Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys

Adam J. Berinsky Massachusetts Institute of Technology
Michele F. Margolis Massachusetts Institute of Technology
Michael W. Sances Massachusetts Institute of Technology

Good survey and experimental research requires subjects to pay attention to questions and treatments, but many subjects do not. In this article, we discuss "Screeners" as a potential solution to this problem. We first demonstrate Screeners' power to reveal inattentive respondents and reduce noise. We then examine important but understudied questions about Screeners. We show that using a single Screener is not the most effective way to improve data quality. Instead, we recommend using multiple items to measure attention. We also show that Screener passage correlates with politically relevant characteristics, which limits the generalizability of studies that exclude failers. We conclude that attention is best measured using multiple Screener questions and that studies using Screeners can balance the goals of internal and external validity by presenting results conditional on different levels of attention.

Good survey and experimental research requires subjects to pay attention to questions and treatments, but not all people pay close attention all of the time. When respondents do not read questions carefully, their responses on related survey items can appear to be unrelated; when subjects do not pay attention to experimental treatments, replications of classic experiments can produce null results. As self-administered surveys—both online and in the lab—continue to grow in popularity, problems arising from inattentive respondents will also grow. Researchers must consider how best to identify and handle inattentive respondents.

Instructional Manipulation Checks (IMCs), or "Screeners," are a potential solution to this problem and are increasingly common in political science and psychology (Oppenheimer, Meyvis, and Davidenko 2009).[1] Screeners work by instructing subjects to demonstrate that they are paying attention by following a precise set of instructions when choosing a survey response option. In Figure 1, we present an example of a Screener. The first sentence of the question suggests that respondents are being asked about news consumption habits: what news source do they turn to when a big story breaks? However, if subjects continue reading, they will notice that they are actually being asked to demonstrate they are paying attention by selecting both "ABC News" and "The Drudge Report" as their responses rather than answering truthfully. By recording who responds with the specified answers, we can identify those respondents who are paying attention at a specific point during the survey. As we will discuss below, a great number of people—between a third and a half of our respondents from

Adam J. Berinsky is Professor of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E53–457, Cambridge, MA 02139 (berinsky@mit.edu). Michele F. Margolis is a Ph.D. candidate in Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E53–470, Cambridge, MA 02139 (margolis@mit.edu). Michael W. Sances is a Ph.D. candidate in Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E53–470, Cambridge, MA 02139 (mwsances@mit.edu).

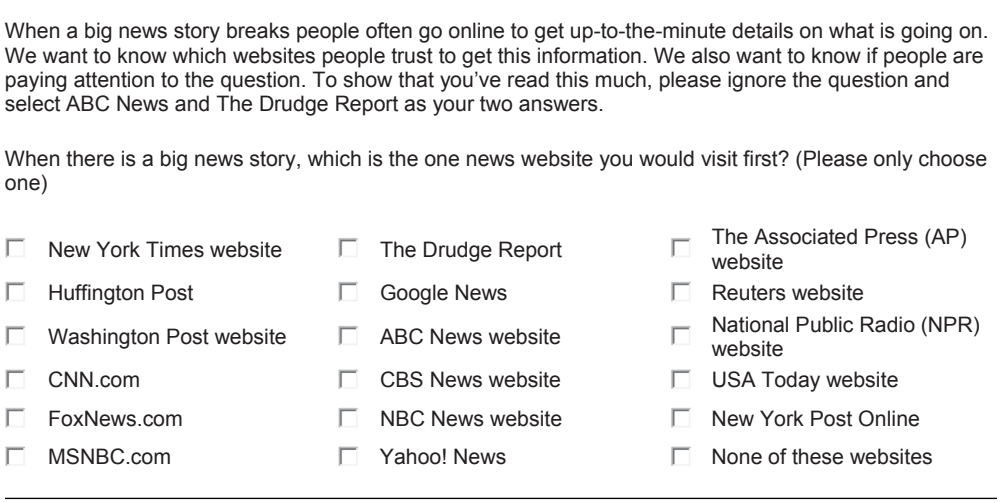
We thank Devin Caughey, James Dunham, Dan Hopkins, Greg Huber, Jennifer Jerrit, Cindy Kam, Krista Loose, Neil Malhotra, Tali Mendelberg, Spencer Piston, Michael Tesler, Chris Warshaw, Tess Wise, and seminar participants at the 2012 Midwest Political Science Association meeting, the 2012 NYU Center for Experimental Social Science conference, and the Fall 2011 MIT Experimental Working Group for helpful comments on previous drafts. We also thank Seth Dickinson and Hari Ramesh for valuable research assistance. Funding for this project came in part from the MIT Political Experiments Research Lab (PERL). Replication data for this study will be made available at http://thedata.harvard.edu/dvn/dv/ajps.

[1] As of July 2013, we found 40 published or forthcoming articles that employ Screener questions since 2006. Details about these studies are available in Table S1 of the online supporting information.

American Journal of Political Science, Vol. 00, No. 0, xxxx 2013, Pp. 1–15

 DOI: 10.1111/ajps.12081

FIGURE 1 An Example of a Screener Question



national samples—fail to properly answer these questions.

While Screeners may be powerful tools for filtering out people who are not paying attention, applied researchers concerned about inattentive respondents have few guidelines for their use. There is little basic research on the measurement properties of these items (Oppenheimer, Meyvis, and Davidenko 2009). Moreover, the ways that Screeners are used in practice have not been evaluated. To date, most researchers using Screeners simply exclude inattentive respondents, often measured from a single Screener, from their analysis.[2]

In this article, we provide guidance for the use of Screeners on political science surveys, drawing on a series of studies conducted through Internet samples of online survey panelists. While all our data come from Internet surveys, our results are applicable to any self-administered survey, such as laboratory studies in psychology and political science. We first demonstrate the benefits of Screeners, showing how they separate attentive "worker" respondents from inattentive "shirker" respondents, by replicating some well-documented survey and experimental findings in political science.

After presenting evidence about the benefits of Screeners, we examine important but understudied questions about Screeners that relate to their proper use. We first show that Screeners do not induce social desirability bias or otherwise affect subjects in undesirable or harmful ways. We next explore how best to measure attentiveness using different Screener questions. We show that while Screeners reduce noise, they are survey questions like any other. As survey items, Screeners are not immune to measurement error. We therefore show that using a single Screener is not the most effective way to parse workers from shirkers. Instead, as with many other psychological constructs, we recommend the use of multiple measures. We also find troubling evidence that Screener passage correlates with politically relevant characteristics, such as education and race. We therefore advise that researchers do not restrict their analyses to only those respondents who pass Screeners. We conclude our article with a series of recommendations and an applied example.

Screeners are a valuable tool for social scientists, but they must be used with care. On one hand, if we do not employ Screeners, we run the risk that our surveys will attenuate substantively meaningful correlations on related items and yield false negatives in experiments. On the other hand, culling the sample based on a single Screener question—as is often done in psychology and political science—will cause us to drop a large and nonrandom portion of the sample, leading to selection bias in our survey and experimental research. Using multi-item scales to measure attentiveness, showing the politically relevant predictors of Screener passage in specific applications, and presenting results stratified by levels of attentiveness can improve both data quality and transparency.

