRs and kappas are sample estimates, and we ran the models using both unweighted and weighted least squares; in the weighted models, the weights were the inverse of the variance of the GDR or kappa. We made the standard assumptions about the distribution of the residuals (that is, they were normally distributed with a mean of zero).

Next, we examined respondent characteristics related to reliability. For these analyses, we computed the percentage of items for which the respondent gave the same answer in both interviews and determined what respondent characteristics were related to this respondent-level reliability measure:

$$p_{k(m)} = \beta_0 + \sum_j \beta_{1j} r_{jk(m)} + \varepsilon_k,$$

in which $p_{k(m)}$ is the proportion of items that respondent k from PSU m answered the same way in both interviews, and r_{jk} is the j^{th} characterstic of that respondent. In these analyses, we used SAS PROC SURVEYREG, treating the PSU from which the respondent had been selected as a cluster variable. We fit these models both with and without the sampling weight and assumed the residuals were normally distributed with a mean of zero.

Then, we present the results from cross-classified, multilevel models that examined item and respondent characteristics at the same time. These models predicted the log odds that a particular respondent would give the same answer to a given item in both interviews:

$$ln\left(\frac{P_{y_{ij}=1}}{P_{y_{ij}=0}}\right) = \beta_0 + \sum_k \beta_{2k} r_{ik} + \sum_l \beta_{1l} x_{jl} + \mu_i + \mu_j,$$

in which the log odds that respondent i gives the same answer to item j in both interviews depends on the respondent's characteristics (r_{ik} represents the k^{th} characteristics of respondent i) and the item's characteristics (x_{jl} is the l^{th} characteristic of item j). We used SAS PROC GLIMMIX to fit cross-classified models and did not use weights.

Finally, we examined the issue of the relationship between three potential indicators of difficulties in answering a question—the two reliability measures (GDR and kappa), the item nonresponse rate, and the median response time for the item.

3.1 Item Reliability and Item Characteristics

Table 3 presents the results of the item-level models.[2] Among the item characteristics, the centrality variable reflected the degree to which the topic was important, salient, or familiar to one's life. For the social desirability coding, we used a three-level scheme where an item that was sensitive or invoked social desirability concerns for most people was coded as three. An item that was sensitive or invoked social desirability concerns for *some* people or on *some* occasions was coded as two. Items that were not sensitive and did not invoke social desirability concerns were given a code of one.

After we dropped some outliers (items with large residuals or large absolute values of Cook's D), the models explain most of the variation across items in both GDR's (R^2 of 0.72 for the adults and 0.68 for the youths) and kappas (R^2 of 0.72 for the adults and 0.71 for the youths).

In the models for both questionnaires, the type of item makes a difference. Attitudinal items elicit less reliable answers than factual items, and demographic items receive more reliable answers than other types of questions. If we separate the items into three categories—attitudinal, behavioral, and demographic—the average GDRs were 0.31, 0.09, and 0.01 for the adults and 0.31, 0.14, and 0.01 for the youths for the three types of items; similarly, the

2. The item characteristics were first coded by one coder. A second coder double-coded 10 percent of the items. Then both coders met with two of the authors to resolve any discrepancies. The first coder then applied the same resolutions to all applicable items. Then one of the authors coded all items again on centrality and social desirability concerns. The agreement with the final coding from the first coder was high and any remaining disagreements were resolved after discussing with another author.

Table 3. Linear Regression Coefficients for Item-Level Models of GDR and Kappa, By Sample

Adults	GDR			Kappa		
Question characteristic	Unweighted	Weighted	Weighted, problem items dropped	Unweighted	Weighted	Weighted, problem items dropped
Attitudinal (vs. factual)	0.13***	0.08***	0.10***	−0.05	−0.06	−0.03
Demographic (vs. other)	−0.13***	−0.08***	−0.11***	0.25***	0.26***	0.16***
Number of sentences	0.00	0.01†	0.01†	−0.01	0.00	−0.02*
Number of words per sentence	0.002*	0.001†	0.001	−0.003	−0.01*	−0.004***
Number of response options	0.04***	0.03***	0.02***	−0.01	−0.01	−0.01
Position in the questionnaire	0.01*	0.01†	0.00	−0.02†	−0.03**	−0.02***
Not a scale (vs. a response scale)	−0.06**	−0.07***	−0.09***	0.16**	0.12*	0.26***
Frequency scale (vs. other scales)	0.06*	0.02	−0.01	−0.05	0.03	−0.04
Extent of social desirability concerns	−0.01	−0.01**	−0.01**	0.09***	0.06*	0.04***
Central topic	0.05***	0.02***	0.03***	−0.02	0.02	−0.01
Present/future reference period (vs. past)	0.01	0.01*	0.01†	0.00	0.03	0.00
Flesch score	0.00	0.00	0.00	0.01	0.01	0.01
Number of items	426	397	389	419	389	375
R-squared	0.69	0.60	0.72	0.32	0.34	0.72

NOTE.—In the weighted models, the weights were the inverse of the variance of the GDR or kappa.
***$p < 0.0001$; **$p < 0.01$; *$p < 0.05$; $^{\dagger}p < 0.10$.

