Journal of Experimental Social Psychology 66 (2016) 20–28

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

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

Article history:
Received 4 September 2014
Revised 21 September 2015
Accepted 22 September 2015
Available online 24 March 2016

Keywords:
Attention checks
Surveys
Validity
Screener
Noise

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.

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, 2016-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, 2016-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

* Corresponding author.

http://dx.doi.org/10.1016/j.jesp.2015.09.010

people are more likely to pay attention and respond to an experimental 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.

Table 1
Study details.

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

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., 2016–this issue; Stroebe, 2016-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.

> actually take the time to read the directions; if not, some results may not tell us very much about decision making in the real world. To show that you have read the instructions, please ignore the question below about how you are feeling and instead check only the "none of the above" option as your answer. Thank you very much.
>
> Please check all words that describe how you are currently feeling.

A second Screener asked respondents about the news websites they visit. This question read,

> When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select ABC News and The Drudge Report as your two answers.
>
> When there is a big news story, which is the one news website would you visit first? (Please only choose one).

A third Screener asked respondents about their favorite color. This question read,

> We would like to get a sense of your general preferences.
>
> Most modern theories of decision making recognize that decisions do not take place in a vacuum. Individual preferences and knowledge, along with situational variables can greatly impact the decision process. To demonstrate that you've read this much, just go ahead and select both red and green among the alternatives below, no matter what your favorite color is. Yes, ignore the question below and select both of those options.
>
> What is your favorite color?

Finally, we embedded a fourth Screener in a question about the respondent's interest in politics and current events. This question read,

> People are very busy these days and many do not have time to follow what goes on in the government. Some do pay attention to politics but do not read questions carefully. To show that you've read this much, please ignore the question below and just hit the k key on your keyboard. That's right, just press the k key and ignore the choices below.
>
> How interested are you in information about what's going on in government and politics?

In this study, we built on Oppenheimer et al. (2009) who "trained" respondents to pay attention by not allowing participants who failed the IMC question at the beginning of the survey to continue until they passed the IMC. To directly test the effect of training respondents, we randomly assigned half the respondents to a training condition at the outset of each survey wave. Those in the training condition who failed the initial Screener received the following message: "There was a problem with your response. Please try again" and were re-asked the Screener – multiple times if necessary – until they passed. In addition to the initial Screener and training, we asked respondents the three remaining Screener questions at regular intervals throughout the survey. To measure noise reduction, we also replicated Tversky and Kahneman's (1981) "Asian Disease Problem" framing experiment and asked a battery of economic liberalism questions from the American National Election Study (ANES). Both will be described in detail in the sections below.

Study 2 took place in October 2012, using sample collected by SSI. The purpose of the study was to test multiple different strategies for inducing attention: training, warning, and thanking. In the training condition, respondents were given the same training task described in Study 1. Respondents in the warning condition were warned at the beginning of the survey that the researchers check responses carefully to make sure they read the instructions and responded carefully. The exact wording of this message is as follows:

> It is essential that you pay attention over the course of the survey. We will check each of your responses closely in order to make sure that you have read the instructions for the task and responded carefully. We will only accept your responses if you clearly demonstrate that you have read and understood the survey. Again, there will be questions that test whether you are reading the instructions.[4]

Respondents in the thanking condition received a gentler introduction which thanked respondents for their time and close attention. The thanking message read:

> Thank you very much for participating in our study. We hope that you will pay close attention to the questions on our survey. The thoughtful responses you provide to our questions are essential to advancing our scientific research, and all respondents can help us reach our goal. We would not be able to conduct our research without your thoughtful and careful answers to our questions. We know you are busy and greatly appreciate your time.

Finally, respondents in a control condition did not receive any of the interventions. All respondents answered the same four Screener questions, evenly spaced throughout the survey, as well as the same "Asian Disease" experiment and the ANES economic liberalism questions from Study 1.[5]

Study 3 took place in September 2012, using an online convenience sample, recruited through Amazon Mechanical Turk (Berinsky et al., 2012; Buhrmester et al., 2011; Paolacci et al., 2010).[6] The purpose of the study was to test whether we could induce attention by making the survey more interesting to respondents. At the outset of the survey, we randomly assigned respondents to an "interests and hobbies" condition. Respondents in this condition were asked at the beginning of the study: "We are interested in learning more about your hobbies and interests. Which one of the following categories interests you most?" The response options included: Sports, fashion, movies, cars, video games, music, and food. In the middle of the survey, respondents in the "interests and hobbies" condition answered a series of questions related to their previously stated interests directly before answering a Screener question. A description of the specific interests and hobbies questions is available in the Appendix.