[2]There are certain types of experiments—those involving subliminal priming, for example—where acute attention is not necessary, and other cases where the researchers actually want respondents to be distracted from the task at hand. While Screener questions may be less applicable in these cases, they can still be beneficial. For instance, while experimenters may want distracted respondents for experimental purposes, researchers may still want good measures of respondent opinion after the treatment, which requires people to read the question at hand.

Data Collection

Between June 2011 and April 2012, we conducted two Internet studies using samples collected by Survey Sampling International (SSI), an Internet survey company.[3] These studies included a variety of survey questions and experiments, some of which were conducted for other purposes. Both studies enable us to assess the general measurement properties of Screener questions, but each study also allowed us to target specific questions about Screeners.

Study 1 consisted of a two-wave panel in June–July 2011, with about two weeks between waves. There were 1,227 and 728 respondents in Wave 1 and Wave 2, respectively, and the AAPOR COOP1 cooperation rate for Study 1 was 80%. In each wave, we asked four Screener questions spaced evenly throughout the survey. The Screeners ostensibly asked respondents about their favorite color, how they were currently feeling, websites they visit, and their interest in politics and current events. The Screeners are described in full in Section 2 of the online supporting information. The four Screeners were presented in a random order for each subject. The design allows us both to measure the passage rates for different types of Screeners and assess the variability in passage rates between subjects. We then repeated the exact same Screener questions in the second wave to see if passage rates for the same individual change over time. In addition to the Screener questions, we replicate Tversky and Kahneman's (1981) "Asian Disease Problem" framing experiment and ask a battery of economic liberalism questions from the American National Election Study (ANES). Both will be described in greater detail in the next section.

Study 2 consisted of a single-wave survey in April 2012, which included 1,255 respondents and had an AAPOR COOP1 cooperation rate of 66%. The purpose of the study was to test whether receiving a Screener question changes the response pattern or completion rate for subjects. As such, half the respondents received a Screener question before the substantive questions on the survey, while the other half received the Screener question at the very end of the survey.[4] We asked everyone a series of sensitive items about drug use and racial resentment.

[3] 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 the premeasured profile characteristics of the respondents). The resulting sample is not a probability sample but is a diverse national sample. It should be noted that SSI samples have been used in a number of recent publications in political science (Kam 2012; Malhotra and Margalit 2010; Malhotra, Margalit, and Mo 2013).

Table 1 An Example of How Screeners Enhance Data Quality: The Tversky and Kahneman (1981) Framing Experiment

	Mortality Frame	Save Frame
All Respondents		
Probabilistic	62	39
"Sure"	38	61
Passed Screener		
Probabilistic	64	37
"Sure"	36	63
Failed Screener		
Probabilistic	53	49
"Sure"	47	51

Note: Cell entries are column percentages. N = 376 for all respondents, N = 301 for passed Screener, and N = 75 for failed Screener. *Data Source*: Study 1.

Screeners Identify Inattentive Respondents and Reduce Noise

A key reason to use Screeners in surveys and experiments is to uncover satisficing behavior (Krosnick 1991) and identify respondents who offer careless and haphazard survey responses (Huang et al. 2012; Meade and Craig 2012). Oppenheimer, Meyvis, and Davidenko (2009) note that participants who satisfice will not bother to closely read questions or instructions on a survey. Because attention is a prerequisite for receiving the treatment in most survey experiments, Screeners effectively reveal who receives the treatment and who does not. In Table 1, we show how treatment effects vary for those who do and do not pass the Screener, using a much-replicated framing experiment, the "Asian Disease Problem" reported in Tversky and Kahneman (1981) (these data come from Wave 2 of Study 1, described above). In this experiment, all respondents are initially given the following scenario:

[4] Several of our studies included experimental conditions that attempted to induce respondents to pay greater attention by forcing them to repeat the Screener until they passed. These results are discussed below. To avoid contaminating the results presented in these sections, however, we limit the analysis to subjects who were not assigned to this training condition.

> Imagine that your country is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative 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 1/3 probability that 600 people will be saved, and 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 1/3 probability that nobody will die, and 2/3 probability that 600 people will die."

In both of these conditions, respondents are asked to choose between two policy options. The first is a program with certain consequences, and the second is a program where the outcome is probabilistic. The scenarios in both conditions are exactly the same in their description of the expected consequences of each program, 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. Framing the outcomes in positive terms therefore produced a reversal of participants' preferences for the two programs compared to when it was presented in negative terms.

Without differentiating between attentive and inattentive respondents, we replicate the familiar result. As Table 1 shows, 62% of subjects in the "Mortality Frame" condition prefer the Probabilistic Program versus 39% in the "Lives Saved Frame" condition. But when we divide the sample into those who do and do not pass the Screener, large differences emerge. Among those who passed, 64% of subjects in the "Mortality Frame" condition prefer the Probabilistic Program, compared to 37% in the "Lives Saved Frame." However, among those who failed the Screener directly before the experiment, there is essentially no treatment effect: support for the Probabilistic Program is 53% in the "Mortality Frame" condition versus 49% in the "Lives Saved" condition. Thus, respondents who fail the Screener question contribute a great deal of noise to our data.[5]

Screeners can also reduce noise in a nonexperimental setting when question wordings require close reading. For the last four decades, the ANES has 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 health 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?

(The question wordings for the other ANES questions are available in Section 3 of the online supporting information.) While these three questions tap into the same underlying set of beliefs—support for social welfare programs (Jacoby 2000)—the response options differ in subtle ways. For two of the questions, a low response (1) represents a conservative position while a high response indicates a liberal position (7). On the third question, the scale is reversed; the highest response (7) is a conservative position, and the lowest response (1) is a liberal position. By varying the response options on similar questions, researchers can detect satisficing behavior by comparing the correlations of the questions with reversed scales. Table 2 presents the correlations between responses to these questions, which we asked on Wave 1 of Study 1. All variables have been recoded so that high numbers indicate more conservative responses. The positive correlation in the full sample is not surprising, given the nature of these questions. However, when looking specifically at those who failed the Screener directly before the questions were asked, it is clear that some subjects failed to notice the scale reversal. Compared to the correlation between the items that share the same scale ($\rho = 0.48$), there appears to be no relationship between the questions with the reversed scales ($\rho = 0.06$ for both pairs).[6]

[5] In a bivariate regression, the coefficient on the "Mortality Frame" is 0.27 with a standard error of 0.06 for respondents who passed the Screener. For respondents who failed the Screener, the coefficient is 0.04, and the standard error is 0.12.

