Can we turn shirkers into workers?

Adam J. Berinsky [a,*], Michele F. Margolis [b], Michael W. Sances [c]

[a] Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E53-457, Cambridge, MA 02139, United States
[b] Department of Political Science, University of Pennsylvania, United States
[c] Department of Political Science, University of Memphis, United States

ARTICLE INFO

ABSTRACT

Survey researchers increasingly employ attention checks to identify inattentive respondents and reduce noise. Once inattentive respondents are identified, however, researchers must decide whether to drop such respondents, thus threatening external validity, or keep such respondents, thus threatening internal validity. In this article, we ask whether there is a third way: can inattentive respondents be induced to pay attention? Using three different strategies across three studies, we show that while such inducements increase attention check passage, they do not reduce noise in descriptive or experimental survey items. In addition, the inducements cause some respondents to drop out of the survey. These results have important implications for applied research. While scholars should continue to measure inattention via attention checks, increasing the attentiveness of "shirker" respondents is not as easy as previously thought.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

As self-administered survey instruments become more common in social science research, the quality of survey responses is increasingly important. To ensure that respondents pay attention to questions before answering them, researchers now frequently use Screeners to identify inattentive subjects (Berinsky, Margolis, & Sances, 2014; Meade & Craig, 2012; Nelson & Simmons, 2007; Oppenheimer, Meyvis, & Davidenko, 2009; Peer, Vosgerau, & Acquisti, 2014; for a discussion of other approaches, see Curran, in this issue). These questions are also sometimes referred to as Instructional Manipulation Checks or IMCs. Between 2006 and 2013, Berinsky et al. (2014) found 40 articles that employed screener questions, and dozens of additional papers have continued to use Screener questions to assess the quality of survey and experimental data since then.

Screener questions instruct respondents to demonstrate that they are paying attention to the stimulus, but disguise these instructions as a typical survey question. For instance, the first sentence of such a question may ask subjects about their favorite color; those who read the question closely, however, will notice that the final sentence of the question instructs respondents to choose a particular combination of colors to demonstrate that they are paying attention. By telling the respondent to ignore the rest of the question and choose an otherwise nonsensical answer, researchers can differentiate between attentive "worker" and inattentive "shirker" respondents.

This strategy is an effective way to identify inattentive respondents: numerous studies have found that passing a Screener is associated with considerable noise reduction, both in terms of experimental treatment effects (Berinsky et al., 2014; Oppenheimer et al., 2009) and construct validity (Berinsky et al., 2014). For example, Berinsky et al. (2014) find that subjects who pass screeners exhibit a strong preference for risky choices in the domain of losses in the context of Tversky and Kahneman's (1981) famous "Asian Disease Problem." Those who fail the Screener exhibit no such preference. Screeners are therefore powerful tools for identifying and classifying inattentive respondents, a task of growing importance with the rise of self-administered surveys conducted over the Internet (see Curran, in this issue, for further discussion).

After Screeners successfully identify inattentive respondents, however, researchers face a choice. They must either keep the inattentive respondents in the sample, thereby weakening the results by introducing noise, or exclude the inattentive respondents from the analysis, thereby decreasing sample size and altering its composition. The latter option, commonly implemented by researchers, threatens a study's validity, as "workers" and "shirkers" differ on a host of observable characteristics, such as age, education, and race (Berinsky et al., 2014). These systematic differences present a problem for both external and internal validity. First, by excluding inattentive respondents – who are disproportionately younger, less educated, and non-white – researchers lose one of the main benefits of collecting data online, namely the ability to collect a diverse sample. Second, if attention is correlated with characteristics that interact with the treatment, the culled sample could produce an artificially large (or small) treatment effect. For example, if highly educated people are more likely to pay attention and respond to an experimental

* Corresponding author.

treatment, then running the analysis on the attentive sample will produce inflated results.

Because neither including nor excluding inattentive workers is ideal, researchers might be tempted to pursue a third strategy: make otherwise inattentive respondents act as if they are attentive. Several studies have suggested different ways that inattentive respondents could be induced to pay attention to the survey. For example, researchers might prevent respondents from proceeding in the survey until they pass the Screener (Oppenheimer et al., 2009), or warn respondents at the beginning of the survey that their responses will be checked in order to encourage attentiveness (Berinsky et al., 2014; Clifford & Jerit, 2015). These strategies are meant to improve data quality without dropping inattentive respondents, thereby maintaining the sample composition—in essence turning "shirkers" into "workers". To date, however, there have been no systematic tests exploring the consequences of strategies aimed at increasing respondent attention. A host of unanswered questions remain. Do such strategies cause respondents to pay attention to the survey or do they, at best, simply make them better at answering Screeners? And will subjects react to these potentially obtrusive methods by exiting the survey?

