Political Science Research and Methods (2021), **9**, 430–437
doi:10.1017/psrm.2019.53

RESEARCH NOTE

Using screeners to measure respondent attention on self-administered surveys: Which items and how many?

Adam J. Berinsky[1], Michele F. Margolis[2*], Michael W. Sances[3] and Christopher Warshaw[4]

[1]Department of Political Science, MIT, Cambridge, USA, [2]Department of Political Science, University of Pennsylvania, Philadelphia, USA, [3]Department of Political Science, Temple University, Philadelphia, USA and [4]Department of Political Science, George Washington University, Washington, USA
*Corresponding author. Email: mmargo@sas.upenn.edu

(Received 15 February 2019; revised 6 August 2019; accepted 2 September 2019; first published online 12 November 2019)

Abstract

Inattentive respondents introduce noise into data sets, weakening correlations between items and increasing the likelihood of null findings. "Screeners" have been proposed as a way to identify inattentive respondents, but questions remain regarding their implementation. First, what is the optimal number of Screeners for identifying inattentive respondents? Second, what types of Screener questions best capture inattention? In this paper, we address both of these questions. Using item-response theory to aggregate individual Screeners we find that four Screeners are sufficient to identify inattentive respondents. Moreover, two grid and two multiple choice questions work well. Our findings have relevance for applied survey research in political science and other disciplines. Most importantly, our recommendations enable the standardization of Screeners on future surveys.

Keyword: Survey methodology

1. Introduction

In order to ensure that respondents pay attention on self-administered surveys, researchers frequently use "Screener" questions to identify inattentive respondents (Oppenheimer *et al.*, 2009; Meade and Craig, 2012; Berinsky *et al.*, 2014). By instructing respondents to select a specific, otherwise atypical response to demonstrate their attention, these questions effectively reveal the proportion of respondents who do not read questions carefully. Using this method, Berinsky *et al.* (2014) show that as many as 40 percent of respondents will fail Screener questions, and that attentive and inattentive individuals respond to the same stimuli in very different ways.[1]

While Screeners hold great potential for identifying inattentive respondents, questions remain regarding their implementation. First, what is the optimal number of Screeners for identifying inattentive respondents? Berinsky *et al.* (2014) present evidence that a single Screener measures attention with error and ultimately argue for an "additive scale based on multiple measures" (747). Thus, multiple questions are needed. However, it is currently unclear just how many questions are necessary—and thus how much survey time researchers should allocate—for a useful scale.

Second, what types of Screener questions best capture inattention? Existing work offers a plethora of potential Screeners that vary both in content—for instance, questions about a respondent's favorite color, current mood, or interest in politics—and form—such as stand-alone questions that instruct respondents to choose a given option or perform a specific task (Oppenheimer *et al.*, 2009; Berinsky *et al.*, 2016) or attention checks that appear within a grid or among a battery

[1]While others refer to these sorts of questions as Instructional Manipulation Checks, or IMCs (Oppenheimer *et al.*, 2009; Hauser and Schwarz 2015), we will refer to questions that measure attentiveness as Screeners.

of questions (Kung *et al.*, 2018). The dozens of political science articles that have been published since 2014 that use Screeners have employed them in an ad hoc way, raising concerns about generalizability and replicability.

In this paper, we examine how to best capture survey attentiveness using a relatively small set of survey questions. We provide general guidance for the kinds of 10–20 min self-administered Internet surveys now common in political science research. We show it is possible to accurately capture survey attentiveness using only two stand-alone multiple choice Screener questions and two simpler true/false questions within a grid. Moreover, our results highlight that while stand-alone Screeners are well equipped to distinguish between respondents at the top of the attentiveness spectrum, grid Screeners are better able to do so among respondents with low levels of attention. Finally, we make general recommendations for applied researchers interested in using a standard attentiveness scale. Though this advice is primarily aimed at scholars using a 10–20 min online survey, these guidelines can be adapted to other surveys as well. Our purpose here is to advance a measurement approach to gauge attentiveness reliably in as short a scale as is feasible.