Youths	GDR			Kappa		
Question characteristic	Unweighted	Weighted	Weighted, problem items dropped	Unweighted	Weighted	Weighted, problem items dropped
Attitudinal (vs. factual)	0.05	0.02	0.04	0.02	−0.03	−0.07*
Demographic (vs. other)	0.02	0.10***	0.07*	0.23***	0.26***	0.27***
Number of sentences	0.00	−0.01	−0.02	−0.01	0.04†	0.01
Number of words per sentence	0.00	0.00	0.00	0.01†	0.00	0.00
Number of response options	0.04***	0.03***	0.05***	−0.01	0.00	−0.01
Position in the questionnaire	0.02†	0.01	0.00	0.01	−0.02	0.01
Not a scale (vs. a Response scale)	−0.08*	−0.13***	−0.05	0.21***	0.17**	0.16***
Frequency scale (vs. other scales)	−0.10*	−0.05	−0.05	0.01	−0.06	0.02
Extent of social desirability concerns	−0.03*	−0.01	−0.01	0.06**	0.04*	0.08**
Central topic	−0.06***	−0.10***	−0.10***	−0.04	−0.01	−0.01
Present/future reference period (vs. past)	−0.10***	−0.09***	−0.09***	0.08	0.04	0.13
Flesch score	−0.01	−0.01**	−0.02***	0.02*	0.02**	0.01
Number of items	224	198	196	213	187	182
R-squared	0.67	0.63	0.68	0.54	0.53	0.71

NOTE.—In the weighted models, the weights were the inverse of the variance of the GDR or kappa.
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; †$p < 0.10$.

average kappas were 0.46, 0.63, and 0.88 for the adults and 0.42, 0.56, and 0.92 for the youths. Only the advantage for demographic questions is significant for the youths. We note that one of the few items in either questionnaire where the answers were perfectly reliable was the question asking the respondent's sex. Question length—both the number of sentences and the number of words per sentence—reduced reliability, but the effect of question length is apparent only for the adults. It is almost inevitable that the more response options the lower the GDR, and that effect is apparent with both adults and youths. With more options, it is more likely that respondents will choose a different answer in the second interview. For both adult and youth respondents, the effect of the number of options was no longer significant in the models for kappa. The other item characteristics were significant predictors of reliability in at least some of the models: whether the item featured a response scale (which lowered reliability), the extent to which it raised social desirability concerns (which increased reliability, though only for the kappa statistic in the youth sample), the position of the item in the questionnaire, readability as measured by the Flesch reading-ease score, whether it touched on a central topic, and whether it concerned the present or future rather than the past. Later items were found to have lower reliability than earlier items, but this effect was only statistically significant for the adult kappa model. For youths, easier items as indicated by higher Flesch reading-ease scores produced higher reliability than harder items. The effect of topic centrality was statistically significant only for the GDR statistic in both samples, and it lowered reliability for adults but improved reliability for youths. Retrospective questions produce more reliable answers than questions about the present or future, although this effect was more consistent in the models for the GDR in youth samples. (In both questionnaires, only a few items asked about the future—three in the adult questionnaire and five in the youth questionnaire.)

Using the same variables included in our item-level model, we coded ninety-one items from the NSDUH reliability study for which estimates of kappa had been published. We used the final model estimates (combining the PATH-RV data for adults and youths, because these were not separated in the NSDUH reliability study) to predict kappas for the NSDUH items. The overall correlation between our predictions, which were based solely on the PATH-RV study and the published NSDUH kappa values, was 0.58. Our item-level model consistently underpredicted the kappas for the NSDUH items; the average deviation between the predicted and actual values was −0.05, and the root mean square error of prediction was 0.15. We also obtained predicted reliabilities from Survey Quality Prediction (SQP) for these ninety-one items and compared SQP predictions with the published NSDUH kappas. The overall correlation between the two was 0.34. In addition, SQP predictions underestimated the reliability of the NSDUH items; the average deviation between SQP predictions and the published NSDUH kappa values was −0.16, and the root mean square error of prediction was 0.55.