Study 3 is designed to test the hypothesis that engaged respondents will be more attentive throughout the survey as a whole. If respondents were excited by the questions asked in the survey, they might pay closer attention to other questions on the survey as well. For this study, all respondents answered two Screener questions, one at the beginning of

[4] While we did not actually penalize respondents for inattention, respondents might have believed that low-quality responses would cost them the participation credit that they typically receive for answering SSI surveys. This treatment thus appeals both to respondents' desires to avoid looking careless to the researcher, as well as their desire to earn their participation credit.

[5] The three intervention conditions were also fully crossed with each other producing 7 treatment conditions: training, warning, thanking, training and warning, training and thanking, warning and thanking, training and warning and thanking. We present the results for the individual interventions against the global control condition in the paper, and the crossed results are available on request. The strategy is appropriate because we randomized the different treatments (thanking, warning, and training) in an orthogonal manner.

[6] We determined the sample size of Study 3 would be 1000 respondents before data collection began. We paid Mechanical Turk in advance for the 1000 respondents. We did not analyze any data until all 1000 respondents had taken part in the experiment.

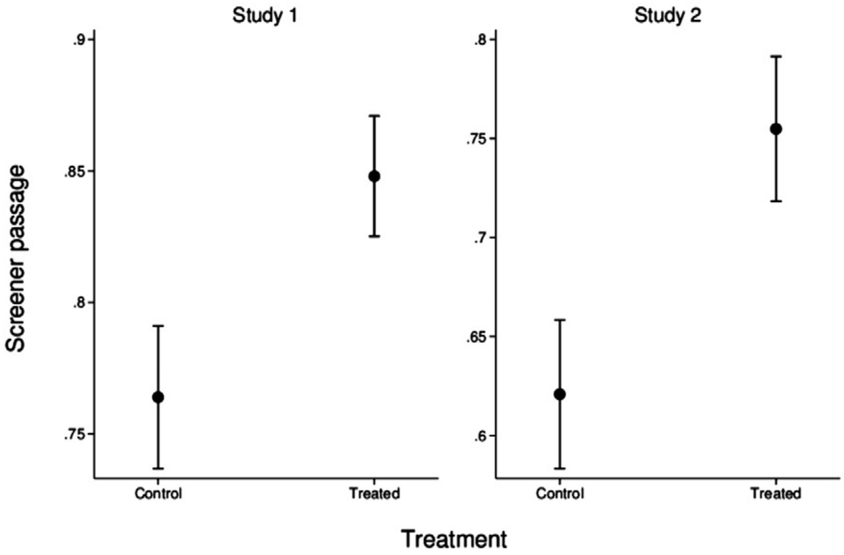

Fig. 1. Training increases Screener passage.

the study and one in the middle of the survey. The first Screener asked about news web sites, as in Studies 1 and 2. The second Screener asked about newspaper sections, and read:

> We are also interested in what sections people like to read in the newspaper. What people read in the paper might affect their opinions on current events. We also want to see if people are reading the questions carefully. To show that you've read this much, please mark both the classified and none of the above boxes below. That's right, just select these two options only.
>
> Regardless of how frequently you read the newspaper, what would you say are your favorite newspaper sections to read? (Please check all that apply).

3. Manipulations increase Screener passage

Knowing the potential benefits and drawbacks of using Screener questions, we tested the effectiveness of four manipulations designed to increase attention on surveys. We first test whether we can increase respondents' passage rates on Screener questions. Fig. 1 plots the total proportion of correct Screeners on the y-axis.[7] The x-axis identifies whether respondents were assigned to the training condition. In three separate trials, respondents assigned to the training condition performed better on subsequent Screeners than respondents assigned to the control condition. Looking at the left panel of Fig. 1, the total proportion of correct Screener responses is 0.76 among those in the control condition. Among respondents in the trained condition, however, the proportion of Screeners passed is to 0.85 ($t\,(1193) = 4.59$, $p < 0.01$), or an increase of ten percentage points over the base passage rate.