2. Data and methods

We use the two-parameter item response theory (IRT) model (Clinton *et al.*, 2004; Van der Linden, 2005) to measure respondents' latent attentiveness on surveys.[2] This model characterizes each Screener response $y_{ij} \in \{0, 1\}$ as a function of subject i's latent *attentiveness* (θ_i), the *difficulty* (α_j), and *discrimination* (β_j) of item j, where

$$\Pr[y_{ij} = 1] = \Phi(\beta_j \theta_i - \alpha_j) \tag{1}$$

where Φ is the standard normal CDF (Jackman, 2009; Fox, 2010).[3] We estimate the IRT model using the `ideal` function in the `pscl` R package (Jackman, 2010). We identify the attentiveness estimates by post-processing them to have a standard normal distribution.

While a greater number of screener items facilitate more accurate measures of attentiveness (Berinsky *et al.*, 2014), researchers are rarely able to include a large number of screeners in their surveys. In order to evaluate an optimal set of screener items to measure attentiveness, we draw from optimal test theory (van der Linden, 1998; Tausanovitch and Warshaw, 2012; Montgomery and Cutler, 2013). Specifically, we seek to maximize Fisher's Information for a given scale. Under this framework, the contribution of a given item to our level of certainty at a particular value of attentiveness, θ_i, can be determined by evaluating Fisher's Information for the item at that value (Bimbaum, 1968; Van der Linden, 2005):

$$\mathrm{IIF}_j(\theta) = \beta_j^2 * p * q \tag{2}$$

[2]A three-parameter IRT model would also be a reasonable way to fit models of attentiveness. This could account for guessing behavior and confusion about the instructions for an item, as opposed to inattentiveness, on individual items. However, we obtain similar results using a two-parameter IRT model and a three-parameter one. Due to its greater simplicity and the fact that the results are very similar across models, we focus on the two-parameter IRT model in the remainder of our analysis. Also, following past literature, our IRT model measures attentiveness on a uni-dimensional scale. We evaluated the validity of this model using exploratory factor analysis. We found that there is a clear drop-off in explanatory power between the first principal component and higher-order ones. This suggests that it is reasonable to summarize attentiveness with a single latent trait.

[3]An IRT model has a number of advantages over additive and factor-analytic models. First, unlike these approaches, an IRT model allows users to characterize measurement error in their estimates (Treier and Hillygus, 2009). Second, conventional factor analysis can produce biased estimates of latent variables with binary indicators (Kaplan, 2004). Third, IRT models can easily handle missing data if a particular respondent does not answer all the screener questions. However, a factor-analytic model will generally yield similar estimates of attentiveness as our model for respondents with non-missing Screener responses.

where $p = \Phi(\beta_j\theta - \alpha_j)$ and $q = 1-p$. This is referred to as the Item Information Function (IIF). The Test Information Function (TIF) for a set of items is simply the sum of the individual IIFs (Van der Linden, 2005, 16–17). We use the IIF as means of selecting items and the TIF as a way of comparing sets of items.

Scholars may want to maximize information across the entire range of attentiveness. Van der Linden (2005) shows that this can be done by maximizing the TIF for a small set of uniformly distributed points in the range of attentiveness, θ. Since the TIF is an additive function of the IIFs, this requires only that we calculate the values of the IIF at each of these points, and choose the items with the highest sum of these values.

Alternatively, we may also want to discriminate between low and medium/high attention respondents—that is, between shirkers and workers (see Van der Linden, 2005, 22–22). For example, we might want to just separate respondents in the bottom quartile of the range of attentiveness from the rest of the respondents. To do this, we can maximize the TIF at a value in the lower end of the attentiveness spectrum. This gives the optimal set of items to separate low attentiveness respondents—aka, shirkers—from the rest of the respondents.

To examine how to best capture attentiveness using a small set of survey questions, we conducted a nationally diverse online survey of 2,526 Americans via Survey Sampling International (SSI) in August 2016. The survey included eight Screeners. Following Berinsky *et al.* (2014), four of these items were Screeners asking about favorite colors, the most important problems facing the country, news web sites, and newspaper sections. Each of these Screeners is stand-alone—the Screener question is the only question to appear on the page—which has been the traditional way of asking Screener questions to date. We show screenshots of these questions in the online Appendix A.

