*Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys*



**Citation:** Read, B., Wolters, L., & Berinsky, A. (2021). Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys. Political Analysis, 1-20

**As Published:** 10.1017/pan.2021.32

**Persistent URL:** https://hdl.handle.net/1721.1/145319

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

# Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys*

Blair Read† Lukas Wolters‡ Adam J. Berinsky§

July 9, 2021

**Abstract**

Internet-based surveys have expanded public opinion data collection at the expense of monitoring respondent attentiveness, potentially compromising data quality. Researchers now have to evaluate attentiveness ex-post. We propose a new proxy for attentiveness – **response-time attentiveness clustering** (RTAC) – that uses dimension reduction and an unsupervised clustering algorithm to leverage variation in response time between respondents and across questions. We advance the literature theoretically arguing that the existing dichotomous classification of respondents as *fast* or *attentive* is insufficient and neglects *slow* and inattentive respondents. We validate our theoretical classification and empirical strategy against commonly used proxies for survey attentiveness. In contrast to other methods for capturing attentiveness, RTAC allows researchers to collect attentiveness data unobtrusively without sacrificing space on the survey instrument.

**Key Words: Response time, survey attentiveness, Gaussian mixture model**

*We thank Jacob Montgomery, Lucas de Abreu Maia, Minh Trinh, Leah Rosenzweig, participants at the 2019 Midwest Political Science Conference, and members of the the MIT Quantitative-Works-in-Progress, Kim Research Group, and Political Experiments Research Lab who provided useful comments and feedback. James Dunham, Kathryn Treder, Anna Weissman, and Laurel Bliss provided excellent editorial assistance.

†PhD candidate, Department of Political Science, Massachusetts Institute of Technology, Cambridge MA 02139. Email: bmread@mit.edu

‡PhD candidate, Department of Political Science, Massachusetts Institute of Technology, Cambridge MA 02139. Email: woltersl@mit.edu

§Mitsui Professor of Political Science, Department of Political Science, Massachusetts Institute of Technology, Cambridge MA 02139. Email: berinsky@mit.edu

# 1 Introduction

Answering survey questions can be hard. Respondents have to carefully read and comprehend the questions they are asked, retrieve the information associated with the questions, judge this information to form an answer, and express that answer (Tourangeau, Rips and Rasinski, 2000). All of these tasks are cognitively taxing. As a consequence, some survey respondents may try to avoid exerting the effort necessary for these tasks, and instead choose the first minimally acceptable alternative that comes to mind, a process which Krosnick (1991) called "satisficing."

Crucially, satisficing behavior and survey attentiveness do not vary randomly but rather systematically across individuals. For instance, the increasing number of professional survey-takers, experienced respondents who seek out large numbers of surveys for the rewards offered (Baker et al., 2010), are more likely to satisfice and systematically show lower levels of political knowledge, interest, and ideological extremism (Hillygus, Jackson and Young, 2014). Moreover, survey attentiveness could also be correlated with subjects' gender, age, and race (Berinsky, Margolis and Sances, 2014; Alvarez et al., 2019). Ignoring whether respondents pay attention to survey questions can introduce substantial non-sampling bias in survey results. But how do we identify those respondents who are taking the time to answer the questions thoroughly, and those who are not?

We propose a new method of identifying inattentive respondents: response-time attentiveness clustering (RTAC). RTAC is a viable alternative to existing measures that use manipulation checks or screener questions. Rather than add new items to a survey questionnaire, we instead propose using per-question response time as a proxy for attentiveness. Question-by-question response time can be collected unobtrusively on popular online survey platforms.

We provide researchers with a step-by-step process for how to use response time data to ascertain which respondents are paying attention. After fielding a survey, researchers can collect data on how long each respondent spent on each survey question.[1] Yet, for reasons we discuss below, these data are not immediately useful. RTAC is a two-step process that enables the extraction of a single measure of attentiveness for each respondent from these multidimensional data. We begin by reducing the high dimensionality of these response time data on every question through principal-component analysis, isolating the signal of attentiveness generated by differences in response times from the noise inherent in such data. We take these transformed data and fit

[1]These data, also called paradata, comprise a matrix the length of the number of respondents and width of the number of questions. Each observations contains the number of seconds respondents take to answer each question.

a Gaussian mixture model through expectation maximization to estimate latent attentiveness. At the end of the process, we obtain a single measure of respondent attentiveness that assigns respondents to one of three attentiveness clusters based on their survey-taking behaviors. To assist researchers in using RTAC to analyze their own surveys, we provide documentation and an accompanying vignette that take users through each step of the process. This vignette is in Appendix D.