In this article, we test four strategies meant to increase attention in self-administered online surveys—an increasingly common strategy for conducting research in psychology and related disciplines (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Chang & Krosnick, 2009; Couper, 2000; Paolacci, Chandler, & Ipeirotis, 2010). While some strategies we examine have been advanced by previous researchers, some are of our own design. Despite the use of multiple strategies and data sets, our results lead to a common conclusion: methods aimed at encouraging attention *do* improve passage rates on Screener questions on these surveys, but these increased levels of Screener passage *do not* carry over to increased attentiveness on the rest of the survey. Thus, while we can induce respondents to pass Screeners, the strategies do not, according to our measures, improve the overall data quality. In addition to illustrating the apparent lack of benefits, we also show the costs associated with attempting to induce attention. For example, we find that "training" respondents – preventing them from continuing the survey until passage – causes many respondents to drop out of survey, and makes it less likely that respondents will participate in a follow-up study.

These results have important implications for survey researchers. While scholars should continue to measure inattention via Screeners, increasing the attentiveness of "shirker" respondents has thus far proven difficult. The tradeoffs between quality data and representativeness that confront researchers using Screeners are not as easily avoided as previously thought. We therefore encourage researchers to use Screeners to identify inattentive respondents in their data, but to also be transparent in their analyses. This includes presenting empirical results for both attentive and inattentive respondents and considering how differences between the subsamples may affect the results. We offer more detailed suggestions for implementing and using Screeners in Section 6.

Beyond the outcomes of individual studies, our results may speak to broader concerns about the replicability of survey experiments. By adding noise to survey data, inattentive respondents could make it harder to detect real treatment effects. If attention varies across samples, an attempt to replicate a negative result from an inattentive sample may find a positive result on a more attentive sample.

More troubling, it could be the case that respondents pay less attention to surveys when taking them in the privacy of their own homes, as compared to a lab, monitored by a researcher. If indeed inattention is higher on self-administered online surveys than in the lab, attempts to replicate established results from psychology experiments will face an uphill battle. Some researchers have raised this concern when noting that Internet survey respondents may put less time and effort into an online survey compared to respondents in the laboratory (Williams, Cheung, & Choi, 2000) and that researchers are less able to monitor and control respondents in an online setting (Kraut et al., 2004). The lower level of attention online may even help to explain why certain results, initially found in laboratory studies, have been found not to replicate on online surveys (Crump, McDonnell, & Gureckis, 2013; Klein et al., 2014; see also Schweinsberg et al., this issue; Stroebe, in this issue). To avoid underestimating – or failing to detect entirely – effects that have previously been found in the lab, online replications must exercise care with regard to respondent attentiveness. This means not only measuring attentiveness, but also avoiding heroic attempts at manipulating it. As we show in this paper, such attempts may succeed in increasing passage on Screeners, but these correct responses do not represent increases in attention to other, more important sections of the survey.

2. Methods

Between June 2011 and October 2012, we conducted three self-administered studies using Internet samples. An overview of the details of the study is presented in Table 1.[1] Study 1 was collected in June–July 2011, by Survey Sampling International (SSI), an online panel company that produces samples which are typically more diverse than those collected through online convenience samples, such as Amazon's Mechanical Turk website.[2] Respondents were asked four Screener questions throughout the survey, with the presentation order randomized. These screeners were modeled on the IMCs employed by Oppenheimer et al. (2009). We followed the advice of those authors to make the screeners similar in format and theme to other questions asked on our broad, omnibus survey about decision making and politics.[3] One Screener asked respondents how they were currently feeling. This question read,

> Before we proceed, we have a question about how you're feeling.

> Recent research on decision making shows that choices are affected by context. Differences in how people feel, their previous knowledge and experience, and their environment can affect choices. To help us understand how people make decisions, we are interested in information about you. Specifically, we are interested in whether you

---

[1] For Studies 1 and 2, we employed a two-wave panel because we were interested in exploring the over-time stability of the screener questions (see Berinsky et al., 2014 for details). The main results presented in the paper come from the first wave of each study. However, we use both waves of the studies when testing whether the attempts aimed at increasing attention affected attrition between survey waves.

[2] SSI recruits participants through various online communities, social networks, and website ads. SSI makes efforts to recruit hard-to-reach groups, such as ethnic minorities and seniors. These potential participants are then screened and invited into the panel. When deploying a particular survey, SSI randomly selects panel participants for survey invitations. We did not employ quotas but asked SSI to recruit a target population that matched the (18 and over) census population on education, gender, age, geography, and income (based on premeasured profile characteristics of the respondents). The sample sizes of these studies (1200 in each study) were contracted in advance with SSI. We collected the number of cases specified in the contract and did not analyze any data until all 1200 respondents had taken part in the study.

[3] Specifically, we followed Oppenheimer et al.'s (2009) advice to include Screener questions that are similar in format to other questions on the survey. We do this by creating Screeners with multiple choice and ordinal response options, both of which are present throughout the survey, as well as by including questions that are similar in theme – in this case politics and decision making – to other questions on the survey.

Table 1
Study details.

Study	Sample	# Screeners	Strategies
1	Survey Sampling International	4	Training
2	Survey Sampling International	4	Training Warning
3	Mechanical Turk	2	Thanking Interests