Were the Benchmarks Really Wrong?

By David S. Yeager and Jon A. Krosnick, Stanford University

December, 2009

In a recently posted essay, Humphrey Taylor of Harris Interactive offered a surprising interpretation of data we reported in a recent paper (Yeager et al., 2009).[1] Our paper compared the accuracy of an RDD telephone survey with various surveys done via the Internet. One of our conclusions was that a Harris Interactive survey we commissioned was less accurate than an RDD telephone survey with the same content that was administered at about the same time. Mr. Taylor views our data as supporting the opposite conclusion: that "the Harris Interactive data used by Yeager et al. is generally more accurate than the RDD sample."

His argument goes as follows:

1) We evaluated a series of surveys by comparing six of their measurements (e.g., the percent of people who smoke cigarettes) with assessments of the "true" values of these variables either from federal surveys administered by live interviewers, or from official government statistics.

2) Answers to the "benchmark" questions in the federal interviewer-administered surveys were distorted by social desirability response bias.

3) The same bias was present in the RDD telephone survey we commissioned.

4) Because of this shared bias, any resemblance of the RDD telephone survey to the federal surveys in the measures contaminated by social desirability bias is due to shared measurement error, not evidence of accuracy of the telephone survey.

5) Internet surveys are less subject to social desirability response bias than are interviewer-administered surveys.

6) Harris Interactive surveys typically elicit more reports of socially undesirable attitudes and behaviors than do interviewer-administered surveys, which is evidence of the superior accuracy of the former.

7) The Harris Interactive survey we commissioned collected more reports of cigarette smoking and alcohol consumption than did the RDD telephone survey we commissioned and the federal benchmark surveys, which is evidence of the superior accuracy of the Harris Interactive survey, due to its reduced susceptibility to social desirability response bias.

8) The Harris Interactive survey yielded results that were closer to the two official government statistics benchmarks (which were not contaminated by social desirability response bias) than were the telephone survey's numbers, again showing the superior accuracy of the former.

9) Therefore, the Harris Interactive survey was "generally more accurate."

In this essay, we evaluate the plausibility of this argument and demonstrate how the available data disconfirm most of these assertions, including the overall conclusion.

Specifically we show:

1) Internet surveys yield results less distorted by social desirability bias than do interviewer-administered surveys.

2) But the measures of smoking and drinking we examined were not contaminated by social desirability bias.

3) A careful reanalysis of the available data underlying the above claims leads to different conclusions.

4) When examining all 19 benchmarks that we discussed in our paper and subsets of them, the Harris Interactive survey manifested more error than the RDD telephone survey we commissioned.

We conclude that computers are terrifically helpful for survey data collection, partly because they can help to minimize social desirability pressures. But honest reporting is only one of the conditions necessary for a survey's results to be accurate. The sample must also be representative of the population. Our evidence suggests that people who opt in to do Internet surveys are not as representative of the population as are probability samples, which explains the superior accuracy of surveys of the latter.

Where We Agree

We agree that Internet surveys are less subject to social desirability bias than are surveys involving live interviewers. In three papers that will soon appear in Public Opinion Quarterly, we provide evidence of such effects (Chang & Krosnick, 2009; Chang & Krosnick, in press; Holbrook & Krosnick, in press).

One paper (Chang & Krosnick, in press) reports a laboratory experiment in which half the respondents completed a questionnaire privately on a computer, and the other half completed the same questionnaire that was administered orally by a live interviewer who was down the hall via a telephone-like electronic communication system. White respondents reported more opposition to government support for African Americans on the computer than they uttered aloud to the interviewers.[2]

The second paper (Chang & Krosnick, 2009) reports a field experiment in which the same questionnaire was administered to a national probability sample via RDD telephone interviewing and to a national probability sample who completed the questionnaire via the Internet. There, too, white Internet respondents expressed more opposition to government help to African Americans than did whites interviewed by telephone.