3.2 Reliability and Respondent Characteristics

Table 4 presents the results for the respondent-level models. The dependent variable in these analyses was the percentage of items for which the respondent provided the same answers in both interviews. We display both weighted and unweighted results (the weight was the final sampling weight for the respondent, which adjusted for unequal selection probabilities and differential nonresponse and was raked to population totals), and the analyses take into account the clustering by PSU. For the youth sample, age and education are highly confounded, and we dropped the latter variable from the model.

For the adult sample, six respondent characteristics were significantly related to the percentage of identical answers. As we expected, conscientiousness was positively related to this variable; respondents with a higher level of conscientiousness provided a higher percentage of identical answers than those with a lower level of conscientiousness. However, adult respondents with high school or less education, respondents who report using tobacco (either every day or some days), males, and respondents from households with a household income less than $50,000 provided fewer identical answers across the two interviews. In addition, there was a significant effect for the number of days between the interview and reinterview; respondents gave somewhat fewer identical answers the more time that elapsed between the two interviews.[3]

Among the youths, some day tobacco users, Hispanic respondents, and non-Hispanic white respondents were significantly less likely to provide the same answers in both interviews.

Respondents whose saliva test results agreed with their survey responses were only slightly more reliable on average than those whose saliva test results disagreed with their survey responses. The average proportion of identical answers across interviews was 0.85 for adults whose saliva test agreed with their reported tobacco use versus 0.84 for those who saliva test disagreed with their reported tobacco use. The corresponding figures in the youth sample were 0.81 and 0.80.

3.3 Cross-Classified Multilevel Models

The results of the multilevel cross-classified logistic regression models are shown in tables 5 and 6. In these models, the outcome variable, whether a given respondent gave the same answer to a given item in both interviews, is dichotomous. It is clear from the unconditional models (which incorporate

3. We were somewhat surprised by this finding. The median time between interviews was twelve days. Adult respondents reinterviewed within twelve days gave the same answer to 85 percent of the questions on average; the figure was identical for adults reinterviewed after thirteen to twenty-four days had elapsed. Similarly, youths reinterviewed within twelve days gave identical answers to 82 percent of the questions on average versus 81 percent for those reinterviewed thirteen to twenty-four days after their initial interview.

Table 4. Regression Coefficients for Respondent-Level Models, by Sample

Respondent characteristic	Adults			Youths		
	Unweighted	Weighted	Weighted, problem cases dropped	Unweighted	Weighted	Weighted, problem cases dropped
60 or older (vs. younger than 60) for adults; continuous for youth	0.00	0.00	0.00	−0.01	−0.01†	−0.01†
High school or less (vs. more than HS) for adults; omitted for youth	−0.02**	−0.03**	−0.02**	—	—	—
Conscientiousness	0.005*	0.008*	0.006*	−0.01	−0.01	−0.01
Need for cognition; omitted for youth	0.00	0.00	0.00	—	—	—
Every day tobacco user (vs. nonuser) for adults; last 7 day user (vs. nonuser) for youth	−0.03***	−0.03*	−0.04***	−0.04	−0.04	−0.04†
Some day tobacco user (vs. nonuser)	−0.02***	−0.02	−0.02**	−0.06***	−0.05*	−0.07***
Male (vs. female)	−0.02***	−0.03**	−0.02***	−0.01	−0.01	−0.02
Hispanic (vs. all others)	−0.01	−0.01	−0.01	−0.06*	−0.07**	−0.05*
Non-Hispanic black (vs. all others)	−0.01	−0.01	0.00	−0.01	0.00	−0.01
Non-Hispanic white (vs. all others)	0.01	0.02	0.02	−0.06*	−0.06*	−0.06*
Households receiving income assistance	−0.02**	−0.01	−0.01	−0.01	0.00	−0.01
Households < $50k	−0.02***	−0.02†	−0.02***	0.00	0.02	0.01
Number of elapsed days	−0.001†	−0.002	−0.002***	−0.002	−0.001	−0.003
Number of respondents	407	407	383	111	111	110
R-Squared	0.35	0.35	0.51	.019	0.26	0.30

NOTE.—***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; $^{\dagger}p < 0.10$. The need for cognition items were omitted from the youth questionnaire. Except for the number of elapsed days, which was derived from paradata, the respondent characteristics reflect respondent self-reports in the initial interview. The weights used in the weighted models are the final sampling weights for the respondents.