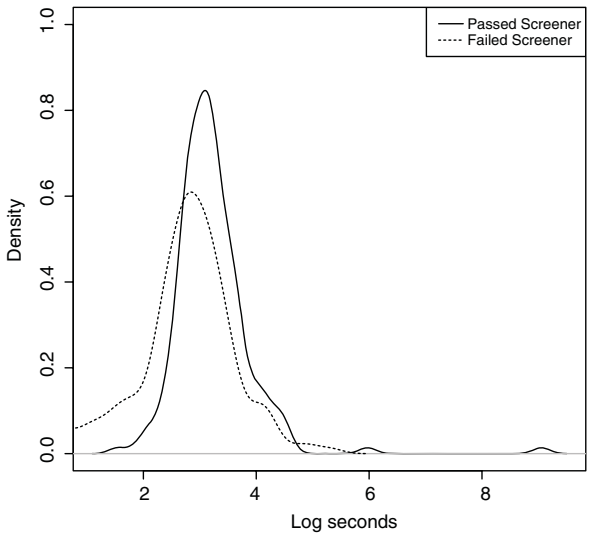
[6] The general pattern of high correlations among passers and zero correlations among failers holds regardless of educational

TABLE 2 An Example of How Screeners Enhance Data Quality: Correlations among Responses to American National Election Survey (ANES) Ideology Questions

	Std of Living	Spending	Income Ineq
All Respondents			
Std of Living	1		
Spending	0.31	1	
Income Ineq	0.53	0.35	1
Passed Screener			
Std of Living	1		
Spending	0.41	1	
Income Ineq	0.55	0.46	1
Failed Screener			
Std of Living	1		
Spending	0.06	1	
Income Ineq	0.48	0.06	1

Note: Cell entries are pairwise correlations. N = 668 for all respondents, N = 483 for passed Screener, and N = 185 for failed Screener. *Data Source*: Study 1.

FIGURE 2 Screener Passers Spend More Time Reading Questions



Note: Lines are densities of the amount of time respondents spent reading the four American National Election Survey (ANES) questions. The number of seconds is first averaged across the four items and then logged. *Data Source*: Survey Sampling International (SSI) January 2012.

Our analyses suggest that these effects arise because people who fail the Screeners indeed read the questions less carefully than those who pass. In particular, Screeners detect respondents who spend less time reading questions. In Figure 2, we show density plots of the (logged) number of seconds respondents spent on the ANES questions.[7] The solid lines in the figures are the densities for respondents who passed the Screener, and the dashed lines are for respondents who failed. Though the patterns we show here hold across our experiments, for illustrative purposes, here we present these densities conditional on passing the Screener asked immediately before the ANES items. We see that those respondents who pass the Screener also spent more time reading that Screener. Looking first at means, the differences in all four of these plots are significant at the 0.05 level. Tests for equality of distributions also indicate that the distributions are significantly different.[8] Our results corroborate work in both political science and psychology that uses amount of time spent on a survey page as a measure of respondent effort (Huang et al. 2012; Malhotra 2008; Wise and DeMars 2006; Wise and Kong 2005). While response times can be a proxy for satisficing behavior, the goal of Screener questions is to more accurately measure an individual's attention, or lack thereof, on a survey.

Comparison to Traditional Manipulation Checks

It should be noted that Screeners are in many ways similar to "manipulation checks" that are often used in political science and psychology. In survey experiments, manipulation checks typically assess whether or not the subject was exposed to the treatment by asking a question that could only be answered by reading the treatment or testing whether an intervening variable varies by condition. For example, in the Tversky and Kahneman (1981) experiment, we might ask subjects to recall the exact estimates of the number of lives saved or lost.

However, manipulation checks can pose problems for researchers because they are directly tied to the

background. When we segment the sample into high and low education groupings, the difference between the passers and failers is the same for both groups.

[7]These results come from a supplementary SSI study conducted in January 2012, in which we asked the newspaper section and political interest Screener questions. We present these results because there were not timers on the ANES questions in Study 1. We replicated Study 1's pattern of correlations among the ANES questions with this new study.

[8]In Section 4 of the online supporting information, we show that Screener passage is associated with greater time spent on additional questions. We also show that those who pass Screeners think more deeply about a cognitive processing task.

experimental treatment. First, by including a manipulation check between the experimental treatment and the dependent variable, the researcher adds an additional event in the subject's experience. By including such an event, the researcher may inadvertently change the very process that she is trying to capture with the experiment (Ellsworth 2010). For instance, manipulation checks run the risk of priming respondents about the treatment they just experienced, in effect treating them for a second time. Conversely, asking the dependent variable before the manipulation check may change responses on that manipulation check—the very measure a researcher needs to identify who is paying attention. In sum, a manipulation check may affect responses on key questions, regardless of its placement.

Screener questions, on the other hand, may be asked before the treatment, avoiding concerns about introducing posttreatment confounders (King and Zeng 2006). Furthermore, these checks cannot be easily used to measure attention on nonexperimental tasks, such as the ANES questions discussed above. Put simply, Screeners are more flexible and less prone to inducing bias than typical manipulation checks.

Do Screeners Affect Subjects in Other Ways?

One concern with Screener questions is they may signal to respondents that their answers are being monitored, which could alter respondent behavior. To test whether this is the case, we designed Study 2 so that half the respondents received a Screener question at the outset of the survey, while half received a Screener at the very end. We can compare respondents who did and did not receive a Screener question on two dimensions: attrition rates and responses to sensitive questions.

First, we examined patterns of attrition, that is, respondents who exit the survey prematurely. While a certain amount of attrition is natural for online surveys, this tendency is uncorrelated with receiving a Screener question at the beginning of the survey. Receiving the Screener at the beginning of the survey does not affect the likelihood that respondents will complete the survey (81% of those who receive the Screener, 82% of those who do not; $t = 0.51$, two-sided $p = 0.61$).[9]

Second, we included two sets of sensitive items on the idea that respondents might answer these questions differently if the inclusion of a Screener question makes them believe they are being "watched" by the researcher. First, we asked respondents two questions about race relations; question wordings are available in Section 5 of the online supporting information. Previous work (e.g., Fazio et al. 1995) found that these questions can be subject to socially desirable reporting.[10] However, we find no difference in the distribution of responses between those who received the Screener and those who did not, nor do we find a difference in the percentage of respondents who did not answer the questions.[11] Second, we asked about drug use in the past 12 months (Tourangeau and Yan 2007). Again, we find no difference in the rates of reported drug use between those who do and do not receive the Screener question.[12] Together these results bolster our confidence that the presence of Screeners does not alter other responses on the survey.[13]

[9]We code completion as whether the respondent answered a question about party identification or not. This question comes near the end of the survey, but before respondents in the "end" condition are exposed to the Screener question (which is the very last question on the survey). The design of this survey prevents us from using actual completion, because by the end of the survey, all subjects have been exposed to a Screener, which would invalidate the comparison. If we do use the actual completion variable, those who received the Screener at the beginning completed the survey 81% of the time, and those who received the Screener at the end also completed the survey 81% of the time ($t = 0.17$, two-sided $p = 0.86$).

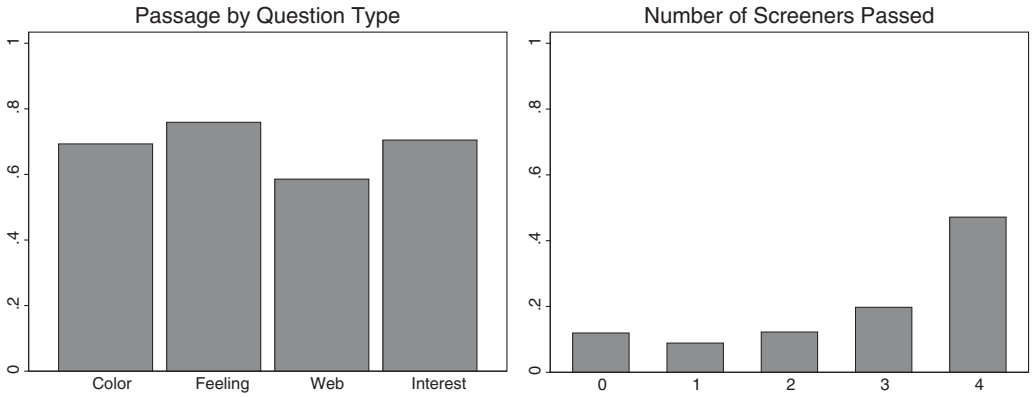
[10]Though these racial resentment questions were designed to be unobtrusive measures of racial attitudes, evidence suggests that since the 1990s, racial resentment questions have not been empirically distinct from traditional prejudice scales (Swim et al. 1995). It appears that the racial resentment items we ask are therefore subject to socially desirable reporting.