Other strategies have mixed success. In addition to training respondents, we also tried to increase attention in Study 2 by warning and thanking respondents at the outset of the survey. The left and middle panels of Fig. 2 show that both strategies marginally improved passage rates compared to respondents who went throughout the survey without an intervention to increase attention, with increases of between five and eight percentage points (Warning: $M = 0.69$ and $M = 0.61$, $t\,(784) = 2.46$, $p < 0.05$. Thanking: $M = 0.67$ and $M = 0.62$, respectively, $t\,(788) = 1.90$, $p < 0.10$). In Study 3, we also attempted to induce attention by making the survey more interesting. As shown in the right panel of Fig. 2, this strategy had a positive but insignificant and substantively small effect on passage, about three percentage points ($M = 0.8$ and $M = 0.77$, $t\,(1000) = 1.29$, $p = 0.20$).

Across these studies, we find evidence that researchers can induce attention by training respondents who initially fail a Screener, warning respondents that the researchers pay close attention to the responses, and thanking respondents for their time and thoughtful responses. Of the three strategies that actually affected passage, training appears to produce the strongest effects.

4. Manipulations do not reduce noise

While increasing passage appears to be a positive result, the intent of Screeners is to reduce noise on other questions. Prior research has found that Screeners do serve this purpose: when researchers drop respondents who fail Screeners, measurement error is reduced and experimental effects become larger and more precise (Berinsky et al., 2014; Oppenheimer et al., 2009). We replicate these results in the Appendix, showing that in our studies, Screener passage is associated with large reductions in measurement error. Indeed, the larger purpose of these manipulations is to induce a general increase in attentiveness across the survey; those who propose training as a strategy also cite evidence that training respondents produces less noisy results (Oppenheimer et al., 2009).

To test whether our interventions had an impact on attention to non-Screener items, we replicated a variety of decision-making paradigms. Most notable among these was the well-known framing experiment of Tversky and Kahneman (1981). We chose this experiment because it has been replicated many times using many different types of samples, including student samples, such as those originally employed by Tversky and Kahneman, as well as national probability samples (Kam & Simas, 2010) and samples collected on Mechanical Turk (Berinsky et al., 2012). In this experiment, all respondents are initially given the following scenario:

> Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative

[7] Studies 1 and 2 use three Screener questions to create the proportions presented in Fig. 1. We discard the initial Screener from this analysis, as it is impossible for the training to affect this question.

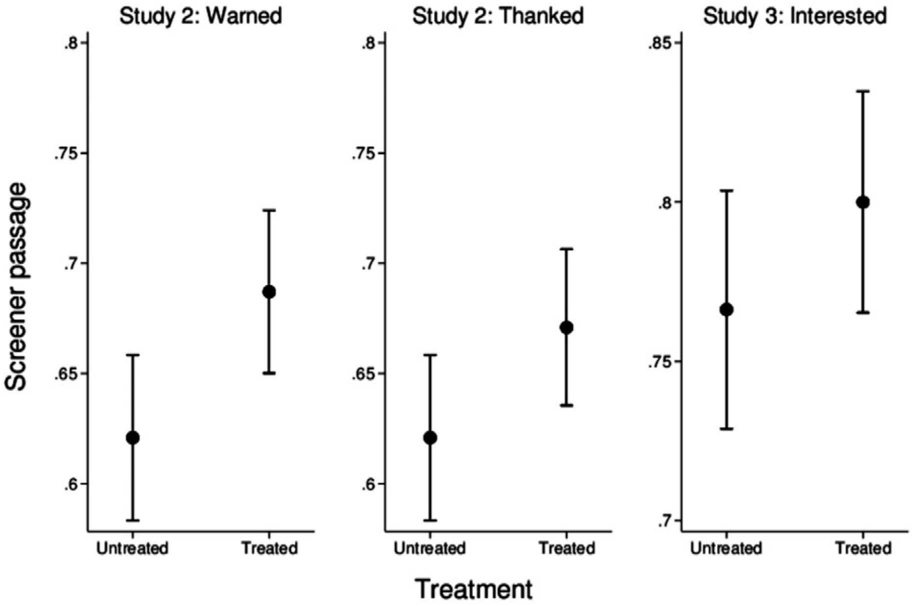

Fig. 2. Other interventions increase Screener passage.

programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

Subjects are then randomly assigned to one of the two following conditions:

> Condition 1, Lives Saved Frame: "If Program A is adopted, 200 people will be saved. If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved."

> Condition 2, Mortality Frame: "If Program A is adopted, 400 people will die. If Program B is adopted there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die."

The scenarios in both conditions are the same, but the conditions differ in their framing of the alternatives. Tversky and Kahneman (1981) report that when the problem was framed in terms of "lives saved," respondents were more likely to pick the certain choice. When it was framed in terms of lives lost, as in the "mortality frame," respondents were more likely to pick the risky choice. We expect that the difference between these two conditions will be greatest for those paying the closest attention to the survey and we assess whether training respondents increases the size of this difference.