Table 5. Logistic Regression Coefficients for Multilevel Model of the Likelihood of Giving the Same Answer in Both Interviews: by Sample

	Adults		Youths	
	All cases	Problem cases dropped	All cases	Problem cases dropped
Item characteristic				
Attitudinal (vs. factual)	−0.79***	−1.11**	−0.39	−1.58**
Demographic (vs. other)	2.95***	5.01***	1.39***	0.96
Number of sentences	0.01	−0.01	0.01	0.58†
Number of words per sentence	−0.02*	−0.07***	−0.03†	−0.04*
Number of response options	−0.28***	−0.40***	−0.29***	−0.49***
Position in the questionnaire	−0.05	−0.13	0.02	−0.33†
Not a scale (vs. not a response scale)	0.91***	1.96***	0.96***	2.13***
Frequency scale (vs. other scales)	−0.26	−0.01	1.02**	1.67*
Extent of social desirability concerns	0.42***	0.75***	0.30**	0.63**
Central topic	−0.80***	−1.05***	0.21	0.15
Present/future reference period (vs. past)	−0.19	−0.39	0.98***	2.43***
Flesch score	−0.02	−0.10	0.07	0.04
Respondent characteristic				
60 or older (vs. younger than 60) for adults; continuous for youth	0.00	0.00	−0.04	−0.06
High school or less (vs. more than HS) for adults	−0.17***	−0.17***	—	—
Conscientiousness	0.05*	0.07**	−0.04	−0.03
Need for cognition	−0.01	−0.02	—	—
Every day tobacco user (vs. nonuser) for adults Last 7 day user (vs. nonuser) for youth	−0.23***	−0.29***	−0.30	−0.24
Some day tobacco user (vs. nonuser)	−0.12†	−0.15†	−0.45**	−0.36†
Male (vs. female)	−0.14***	−0.13**	−0.07	−0.07
Hispanic (vs. all others)	−0.05	0.04	−0.44†	−0.45
Non-Hispanic black (vs. all others)	−0.03	0.01	−0.10	−0.09
Non-Hispanic white (vs. all others)	0.13	0.19†	−0.47†	−0.51

Continued

Table 5. *Continued*

	Adults		Youths	
	All cases	Problem cases dropped	All cases	Problem cases dropped
Households receiving income assistance	−0.22***	−0.21**	−0.05	−0.09
Households <50k	−0.21***	−0.21***	−0.01	−0.04
Elapsed time between interviews	−0.01*	−0.01*	−0.02	−0.03†

NOTE.—Youths did not receive the need for cognition questions; problem cases were those with values of Cook's D > 3 or absolute Studentized residuals > 3. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; †$p < 0.10$.

Table 6. Variance Components and Intraclass Correlation Coefficients (ICCs) from Cross-Classified Models

	Adults		Youth	
	Variance	ICC	Variance	ICC
Unconditional				
Respondent	0.260	0.024	0.368	0.036
Question	7.420	0.677	6.486	0.640
Conditional				
Respondent	0.190	0.026	0.365	0.051
Question	2.792	0.386	2.083	0.288

only random effects for respondents and items but none of the predictors) that most of the variation is due to items rather than respondents (see table 6). The intraclass correlation coefficients (ICCs) for respondents were 0.02 for the adult sample and 0.04 for the youths. By contrast, the ICCs for the items were 0.68 and 0.64 for the adults and youths, respectively. We used the "covtest" statement in PROC GLIMMIX to test whether the variance components were significant and found that all the variance components reported in table 6 were significantly larger than zero.

Most of the same variables that were significant in the earlier models were significant in the cross-classified model, as well. In the adult sample, respondents were significantly more likely to give the same answer in both interviews

to factual items (especially demographic items) than to attitudinal items; questions with fewer words per sentence were more likely to elicit identical answers in both interviews than longer questions; and questions with fewer response options were more likely to receive identical answers than questions with more response options (although, again, this is almost inevitable for this outcome variable). In addition, items that do not use response scales and those with a higher level of social desirability concerns were more likely to elicit the same answers in both interviews, and those that concerned a central topic were less likely to elicit the same answer. Again, it is surprising that social desirability predicts *greater* reliability. Saris and Gallhofer (2007a, 2007b) also found a positive impact of social desirability on reliability, although the impact didn't reach statistical significance. Similarly in the adult sample, the respondent's education, conscientiousness, tobacco use, sex, and household income were all significantly related to the likelihood the respondent would give the same answer to a given item in both interviews; and again, there is a significant effect of the elapsed time between interviews.

For the most part, the results for the youth sample are similar to those for the adults, with a few exceptions. For the youth sample, the difference between demographic and other items was not significant (although it was in the same direction as for the adults); in addition, the effects of the number of words per sentence and topic centrality were no longer significant. Frequency scales significantly increased the chance that respondents in the youth sample would give the same answer in both interviews, and those that asked about the present or future were more likely to elicit the same answer in both interviews. By contrast, Saris and Gallhofer (2007a, 2007b) found that questions asking about the past were more reliable than those asking about the present or future. Among the youths, the only respondent characteristic that was statistically significant was smoking status, and this effect was only marginally significant in the final model. Given the restricted range on the age variable in the youth sample, it is not surprising that that variable was not significant in the models for the youths.