RTAC improves on existing methods using response time in two ways. First, we retain data on response time for each question per respondent rather than focusing solely on either respondent- or question-level aggregate measures. With PCA, we take advantage of the fact that some questions are more discriminating than others, leveraging those questions where respondent behavior varies the most, while accommodating the sparse nature of this type of data. Second, we introduce a new framework to characterize attentiveness. Whereas previous work assumes inattentive respondents rush through surveys, we note that inattentive subjects may also be distracted, focusing on other tasks other, and thus exhibit longer response times. We therefore propose a threefold classification of survey takers, with both fast- and slow-inattentive respondents as well as attentive respondents, allowing for non-monotonicity in the relationship between response time and attentiveness.

We systematically validate the use of response time as a proxy for attentiveness by comparing it to other commonly-used measures of attentiveness. Data from an internet-based survey designed to reflect the census distribution of key variables show that RTAC is consistently able to identify survey respondents who are less likely to pass instructional manipulation checks (IMC), less likely to pay attention to the direction of a Likert scale, exert less effort in open-ended questions, and produce significantly weaker experimental treatmentment effects. We replicate these results with other internet surveys, reflecting different survey vendors and respondent recruitment methods. In short, we show not only that RTAC is able to identify inattentive respondents as well as or more effectively than IMCs, which are the current standard practice, but also that it captures attentiveness more effectively than the current practices for using response time data.

The paper proceeds as follows: We first give an overview of existing methods to estimate respondent attentiveness through response time, and highlight their limitations by exploring what response times actually look like in survey data. In the next section, we extend the theoretical framework of previous work that relied on a simple "fast and inattentive" and "average and attentive" dichotomy by introducing a third category of survey respondents: "slow and inattentive." Next, we describe the two statistical techniques we use to estimate response-time attentiveness clusters. Principal-component analysis (PCA) extracts the maximum possible information from the data by accounting for the fact that response times to many questions are similar and thus provide little information concerning the respondent's attentiveness. We then apply an unsupervised clustering algorithm to those PCA weights, which frees us from the need of making any ad hoc decisions about what counts as a "fast" or "slow" response. We next validate this measure against other commonly used measures of attentiveness, including IMCs. We conclude by discussing the usefulness of the measure and modifications that researchers may make.

# 2 The Limitations of Existing Methods

Researchers have a number of tools to assess attentiveness on self-administered surveys. Among those most often used are IMCs or screener questions (e.g. Oppenheimer, Meyvis and Davidenko 2009; Berinsky, Margolis and Sances 2014). These questions mirror other regular survey questions in length and format but ask the respondent to ignore the standard response format and instead confirm they have read the question by providing specific answers. Researchers can then analyze participants' responses to this question to identify those who carefully read and followed the hidden instruction. However, introducing screener questions comes at a cost. Scholars recommend the inclusion of multiple such questions to measure attentiveness accurately (Berinsky et al., 2019), which means increased questionnaire length and completion time. This strategy might also result in greater respondent fatigue, and thus influence responses to later questions (Alvarez et al., 2019). Moreover, there are no standardized screener questions, and question length can be a powerful predictor of a respondent's likelihood to fail the screener (Anduiza and Galais, 2016).

An unobtrusive alternative is the use of survey response time to assess respondent attentiveness. Response time (or response latency) is the amount of time a respondent takes to answer a question, measured as the number of seconds between the respondent's first click onto a page and last click leaving a page. Response times are an example of paradata, like mouse-clicking or eye-tracking patterns, that provide insight into how respondents are taking surveys (Yan and Olson, 2013). Response time correlates with other proxies of response quality, such as self-reported effort (Wise and Kong, 2005), attention to detail (Börger, 2016), response consistency throughout the survey (Wood et al., 2017), straight-lining (Zhang and Conrad, 2014), and susceptibility to survey design effects (Malhotra, 2008).

Yet how to effectively use response time data remains an active area of research. As Fazio (1990, 89) writes, "there may be nothing scientifically less meaningful than the simple observation that subjects responded in X milliseconds." Making these data meaningful involves three steps. First, researchers must decide whether to only use response time indicators for some survey items, or to aggregate all items together into one response time metric. Second, they must decide which response times represent markedly fast and slow answers, and determine cut-off thresholds for categorization. Finally, researchers must determine how respondents' response time metric maps onto the latent measure of survey-taking behavior that researchers wish to examine. For example, are slow respondents distracted or confused?