We further test whether training respondents improves the quality of nonexperimental data when question wordings require close reading. For the last four decades, the American National Election Studies have asked a series of three questions on economic liberalism. As an example, one of the questions asks about the trade-off between government spending and services:

> Some people think the government should provide fewer services, even in areas such as healthcare and education, in order to reduce spending. Suppose these people are on one end of the scale, at point 1. Other people feel that it is important for the government to provide many more services even if it means an increase in spending. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between. Where do you place YOURSELF on this scale?

While these three questions tap into the same underlying set of beliefs – support for social welfare programs – the response options differ in subtle ways. For two of the questions, a low response (1) represents a liberal position while a higher response indicates a conservative position.[8] On the third question, displayed above, the scale is reversed; the highest response (7) is a liberal position, and the lowest response (1) is a conservative position. This case is an ideal setting for exploring the impact of inattentiveness because the question-wording effect is so clear. By comparing the correlation of the questions with reversed scales, researchers can detect satisficing behavior, which is the act of selecting the first minimally acceptable alternative that comes to mind rather than attempting to find an optimal solution to a problem (Krosnick, 1991). All variables have been recoded so that higher numbers indicate more conservative responses.

Fig. 3 shows that although our training increased Screener passage rates, contrary to the findings of Oppenheimer, Meyvis, and Davidenko, we do not find any difference in the quality of our experimental or non-experimental data. In the top row of the figure, we plot the experimental treatment effects, calculated by taking the difference in means between being assigned to the "mortality frame" (1) rather than the "save frame" (0) and choosing the probabilistic outcome (1) rather than the certain outcome (0). The same results appear in both conditions and across both samples: when confronted with a situation in which people might die, the likelihood of choosing a probabilistic option is much higher than when the same scenario is discussed with the

[8] The first economic liberalism question with this response scale reads: "Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising taxes of wealthy families or by giving income assistance to the poor. Suppose these people are on one end of the scale, at point 1. Others think that the government should not concern itself with reducing this income difference between the rich and the poor. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between. Where would you place yourself on this scale?" The second economic liberalism question with this response scale reads: "Some people feel that the government in Washington should see to it that every person has a job and a good standard or living. Suppose these people are at one end of a scale, at point 1. Others think the government should just let each person get ahead on his/their own. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between. Where would you place yourself on this scale?"

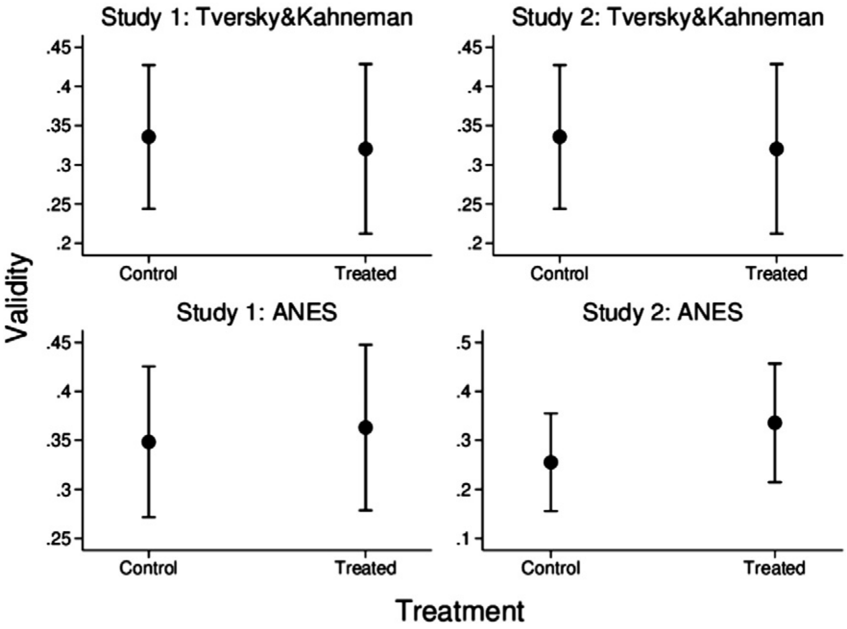

Fig. 3. Training does not reduce noise.

potential for lives being saved. Importantly, however, the training interventions did not increase the magnitude of these experimental effects. The top left panel of Fig. 3 shows that the experiment effect was identical – 0.23 – for respondents in both the training and control conditions (t (738) = 0.02, p = 0.987). The top right panel finds similarly null effects in Study 2 (difference = −0.2, t (708) = 0.21, p = 0.832.)