The item variables reduced the item-level ICCs by from 0.68 to 0.39 in the adult sample and 0.64 to 0.29 in the youth sample, suggesting the item characteristics included in the models explain much of the variation across items. We note that there were fewer predictors available for the youth respondents than for the adults, and that some variables such as age did not vary as much within the youth sample.

3.4 Reliability, Item Nonresponse, and Response Times

If the same cognitive difficulties that give rise to unreliable answers also engender slow response times and missing data, we might expect these outcomes to be related to each other. Table 7 shows the intercorrelations among six

Table 7. Intercorrelations among Measures of Item Difficulty

	GDR	Kappa	Item NR—1	Item NR—2	RT—1	RT—2
Adult sample						
Gross discrepancy rate (GDR)	—	−0.498*** (419)	−0.059 (426)	0.029 (426)	0.371*** (235)	0.329*** (235)
Kappa		—	−0.030 (419)	−0.173*** (419)	−0.128† (235)	−0.092 (235)
Item NR—1st interview			—	0.558*** (426)	0.114† (235)	0.075 (235)
Item NR—2nd interview				—	0.184** (235)	0.193** (235)
Median response time—1st interview					—	0.934*** (235)
Median response time—2nd interview						—
Youth sample						
Gross discrepancy rate (GDR)	—	−0.653*** (213)	0.347*** (224)	0.406*** (224)	0.292*** (156)	0.258** (156)
Kappa		—	−0.119† (213)	−0.192** (213)	−0.187* (153)	−0.126(153)
Item NR—1st interview			—	0.829*** (224)	0.230** (156)	0.146† (156)
Item NR—2nd interview				—	0.327*** (156)	0.259** (156)
Median response time—1st interview					—	0.906*** (156)
Median response time—2nd interview						—

NOTE.—***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; $^{\dagger}p < .10$. For the few items with marginal proportions of zero for some responses, kappa cannot be calculated; response times are available only for items that appeared on their own screen.

item-level variables—the GDR, kappa, item nonresponse rates in the initial interview and reinterview, and the median response times for the item in the two interviews. We examined these correlations for both the adult and the youth samples. The two reliability measures are strongly, though inversely, related to each other (adult sample: $r = -0.498$, $p < 0.001$; youth sample: $r = -0.653$, $p < 0.001$). The GDR is weakly related to the item nonresponse rates in the adult sample; neither correlation is significant. But the same correlations were statistically significant for the youth sample ($r = 0.347$, $p < 0.0001$, and $r = 0.406$, $p < 0.0001$). The GDR was positively related to the median response times (adult sample: $r = 0.371$ and $r = 0.329$ in the first and second interview, respectively, both $p < 0.001$; youth sample: $r = 0.292$, $p = 0.0002$ in the first interview and $r = 0.258$, $p = 0.001$ in the second interview). Kappa was significantly related to both item nonresponse (but only in the second interview; $r = -0.173$, $p < 0.001$ for the adult sample, and $r = -0.192$, $p = 0.005$ for the youth sample) and median response times (only in the first interview; $r = -0.128$, $p = 0.05$ for adults, and $r = -0.187$, $p = 0.02$ for youths).

All of the relationships in table 7 were in the expected directions, but those involving the item nonresponse rates were weak. This may be because the item nonresponse rates were very low and did not vary much across items. In both interviews, the adult respondents skipped fewer than 1 percent of the questions they were supposed to answer, and youth respondents skipped slightly more than 1 percent of the questions (1.6 percent for the first interview and 1.3 percent for the second). Still, the item nonresponse rates were strongly correlated across interviews (adult sample: $r = 0.558$, $p < 0.0001$; youth sample: $r = 0.829$, $p < 0.0001$), as were the median response times (adult sample: $r = 0.934$, $p < 0.0001$; youth sample: $r = 0.906$, $p < 0.0001$).

We also reran the item-level models presented in table 3 but included the two-item nonresponse rates and the two response times as additional predictors. For the adult sample, only the second interview nonresponse rate was significantly related to the GDR (and marginally related to kappa). For the youth sample, the median response time in the second interview was significantly related to GDR and kappa, and the item nonresponse rate in the second interview was significantly related to kappa.