To address the first issue of how to aggregate response time patterns across questions, many studies assess response time globally (Malhotra, 2008), looking at the raw total response time for survey completion. Other researchers develop attentiveness measures that look at single question response times (Zandt, 2002) or compare response times within or between specific modules (Vandenplas, Beullens and Loosveldt, 2019) and experimental conditions (Fazio, 1990). From there, researchers will often calculate an aggregate score that indicates the proportion of survey questions which the respondent answered above or below a time threshold (see, for instance, Wise and Kong 2005; Barge and Gehlbach 2012; Greszki, Meyer and Schoen 2015; Yan et al. 2015).

However, there is no consensus as to what such a response time threshold should look like; whether it should be a common threshold, or dependent on the question content or the distribution of response times (Kong, Wise and Bhola, 2007; Huang et al., 2012). Even if researchers do not pick a threshold and use the raw data, they still must decide whether attentiveness is increasing or decreasing with response times, or if the relationship is curvilinear.

This brings us to our third conceptual decision. Researchers need to determine what it means substantively when respondents rush through or drag throughout surveys. Here, researchers are divided in their interpretation of response time. One cohort of scholars argue that response time is a measure of attitude accessibility (Huckfeldt et al., 1999; Mulligan et al., 2003; Johnson, 2004) or the clarity of the survey instrument itself (Bassili and Scott, 1996; Olson and Smyth, 2015). For these researchers, long response times indicate either that respondents are struggling to connect attitudes to questions, or that the survey instrument is impenetrable. Notably, these scholars largely focus on response time in interviewer-administered surveys (face-to-face or phone surveys), where respondents are less able to multi-task than with self-administered web surveys.

Researchers examining response time in *web-based* surveys, however, are usually concerned with the fast end of the spectrum, suggesting that when respondents answer very quickly, they have not taken enough time to understand the question and provide an accurate answer. Instead, they are satisficing (Callegaro et al., 2009; Greszki, Meyer and Schoen, 2015). The assumption here is that rushing respondents are inattentive. We join these scholars in arguing that response times are indeed a clear indicator of attentiveness, rather than respondent comprehension. Our

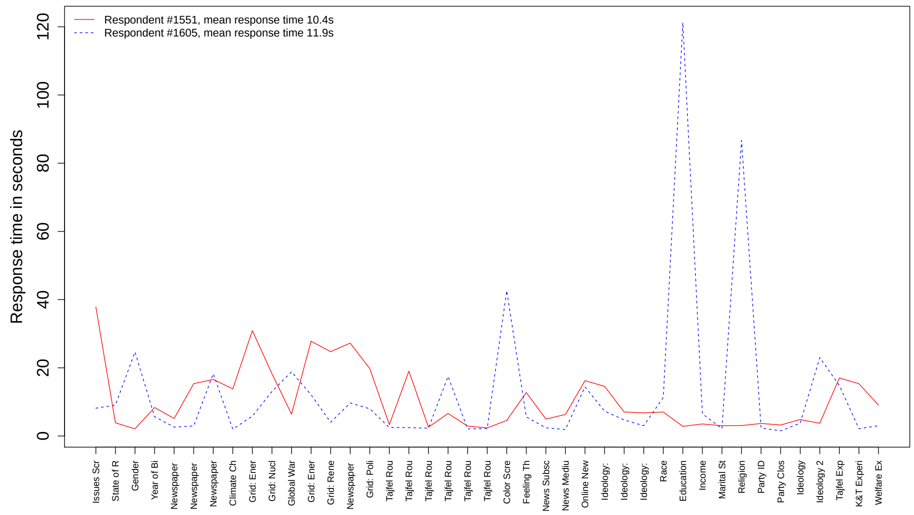



Figure 1: **Average response time conceals important cross-question variation:** This figure shows the per-question response time for two respondents who have similar average question and global survey response times. Respondent #1551 has an average question response time of 10.4 seconds, 1.5 seconds less than respondent #1613, a difference of less than a one-hundredth of a standard deviation. Despite similar average response times, these two respondents behave very differently. Respondent #1551 spends more time on complex grid and ranking questions (`Climate Ch` and `Grid`), whereas respondent #1605 dwells on questions that ask about the respondent's gender, favorite color, educational background, religion, and ideology.

validation exercises show that this is likely the case. While we generally agree that response time can be an indicator of respondent attentiveness, we argue that the relationship between the two variables is not strictly monotonic. In particular, respondents' ability to multi-task and propensity to be distracted in web-based surveys means that very long response times, not just particularly short ones, can also be indicative of inattentiveness.

To highlight these points, we visualize trends in actual response time behavior. We present data on response time from a survey fielded in 2016 using Survey Sampling International (SSI)[2] to illustrate response time fluctuation throughout a survey, and the degree to which different questions provide vastly different amounts of information about survey-taking behavior.