Other papers have also reported evidence consistent with the claim that computer administration of questionnaires minimizes social desirability bias, including a paper that Mr. Taylor mentioned in his essay:

> "Our online surveys have always found substantially more people than our telephone surveys who tell us they are gay, lesbian or bisexual (by a 3-to-1 margin). Our online surveys also find fewer people who claim to give money to charity, clean their teeth, believe in God, go to religious services, exercise regularly, abstain from alcohol, or drive under the speed limit."

Such evidence is consistent with a 2005 article in Public Opinion Pros (Taylor, Krane, & Thomas, 2005). That paper compared an RDD telephone survey with an opt-in Internet survey that asked respondents about topics such as: doing volunteer work, gambling, feeling sexually attracted to someone of the same sex, and being diagnosed with depression. The percentage of opt-in Internet survey respondents reporting socially undesirable behaviors and beliefs was higher than in the telephone survey.

Such findings demonstrate that computer-administered questionnaires elicited more reports of socially undesirable behaviors and beliefs than did live interviewers. But none provides direct evidence that computer administration of questionnaires yielded more accurate measurements. It could be that the people who choose to participate in the Internet surveys genuinely possessed socially undesirable attributes more than people who participate in RDD telephone surveys, so the enhanced reporting of such characteristics in the former may not be evidence of greater honesty in Internet surveys. Or computer self-completion of questionnaires could lead to more accidental misreading and mistyping, yielding inaccurate reports of socially undesirable attributes. A more direct test is therefore required to demonstrate that higher rates of reporting socially undesirable attributes in Internet surveys is due to increased accuracy.

Taylor, Krane, and Thomas (2005) reported just such a test, but their results disconfirmed the social desirability hypothesis. The test was based on the notion that the more undesirable a belief or behavior is, the more respondents will hide it during a telephone interview. So Taylor, Krane, and Thomas (2005) asked a separate group of respondents to rate the social desirability of each behavior and belief. The social desirability ratings explained only 4% of the variance in the differences between reports of the beliefs and behaviors in the telephone and Internet surveys, and explained only 8% of the ratios of the two surveys' estimates (we computed these R^2s using figures from Taylor et al.'s [2005] Table 3). Thus, social desirability did very poorly at explaining the

differences between the telephone and Internet surveys and therefore seems unlikely to account for those differences. In fact, the authors agreed with this conclusion: "an item's rating along the scale of 'good' to 'bad' … was not particularly informative concerning the size of the [mode] difference."

More likely, some other difference(s) between the telephone and Internet surveys explain the differences in answers. For example, perhaps the people who agreed to participate in the opt-in Harris Interactive Internet survey genuinely possessed the studied undesirable attributes at higher rates than did the respondents in the RDD telephone survey. Therefore, the evidence reported by Taylor et al. (2005) should not be viewed as evidence of greater honesty in Internet surveys than in telephone surveys.

However, several studies have produced more direct evidence showing that Internet surveys can acquire reports that are less distorted by social desirability pressures than telephone surveys. For example, Holbrook and Krosnick (in press) compared direct reports of voter turnout in the 2004 national election with measurements of turnout acquired using the Item Count Technique (ICT, see e.g., Droitcour et al., 1991).[3] When implementing the latter, respondents report the number of behaviors on a list that they have performed. Holbrook and Krosnick (in press) gave some respondents (chosen randomly) a list that did not include "voted in the 2004 election", and other respondents were given the same list plus "voted in the 2004 election." By subtracting the average number of behaviors reported by the former group of respondents from that reported by the latter group, Holbrook and Krosnick (in press) gauged the percent of respondents who confidentially reported that they had voted.

Among respondents in an RDD telephone survey who were asked a direct question about turnout, 72% claimed to have voted, but only 52% of a different group said they did so when asked confidentially using the ICT, a highly significant difference. Thus, 20% of the telephone respondents may have intentionally claimed to have voted when they knew they did not. The ICT approach in the RDD survey closely matched the federal government's reported rate of turnout in that election.

Social desirability-driven intentional over-reporting did not appear in a series of probability and non-probability sample Internet surveys (Holbrook & Krosnick, in press). Respondents who were asked direct questions about turnout answered affirmatively at the same rate as did respondents who reported turnout via the ICT. Thus, Internet turnout reports appear not to have been distorted by social desirability bias.