The second row of Fig. 3 plots the relationship between the two same-scale ANES questions and the third ANES question with the reversed scale.[9] In each of the different conditions, the correlations between the two same-scale questions range from 0.52 to 0.59. These strong associations are unsurprising given that the questions tap into the same underlying preference. If respondents read the third question carefully, we would expect a similarly strong relationship. Indeed, previous work has found that these correlations are higher among those who pass Screeners, and we replicate this result in the Appendix.

In contrast, attempts to increase attention do not increase these inter-item relationships. In the various control conditions, the relationship between the reverse item and the averaged same-item scale ranged between 0.25 and 0.35, which is far below the same-item relationship. Moreover, training does not increase the association. In the bottom left panel of Fig. 3, we show that the correlation between the reverse item and averaged same-item scale is essentially the same in both conditions: 0.36 in the training condition and 0.35 in the control condition (t (1247) = 0.25, p = 0.802). We did find a slight increase in the correlation in Study 2, but the effect was small and statistically no different from zero (effect = 0.34 and effect = 0.26, t (771) = 1.01, p = 0.313.)

Fig. 4 replicates the results for the warning and thanking conditions from Study 2. As with the training condition, neither of these interventions changed the Tversky and Kahneman experiment treatment effect. (Warning: effect = 0.36 and effect = 0.34, t (813) = 0.31, p = 0.754. Thanking: effect = 0.33 and effect = 0.34, t (833) = 0.09, p = 0.931) or the ANES reverse-item correlation (Warning: effect = 0.32 and effect = 0.26, t (885) = 0.89, p = 0.375. Thanking: effect = 0.26 and effect = 0.26, t (902) = 0.13, p = 0.897).

It should be noted that the other decision tasks we employed yielded results similar to those found here. The full results are presented in the Appendix, but are easily summarized here. In particular, the training did not increase the size of the treatment effect in the "sunk cost" experiment (Thaler, 1985) utilized in Oppenheimer et al. (2009) or the welfare question wording experiment (Rasinski, 1989; Smith, 1987), nor did it increase the successful completion of the "bat and ball" cognitive processing task (Kahneman & Frederick, 2005) or the time spent reading survey questions (Huang, Curran, Keeney, Poposki, & DeShon, 2012).

In sum, although we were able to successfully train and encourage respondents to pass Screener questions, these strategies do not improve the quality of our data; we were unable to turn shirkers into workers. We note this is a surprising result, given that Screener passage in the absence of these manipulations is strongly associated with increased validity. We discuss possible explanations in the concluding section.

5. The costs of training

In addition to measuring the potential benefits of interventions to induce attention, we also assess costs in terms of survey attrition. The results presented above for Studies 1 and 2 both came from the first wave of two-wave panel studies. In this section we move beyond the first wave analysis to explore the second wave results. Specifically, we can see if being trained in the first wave of the study affects the likelihood of returning for the second wave of the study. In both studies 1 and 2, training respondents produced higher levels of respondent drop-out rates within a given wave. We show these results in Fig. 5. In Study 1, 73% of respondents assigned to the training condition at the survey's outset completed the survey, while 83% of respondents who were not assigned to the training condition completed the survey, a decrease of ten percentage points (t (1541) = 4.58, p < 0.01).[10] The higher attrition rate may be because respondents think that the survey has malfunctioned once they reach the training component of the study, or that respondents do not like being corrected. In Study 2, drop-out rates increased by fourteen points with training (M = 0.49 and M = 0.63, t (1175) = 4.86, p < 0.01). In contrast, warning and thanking respondents actually increased respondents' likelihood of continuing with the survey by between six and eight points. 69% in warning condition completed the survey, compared to 63% of those not warned, (t

[9] We construct this measure by regressing the average of the two same-scale items on the reverse-scale item, then taking the coefficient on the reverse-scale item.

[10] When comparing respondents who failed the initial Screener, 52% in the training condition went on to complete the survey while 77% in the control condition did so (t (618) = 6.55, p < 0.01).