The first important trend that emerges from the data is that global measures of response time obscure important within-respondent variation. Response time behavior often fluctuates throughout the course of a single survey. Figure 1 shows the per-question response time for two illustrative survey takers. Both respondents have similar average and global response times; the difference between their aggregate response times is less than a one-hundredth of a standard

[2]SSI constructed a target population that matched the census population on education, gender, age, geography, and income for a sample of 2,952 respondents. We provide an overview of the surveys we use in this analysis in Appendix C. We also present replications of these findings using other data sources.

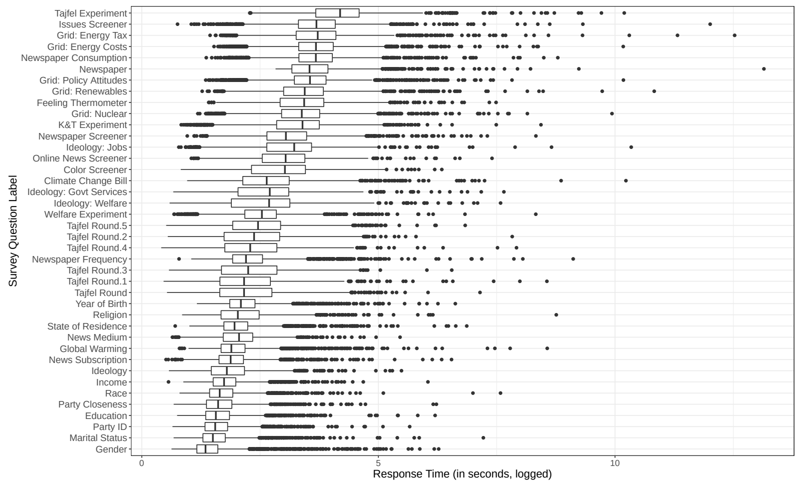



Figure 2: **Questions with larger response time variation provide more information on attentiveness:** This figure displays a boxplot for each question contained in our data with the distribution of response time across respondents. We can see that for some questions, e.g. `Party ID`, the bulk of the respondents are relatively fast. On other questions, however, there are respondents who are both much faster and much slower than the middle 50% of the response times (e.g., `Grid: Energy Tax`). These latter types of questions are likely to better distinguish attentive from inattentive respondents than the former type.

deviation. Yet they behave very differently. One respondent takes considerably more time to answer complex ranking and grid questions. The other spends significantly more time on basic factual questions about the respondent's gender, education, religion, and ideological affinity. These important differences are concealed under any approach that looks at aggregated response times.

Moreover, such an approach inherently assumes that each question provides the researcher with the same amount of information about the respondent's overall attentiveness. This assumption seems problematic. When the question is short, the answering process is straightforward. Both attentive and inattentive respondents should take similarly answer quickly. It is when respondents encounter longer and more complex questions that actually require respondents to process the question setup and think carefully about their answers that we should expect response times to be a good indicator of a survey-taker's attentiveness. To illustrate, Figure 2 shows boxplots for the response times for all questions in our 2016 survey. The distributions highlight that there is a great deal of variation in the amount of information contained in each question. Importantly, for some questions, most respondents are relatively fast. On other questions, however, there are respondents who are both much faster and much slower than the middle 50% of the response times. These latter types of questions are likely to offer us more information on respondent attentiveness than the former type. Figure 2 clearly suggests that it is in more complicated questions, such as experiments (`Tajfel` and `KT Experiment`), that we can find the most response time variability.

Figures 1 and 2 also highlight a final important point: some respondents can take a very long time to complete survey questions. The existing literature often overlooks this type of slow respondent. Focusing primarily on "rushing" respondents, existing measures assume that while extremely short response times are indicative of satisficing. That is, past a certain extremely short threshold time response times do not provide much information about attentiveness or satisficing. Yet there is a substantial amount of variation to explore at the *slower* end of the distribution. Simply put, if rushing were the only deviation from typical survey-taking behavior, we would see more outliers towards the fast end of the scale, and fewer towards the slow end. Instead, modeling respondent behavior on self-administered online surveys needs to take into account internet browsing behavior, which may include switching between different tabs and engaging in several activities at the same time. Using response time as a proxy for respondent attentiveness could therefore benefit empirically from incorporating all respondent data, and theoretically from introducing slow respondents as a third category of inattentive respondents.