We embedded the four remaining Screeners in question grids alongside other questions. The purpose of these grid Screeners was to explore the feasibility of increasing the total number of Screeners asked while taking up less space. We presented subjects with two grids of questions over the course of the survey. For each row in the grid, a respondent was presented with a (randomly ordered) statement with which they could agree strongly, agree, neither agree nor disagree, disagree, or disagree strongly. Along with sincere attitudinal questions such as whether the federal government should guarantee health insurance and whether gays and lesbians should have the right to marry, the first grid included two Screener statements that have a single right answer: that World War I came after World War II; and an instruction to "Please check 'neither agree nor disagree"'. The second grid similarly contained two Screener statements—"Obama was the first president" and "Two is greater than one"—amid the sincere attitudinal statements.

3. Results

As a benchmark, we first measure attentiveness using all eight items. Table 1 shows the results. First, it shows the percentage of people that got each item right. It also shows the "discrimination" parameter for each item, β_j, which captures the degree to which respondents' latent attentiveness affects the probability of a correct answer on each question. If β_j is 0, then question$_j$ tells us nothing about attentiveness. In addition, it shows the difficulty parameter for each item, α_j, which indicates how hard an item is to get right. Finally, it shows how much information each individual Screener item provides about the full attentiveness scale, as well as for high and low attentiveness respondents.[4]

The top four survey items in Table 1 are traditional, stand-alone Screeners. These items have relatively low passage rates, ranging from 25 to 58 percent, and the high difficulty parameter

[4]We use Equation (2) to estimate the level of information for low-attention respondents (one standard deviation below the mean), and high-attention respondents (one standard deviation above the mean). Online Appendix D graphically displays the IIFs for each item.

Table 1. Item parameters

Type	Item	Pass rate	Difficulty Param.	Discrim. Param.	IIF (Full Dist.)	IIF (High Atten.)	IIF (Low Atten.)
Stand-alone	Websites	0.39	0.65	1.76	3.93	0.36	0.02
Stand-alone	Most important problem	0.32	0.84	1.28	2.65	0.36	0.03
Stand-alone	Favorite color	0.58	−0.33	1.23	2.68	0.08	0.23
Stand-alone	Section of newspaper	0.25	1.65	1.89	4.03	0.86	0.00
Grid	World War I came after World War II	0.61	−0.37	0.77	1.37	0.07	0.13
Grid	Please check "neither agree nor disagree"	0.90	−1.96	0.98	0.76	0.00	0.13
Grid	Obama was the first president	0.72	−0.81	0.90	1.54	0.03	0.20
Grid	Two is greater than one	0.76	−0.88	0.63	0.77	0.02	0.09

values for these questions suggest even relatively attentive respondents failed some of these Screeners. That said, these questions all discriminate well on the latent scale, and they each contribute a good deal of information to the full scale. The high difficulty of stand-alone Screeners means they do a good job of discriminating between those with moderate and high levels of attention but are unable to distinguish among respondents at the bottom range of attentiveness (see online Appendix D).

In contrast, the four grid items all have relatively high passage rates—ranging from 61 to 90 percent. The low difficulty parameters confirm that only inattentive people failed many of these Screeners. While these items do not contribute as much information to the full attentiveness scale (or at the top end of the range of attentiveness) as the stand-alone Screeners, they do discriminate very well between people at the low end of the scale (since these are the people that tend to fail the grid items). Examining the IIF for the low-attentiveness respondents in the last column, we see that all four grid Screeners contribute more information at the low end of the scale than the website, most important problem, and section of the newspaper stand-alone Screeners.[5]

Next we evaluate the validity of various scales that combine multiple screener items. First, we evaluate the full scale with all eight Screener items. Next, we evaluate scales that use the four traditional Screeners or the four grid Screeners. The scale with only the four stand-alone Screeners is likely to do a good job discriminating among high attention respondents, but a poor job at discriminating among low attention respondents. Conversely, the scale with only the four grid Screeners is likely to do a good job discriminating among low attention respondents, but a poor job at discriminating among high attention respondents. Finally, we evaluate a mixed attention scale that combines two grid Screeners that provide information about the attentiveness of low attention respondents and two stand-alone Screeners that provide information about the attentiveness of high attention respondents.[6]

In Figure 1, we follow the model of Berinsky *et al.* (2014) and evaluate how each of four attentiveness scales fares at predicting respondents' performance on Tversky and Kahneman's (1981) unusual disease framing experiment (see online Appendix B). The y-axis represents the framing treatment effect and the x-axis is the attentiveness scale with larger numbers indicating greater levels of attention. Each figure includes points that represent quintiles along the attentiveness scale as well as a loess line and 95 percent confidence bands, which use 40 binned groups.