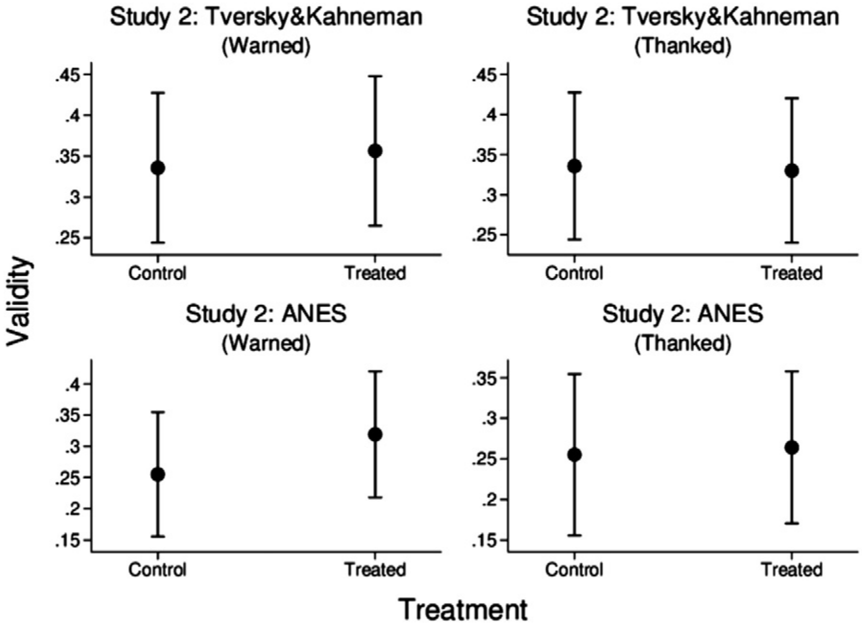

Fig. 4. Other interventions do not reduce noise.

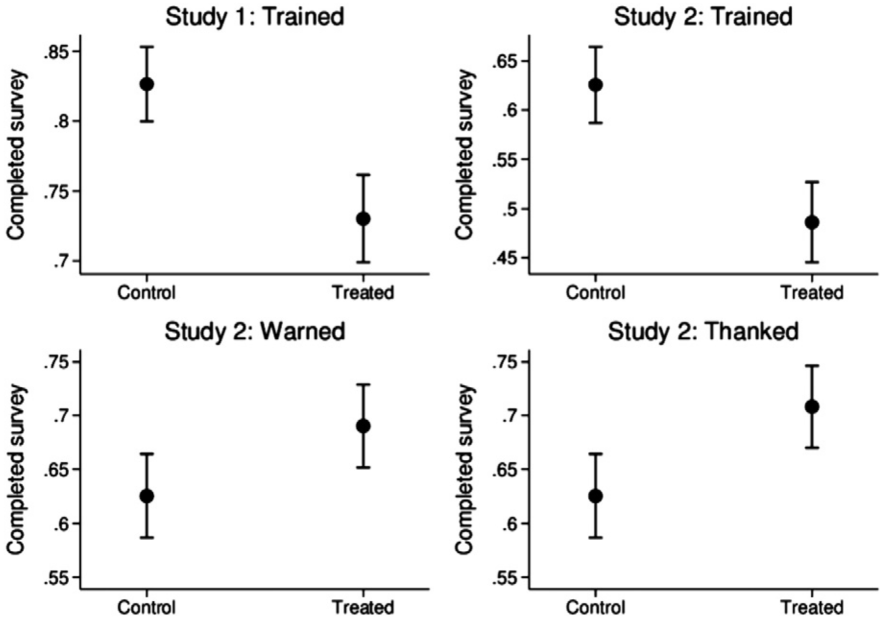

Fig. 5. The costs of Screener conversion attempts: survey completion.

$(1151) = 2.31$, $p < 0.05$), while 71% of respondents in the thanking condition completed the study compared to 63% who were not ($t\,(1144) = 2.97$, $p < 0.01$).

We also find that training leads to higher attrition between survey waves in our multi-wave studies. These results are presented in Fig. 6. Fifty-six percent of respondents assigned to the control condition in the first wave of Study 1 took part in the second wave of the study two weeks later, while only 46% of respondents assigned to the training condition did so, a decrease of ten points ($t\,(1537) = 3.92$, $p < 0.01$).[11] In Study 2, this effect was fifteen points ($M = 0.31$ and $M = 0.46$, $t\,(1034) = 5.02$, $p < 0.01$). In contrast, there are no differences between the thanking and warning conditions on continuation into the second wave (Thanking: $M = 0.50$ and $M = 0.46$, $t\,(1030) = 0.97$, $p = 0.33$. Warning: $M = 0.46$ and $M = 0.46$, $t\,(1052)$, $p = 1.00$).