# 3 A New Approach to Estimating Response Time

The first step in our approach lies in conceptualizing attentiveness as a latent variable.[3] It is not a single observable variable but a multi-dimensional concept that incorporates how closely respondents read questions, how deeply they understand the text, how thoughtfully they answer the questions, and how well they are able to remember survey content over several pages of questions. None of these metrics, however, can be directly observed and recorded by the researcher; thus, researchers rely on observable behaviors to proxy for respondent attentiveness. To translate response time paradata to attentiveness, we introduce response time attentiveness clustering (RTAC) to proxy for attentiveness.

## 3.1 The theoretical framework: Three types of respondents

When one thinks of inattentive respondents, people typically envision a respondent who rapidly clicks through each question without taking the time to even read the question text. Yet individual internet behavior is often less focused than this approach assumes. Fast respondents may rush through each question, but answering *all* questions quickly means that respondents are focused

[3] More attentive respondents can increase internal validity. These respondents take the time to read the questions, read newspaper articles or watch YouTube videos that comprise our survey content. We do not, however, assume that these are *higher quality* respondents, and make no conclusions regarding a survey's external validity and respondent attentiveness. We can think of attentive respondents as experimental compliers which affects our estimation of the average treatment effect, but says little about the generalizability to the general population on the basis of attentiveness.

on the survey itself, even if they are not paying attention to the content.

Although "fast" respondents are probably not paying close attention to the survey, slow respondents might be distracted as well. It is easy to imagine two different types of inattentive respondents who exhibit very different response-time behavior. The first respondent rushes through each question, finishing the survey as quickly as possible. The second respondent flips back and forth between texts with friends, emails, and the survey. On some questions, this respondent is very fast, clicking through as quickly as possible. Yet on other questions, the respondent is unusually *slow*, as he loses focus on the survey and his attention moves to other tasks. Any method to distinguish between attentive and inattentive respondents must account for the behavior of both types of distracted respondents, not just the consistently fast ones.

Multi-tasking is indeed prevalent in online surveys. Ansolabehere and Schaffner (2015) found that between 25% and 50% of respondents engaged in at least one non-survey task during a survey. Respondents who reported distractions took longer than those who did not. Sendelbah et al. (2016) found that 62% of respondents multi-tasked during a survey. Half of those respondents both took an unusually long time to complete survey questions *and* clicked away from the survey, while 32% exhibited long response times without clicking away from the browser.[4]

Theoretically, if multi-tasking respondents are slow when distracted but rushing when focused on the survey, distracted respondents should exhibit high variance in their response times. High variance and low-quality responses should distinguish inattentive slow respondents from those who take more time to read and comprehend questions. In that case, respondents might be just as attentive and engaged with the survey as others, yet might face difficulty accessing attitudes or understanding questions. Table A.1 summarizes this theoretical framework, and provides a typology that links exhibited response time behavior to assumptions about survey-taking behavior. Our validation strategy is designed to test the assumption that slow-classified respondents are inattentive; we show that slow inattentive respondents provide very short answers to an open-ended question, illustrating that they are not providing high-quality responses.

[4]The presence of distracting activities that take place away from the computer makes it difficult to capture such inattentiveness with the paradata used most often to to study multi-tasking behavior, such as mouse-click and browser tracking. Indeed, while Ansolabehere and Schaffner (2015) and Sendelbah et al. (2016) show that up to 62% of respondents report distractions, a relatively smaller share of just 15% actively switch from the survey to a different website (Höhne et al., 2020). We therefore opt to use only response time data, which captures on- and off-line multi-tasking behavior and is very easy to measure in web-based surveys.

## 3.2 The empirical framework

Thus, conceptually, there are not only different types of inattentive respondents, but also respondents who take surveys at similar paces might be different. Response time, as a measure of attentiveness, is therefore a multi-dimensional concept. The average response time across all questions tells us one thing, but the variation in how long respondents spend on each question tells us another crucial piece of information about attentiveness. Figure A.2 in the Appendix provides empirical support for this interpretation, showing that there is a strong and positive correlation between response time and variance for respondents across all questions, suggestive of a different data generating process for very fast and very slow respondents. RTAC takes both these patterns – average time across all questions, and fluctuations in timers across all questions – into account.

Building on previous research on response time, and attempting to account for some of its shortcomings, RTAC needs to perform several specific functions. First, it needs to use all the response time data in a parsimonious fashion to avoid making arbitrary decisions about which variables to include or exclude. Second, it should include a tri-fold classification of survey attentiveness to allow for the inclusion of slow respondents and account for non-monotocity in the relationship between response time and attentiveness. Third, it should provide a disciplined way of identifying who is slow or fast that is based on similarities in response time behaviors across respondents, rather than subjective cut-off thresholds for response time.