[11]For the first racial resentment question, the p-value for a two-tailed t-test of the difference in means is 0.73; the p-value from a Kolmogorov-Smirnov test against the null of equal distributions is 0.98. For the second racial resentment question, the corresponding p-values are 0.28 and 0.86. Those who received the Screener were also no less likely to answer the question: regressing two indicators for missingness on these two racial resentment questions on a dummy for being shown the Screener yields coefficients of 0.013 ($SE = 0.019$, $p = 0.476$) in both regressions.

[12]The p-value for a two-tailed t-test of the difference in means is 0.33; the p-value from a Kolmogorov-Smirnov test against the null of equal distributions is 1.00. Regressing an indicator for missingness on the drug use question on an indicator for receiving the Screener yields a coefficient of 0.012 ($SE = 0.020$, $p = 0.545$). Admittedly, the base rate of reported drug use is so low that we might be encountering floor effects.

[13]As one additional piece of evidence, at the end of each of our surveys we asked respondents for feedback on the survey. Across all the studies that have employed Screeners, the comments are overwhelmingly positive. In Study 1, 27% of respondents provided some form of feedback, and we coded these comments. Of the 302 comments, 7% mentioned enjoying the "trick" questions, 68% provided generally positive comments about the survey, but they did not mention the Screeners explicitly. In contrast, only 0.3% (one respondent) gave negative feedback about the Screener

FIGURE 3 Screener Passage by Question Type and Total Screeners Passed



Data Source: Study 1.

Screeners Measure Attention with Error

Knowing that Screeners are useful tools is one thing. But knowing how to best use them is another. To date, surprisingly little research has been conducted on this question. We therefore sought to explore how best to measure attention with Screeners. By asking multiple Screeners on a single survey, we compare how well a single Screener measures attention as compared to a scale constructed from a series of items.

We begin by examining passage rates across questions and subjects. In the left panel of Figure 3, we show the aggregate passage rates for four Screener questions in Study 1 (we include means and standard deviations for our Screener passage variables in Section 6 of the online supporting information). This figure shows that the passage rates on Screeners vary greatly, ranging from as low as 59% on the website Screener to as high as 76% on the feeling Screener. In the right panel of Figure 3, we show the distribution of the number of Screeners passed. These figures show that only 47% of the sample answers all Screeners correctly, while 12% of the sample fails all the Screener questions. The rest of the sample falls somewhere in between. These passage rates are comparable to those found by researchers who use students in a lab setting; Oppenheimer, Meyvis, and Davidenko (2009) found that 54% of their subjects passed their Screener question. Similarly, Clifford and Jerit (2013) used two Screener questions on a nationally representative sample and found that 38% passed their first item, and 62% passed their second question.[14]

question, and another 5% were generally negative comments about the survey. The remaining 20% were other types of comments.

TABLE 3 Correlation among Passage Rates

	Web	Interest	Color	Feeling
Web	1			
Interest	0.46	1		
Color	0.46	.046	1	
Feeling	0.38	.043	.041	1

Note: N = 1,227. *Data Source*: Study 1.

It is not surprising that there is a great deal of variation in Screener passage rates across subjects. What may be surprising, however, is that there is also great variability in Screener passage rates *within* subjects. Table 3 presents the correlations among the four Screeners asked in Wave 1 of Study 1. The correlations range between 0.38 and 0.46, suggesting that passing a Screener at one point in the survey is a poor predictor of passing a Screener later on.

We also find marked instability within subjects across time. Recall that we used the exact same four Screeners in both waves of Study 1. Every respondent, therefore, answered each Screener twice—first on the initial survey and next on the follow-up survey deployed two weeks later.

[14]In additional studies, we have found that passage rates are typically higher when recruiting subjects from Amazon.com's Mechanical Turk platform (Berinsky, Huber, and Lenz 2012). For example, whereas 69% of the SSI sample passed the color Screener, 91% of Mechanical Turkers passed in a May 2011 study. Likewise, 70% of a September 2012 Mechanical Turk survey passed the news Screener that only 59% of the SSI sample passed. We attribute these higher passage rates to the MTurk population being accustomed to performing nonsurvey tasks where payment is conditional upon attention to detail.

TABLE 4 Correlation of Passage across Panel Waves

Question Type	Cross-Wave Correlation
Web	0.36
Interest	0.39
Color	0.39
Feeling	0.33

Note: N = 742 for the Web correlation, N = 742 for the Interest correlation, N = 737 for the Color correlation, and N = 737 for the Feeling correlation. *Data Source*: Study 1.

As Table 4 demonstrates, there is a great deal of variation in passage across waves, for the exact same respondents with the *exact same* Screener. Specifically, passing a specific Screener in Wave 1 correlates with passing the same Screener in Wave 2 at only about 0.4.

Altogether, these results indicate there is great variability in responses to Screeners. We find that Screener passage on any single item does not perfectly predict passage on other Screeners on the same survey. Furthermore, passage of a given Screener on a given survey is only weakly correlated with passage of that same item only two weeks later.

The pattern of results we find in the data could arise for one of two reasons. First, the instability in Screener passage rates could occur because attention truly waxes and wanes across the course of a survey. In this view, Screeners accurately measure attention at one point in time. It is the variation in respondent attentiveness that causes instability in Screener passage rates. Alternatively, the Screeners could imperfectly measure true attentiveness. From this point of view, we can think of attentiveness as a latent variable, measured by each Screener question with some measurement error. Such a view harkens back to classic debates over the meaning of instability in responses to policy questions in opinion polls (Achen 1975; Converse 1964). Whether this error arises from imperfect questions, imperfect respondents, or the combination of the two, the result, as Zaller (1992) points out, is the same; survey questions always measure the underlying concepts we are interested in with some error. Thus, Screener questions are imperfect measures, like any other survey question.

These two distinct explanations for the apparent instability in the performance of Screeners have important implications for how researchers should employ Screeners in surveys or experiments. If attention truly waxes and wanes over the course of a survey, then we must employ a targeted approach to measure attention at the moment of an experiment or a survey question of interest. However, if Screeners are survey questions plagued by measurement error, researchers should employ multiple measures of attentiveness in a given survey. As Ansolabehere, Rodden, and Snyder (2008) note, the standard measurement error model assumes a true underlying trait and an additive random error in the response. The best way to measure an underlying concept in this situation is to create scales from multiple measures (Carmines and Zeller 1979; McIver and Carmines 1981).[15]

A closer look at the data can help us distinguish between these explanations. We conducted two sets of analyses to attempt to distinguish the measurement error approach from the hypothesis that true attention varies over the course of the survey. We first examined whether the proximity of a Screener to a given experiment affected the ability of the Screener to distinguish the workers from the shirkers. If attention varies over the course of a survey, the proximity of the Screener to the questions of interest matters; a Screener that comes immediately prior to an experimental treatment should do a better job at weeding out inattentive respondents than a Screener that appears much earlier in the survey. However, if the measurement error approach is correct, either of these Screeners should equally (and imperfectly) measure attentiveness.