Such evidence supports the conclusion that Internet surveys *can* elicit more honest reports of sensitive matters. But the direct test reported by Taylor, Krane, and Thomas (2005) shows that more frequent reports of socially undesirable attributes in non-probability sample Internet surveys than in telephone surveys cannot always be attributed to superior accuracy of those surveys.

In fact, despite their apparent immunity to social desirability pressures when measuring turnout, non-probability sample Internet surveys are strikingly inaccurate when measuring turnout for other reasons, as we will explain below.

Where We Disagree

Mr. Taylor provided no direct evidence to support his assumption that admitting to smoking cigarettes some days or every day and admitting to drinking alcohol are socially embarrassing. In fact, a variety of studies show that this assumption is incorrect.

Our Evidence

First, if reports of smoking and drinking are distorted by social desirability pressures, and if Internet surveys are less subject to such pressures than telephone surveys, then we should have found more reports of such behaviors in the probability sample Internet survey we commissioned than in the RDD telephone survey we commissioned. Since both surveys used probability sampling, comparing them constitutes the cleanest test of the social desirability hypothesis.

In fact, the two probability sample surveys yielded essentially identical results for smoking and drinking. 77% of the telephone survey respondents said they were non-smokers, as compared to 75% of the probability sample Internet survey respondents. This is not a statistically significant difference. Similarly, 85% of the telephone survey respondents reported having consumed 12 drinks of alcohol in their lifetimes, as compared to 87% in the probability sample Internet survey. Again, this difference is not statistically significant. And the telephone survey found that 40% of respondents consumed one drink on days when they drank, the exact same percentage produced by the probability sample Internet survey.

Aguinis et al.'s Evidence

Dozens of past experiments also refute the assertion that reports of smoking and drinking by adults are biased by social desirability concerns. Aguinis, Pierce, and Quigley (1993, 1995) reported meta-analyses of these studies, which used a well-established method to assess intentional bias in reports of smoking and drinking: comparisons of direct self-reports with reports made using the Bogus Pipeline Technique.

This technique entails collecting a biochemical marker (e.g., saliva or exhaled air) from respondents and telling them that analysis of the marker will reveal whether they are a smoker or drinker. Collecting this marker before collecting direct self-reports of smoking and drinking enhances pressure on respondents to report honestly and does in fact increase reports of cigarette smoking among children (see Aguinis et al., 1993: 363). In their meta-analyses of many past studies, however, Aguinis et al. (1993, 1995) found that the Bogus Pipeline Technique did not increase reports of smoking or drinking among adults.

Evidence from the NHANES

Overview. To argue that adults' smoking self-reports in face-to-face surveys are biased due to social desirability, Mr. Taylor described findings from the National Health and Nutrition Examination Survey (NHANES), a face-to-face survey conducted regularly by the federal government. He said:

> "The NHANES study reported that 24.9 percent of adults said they were smokers but that blood tests showed that an additional 4.5 percent had smoked in the previous 24 hours but had not reported it when asked by an interviewer. The resulting NHANES estimate of 29 percent is closer to our estimate of 28 percent than … the RDD sample's 24 percent."

These numbers appear to have been taken from an article by Klein, Thomas, and Sutter (2007).

However, comparing the "corrected" NHANES figure of 29% to measurements from the telephone and Harris Interactive surveys that we commissioned is inappropriate for a variety of reasons:

- Data collected in one year were compared with data collected two to three years earlier.

- Self-reports of cigarette smoking were compared with analysis of blood samples collected weeks or months later.

- Self-reports of only cigarette smoking were compared with blood test results that reveal cigarette smoking plus all other nicotine consumption and more.

- Self-reports of cigarette smoking from one group of people were compared with blood samples drawn for only a subset of the self-reporters.

- Well-documented measurement error in blood test results reduced their validity.

We explain each of these next.

Comparing data collected in one year with data collected two to three years earlier. The Harris Interactive survey we commissioned was conducted in 2004, whereas Klein et al. (2007) reported figures from the 2001-2002 NHANES survey. Thus, to compare NHANES data to the surveys we commissioned in 2004, the 2003-2004 NHANES should have been used.