For this, we rely on two methods: principal-component analysis, which extracts the maximum amount of variation from the data to provide an optimal low-dimensional representation of the data and captures multiple dimensions of variation (e.g. total time and variation in timing between questions), and expectation-maximization, which fits a Gaussian mixture model to cluster our response time data into categories based on underlying similarities in the data. Together, these methods meet the criteria for RTAC expressed above.

Figure 3 provides an overview of the process. After transforming the response time data to avoid over-fitting the clusters of respondent attentiveness, we perform two empirical steps. First, we pre-process the response time data with PCA to reduce its dimensionality. This reduces the number of dimensions of response time that we need to incorporate into our model from a dimension for each question to a smaller, parsimonious set of uncorrelated dimensions. If we believe that each per-question response time indicator does not capture wholly different information about how a respondent takes a survey, then PCA will collapse the response time data into a set of orthogonal dimensions that combine similar features of the data.

We use the transformed response time data to cluster respondents into three attentiveness categories. To do this, we use the EM algorithm to estimate membership in a Gaussian mixture model. This type of model assumes that the data at hand are not drawn from the same one distribution, but rather from multiple distinct normal distributions, each with its own parameters (mean and variance). The algorithm then estimates both the shape of those normal distributions, and calculates the probability that a data point belongs to each of those distributions. Under our theoretical framework, we assume there to be three distinct groups of survey-takers, and each is drawn from a distribution of response times with a distinct mean and variance: fast/inattentive, slow/inattentive, and baseline/attentive. The algorithm we employ therefore calculates the probability that a respondent belongs to one of three clusters, each representing a different type of survey-taking behavior. From here, we explain the methodology in greater detail. Readers who are less interested with the technical details behind the method should skip to Section 3.3.

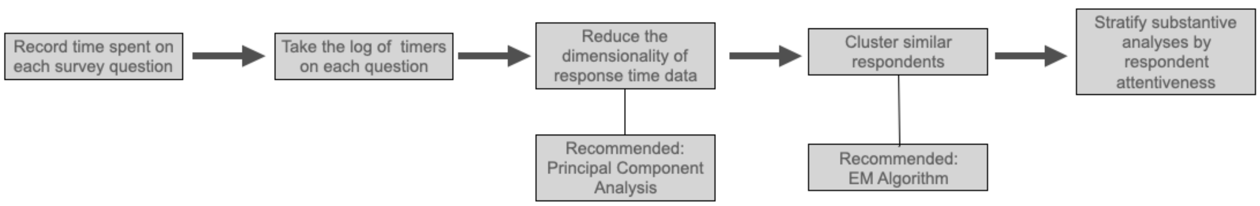



Figure 3: **Overview of Estimation Process**

The first step of our approach is pre-processing: taking high dimensional data – response time for each question – and condensing them such that the data are parsimonious while still capturing sufficient variation to characterize respondents. This focuses our analysis on those parts of the data where we can find the most information about respondents' survey behavior. As highlighted in Figure 2, many survey questions contain very little information about how long different respondents take to answer the questions (i.e. have low variance), while others are much more discriminating. Therefore, the full matrix of per-question response times is likely to have a high noise to signal ratio. To reduce such noise, researchers might be tempted to subset the data to only those questions with high variance in response times; however, this would require a subjective evaluation of what "high variance" means. We therefore prefer to use principal-component analysis (PCA) to extract the meaningful variance from the data.

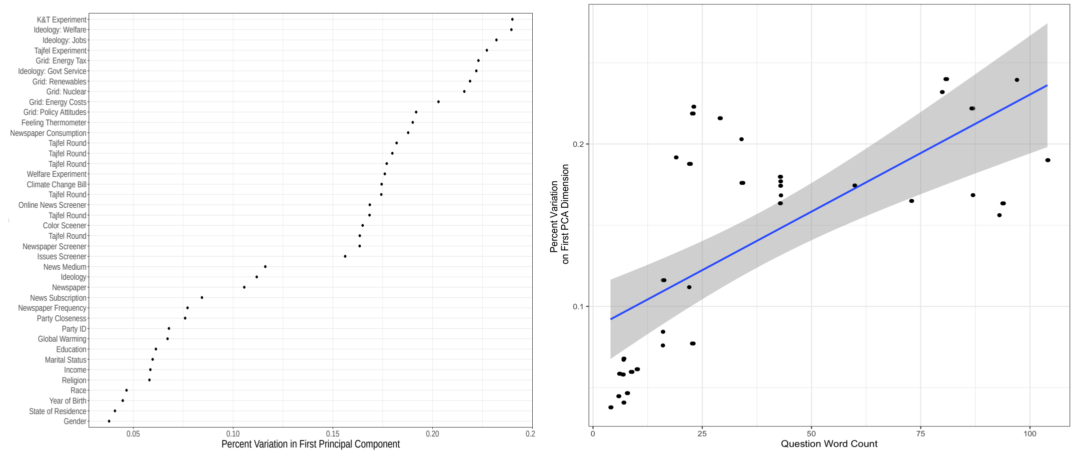
PCA transforms highly correlated variables into a smaller set of uncorrelated variables for the purpose of dimension reduction, allowing researchers to use a parsed-down number of variables to represent variation in the original data. By projecting data onto a lower-dimensional space, the PCA algorithm identifies directions in which the data vary. The first principal component contains the most variation, while the second principal component contains the most variation *orthogonal* to the first component. The process continues until it has projected the data onto a $k$-dimensional space, where $k$ is the number of variables in the original dataset.