4. DISCUSSION

Some questions are clearly easier for respondents to answer consistently than others. Factual questions, especially demographic ones, produce more reliable question than subjective questions, a point also made by Alwin (2007; see Saris and Gallhofer 2007a, 2007b for similar findings). Simpler questions—those with fewer words per sentence and fewer response options—also seem to produce more reliable data than more complicated ones (cf., Alwin 2007; Revilla et al. 2014; Alwin et al. 2018). Items with ordinal response scales

produce lower reliabilities than other items, presumably because they require respondents to make graded judgments, which can present cognitive challenges and also make it harder to map the underlying judgment onto the response scale. Paradoxically, items that raised social desirability concerns seemed to elicit more reliable responses. If respondents are editing their answers before they report them, they are doing it consistently. For the youth sample at least, questions about the future or present are also easier than retrospective questions, which presumably require more difficult retrieval processes. Earlier studies have found that people are particularly prone to inconsistency when they are reporting about events, such as the onset of substance use, that happened long ago (e.g., Johnson and Mott 2001). Saris and Gallhofer (2007a, 2007b) found questions about the past to be more reliable than questions about the present or future. Still, their work is mostly based on surveys administered to people age 18 or older, and our finding is based on younger respondents.

And some respondents give more reliable answers than others. Adult respondents with more education gave more reliable answers to the questions than less educated respondents (see also Alwin 1989, 2007; Alwin and Krosnick 1991; Saris and Gallhofer 2007a,b); it is likely this finding reflects the greater (or better developed) cognitive skills of those with more education. Conscientiousness was also related to reliability. Additional analyses indicate that none of the other Big Five traits were associated with reliability. It makes sense that respondents who habitually do things carefully will answer survey questions more consistently than those who do not. Females and respondents from more prosperous households were also more likely to give reliable answers than males or respondents from poorer households. One surprising finding was that tobacco users were less likely to give reliable answers than nonusers. We were concerned that this might reflect the fact that tobacco users had to answer more (and different) questions than the nonusers, but when we restricted the analysis to the questions that every respondent got, regardless of whether they used tobacco or not, the tobacco users were still less reliable. It may be the sheer number of questions they had to answer led to their lower reliability. The median interview length for the first adult interview was sixty-one minutes for nonusers, but eighty-three minutes for everyday adult tobacco users. Similarly, the median interview length for the second adult interview was sixty minutes for the nonusers and eighty-six for the everyday tobacco users. A similar pattern was apparent for youth interviews, as well; tobacco users got more questions than nonusers and took longer to complete the interview.

We were surprised to see that the coefficient for the number of elapsed days between the interviews was significant in the respondent and cross-classified models, though only for adults. It is hard for us to believe that this is a memory effect. Both youth and adult respondents answered literally hundreds of questions in their initial interviews. It is not clear why they would remember such

low salience behaviors. An alternative interpretation of this finding is that as more time elapsed between interviews, more actual changes occurred. A previous analysis of the discrepancy probes (which asked respondents why their answers had changed) found that adults attributed 11 percent of their discrepant answers to true change between interviews; youth respondents attributed an even higher proportion of the discrepancies—44 percent—to true change (Tourangeau et al. 2018). Thus, at least in part, the effect of the passage of time may reflect true variability over time in some of the phenomena of interest.

Although many of the findings from the PATH-RV study replicate earlier findings about the correlates of reliability, we believe the study is still valuable because it uses a simpler, less model-dependent approach for estimating reliability than the earlier studies. Most of the work done by Alwin and his colleagues rests on a particular model, the quasi-simplex model for three-wave panel data, and a set of auxiliary assumptions needed to make the parameters of that model identifiable. Thus, it is reassuring that our findings, based on reinterview data, come to similar conclusions. Similarly, the work of Saris and his colleagues depends heavily on analyses of multitrait multimethod experiments, which rest on quite different assumptions from those of a reinterview study. Still, if Saris and Gallhofer's emphasis on the impact of methods effects is warranted, our reliability estimates may be inflated by such effects.

Our study also included some new variables, not examined in the prior literature. We found a significant effect for conscientiousness, at least in the adult sample. Not surprisingly, conscientious respondents were more likely to give consistent answers across interviews. In addition, our conjecture that reliability is related to item nonresponse and response latencies received some support (see table 7). We thought that many of the same variables that lead to unreliable answers also lead to item nonresponse and slower response times. Respondents tended to take longer on questions they did not end up answering, suggesting that these nonresponses were not the product of satisficing but reflected respondents' genuine inability to come up with acceptable answers. They took longer not because they took a cognitive shortcut but because they tried to answer the question and then gave up.