Comparing self-reports of cigarette smoking with analysis of blood samples collected weeks or months later. According to Klein et al. (2007), 24.9% of the 2001-2002 NHANES respondents reported that they smoked cigarettes every day or some

days. And Mr. Taylor said that blood tests from these respondents indicated that an additional 4.5 per cent had smoked during the previous 24 hours. But in fact, no blood tests were conducted to document smoking during the 24 hours prior to when those self-reports were made. Instead, the blood samples were taken from the survey respondents weeks or months after the self-reports were provided.

The self-reports that Klein et al. (2007) examined were collected during NHANES face-to-face interviews in respondents' homes. At the end of their face-to-face interviews, respondents were asked whether they would be willing to later visit NHANES's Mobile Examination Center (MEC) and participate in a physical exam. For most respondents, this visit occurred between 2 and 9 weeks later, at which time the blood samples were taken. Thus, the blood samples cannot be used to assess cigarette smoking during the 24 hours prior to the face-to-face interviews, because people's smoking behavior may have changed during the 2 to 9-week interim period.

Comparing self-reports of only cigarette smoking with blood test results that reveal cigarette smoking plus all other nicotine consumption and more. Although Mr. Taylor said that "4.5 percent had smoked in the previous 24 hours," in fact the blood test did not document only cigarette smoking. Instead, the blood test measurements of cotinine would have been elevated by pipe smoking, cigar smoking, consumption of chewing tobacco, snuff, nicotine gum, patches, or inhalers, or nicotine-containing medications (see Benowitz, 1988; Davis et al., 1991). Furthermore:

> "Some immunoassays overestimate cotinine concentrations because of cross-reactivity with other nicotine metabolites (Anderson, Proctor, & Husager, 1991; Schepers & Walk, 1988; Zuccaro et al., 1997)" (SRNT Subcommittee on Biochemical Verification 2002: 150).

According to the 2003-2004 NHANES household survey, 3.63% of people said they smoked cigars and/or smoked pipes and/or used dip or snuff and/or chewed tobacco, but said they did not smoke cigarettes. Thus, their cotinine levels could have been elevated as compared to their cigarette smoking self-reports for all of these reasons. If people who used cotinine-enhancing medications were added to this number, it would grow even larger.

Comparing self-reports of cigarette smoking from one group of people with blood samples drawn for only a subset of the self-reporters. Klein et al. (2007) compared the NHANES self-reports of smoking with the MEC blood test results, but different groups of people produced these two sets of measurements. 89% of the 5,033 respondents who were asked about smoking in the face-to-face in-home survey completed the MEC visit. Therefore, it would have been more appropriate to compare the blood test results from these 4,469 people with those same people's self-reports from the face-to-face interviews.

Measurement error in blood test results. Klein et al. (2007) assumed that people with nicotine levels of 15 ng/mL or greater were smokers, consistent with the recommendation by the SRNT Subcommittee on Biochemical Verification (2002). Yet

the SRNT subcommittee's report (2002) acknowledged that this cutoff does not avoid all measurement error and misclassifies some smokers as non-smokers and misclassifies some non-smokers as smokers (p. 151).[4] Furthermore, different investigators have used different cut points (e.g., 10 ng/mL by Klebanoff et al., 1998, and 20 ng/mL by Luepker et al. 1989), showing that there is some arbitrariness to the choice of cut point.

In fact, using a single cut point is not optimal, because the half-life of cotinine varies according to the quantity of nicotine the respondent has ingested and his or her metabolism. For example, many daily smokers manifest elevated cotinine levels up to seven days after quitting, whereas many light smokers manifest elevated cotinine levels for only two days after quitting (SRNT Subcommittee on Biochemical Verification 2002, p. 152). Also:

> "For some populations, such as African-Americans or pregnant women, nicotine and cotinine metabolism differ from the general population, and optimal cut-points are likely to differ as well (SRNT Subcommittee on Biochemical Verification 2002, p. 151)."

All this suggests that blood cotinine levels above and below 15 ng/mL do not constitute perfect indicators of nicotine exposure and non-exposure, respectively.