We proceed by performing a PCA on a matrix of logged response times[5] for each question and each respondent.[6] Often, researchers use PCA to develop an index measure constructed of several correlated measures. For example, if someone wanted to measure ideology using a battery of questions concerning political beliefs, they might run a PCA, extract the first two dimensions, and name them the "economic" and "social" dimensions of opinion. However, we use PCA for a different purpose. We employ PCA to weight the response time indicators according to how much information each indicator contains about overall response time behavior. Each variable input to the PCA measures the same thing: how respondents interact with the survey instrument. The PCA extracts the maximum amount of information from that data and provides us with appropriate weights that allow us to glean as much information as possible from response time without having a very sparsely populated dataset. Thus, unlike most PCA users, we are not interested in interpreting what each PCA dimension means.

That said, examining the PCA loadings can help illustrate how the PCA is working in practice. We show the first component's loadings - i.e., the weights that indicate the relationship between each variable and principal component - in Figure 4. As expected, we observe that longer and more complex questions, including those of two survey experiments (`KT Experiment` and `Tajfel Experiment`), those where respondents had to evaluate their own position on a scale (questions starting with `Ideology:`), and large grid questions (those starting with `Grid`) are the main variables loading onto the first dimension. These variables therefore contribute a great deal of the variation contained in the first component. Indeed, we would expect longer and more complex questions to exhibit the greatest amount of variation as attentive respondents will take the time to read a block of text, whereas fast respondents will continue rushing through. In contrast, easy

[5] We still log the matrix of response times to transform the values because of the skewed nature of the data. This ensures that the extreme right-skew will not dominate the variation observed in the PCA and favor explaining the variation primarily among slow respondents. Figure A.1 in the Appendix presents the distributions of both the raw and transformed data. As this figure shows, the raw data have a few points (very slow respondents) that are very distant from any other points. Mechanically, when we use an EM algorithm to fit a Gaussian mixture model using the raw data, the algorithm will try to fit a mixture around a single outlier point. This distribution, consisting of only one point, will have zero variance, and the log-likelihood function goes to zero (Bishop, 2006, 433). More generally, normalizing the data addresses the common problem of over-fitting in EM-estimated mixture models, of which the singularity problem is an example.

[6] Because PCA requires a matrix of complete data, we are forced to drop respondents who skipped over some questions or exited the survey early. Researchers could consider estimating survey attentiveness for only parts of the survey if there are concerns about attrition.




(a) (b)

Figure 4: **Variable Loadings on First Principal Component:** This plot shows the degree to which question response times are included in the first component of the PCA, and the complexity (word count) of those questions. Complex questions, including experimental questions (`KT Experiment` and `Tajfel Round`), grid questions (those starting with `Grid:`), and scale questions (those starting with `Ideology`) contribute the most to the first component.

questions that require very little processing and recall, like questions about the respondent's race, year of birth, and gender, contribute the least amount of the variation to the first component.

We confirm this relationship in Figure 4b, showing that there is a positive trend between word count and contribution to the PCA's first dimension. In plain English, this means that longer questions have the most variation in response time, providing the PCA a good deal of variation with which to work. Longer questions are most informative. However, this does not mean we should only include the first principal component in our analysis. While the first dimension captures variation in response times for long and complex questions, subsequent dimensions capture variation in shorter questions and within-respondent differences across questions (e.g., respondents who take a long time to answer grid questions but are quicker on scale questions). Those dimensions therefore contain meaningful information about survey-taking behavior that should not be discarded, as is often done in response time aggregation methods.