6. Discussion

Screeners are powerful tools for measuring attention, but can researchers effectively convert inattentive “shirker” respondents into attentive “workers”? Though being able to identify inattentive respondents marks an improvement in survey and experimental research, it would be even better to avoid difficult tradeoffs between high-quality data and an unrepresentative sample. Using several different surveys and interventions for online surveys, we found consistent results. We

[11] When comparing respondents who failed the initial Screener, 35% in the training condition who completed the first survey took part in the second wave, while 40% in the control condition did so ($t\,(615) = 3.66$, $p < 0.01$).

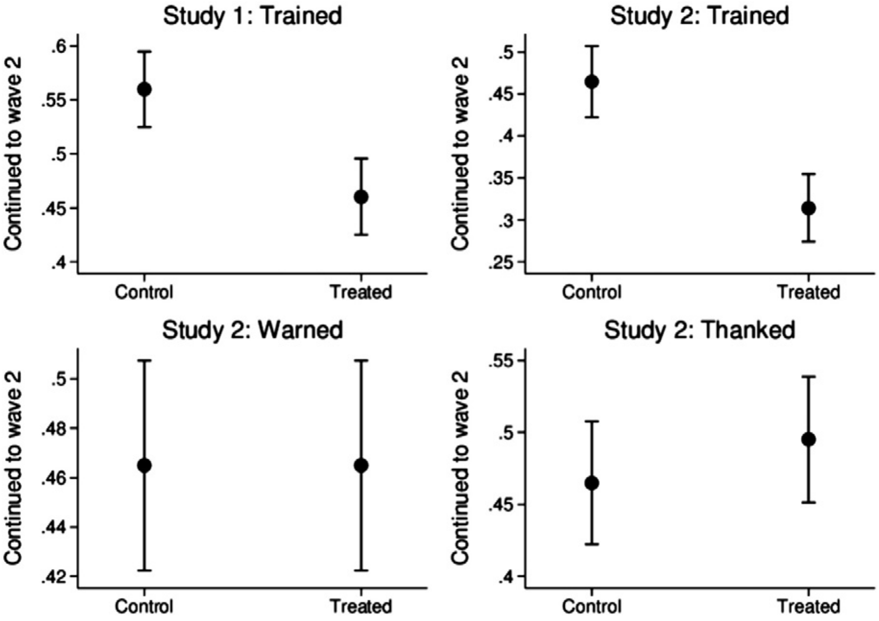

Fig. 6. The costs of Screener conversion attempts: panel attrition.

are able to increase Screener passage rates, but these increases did not translate into higher-quality data. Put another way, while we were able to make shirker respondents look like workers in a superficial sense, these behaviors did not carry over into increased attention on substantive questions we care about. Even worse, we found that one particular intervention – training – reduces survey completion and increases panel attrition.

Most notably, we were unable to replicate Oppenheimer et al.'s (2009) findings that respondents who fail a Screener can be trained to pay attention. The differences in our samples, experimental design, and mode of data collection may contribute to the different results. Although all the studies were self-administered, Oppenheimer, Meyvis, and Davidenko ran their study on college students in a lab. In contrast, our studies were done on diverse national samples over the Internet. The difference in results may indicate that training respondents is a useful strategy in the lab – where respondents can ask for assistance – while the same strategy over the Internet – where there is no oversight – proves ineffective. In addition to demographic and monitoring differences, our online respondents may simply be more committed to expending minimal effort. Indeed, one possibility is that our interventions taught inattentive subjects to differentiate Screeners from other questions. After receiving the intervention, subjects would then be able to guess which particular questions they needed to pay close attention to.[12]

We can confidently rule out one potential explanation for this result: that subjects collected through Amazon's Mechanical Turk may be experienced survey takers (Chandler, Mueller, & Paolacci, 2014) who are familiar with Screener questions. Recall that we used subjects from Mechanical Turk for only one of our three studies—Study 3, which contained the “interests and hobbies” manipulation and was most dissimilar to the treatments used by Oppenheimer et al. (2009). For the other two studies – which were most similar in spirit to the Oppenheimer, Meyvis, and Davidenko study – we used subjects recruited by SSI, who are generally unaware of the Screener questions and are properly considered naïve subjects.[13] Thus, we believe that our failure to induce training effects is properly seen as a general issue facing researchers who seek to conduct online research.