Appendix Table 1. Distribution of the PATH-RV, Main PATH, and CPS Samples, by Age, Sex, and Race/Ethnicity

Adults	Screener (n = 855)	PATH-RV Time 1 (n = 442)		PATH-RV Time 2 (n = 403)		Main PATH Wave 3 (n = 28,122)		2016 CPS
		Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Weighted
Age								
18–59	68.7%	69.6%	74.4%	70.9%	76.0%			72.7%
>=60	31.3%	30.4%	25.6%	29.1%	24.0%			27.3%
Median Age	52.0	53.0	37.0	53.0	35.8			47.0
Mean Age	51.2	51.3	44.6	50.7	43.8			47.0
Gender								
Male	47.8%	44.1%	47.6%	43.2%	46.4%	49.0%	48.0%	48.4%
Female	52.2%	55.9%	52.4%	56.8%	53.6%	51.0%	52.0%	51.6%
Ethnicity								
Hispanics		8.2%	15.5%	8.5%	16.2%	18.9%	15.5%	15.7%
Non-Hispanic Whites		74.3%	62.2%	74.9%	62.2%	58.7%	65.6%	64.4%
Non-Hispanic Blacks		8.7%	10.8%	8.7%	11.3%	14.7%	11.2%	11.8%
Non-Hispanic Others		8.9%	11.5%	8.0%	10.3%	7.6%	7.7%	8.0%

Youths	Screener (n=266)	PATH-RV Time 1 (n=129)		PATH-RV Time 2 (n=117))		Main PATH Study Wave 3 (n=11,792)		2016 CPS
		Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Weighted
Age								
12	15.0%	17.1%	18.7%	18.8%	20.8%			16.4%
13	17.7%	18.6%	17.1%	18.0%	17.3%			15.9%
14	17.3%	18.6%	18.7%	17.1%	16.1%			16.1%
15	18.4%	17.8%	19.1%	18.8%	20.3%			16.6%
16	15.4%	14.0%	11.5%	13.7%	11.6%			17.8%
17	16.2%	14.0%	14.9%	13.7%	13.9%			17.1%
Median Age	14.5	14.0	13.8	14.0	13.7			15.0
Mean Age	14.5	14.3	14.3	14.3	14.3			14.5
Gender								
Male	49.3%	56.6%	51.2%	55.6%	49.3%	51.8%	51.4%	50.7%
Female	50.8%	43.4%	48.8%	44.4%	50.7%	48.2%	48.6%	49.3%
Ethnicity								
Hispanics		18.3%	27.7%	15.8%	24.5%	29.8%	23.9%	23.2%
Non-Hispanic Whites		64.3%	48.8%	66.7%	50.7%	47.2%	53.1%	53.8%
Non-Hispanic Blacks		6.4%	9.4%	6.1%	9.5%	13.7%	13.1%	13.8%
Non-Hispanic Others		11.1%	14.1%	11.4%	15.3%	9.3%	10.0%	9.1%

Note. The PATH Study Public Use Files are used for the Wave 3 estimates.

References

Alwin, D. F. (1989), "Problems in the Estimation and Interpretation of the Reliability of Survey Data," *Quality & Quantity*, 23, 277–331.

———. (2007), *Margins of Error: A Study of Reliability in Survey Measurement*, Hoboken, NJ: John Wiley.

Alwin, D. F., E. M. Baumgartner, and B. A. Beattie (2018), "Number of Response Categories and Reliability in Attitude Measurement," *Journal of Survey Statistics and Methodology*, 6, 212–239.

Alwin, D. F., and J. A. Krosnick (1991), "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes," *Sociological Methods and Research*, 20, 139–181.

Andrews, F. (1984), "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach," *Public Opinion Quarterly*, 48, 409–442.

Brener, N. D., J. L. Collins, L. Kann, C. W. Warren, and B. I. Williams (1994), "Reliability of the Youth Risk Behavior Survey Questionnaire," *American Journal of Epidemiology*, 141, 575–457.

Cacioppo, J. T., and R. E. Petty (1982), "The Need for Cognition," *Journal of Personality and Social Psychology*, 42, 116–131.

Cacioppo, J. T., R. E. Petty, and C. F. Kao (1984), "The Efficient Assessment of Need for Cognition," *Journal of Personality Assessment*, 48, 306–307.

Cannell, C. F., P. Miller, and L. Oksenberg (1981), "Research on Interviewing Techniques," in *Sociological Methodology 1981*, ed. S. Leinhardt, pp. 389–437. San Francisco, CA: Jossey-Bass.

Cottler, L. B., W. M. Compton, L. Brown, A. Shell, S. Keating, A. Shillington, and R. Hummel (1994), "The Discrepancy Interview Protocol: A Method for Evaluating and Interpreting Discordant Survey Responses," *International Journal of Methods in Psychiatric Research*, 4, 173–182.

Couper, M. P., and F. Kreuter (2013), "Using Paradata to Explore Item Level Response Times in Surveys," *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 176, 271–286.