We test this prediction using both the Tversky and Kahneman experiment (1981) and ANES ideology questions from Study 1. In Section 7 of the online supporting information, we show the results for respondents who passed the first Screener at the outset of the survey and for those who passed the Screener directly preceding the experiment (Screener 3) and ANES questions (Screener 2), respectively. The results for the different groups are substantively the same: using a Screener at the beginning of a survey will give the same results as a Screener asked immediately before the question of interest. The same difference holds when we look at Screener passage across waves separated in time by two weeks. Specifically, we find similar results on the Tversky and Kahneman experiment when we look at respondents who passed the third Screener in Wave 1 and those who passed the third Screener in Wave 2. In spite of the fact that the Screener in Wave 1 was asked weeks before the experiment and the Screener in Wave 2 was asked directly before it, the two Screeners do a remarkably similar job in separating the "workers" from the "shirkers." Moreover, we find the

[15] Such scales could be created using a variety of methods, including additive scales, factor analysis, or Item Response Theory (IRT) methods. While IRT allows us to compute additional quantities of interest (such as standard errors for the scales), all these methods yield substantive results that are nearly identical. Here we use additive scales, but we also created factor scores. The scales created from the two methods are correlated at 0.999.

same trend for the ANES correlation questions, lending support to the measurement error interpretation.

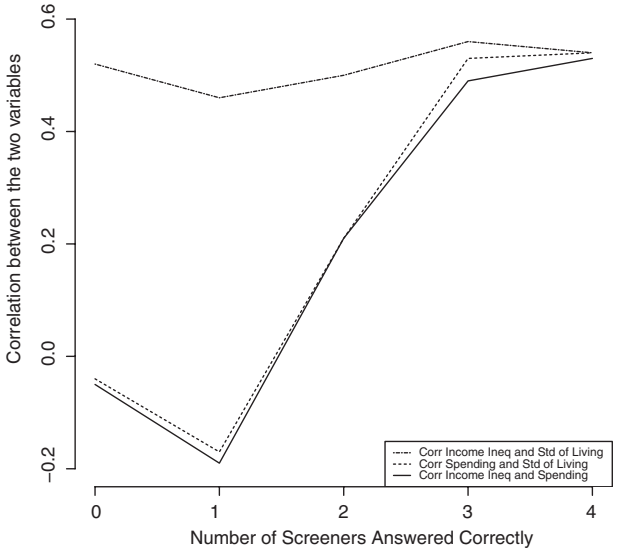
We next look more closely at the over-time correlation of the four Screeners used in Study 1. Recall that the within-subject correlations for the exact same Screeners were relatively low across the waves. But, as with scales of issue items measured with error (Ansolabehere, Rodden, and Snyder 2008), we find that correlation of an additive scale of the items is much larger than any individual item.[16] Where the correlation of the same item across waves did not exceed .39 for any single item, a scale of the four items is correlated at .59.

Interestingly, a factor analysis of the items indicates that each Screener is an equally valid measure of attentiveness, even given the wide variation in passage rates and question topic. We conducted a principle components factor analysis of the four Screeners asked in the first wave of Study 1. The eigenvalue of the first factor was 1.61, and no other eigenvalue was positive, lending strong support to the notion that each question taps a single underlying dimension of attentiveness. The factor loadings for each of the items were extremely similar, ranging from .59 to .66. The results for a factor analysis of the items on the second wave were nearly identical. Thus, an additive scale of the Screener items yields an effective measure of attentiveness, regardless of the content of those items. A scale of the items is also reliable. The alpha of the four item scale was 0.74 in Wave 1 and 0.72 in Wave 2.

Having conducted these two tests, the results of our analysis are now clear. Individual Screener questions imperfectly capture attention on surveys. Thus, attentiveness is more accurately measured on an additive scale based on multiple measures. We demonstrate the effectiveness of this approach empirically in Figure 4, which presents the correlations among the ANES social welfare items, stratified by the respondents' scores on the additive attentiveness scale. This figure shows that the number of Screeners a respondent passed in total appears to have a large impact on the results. While the correlation between the income inequality and standard of living questions—two questions that are on the same scale—is roughly flat across the groups, the correlations on the questions with reversed scales become stronger as attentiveness increases. The proximity of a Screener question to the ANES questions does not yield stronger correlations, yet multiple measures of attention taken together produce the expected strengthening of the results.

[16]Specifically, Ansolabehere, Rodden, and Snyder (2008) analyze data from four different ANES panels and find that in each case multi-item scales are much more highly correlated over time than are individual items.

FIGURE 4 Correlations among Responses to American National Election Survey (ANES) Ideology Questions, Stratified by Respondents' Score on the Additive Attentiveness Scale



Data Source: Study 1.

What Types of Respondents Pass Screeners?

Given that we can measure attentiveness with a scale created from Screener questions, what should we do with inattentive "shirker" respondents? The easiest option is to choose a minimum level of attentiveness and drop respondents who fall below the threshold. Indeed, this is a common practice. As of July 2013, we identified 40 articles in peer-reviewed journals published since 2006 that use Screeners as a tool to identify inattentive respondents. In 32 of these articles, the researchers discard respondents who failed the Screener. In 28 of the articles, the authors purge the sample of respondents on the basis of a single Screener question.[17] However, we do not think this common practice is a good strategy.

By throwing out those who fail Screeners, researchers implicitly assume that subjects may be cleanly partitioned into "worker" respondents, who *always* pay attention, and "shirker" respondents, who *never* pay attention. Thus, practitioners are assuming a deterministic model of survey attention, an assumption that comes with a stark

[17]We include a list of these articles in Section 1 of the online supporting information. We found several other unpublished manuscripts that use Screeners. Here too, the modal practice was to exclude failers.

trade-off between bias and efficiency. Theoretically, using a single Screener to trim the sample should reduce noise. However, even setting aside the fact that our analysis above suggests that Screeners measure attentiveness with error, if attentive and inattentive respondents are different types of people, removing all inattentive respondents may skew the sample. If attention on a survey is a function of the characteristics of respondents—be it via measured factors or unmeasured factors—then discarding respondents who fail the Screener could remove a distinct portion of the population from the sample.[18] For example, if attentive respondents are also wealthier and more educated, this will bias the results of any study that excludes Screener failers. This bias goes beyond the general concern about external validity that comes with all experiments, because there is also the possibility that attention correlates with factors that interact with the treatment. For example, suppose subjects with high levels of education are both more likely to respond to a treatment embedded in a mock news article and more likely to pay attention. In this case, the "attentive" sample will also be the sample most likely to respond to the treatment, inflating estimates of the treatment effect.[19] Excluding failers therefore provides yet another "degree of freedom" (Simmons, Nelson, and Simonsohn 2011) that, unbeknownst to readers, gives the researcher discretion over whether a treatment effect is found at all.