Summary. In sum, the evidence reported by Klein et al. (2007) and Mr. Taylor does not constitute a solid basis for claiming that some respondents' denials of cigarette smoking are intentional lies.[5]

A New Analysis of NHANES Data

To properly assess whether blood tests indicate that some cigarette smokers deny smoking when answering self-report questions, one must compare self-reports of nicotine consumption from all sources to cotinine levels collected from all the same people at the same time. Although some of the problems outlined above cannot be solved with the NHANES data, many of the problems are overcome in the new analysis we conducted.

During the face-to-face interview on the day when the blood samples were drawn, respondents were asked whether during the prior 5 days, they used "any product containing nicotine including cigarettes, pipes, cigars, chewing tobacco, snuff, nicotine patches, nicotine gum, or any other product containing nicotine." Of the respondents who gave a blood sample, 30.7% answered affirmatively.[6]

Although we might expect all of these people to have had cotinine levels of 15 ng/mL or greater, in fact, only 90.1% of them did. It's hard to imagine why someone who was not exposed to nicotine would intentionally claim otherwise, so measurement error in blood test results seem likely to have made them erroneous for 9.9% of these respondents. This is a high rate of false negative blood test results.

Furthermore, of the 69.4% of the MEC respondents who said that they did not use any nicotine product, 98.2% had cotinine levels below 15 ng/mL, meaning that their blood test results matched their self-reports. Thus, for only 1.8% of the respondents who denied being exposed to nicotine, the blood test suggested otherwise. These people constitute 1.3% of the sample of respondents who provided both a self-report and a blood test.

Although it is tempting to think of the 1.3% of people who denied nicotine exposure but had a positive blood test result as having lied, some may have had elevated cotinine levels due to heavy passive smoke exposure in their homes or their places of work or medications they took that enhanced their cotinine levels. Thus, it is impossible to be confident that 1.3% of respondents intentionally denied nicotine exposure due to social desirability pressures.

Conclusion about Social Desirability, Smoking, and Drinking

Taken together, the literature and analysis described above provides no solid evidence that reports of cigarette smoking and alcohol consumption in telephone or face-to-face surveys of adults are distorted downward by social desirability pressures. In fact, this body of evidence disconfirms that claim and suggests remarkable accuracy of self-reports in the face-to-face survey relative to the blood test.

Other Harris Interactive Survey Measurements of Smoking

In this light, it is interesting to note another number in the Klein et al. (2007) paper. Mr. Taylor said that:

> "The resulting NHANES estimate of 29 percent is closer to our estimate of 28 percent than … the RDD sample's 24 percent."

Although the Harris Interactive survey we commissioned did indeed yield an estimate of 28%, a Harris Interactive survey done at the same time with more than 50 times as many respondents (100,000) yielded a lower number, which exactly matched the result produced by the RDD survey we commissioned: 24% (Klein et al., 2007).

This disconfirms the claim that the Harris Interactive methodology always yields higher (and ostensibly more accurate) estimates of the proportion of smokers than RDD surveys do.

Bias in Reports of Voter Turnout

In his essay, Mr. Taylor offered further evidence to support his assertion that face-to-face surveys are contaminated by social desirability bias: "… in-person surveys by the Census Bureau report substantially more people claiming to have voted in elections than actually voted. If there is a better explanation than social desirability bias, I haven't heard it."

In fact, more than a dozen academic studies have yielded evidence supporting other explanations (e.g., Clausen, 1968; Greenwald, Carnot, Beach, & Young, 1987; Kraut & McConahay, 1973; McDonald & Popkin, 2001; Traugott, & Katosh, 1979; Yalch, 1976). For example, the people who choose to participate in surveys in fact vote at higher rates than people who choose not to participate in surveys (Clausen, 1968).

Moreover, in the case of the Current Population Surveys to which Mr. Taylor referred, more than half the turnout reports are proxy reports, meaning that one household member reports whether another household member voted or not. These proxy reports could be inaccurate because of lack of information about what housemates actually did, because proxies give housemates the benefit of the doubt, or because housemates deceive one another about whether they voted.