[5]A concern could be that these grid Screeners are capturing cognitive ability rather than engagement with a survey. Indeed, both Berinsky *et al.* (2014) and Alvarez *et al.* (2019) find Screeners sometimes correlate with education. However, we show in online Appendix E that none of our attentiveness scales are strongly predicted by demographics such as education or age. While political knowledge is a robust and strong predictor, it explains a relatively small portion of the variance in attentiveness. Moreover, exploratory factor analysis indicates that a single latent factor (attentiveness) characterizes the bulk of the variation in the individual Screeners.

[6]There is a 0.94 correlation between these attentiveness estimates using four items and the estimates using all eight items.

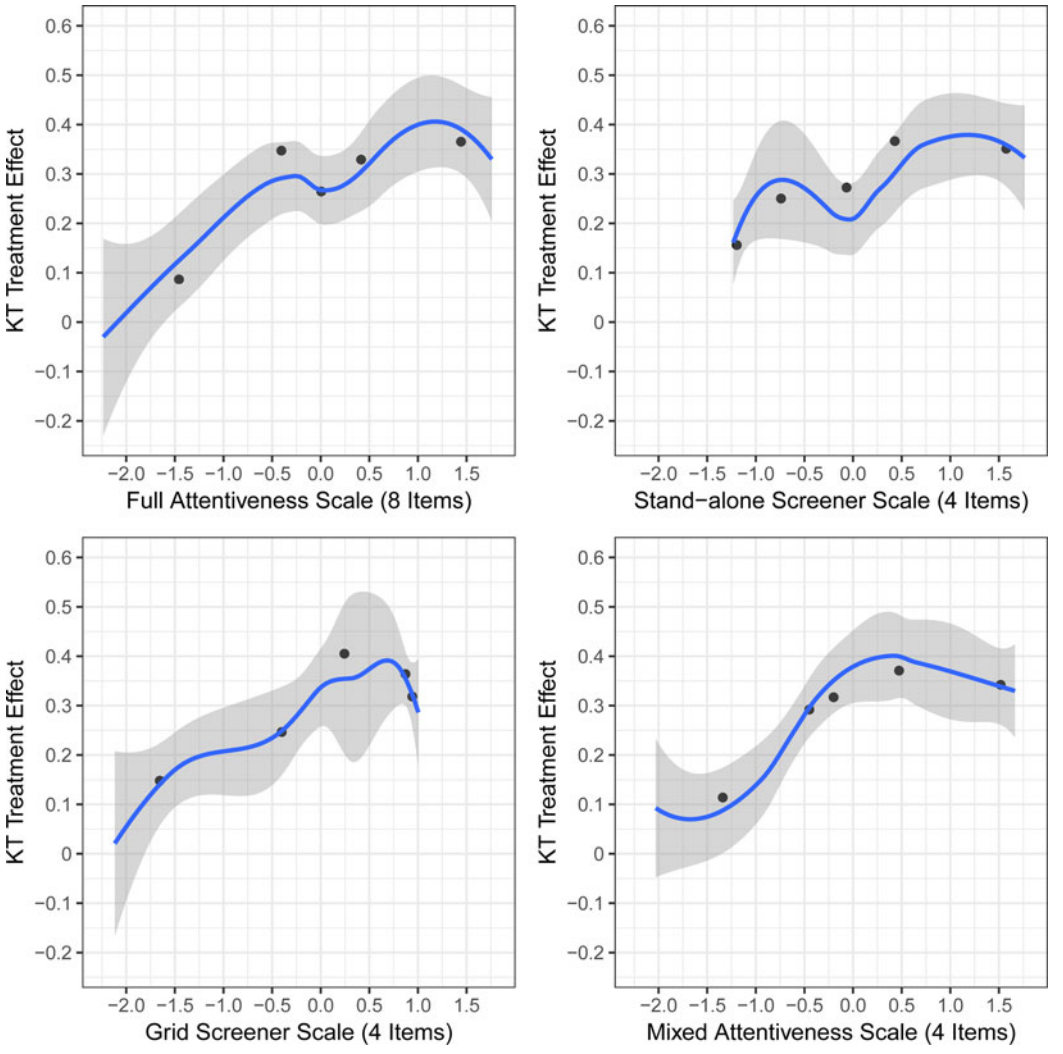

Figure 1. Attentiveness plays a role in detecting experimental treatmenteffects.