An experimentalist might counter these concerns by invoking the importance of internal validity over external validity. Unfortunately, discarding failers can also threaten internal validity as well. As noted in the last section, Screeners measure attention with error. Thus, a single Screener—or even a set of Screeners—will imperfectly measure attention, and the researcher will end up throwing out some of the attentive sample with the inattentive bathwater. The effect of the treatment will be improperly estimated because some of the treated subjects will be discarded. In addition, this strategy will keep in some inattentive respondents who just happened to pass the Screener question, but it will still contribute nothing to the data. Thus, the problem that motivated the researcher to drop Screener failers in the first place—low power leading to a false negative—is not actually solved by dropping failers.

Our results show that, at least on characteristics we can measure, Screener passers look quite different from Screener failers. In Table 5, we show the results of regressions across five surveys, each of which employed at least one Screener.[20] The dependent variable is the number of Screeners answered correctly, rescaled to lie between 0 and 1. The included independent variables differ from study to study, but in each case we strove as best we could to include a common set of predictors. Despite coming from multiple studies, the models show some clear trends. First, older respondents are more likely to pass Screener questions—the coefficients on the main term hover around 0.01, while the quadratic term is negative and between -0.03 and -0.11. The negative sign on the squared term indicates that the positive relationship between age and Screener passage decays for respondents over 60. Women are always significantly more likely to pass Screeners than men—between 6 and 12 percentage points, depending on the study. Finally, racial minorities are less likely to pass Screeners; the coefficients on the race variables range between -0.01 and -0.20 (though their statistical significance varies across the studies). African Americans, for example, are significantly less likely to pass the Screeners compared to white respondents in three of the five studies.[21] Culling the sample based on attentiveness also means restricting the sample on other politically relevant variables. Of course, these factors do not perfectly predict Screener passage. Thus, it is not imbalances in observables alone that drive the differences in experimental effects between those who pass and those who fail them.[22]

[18] Although they find little or no demographic differences between passers and failers in their studies, Oppenheimer, Meyvis, and Davidenko acknowledge "the concern that if an IMC is used to eliminate participants from the sample then the external validity of the study could be harmed" (2009, 871).

[19] This concern can be formalized in the potential outcomes model of regression, where bias can result from a difference in baseline unobservable characteristics and/or an interaction between the treatment and unobservable characteristics (see Morgan and Winship 2007, 78–79). While a "Local Average Treatment Effect" may be identified in this scenario (Imbens and Angrist 1994), the resulting sample of compliers may be an uninteresting subpopulation when the compliers are known to have disproportionately positive treatment effects.

[20] While some of these studies were designed to investigate the properties of Screeners, others were simply surveys fielded for other research projects, but that happened to include Screeners. Column 1 presents the results from an SSI study conducted in May 2010, which asked the favorite color and "state of mind" Screener questions. Column 2 presents the results from an SSI study conducted in March 2011, and Column 3 presents the results from Study 1 that was described above. The results in Column 4 come from an SSI study conducted in January 2012, where we asked the newspaper section and political-interest Screener questions. Finally, Column 5 presents the results from Study 2, described above.

[21] Oppenheimer, Meyvis, and Davidenko (2009) do not find differences in passing the Screeners based on age, race, gender, or need for cognition. The authors note, however, that their list is far from exhaustive. These authors also use a convenience sample of college students, which explains their lack of variability due to age.

[22] In several of our studies, we measured the respondents' level of political knowledge using factual items modeled on Zaller (1992). We found highly significant differences in political

TABLE 5 Screener Passers Differ on Observable Characteristics

	(1) 2010	(2) 2011a	(3) 2011b	(4) 2012a	(5) 2012b
# Screeners	2	3	4	2	1
Some College	0.010	−0.004	0.059	−0.006	0.051
	(0.027)	(0.034)	(0.034)	(0.034)	(0.035)
College or Above	0.104***	0.019	0.031	0.005	0.057
	(0.027)	(0.035)	(0.034)	(0.031)	(0.035)
Age	0.993**	1.563**	1.068*	0.925*	0.728
	(0.334)	(0.522)	(0.473)	(0.465)	(0.483)
Age-Squared	−5.727	−11.209*	−7.836	−7.704	−2.889
	(3.224)	(5.074)	(4.584)	(4.843)	(4.922)
Female	0.101***		0.064*	0.056*	0.120***
	(0.020)		(0.030)	(0.026)	(0.028)
Black	−0.126**	−0.196**	0.002	−0.080	−0.164***
	(0.047)	(0.061)	(0.051)	(0.047)	(0.049)
Hispanic	−0.069	0.034	−0.064	−0.107*	−0.107
	(0.073)	(0.099)	(0.085)	(0.051)	(0.061)
Other Race	−0.035	−0.149**	−0.043	−0.080	−0.141
	(0.049)	(0.055)	(0.061)	(0.053)	(0.075)
Constant	0.220*	0.070	0.329**	0.060	0.236*
	(0.087)	(0.129)	(0.120)	(0.104)	(0.114)
RMSE	0.39	0.39	0.35	0.35	0.48
R-squared	0.07	0.06	0.04	0.03	0.05
N	1,602	802	638	738	1,220

Note: All models estimated using ordinary least squares with robust standard errors in parentheses. Having a high school degree or less is the reference category for education. White respondents are the reference category for race. Age and age-squared are divided by 100 for interpretability.
$*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

We should also note that we find consistent differences in the size of experimental effects between those who pass Screeners and those who fail them. These differences persist when we stratify our sample by the characteristics that predict Screener passage, such as gender and education. Our findings cannot be attributed, for example, to highly educated respondents both passing the Screeners and responding to the experimental stimulus. These results are presented in Section 8 of the online supporting information.

In addition, even given the patterns of results in Table 5, it is important to remember that the Screeners remain more than simply demographic discriminators. First, there is a great deal of unexplained variance, as evidenced by the low R-squared in all the models in Table 5. Second, and more importantly, as our analysis of the data on "time spent reading" questions shows, Screeners do have face validity as independent measures of attentiveness. As one example, throwing out respondents with low levels of education—which we know to be correlated with political interest and engagement—is problematic if researchers want to make claims about the public's attitudes. The pool of people who pass Screeners is different in important ways from the pool of people who fail these questions measuring attentiveness. Researchers must be cognizant of these differences and careful to not fully expunge respondents who fail Screeners from the presentation of results.

information between respondents who passed the Screeners and those who failed. However, it could be the case that these differences are a result of inattentive respondents answering randomly on the information question, not because of real differences in information levels between the two groups. We therefore include this analysis only in the online supporting information (Section 9).

Can We Create Model Respondents?

Given the large differences in the experimental effects between those respondents who pass our screeners and

those who fail, the ideal solution would be to compel everyone to pay attention to the survey. Oppenheimer, Meyvis, and Davidenko (2009) propose to do just that by "training" their subjects; respondents who fail the initial Screener were asked the same question repeatedly until they passed. After their training, participants who initially failed the Screener were indistinguishable from those who initially passed on subsequent experimental tasks. In effect, the training converted "shirkers" into "workers."