Most important, although the ICT test indicates that telephone surveys over-estimate turnout due to social desirability bias, opt-in Internet surveys over-estimate turnout at a much higher rate. For example, the turnout rate in the 2000 U.S. national elections was about 50%, but 88% of respondents in a Harris Interactive Internet survey claimed to have voted, whereas a simultaneous national face-to-face survey yielded a reported turnout rate of only 69% (Malhotra & Krosnick, 2007). Likewise, a national opt-in Internet survey in 2004 obtained a reported turnout rate of 95% for the election held that year, as compared to 80% of respondents in a simultaneous face-to-face national survey (Malhotra & Krosnick, 2007).

How could it be that face-to-face surveys yield more accurate measurements of turnout than do Harris Interactive surveys if the latter are immune to social desirability bias? A likely possibility is that the people who choose to participate in the Harris Interactive surveys genuinely vote at much higher rates than the population and report that accurately. This may be because opt-in Internet surveys tend to attract people who are especially interested in the topic of the survey, which will inflate turnout reports when the survey topic is politics (see Chang & Krosnick, 2009).

<u>The Two Benchmarks from Official Government Statistics</u>

Mr. Taylor said that for the two benchmarks we examined that were non-survey, official government statistics (on possession of passports and driver's licenses), the Harris Interactive results were "closer to the benchmark data than were the findings of the RDD telephone survey."

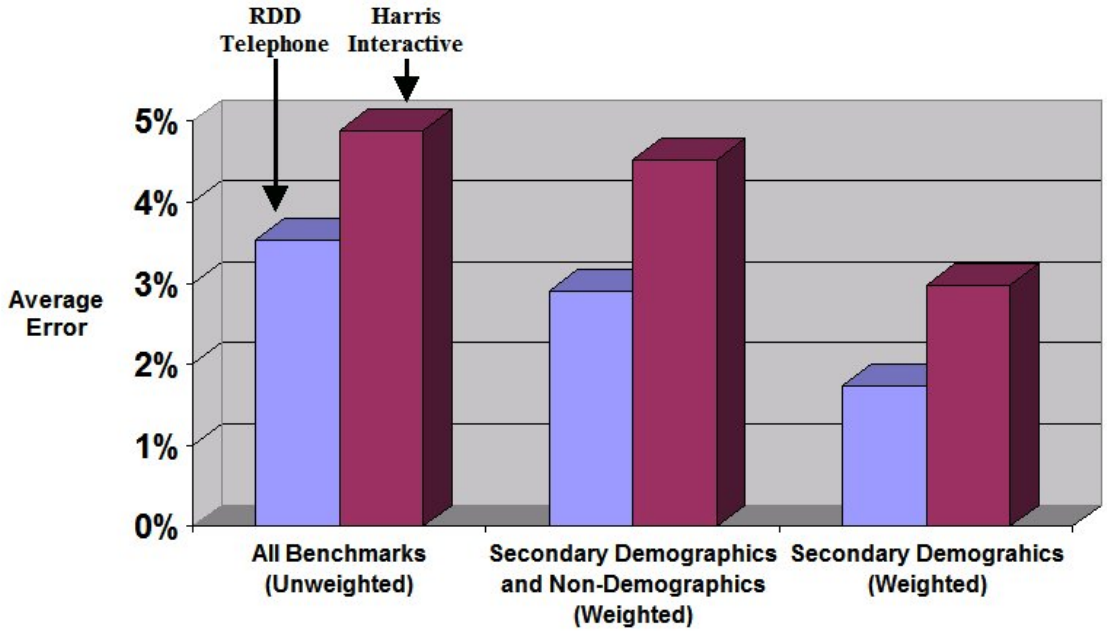
In fact, the results for these two items are: [7]

Variable	Benchmark	RDD Telephone Result	Harris Interactive Result
Have a passport	21.50%	30.50%	27.86%
Have a driver's license	89.00%	92.66%	92.10%

Thus, although the Harris Interactive passport figure is closer to the benchmark than is the RDD figure, the RDD and Harris Interactive numbers for driver's licenses are essentially equal.