Forsman, G., and I. I.Schreiner (1991), "The Design and Analysis of Reinterview: An Overview," in *Measurement Error in Surveys*, eds. P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, pp. 279–302. New York, NY: John Wiley.

Goldberg, L. R. (1992), "The Development of Markers for the Big Five Factor Structure," *Psychological Assessment*, 4, 26–42.

Grant, B. F., D. A. Dawson, F. S. Stinson, P. S. Chou, W. Kay, and R. Pickering (2003), "The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDASIS-IV): Reliability of Alcohol Consumption, Tobacco Use, Family History of Depression and Psychiatric Modules in a General Population Sample," *Drug and Alcohol Dependence*, 71, 7–16.

Grant, B. F., R. B. Goldstein, S. M. Smith, J. Jung, H. Zhang, S. P. Chou, R. P. Pickering, et al. (2013), "The Alcohol Use Disorder and Associated Disabilities Interview Schedule-5 (AUDASIS-5): Reliability of Substance Use and Psychiatric Disorder Modules in a General Population Sample," *Drug and Alcohol Dependence*, 148, 27–33.

Hout, M., and O. P. Hastings (2016), "Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006–2014," *Sociological Science*, 3, 971–1002.

Johnson, T. P., and J. A. Mott (2001), "The Reliability of Self-Reported Age of Onset of Tobacco, Alcohol and Illicit Drug Use," *Addiction*, 96, 1187–1198.

Krosnick, J. A. (1991), "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys," *Applied Cognitive Psychology*, 5, 213–236.

———. (1999), "Survey Research," *Annual Review of Psychology*, 50, 537–567.

Olson, K., and J. D. Smyth (2015), "The Effect of CATI Questions, Respondents, and Interviewers on Response Time," *Journal of Survey Statistics and Methodology*, 3, 361–396.

O'Muircheartaigh, C. (1991), "Simple Response Variance: Estimation and Determinants," in *Measurement Error in Surveys*, eds. P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, pp. 551–574, New York, NY: John Wiley.

Revilla, M., W. E. Saris, and J. A. Krosnick (2014), "Choosing the Number of Categories in Agree/Disagree Scales," *Sociological Methods & Research*, 43, 73–97.

Rodgers, W. L., F. M. Andrews, and A. R. Herzog (1992). "Quality of Survey Measures: A Structural Equation Modeling Approach," *Journal of Official Statistics*, 3, 251–275.

Salthouse, T. A. (1994). "The Aging of Working Memory," *Neuropsychology*, 8, 535–543.

Saris, W. E., and I. Gallhofer (2007a). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*, Hoboken, NJ: John Wiley.

———. (2007b). "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions," *Survey Research Methods*, 1, 29–43.

Saris, W. E., M. Revilla, J. A. Krosnick, and E. M. Schaeffer (2010). "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options," *Survey Research Methods*, 4, 61–79.

Sinclair, M. D. and Gastwirth, J.L. (1996), "On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data," *Journal of the American Statistical Association*, 91, 961–969.

Smith, T. (1980). *Inconsistent People*. NORC: GSS Methodological Report 049.

Soulakova, J. N., A. M. Hartman, B. Liu, G. B. Willis, and S. Augustine (2012). "Reliability of Adult Self-Reported Smoking History: Data from the Tobacco Use Supplement to the Current Population Survey 2002-2003 Cohort," *Nicotine & Tobacco Research*, 42, 952–960.

Stein, A. D., R. I. Lederman, and S. Shea (1993). "The Behavioral Risk Factor Surveillance System Questionnaire: Its Reliability in a Statewide Sample," *American Journal of Public Health*, 83, 1768–1772.

Substance Abuse and Mental Health Services Administration (2010). *Reliability of Key Measures in the National Survey on Drug Use and Health*. Office of Applied Studies (Methodology Series M-8; HSS Publication No. SMA 09-4425). Rockville, MD, 2010.

Tourangeau, R. (1984). "Cognitive Science and Survey Methods," in *Cognitive Aspects of Survey Design: Building a Bridge between Disciplines*, eds. T. Jabine, M. Straf, J. Tanur, and R Tourangeau, 73100 Washington, DC: National Academy Press.

———. (2018). "The Survey Response Process from a Cognitive Viewpoint," *Quality Assurance in Education*, 26, 169–181.

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

Tourangeau, R., T. Yan, H. Sun, A. Hyland, and C. A. Stanton (2018), "Population Assessment of Tobacco and Health (PATH) Reliability and Validity Study: Selected Reliability and Validity Estimates," *Tobacco Control*.

Yan, T., and R. Tourangeau (2008). "Fast Times and Easy Questions: The Effects of Age, Experience, and Question Complexity on Web Survey Response Times," *Applied Cognitive Psychology*, 22, 51–68.