Following Berinsky *et al.* (2014), we expect the treatment effects will be larger among more attentive respondents.

In the top-left panel of Figure 1, we find the full scale with eight items clearly discriminates the most inattentive from everyone else. Indeed, there is essentially no treatment effect among respondents in the lowest quintile of attentiveness (treatment effect = 0.09, SE = 0.04). In contrast, there are clear effects among the remaining 80 percent of the attentiveness scale. To put these results in context, the magnitude of the experimental treatment effect among those who fall in the 20–40th percentile is the same as it is for those who passed any stand-alone Screener.[7] In other words, the traditional, stand-alone Screeners not only require researchers to drop substantial portions of the sample, but the results look virtually identical to those with only moderate levels of attentiveness.

[7]The treatment effect among the second quintile is 0.35 and ranges between 0.34 and 0.35 among passers of each of the stand-alone Screeners.

In contrast, scales that use all traditional Screeners (top-right panel) or all grid Screeners (bottom-left panel) do much worse at discriminating shirkers from workers in this experiment. For the scale that employs only traditional Screeners, there are smaller, but non-null treatment effects in the lowest two quintiles. This result occurs because the stand-alone Screeners do not do a good job distinguishing between those with low and moderate levels of attention. As a result, moderately attentive respondents, who responded to the framing treatment, end up in the bottom quintile of attention. Moreover, only in the top 40 percent of attentiveness on this scale do the treatment effects reach the same strength as using the top 80 percent of attentiveness on the full scale. For the scale with all grid Screeners, the top three quintiles have similar average scores of attentiveness because many respondents answered almost or all the grid screeners correctly. While the grid screener scale can certainly identify true shirkers, it has a more difficult time separating individuals at the higher end of attentiveness. Crucially, however, on both of these scales analysts would have to drop at least 40 percent of the sample in order to clearly separate shirkers from workers, whereas the full eight-item scale can distinguish between shirkers and workers by dropping only the bottom quintile of attentiveness.

While the eight-item scale performs better than using four stand-alone Screeners or four grid Screeners, implementing a survey with eight Screeners is costly. A mixed attention scale with two grid and two stand-alone Screeners performs nearly as well as the full scale (bottom-right panel). The experiment yields small treatment effects among respondents in the lowest quintile of attentiveness (0.11). Once again, however, there is a clear jump in the size of the treatment effects between the bottom and second quintiles, with relatively modest differences across quintiles.[8] Similar to the full eight-item scale, the four-item mixed scale improves upon the strategy of using a single stand-alone Screener by showing that researchers can improve data quality while maintaining a larger proportion of the sample.

These results further show that the framing experiment is not one that requires extreme levels of attentiveness. Respondents need to pay some attention to the treatment—choosing response options randomly will not suffice. But even those individuals who may have only skimmed the experimental stimulus responded to the difference in language between the conditions.

Next, we examine how well the different attentiveness scales do at reducing noise in a non-experimental setting when question wordings require close reading, again following Berinsky *et al.* (2014). For the last four decades, the ANES has asked a series of three questions on economic liberalism (see online Appendix C). For two of the questions, a low response (1) represents a liberal position while a high response indicates a conservative position (7). On the third question, the scale is reversed.

In Figure 2, we examine the difference in (a) the correlation between the reverse-item scale and the one of the two like-coded scales (which should be negative) and (b) the correlation between the two like-coded scales (which should be positive). If respondents are paying attention, the correlation between the same-coded scales should be around 0.5 and the correlation between the reversed scale should be around -0.5, producing a difference of -1. This is exactly what we observe when using the attentiveness scale with all eight items (upper-left panel). Here, we find virtually no difference in the correlations of flipped and non-flipped ANES scales among respondents in the lowest quintile of attentiveness. After a large difference between the bottom and second quintiles of attentiveness (-0.03 versus -0.47), the middle 60 percent of the attentiveness range (second, third, and fourth points) looks similar to one another, whereas those in the top 20 percent in the attentiveness range has a difference correlation of -0.92. Unlike the framing experiment in which respondents in the top 80 percent of the sample all responded similarly to the experimental stimulus, the most attentive people in sample were the most responsive to the relatively long survey questions and subtle change in response options. In other words, for the ANES questions, attentiveness matters at both the top *and* the bottom of the scale.