The Real Accuracy of the Harris Interactive Survey

When discussing our paper's findings on the relative accuracy of the Harris Interactive survey and the RDD survey we commissioned, Mr. Taylor said that we examined only 6 benchmarks. In fact, we described results obtained using 19. Across our many comparisons, using many analytic methods and the full array of benchmarks or subsets of them, the Harris Interactive survey always manifested more error than did the RDD telephone survey (see the Figure below). Even when we omit the non-demographics (that Mr. Taylor says are contaminated by social desirability bias) and the primary demographics (that were used to compute the weights), the Harris Interactive data contained more error.



Furthermore, even after post-stratification to correct for demographic inaccuracies, the largest error we found in the RDD survey data was 9 percentage points, whereas the largest error we found in the Harris Interactive data was 15 percentage points.

Conclusion

Much evidence exists indicating that computers are great for survey data collection. In addition to helping interviewers do their work during face-to-face and telephone surveys, computers allow respondents to complete questionnaires via the Internet, apparently yielding very accurate and honest reports from the participating individuals. But not all sets of people who complete computer-based surveys are

representative. We have shown that *even honest and accurate reports from unrepresentative samples yield inaccurate survey results*. Therefore, assessing the accuracy of an Internet survey requires comparisons of results with trustworthy benchmarks.

The evidence reviewed above indicates that the benchmarks we used were not contaminated by social desirability response bias, and many ways of looking at the totality of our findings lead to the same conclusion: There is no basis for the claim that the Harris Interactive study we commissioned manifested superior accuracy. In fact, it was less accurate than the RDD survey.

References

Aguinis, Herman, Charles A. Pierce, and Brian M. Quigley. (1993). Conditions under which a bogus pipeline procedure enhances the validity of self-reported cigarette smoking: A meta-analytic review." Journal of Applied Social Psychology, 23, 352-373.

Aguinis, Herman, Charles A. Pierce, and Brian M. Quigley. (1995). Enhancing the validity of self-reported alcohol and marijuana consumption using a bogus pipeline procedure: A meta-analytic review. Basic and Applied Social Psychology, 16, 515-527.

Anderson, G., Proctor, C. J., & Husager, L. (1991). Comparison of the measurement of serum cotinine levels by gas chromatography and radioimmunoassay. Analyst, 116, 691–693.

Benowitz, N. L. (1988). The use of biologic fluid samples in assessing tobacco smoke consumption. In Grubowski, J., & Bell, C. S. (Eds.), Measurement in the analysis and treatment of smoking behavior. Washington DC: U.S. Department of Health and Human Services; 1983, 6-26. NIDA research monograph 48.

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. Public Opinion Quarterly.

Chang, L., & Krosnick, J. A. (in press). Comparing oral interviewing with self-administered computerized questionnaires: An experiment. Public Opinion Quarterly.

Clausen, A. (1968). Response Validity: Vote Report. Public Opinion Quarterly, 32, 588–606.

Davis, R. A., Stiles, M. F., de Bethizy, J. D., Reynolds, J. H. (1991). Dietary nicotine: a source of urinary cotinine. Food and Chemical Toxicology, 29, 821–827.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In P. B. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), Measurement Errors in Surveys, pp. 185–210. New York: Wiley.

Greenwald, A. G., Carnot, C. G. Beach, R., & Young, B. (1987). Increasing voting behavior by asking people if they expect to vote." Journal of Applied Psychology, 2, 15-318.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. Public Opinion Quarterly, 67, 79-125.

Holbrook, A. L., & Krosnick, J. A. (in press). Social desirability bias in voter turnout reports: Tests using the item count technique. Public Opinion Quarterly.

Klebanoff, M. A., Levine, R. I., Clements, J. D., DerSimonian, R., & Wilkins, D. G. (1998). Serum cotinine concentration and self-reported smoking during pregnancy. American Journal of Epidemiology, 148, 259-262.

Klein, J., Thomas, R. K., & Sutter, E. J. (2007). Self-reported smoking in online surveys: Prevalence estimate validity and item format effects. Medical Care, 45, 691-695.