Despite it being difficult for researchers to induce attention, Screeners are still useful tools for measuring attention on surveys. Based on our previous research (Berinsky et al., 2014), we offer three points of practical guidance for scholars interested in using Screeners in their own work. First, an accurate measure of attention requires more than one Screener question. Screener questions, like other survey items, measure its underlying construct with error. As such, a scale of attentiveness is preferable to a single measure. Additionally, varying the degree of “difficulty” of the different items – that is, including Screeners that have both high and low passage rates – will offer meaningful variance in the scale. Second, researchers should be aware that respondents who pass and fail Screeners may differ on a host of observable characteristics, such as age, education, and race. As such, researchers should not simply discard inattentive respondents, but rather present results stratified by attention levels. Third and relatedly, researchers should analyze predictors of Screener passage and consider how certain demographics being over or underrepresented in the attentive sample may change how we interpret the results. If correlates of Screener passage—such as race, gender, age, and education—are also related to experimental treatments or issue preferences, then the resultant attentive sample may be skewed.

The Internet has become a boon to researchers interested in collecting individual-level data. Rather than relying on a student or otherwise unrepresentative laboratory sample, researchers may now collect a large, diverse sample relatively cheaply (Berinsky et al., 2012; Buhrmester et al., 2011). But online surveys also heighten concerns about data quality. The anonymous Internet, which offers no participant oversight, makes it very difficult to control whether respondents pay attention and give their best effort when taking online surveys. This potentially higher level of inattention means that studies conducted online may have difficulty replicating, and may explain why certain lab-based results have failed to replicate using this new survey paradigm. Screeners enable researchers to measure attentiveness, but for now we must advise against efforts at correcting the inattentiveness problem. Despite numerous attempts, employing several strategies, we were unable to manipulate general attention. Our failure to convert shirkers in this manner strongly suggests that attentiveness to surveys is outside the realm of factors that researchers can control.

[12] See Hüffmeier, Mazei, and Schultze 2016-(in this issue) for guidance on creating additional replication studies that will allow researchers to more fully understand the relationship between sample, survey mode, and survey attention.

[13] It is also possible that the lack of training was due to the relatively low compensation provided to subjects on Internet surveys. In high-paying lab experiments, perhaps respondents could be better induced to pay attention to the experimental materials.

Acknowledgments

For valuable comments, we thank Evan Apfelbaum, Edith Chen, Greg Miller, and Todd Rogers. We also thank Daniel Guenther and Andres Cambronero Sanchez for research assistance. For funding, we thank the MIT Political Experiments Research Lab.

Appendix A Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2015.09.010.

References

Berinsky, A.J., Huber, G.A., & Lenz, G.S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368.

Berinsky, A.J., Margolis, M.F., & Sances, M.W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*, 739–753.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*, 112–130.

Chang, L., & Krosnick, J.A. (2009). National surveys via RDD telephone interviewing versus the Internet. Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*, 641–678.

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*http://dx.doi.org/10.1093/poq/nfv027.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464–494.

Crump, M.J., McDonnell, J.V., & Gureckis, T.M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*, 1–18.

Curran, P.G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19 (in this issue).

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., & DeShon, R.P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *72*, 81–92 (this issue).

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K.J. Holyoak, & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). New York, NY: Cambridge University Press.

Kam, C.D., & Simas, E.N. (2010). Risk orientations and policy frames. *The Journal of Politics*, *72*, 381–396.

Klein, R.A., Ratliff, K.A., Vianello, M., Admas, R.B., Bahnik, S., Bernstein, M.J., ... Nosek, B.A. (2014). Investigating variation in replicability: A 'many labs' replication project. *Social Psychology*, *45*(3), 142–152.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *Psychological Research Online*, *59*, 105–117.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236.

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.

Nelson, L.D., & Simmons, J.P. (2007). Moniker maladies: When names sabotage success. *Psychological Science*, *18*, 1106–1112.

Oppenheimer, D.M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872.

Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023–1031.

Rasinski, K.A. (1989). The effect of question wording on public support for government spending. *Public Opinion Quarterly*, *53*, 388–394.

Schweinsberg, M., et al. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67 (in this issue).

Smith, T.W. (1987). That which we call welfare by any other name would smell sweeter. An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, *51*, 75–83.

Stroebe, W. (2016). Are most published social psychological findings false? *Journal of Experimental Social Psychology*, *66*, 134–144 (in this issue).

Thaler, R.H. (1985). Mental accounting and consumer choice. *Marketing Science*, *27*, 15–25.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Williams, K.D., Cheung, C.K., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology*, *79*, 748–762.