We sought to pursue this strategy in our research. In Study 1, we randomly assigned half of the respondents to a training condition at the beginning of the survey; the training instrument was modeled after Oppenheimer, Meyvis, and Davidenko (2009). We first extended the Oppenheimer, Meyvis, and Davidenko study to see if the training increased passage rates on subsequent Screener questions. We found that it did. Those who failed the first Screener and were untrained went on to pass 1.53 out of the remaining three Screeners; those who failed the first Screener and were trained passed 2.02 out of the three remaining Screeners ($t = 4.35$, $p < 0.001$).

The more important question, for our purposes, is whether this training induces general attention to survey questions and experimental stimuli. As noted above, Oppenheimer and his colleagues found that training led to more reliable experimental effects on two particular tasks. We attempted to replicate this basic result in our studies using the Tversky and Kahneman (1981) "Asian disease problem" mentioned above. We present the full results of this analysis in Section 10 of the online supporting information, but the bottom line is clear. Unlike Oppenheimer, Meyvis, and Davidenko (2009), we did not find consistent results of training on subsequent experimental effects. Sometimes, the subjects who failed the Screener and were then trained looked like the subjects who passed the initial question. But more often, differences between "passers" and "failers" persisted. We find similarly inconsistent and null training effects on other experiments, including replications of a welfare question wording experiment and the Nelson, Clawson, and Oxley (1997) framing experiment (see Section 10 of the online supporting information for details).

In addition, it appears that training comes with additional costs. We compared the attrition rate between subjects who were assigned to be trained and subjects who were not. Respondents in the training condition were about 23 percentage points more likely to exit the survey after failing the initial Screener ($t = 6.55$, $p < 0.001$). They might drop out due to frustration or they could drop out because they perceive the survey to be "broken." Additionally, respondents assigned to the training condition were 10 percentage points less likely to participate in the second wave of the study conducted two weeks later ($t = 3.92$, $p < 0.001$). Thus, while training may improve attention, that attentiveness comes with the drawback of subject attrition. If attrition is not random—which presumably it is not—the result is not only a smaller sample, but one that may also be biased.

In sum, while training may increase passage rate on subsequent Screeners, any effect of this increased passage on the experimental treatments is inconsistent. In addition, many subjects who were trained still failed subsequent Screeners. Finally, we find that training can increase the dropout rate from a survey. Thus, training is not a cure-all. While it would be ideal to induce attentiveness to surveys, existing strategies (like the training recommended by Oppenheimer and his colleagues) remain imperfect. Instead, we believe that our strategy of measuring attentiveness through multiple Screener questions and presenting the experimental results stratified by attentiveness is currently the best strategy for researchers to follow.

Recommendations for Using Screeners

Throughout this article, we have presented evidence suggesting that Screeners are important tools for applied researchers. Screeners allow us to identify "shirker" respondents and obtain more reliable data. This power comes with some potential drawbacks. As we have shown, Screener passage correlates with politically relevant characteristics, which means that culling the sample using Screeners poses threats to validity. But on balance, we strongly believe that the utility of Screeners outweighs their drawbacks. We therefore conclude by offering some best practices for researchers interested in taking advantage of Screeners in their own work.

First, a single Screener item is insufficient for measuring attention. Screener passage at one point in time does not imply Screener passage at another point in time. Instead, a Screener question, like most survey questions, measures its underlying construct with error. As such, it is preferable to create a scale of attentiveness rather than relying on a single measure.[23] Our factor analysis of the items indicates that each of the Screeners taps underlying attentiveness equally well, but researchers might vary the

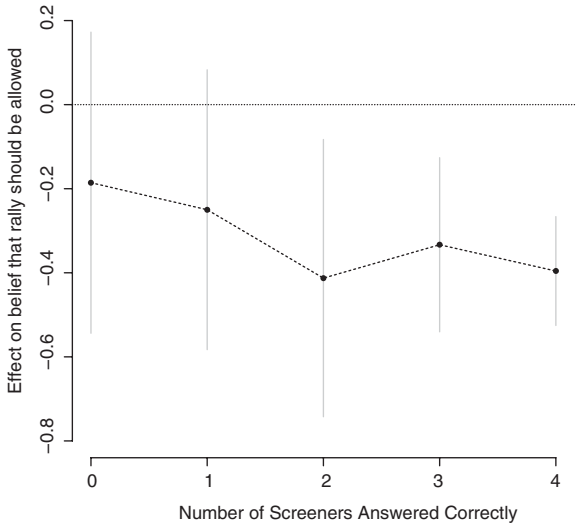
[23] In our work, we scattered these multiple Screeners throughout a survey. However, an alternative strategy would be to ask multiple Screeners in a single block at one point in a survey. Future work could explore the advantages and disadvantages of these different strategies.

"difficulty" of the items included in the scale to induce meaningful variance in the constructed scale.

Second, researchers should present results stratified by attention. Because Screener passage is in part a function of measurable demographic characteristics, researchers should not simply discard respondents who fail Screeners out of hand. By throwing away those who fail a Screener, a researcher may create a sample that overrepresents certain races, ages, and levels of education. However, we have also shown that Screeners can identify respondents who only add noise to the model. The best way to reconcile these two points is to be transparent when presenting results. Stratifying the results by attention allows the readers to easily see how the results change as attention increases. We have already discussed one application of this strategy, in Figure 4, which presented the correlations for ANES social-welfare-spending questions for the different levels of attention.

Third, researchers should analyze the predictors of Screener passage in their sample. Similar to what we present in Table 5, it is important for researchers to know the demographic predictors of Screener passage for a specific sample to gauge whether removing inattentive respondents may skew the sample and induce bias. To model this strategy further, in the next section we present an application of our analytic strategy using a well-known framing experiment from political science.

FIGURE 5 Experimental Effects Vary Based on Attentiveness: Treatment Effects of "Public Safety" Frame in the Replication of Nelson et al. (1997)



Note: Points represent the coefficients from a bivariate regression of support for banning the rally on the "public safety" frame, where the reference group is the group that received the "free speech" frame. Lines span 95% confidence intervals generated using robust standard errors. N = 29 for no Screeners correct. N = 32 for 1/4 Screeners correct. N = 32 for 2/4 Screeners correct. N = 84 for 3/4 Screeners correct. N = 199 for 4/4 Screeners correct. *Data Source*: Study 1, Wave 2.

Applied Example: Framing Effects and Civil Liberties

Drawing on the lessons from above, we replicated the civil liberties framing experiment of Nelson, Clawson, and Oxley (1997) in Wave 2 of Study 1. We chose this study because it requires subjects to pay careful attention to the experimental materials in order to notice the subtle differences in framing. In this experiment, subjects are asked to read a news article about a planned rally by the Ku Klux Klan (KKK) at The Ohio State University. All respondents receive an article with the same core components, but half of the subjects receive a version emphasizing free speech concerns, and half are given a version emphasizing safety concerns. As a dependent variable, we use the response to the question, "Do you think that O.S.U. should or should not allow the Ku Klux Klan to hold a rally on campus?" coded 1 if subjects replied "should allow" and 0 if they replied "should not allow." Thus, the variable measures whether respondents would support allowing the KKK rally.