[8]The magnitude of the treatment effects varies between 0.29 (second quintile) and 0.37 (fourth quintile).

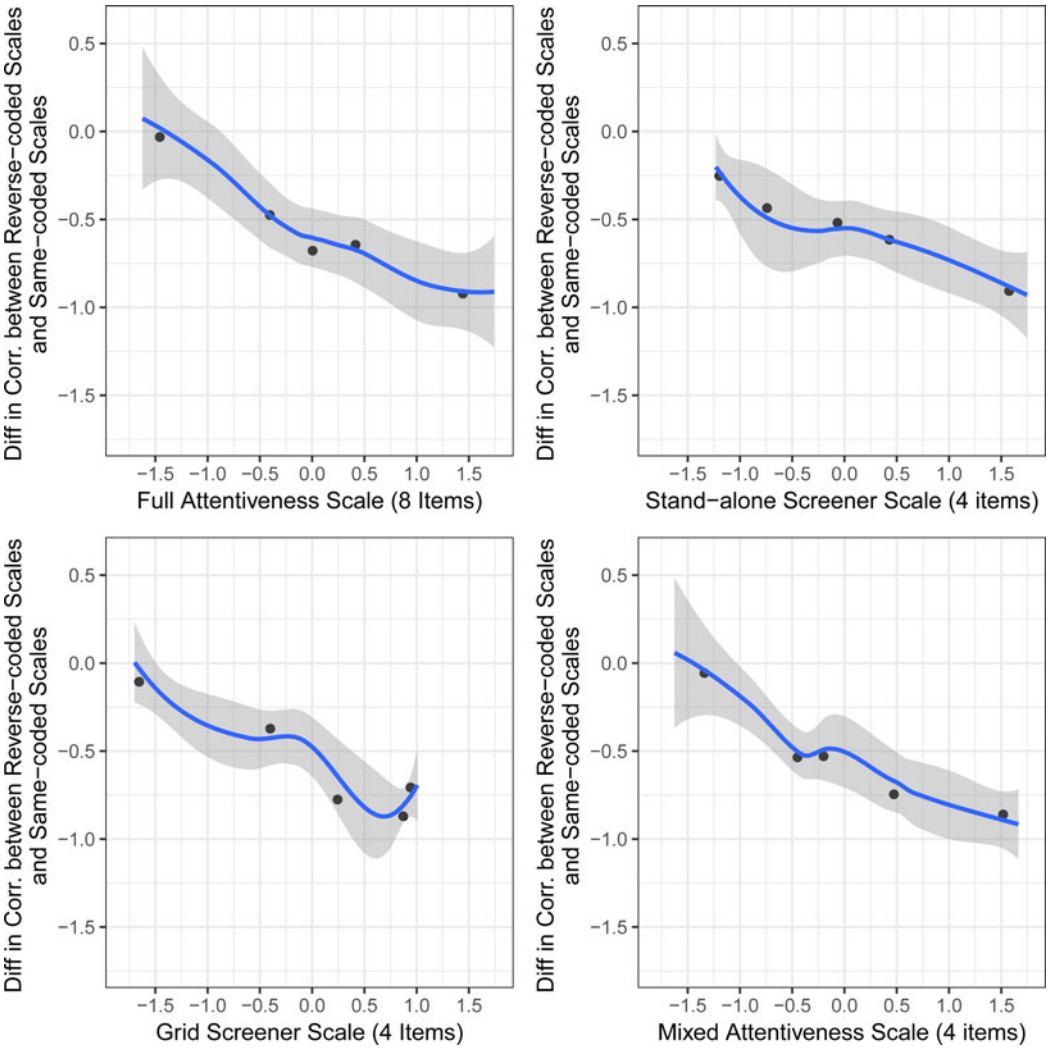

Figure 2. Attentiveness plays a role in non-experimental data collection.

Again, the four-item scale made up of only stand-alone Screeners (top right) does a better job distinguishing among people at the top end of the attentiveness range rather than the bottom, while the four-item scale made up of only grid items (bottom left) does a good job identifying the least attentive respondents but has a more difficult time distinguishing respondents at the top end of the attentiveness spectrum.[9] Again, the four-item mixed attention scale looks quite similar to the eight-item scale, successfully distinguishing between inattentive respondents in the bottom quintile and the rest of the sample. These results indicate that the mixed scale with only four items performs nearly as well as the full scale at detecting inattentive respondents on the ANES scales.