Kraut, R. E., & McConahay, J. B. (1973). How being interviewed affects voting: An experiment. Public Opinion Quarterly, 37, 398–406.

Luepker, R. V., Pallonen, U. E., Murray, D. M., & Pirie, P. L. (1989). Validity of telephone surveys in assessing cigarette smoking in young adults. American Journal of Public Health., 78, 202-204.

Malhotra, N., & Krosnick, J. A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behaviors: Comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. Political Analysis, 15, 286-323.

McDonald, M. P., & Popkin, S. L. (2001). The myth of the vanishing voter. American Political Science Review, 95, 963–74.

Schepers, G., & Walk, R. (1988). Cotinine determination by immunoassays may be influenced by other nicotine metabolites. Archives of Toxicology, 62, 395–397.

SRNT Subcommittee on Biochemical Verification. (2002). Biochemical verification of tobacco use and cessation. Nicotine and Tobacco Research, 4, 149-159.

Taylor, H., Krane, D., & Thomas, R. K. (2005). Best foot forward: Social desirability in telephone vs. online surveys.
http://www.publicopinionpros.norc.org/from_field/2005/feb/taylor.asp

Traugott, M. W., & Katosh, J. P. (1979). Response validity in surveys of voting behavior. Public Opinion Quarterly, 43, 359–77.

Yalch, R. F. (1976). Pre-election interview effects on voter turnout." Public Opinion Quarterly, 40, 331–36.

Zuccaro, P., Pichini, S., Altieri, I., Rosa, M., Peliegrini, M., & Pacifici, R. (1997). Interference of nicotine metabolites in cotinine determination by RIA. Clinical Chemistry, 43,180–181.

Footnotes

1. Here is a brief summary of our paper: We commissioned firms to administer the same survey questionnaire via (1) RDD telephone interviewing with a probability sample of American adults, (2) Internet data collection from a probability sample of American adults, and (3) Internet data collection from samples of American adults who volunteered to do surveys for money or prizes and were not randomly sampled from the American adult population (we refer to the latter as "opt-in" samples). We compared the results of these surveys to benchmark measurements of the same phenomena from federal statistical reports and federal government surveys done face-to-face with extremely high response rates. Our principal findings include: (1) The probability sample surveys done by telephone or the Internet were consistently highly accurate. (2) The opt-in sample surveys done via the Internet were less accurate and were sometimes strikingly inaccurate.

2. Holbrook, Green, and Krosnick (2003) showed that this attitude is viewed as socially undesirable.

3. Holbrook, Green, and Krosnick (2003) showed that saying one voted in an election is viewed as socially desirable.

4. It is interesting to note that researchers have determined whether blood tests yield accurate measurements of nicotine consumption by comparing them with self-reports of nicotine consumption, which are treated as accurate benchmarks.

5. Two other aspects of Mr. Taylor's analysis are also suboptimal. First, although the Harris Interactive survey we commissioned collected data from adults age 18 and older, the NHANES survey that Mr. Taylor examined collected data only from adults age 20 and older, so they are not directly comparable.

Furthermore, Mr. Taylor chose to rely on a federal survey, the NHANES, that is less reliable than the federal benchmark survey we used, the National Health Interview Survey (NHIS). The NHANES 2003-2004 collected smoking self-reports from 5,033 adults in 30 primary sampling units (PSUs). By contrast, the 2004 NHIS collected smoking self-reports from 31,669 respondents in 678 PSUs. Not surprisingly, the confidence interval for the NHANES's smoking estimate (23.1% to 27.9%) was a great deal larger than that for the NHIS (20.7% to 21.6%).

6. All of the numbers in this section were produced after weighting the sample using the NHANES post-stratification and sample design (variable name: wtmec2yr).

7. In our paper, we did not identify the names of the survey companies who provided data to us. Mr. Taylor's memo, however, identified specific percentages in our paper as having come from the Harris Interactive survey we commissioned. Since readers can verify that the percentages he refers to are those we described as coming from "Non-probability sample Internet survey 1," readers now also know that all attributes reported

for “Non-probability sample Internet survey 1” are, in fact, those of Harris Interactive's survey.