Without discerning between attentive and inattentive respondents, we find a treatment effect of −0.35 (SE = 0.05; N = 381). That is, as in the original study, we find that framing the rally as a public safety concern significantly lowers subjects' willingness to tolerate the rally, compared to when the rally is framed in terms of free speech. When we condition on levels of attentiveness, however, differences emerge. Study 1 included four Screener questions, from which we create a 5-point scale of attentiveness. The makeup of this scale highlights the dangers of classifying respondents as "workers" or "shirkers" from a single Screener at the outset of the survey. Forty percent who answered three out of four Screeners correctly and 20% who answered two of the four Screeners correctly did not answer the initial Screener question correctly. Had these respondents been excluded from the analysis based solely on their performance in the first Screener, we would have discarded valuable data.

Next, we stratify our results by attentiveness. In Figure 5, we plot the treatment effect of the "public order" frame on support for the rally, conditional on the number of Screeners passed. Similar to the ANES example in the

previous section, the strength of the effect increases in a roughly linear fashion as we increase the number of Screeners passed. The effects range from a statistically insignificant −0.19 (SE = 0.17) for subjects who fail all four Screeners, to a precisely estimated and significant −0.40 (SE = 0.07) for subjects who pass all Screeners. The highest treatment effect is seen for subjects at the midpoint (B = −0.41, SE = 0.15) who pass two of the four Screeners; however, this estimate is not statistically different from the effect conditional on passing all four. In general, the trend on the size of the effect is linear and negative. By not throwing out Screener failers and stratifying the effects in this manner, we can see that, for example, those who fail all Screeners contribute a great deal of noise to the data; and yet, those who fail only one (the effect for those who pass three of the four Screeners) yield a treatment effect that approximates that of the full sample (B = −0.33, SE = 0.10).

Taken together, our overall strategy allows for transparency in results. Stratification based on attentiveness enables the researcher to determine whether inattentive respondents mainly add noise to the model (as it appears those who fail three of the four Screeners do) or whether there is a signal as well (as there appears to be for those who fail only two Screeners). By assessing how results change, the researcher (and readers) can see which respondents drive the results and what inferences are appropriate. Along similar lines, researchers should be cognizant of and address the external validity of their results after screening out inattentive respondents. By making clear which demographic variables are correlated with Screener passage and considering the implications of the culled sample, the strengths and limits of one's findings will be evident.

Conclusion

Self-administered surveys—especially those conducted on the Internet—have enabled researchers to collect data easily and cheaply. This boon for scholars, however, comes at a cost. Without a researcher monitoring the flow of data, respondents can potentially breeze through the survey without paying attention. Our research—and the research of other scholars—demonstrates that as many as half of all respondents behave in this manner. This lack of attention can result in improper estimates of experimental effects and the attenuation of substantively meaningful correlations. In response to this concern, Screeners have become increasingly popular in political science and psychology as a way to distinguish the attentive from the inattentive; however, researchers thus far have employed these questions without considering the implications of their decisions. By conducting a systematic examination of Screeners in social science research, we have presented new evidence of these consequences.

In contrast to the conventional approach, we find that using a single Screener is not the best way to measure attention. Creating a scale of attentiveness based on multiple Screener items of varying difficulty more accurately captures attention, which cannot be reduced to a dichotomous variable.

We additionally provide advice on how to present analyses of social science data in a transparent manner. The common tactic of simply dropping respondents who fail a Screener is problematic because the resultant skewed sample can produce severely biased estimates. Presenting stratified results and considering how the culled sample affects one's findings allows researchers to benefit from Screener questions while avoiding the drawbacks.

While our guidance should be useful to a variety of researchers, there is more work to be done. We should examine how survey researchers could design tasks that limit the problem of inattention in the first place. Avenues of future research can explore different means of engaging survey respondents in order to create more "workers" and fewer "shirkers." But until we can successfully make everyone a model subject, our strategies to identify and cope with inattentive respondents can assist researchers and improve data quality.

References

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *American Political Science Review* 69: 1218–31.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102: 215–32.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351–68.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment.* Beverly Hills, CA: Sage.

Clifford, Scott, and Jenniffer Jerit. 2013. "Do Attempts to Improve Respondent Attention Increase Social Desirability Bias?" Working paper, Florida State University.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, ed. David E. Apter. New York: Free Press, 206–11.

Ellsworth, Phoebe C. 2010. "The Rise and Fall of the High-Impact Experiment." In *The Scientist and the Humanist: A Festschrift in Honor of Elliot Aronson*, ed. M. H. Gonzales, C. Tavris, and J. Aronson. New York: Psychology Press, 79–106.

Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton, and Carol J. Williams. 1995. "Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?" *Journal of Personality and Social Psychology* 69(6): 1013–27.

Huang, Jason L., Paul G. Curran, Jessica Keeney, Elizabeth M. Poposki, and Richard P. DeShon. 2012. "Detecting and Deterring Insufficient Effort Responding to Surveys." *Journal of Business and Psychology* 27(1): 99–114.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Jacoby, William G. 2000. "Issue Framing and Public Opinion on Government Spending." *American Journal of Political Science* 44(4): 750–67.

Kam, Cindy D. 2012. "Risk Attitudes and Political Participation." *American Journal of Political Science* 56(4): 817–36.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2): 131–59.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3): 213–36.

Malhotra, Neil. 2008. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72(5): 914–34.

Malhotra, Neil, and Yotam Margalit. 2010. "Short-Term Communication Effects or Longstanding Dispositions? The Public's Response to the Financial Crisis of 2008." *Journal of Politics* 72(3): 852–67.

Malhotra, Neil, Yotam Margalit, and Cecilia Mo. 2013. "Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact." *American Journal of Political Science* 57(2): 391–410.

McIver, John, and Edward G. Carmines. 1981. *Unidimensional Scaling*. Beverly Hills, CA: Sage.

Meade, Adam W., and S. Bartholomew Craig. 2012. "Identifying Careless Responses in Survey Data." *Psychological Methods* 17(3): 437–55.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press.

Nelson, Thomas E., Rosalee A. Clawson, and Zoe Oxley. 1997. "Media Framing of a Civil Liberties Controversy and Its Effect on Tolerance." *American Political Science Review* 91(3): 567–84.

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45: 867–72.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11): 1359–66.

Swim, Janet K., Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. 1995. "Sexism and Racism: Old-Fashioned and Modern Prejudices." *Journal of Personality and Social Psychology* 68(2): 199–214.

Tourangeau, Roger, and Ting Yan. 2007. "Sensitive Questions in Surveys." *Psychological Bulletin* 133(5): 859–83.

Tversky, Amos, and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211(4481): 453–58.

Wise, Steven L., and Christine E. DeMars. 2006. "An Application of Item Response Time: The Effort-Moderated IRT Model." *Journal of Educational Measurement* 43(1): 19–38.

Wise, Steven L., and Xiaojing Kong. 2005. "Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests." *Applied Measurement in Education* 18(2): 163–83.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Section 1: Studies that have used Screeners
Section 2: Screener Question wordings
Section 3: ANES question wordings
Section 4: Analysis of Screener passage and cognitive effort on survey
Section 5: Social desirability question wordings
Section 6: Summary Statistics for Screeners
Section 7: Tables testing whether attention varies over the course of a survey
Section 8: Robustness checks on experimental results
Section 9: Replication of Table 5
Section 10: Tables testing the effectiveness of training respondents to pass Screeners