[9]The difference in correlations among those in the bottom quintile of the traditional, stand-alone scale is − 0.25. While still substantially lower than the correlation in the top quintile, it is larger than the full attentiveness scale (which has a correlation of − 0.03), indicating that there are people with moderate levels of attention in the bottom quintile. Similarly, the difference in correlations among those in the top two quintiles of the grid scale is − 0.87 and − 0.71, indicating that there are people with moderate levels of attention in the top quintiles.

4. Discussion and conclusion

Previous research has already shown that using a single Screener question is problematic. In this paper, we show that researchers should use multiple Screeners that vary in difficulty in order to accurately place respondents on an attentiveness scale.

As a general rule, we recommend using a multi-item scale that includes Screeners with both high and low passage rates, similar to our four-item mixed scale. We recommend that researchers use an IRT model to construct this scale. But a simpler factor-analytic model will often suffice. This scaling strategy allows researchers to classify respondents at both the top and bottom ends of the attentiveness spectrum. Figures 1 and 2 make it clear to readers how respondents with different levels of attentiveness behave in the survey. That said, researchers may want to tailor a set of attention checks specific to their research needs. For example, grid Screeners will suffice if researchers want to identify the least attentive respondents. Alternatively, if one has a particularly subtle treatment or complicated experimental design that requires respondents to pay careful attention, stand-alone Screeners would be the best way to distinguish among people at the top end of the attentiveness spectrum.

Supplementary Material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2019.53

References

Alvarez RM, Atkeson LR, Levin I and Li Y (2019) Paying attention to inattentive survey respondents. *Political Analysis* **27** (2), 145–162.

Ansolabehere S, Rodden J and Snyder Jr JM (2008) The strength of issues: using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review* **102**, 215–232.

Berinsky AJ, Margolis MF and Sances MW (2014) Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* **58**, 739–753.

Berinsky AJ, Margolis MF and Sances MW (2016) Can we turn shirkers into workers? *Journal of Experimental Social Psychology* **66**, 20–28.

Bimbaum A (1968) Some latent trait models and their use in inferring an examinee's ability. In Lord FM, Novick MR and Birnbaum A (eds), *Statistical Theories of Mental Test Scores*. Oxford, England: Addison-Wesley, pp. 395–479.

Clinton J, Jackman S and Rivers D (2004) The statistical analysis of roll call data. *American Political Science Review* **98**, 355–370.

Fox J-P (2010) *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer (PDF ebook).

Hauser DJ and Schwarz N (2015) It's a trap! Instructional manipulation checks prompt systematic thinking on 'tricky' tasks. *Sage Open* **5**(2), 1–6.

Jackman S (2009) *Bayesian Analysis for the Social Sciences*. Hoboken, NJ: Wiley.

Jackman S (2010) pscl: Classes and methods for R. Developed in the Political Science Computational Laboratory, Stanford University. Department of Political Science, Stanford University, Stanford, CA. R package version 1.03. 5. http://www.pscl.stanford.edu/.

Kaplan D (2004) *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage.

Kung FY, Kwok N and Brown DJ (2018) Are attention check questions a threat to scale validity? *Applied Psychology* **67**(2), 264–283.

Meade AW and Craig SB (2012) Identifying careless responses in survey data. *Psychological Methods* **17**, 437.

Montgomery JM and Cutler J (2013) Computerized adaptive testing for public opinion surveys. *Political Analysis* **21**, 172–192.

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

Tausanovitch C and Warshaw C (2012) How should we choose survey questions to measure citizens' policy preferences? Working paper, Department of Political Science, Stanford University. Available at http://www.chriswarshaw.com/papers/MeasuringPreferences_Feb142012.pdf.

Treier S and Hillygus DS (2009) The nature of political ideology in the contemporary electorate. *Public Opinion Quarterly* **73**, 679–703.

Tversky A and Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* **211**(4481), 453–458.

van der Linden WJ (1998) Bayesian item selection criteria for adaptive testing. *Psychometrika* **63**, 201–216.

Van der Linden WJ (2005) *Linear Models for Optimal Test Design*. New York: Springer Science & Business Media.

Cite this article: Berinsky AJ, Margolis MF, Sances MW, Warshaw C (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods* **9**, 430–437. https://doi.org/10.1017/psrm.2019.53