There are few clear methods for deciding how many components to retain in dimensional reduction. The central question to ask is how much variation is needed to understand the patterns in the data. There is no correct answer to this question; inherent in this decision is a trade-off between capturing variance (retaining *more* components), and losing parsimony (retaining *fewer* components). Some researchers rely on the visual examination of a scree plot, which shows the eigenvalues of the principal components, to determine the point at which these eigenvalues level off. We prefer to follow the procedure of examining the proportion of variance explained cumulatively by the principal components, and retain those components that explain 80% of the total variation. In our case, this means retaining the 19 first principal components. We examine the robustness of this decision in the section *Sensitivity to Researcher Decisions*, and show empirically in the Appendix that estimation becomes stable once about 80% of the variation is used, as well as provide an example of how users can conduct their own robustness checks on PCA.

The PCA weights provide the pre-processed inputs used to assign attentiveness cluster membership for each respondent. Given the latency of respondent attentiveness as a variable, previous work relies on observable outcomes (choice patterns, IMC passage, response time) to group respondents into different categories of survey attentiveness. Grouping this latent variable into a discrete number of categories allows researchers to have a number of theoretically-driven clusters to categorize respondents, providing comparability across surveys with different distributions of respondent attentiveness, and numbers and types of questions.

Clustering algorithms lend themselves particularly well to the task of estimating latent attentiveness. Generally speaking, these unsupervised machine learning techniques can identify similarities in data points and thus group similar data points together. In our case, we use the response time data weighted by the first 19 principal components as our input data. A clustering algorithm can then group respondents according to similar patterns in their response times, using both response times and their variability within and across respondents. This approach presents a number of advantages over other measurement strategies. First, as an *un*supervised machine learning technique, we are not required to make ad hoc decisions about which response times count as slow, baseline, and fast. Previous literature made idiosyncratic decisions about the answer time thresholds that classified as "too fast" (see Kong, Wise and Bhola (2007) for a discussion). An unsupervised clustering algorithm does not require such decisions. It groups data points according to their multidimensional characteristics. The researcher can then inspect those groups and ascertain which one is faster or slower. Second, the input data for such algorithms can be multidimensional. In other words, instead of limiting ourselves to just using one data point per respondent, such as the global or average question response time or response time during survey experiments (Harden, Sokhey and Runge, 2019), we are able to retain *all* the data and use it to estimate the parameters of multivariate normal distributions.

Many methods for clustering exist, with centroid models (e.g., K-Means clustering) and distributional models like Gaussian Mixture Models (GMMs), which are fit with methods like expectation-maximization, being two main contenders for this task. The K-Means algorithm finds groups by starting with initially random group centers, and then assigns each data point to the group center to which it is closest, improving the clusters in each iteration. This approach suffers from a few drawbacks. First, data points are "hard assigned" to a group; group membership is binary and researchers are unable to see the probabilistic assignment into a particular group. Second, the underlying models from which these data points are generated are assumed to be centroids and not full distributions. This is problematic for our approach because theoretically, we think that slow, distracted respondents would have different distributions of response times across questions, rather than simply different means.

Gaussian mixture models, fit with an expectation-maximization (EM) algorithm, find groups by determining a mixture of normal distributions that best fit the data. More specifically, this model assumes that each group is a Gaussian distribution with a mean and variance-covariance matrix, and each data point has been drawn from the distribution associated with its group. An EM algorithm proceeds iteratively. In the $E$-step, given the current assignment of data points to groups or distributions (initially done at random), each groups' distribution's parameters are updated. In the $M$-step, we estimate the posterior probability of group membership given each group's updated parameters. In other words, the algorithm looks for a better group assignment and reassigns data points to the distributions they are most likely to come from, conditional on the distribution's updated mean and variance-covariance matrix. The algorithm iterates through this process until it converges, and produces as an output the maximum a posteriori (MAP) estimate of each respondent's probability of belonging to each group. This approach presents a number of advantages. First, it allows for three different categories of respondents, each of which behaves very differently in responding to the survey, corresponding to different time distributions. Second, it provides us with a "soft" group assignment that indicates the probability that a respondent belongs to a specific group, rather than a hard assignment that leaves no room for uncertainty. We can thus determine how likely a respondent is to belong to a certain cluster.

In this algorithm, random variable $x$ is the weighted sum of a mixture of $K$ Gaussian distributions or groups. We treat respondent group membership as an unobserved variable $z$ and are then able to calculate the joint distribution of $p(x, z)$ according to the marginal distribution of $p(z)$. With latent variable $z$, respondents take a value of 1 if they are assigned to group $z_k$ and a 0 otherwise. Finally, we are also able to calculate the conditional probability of $z$ given $x$, which allows us to estimate the posterior probability of membership in each of $k$ categories according to the observed $x$. The EM algorithm also allows us to calculate the responsibility for each observation, which is the probability that each observation fits into each of $k$ categories. These probabilities sum to 1.

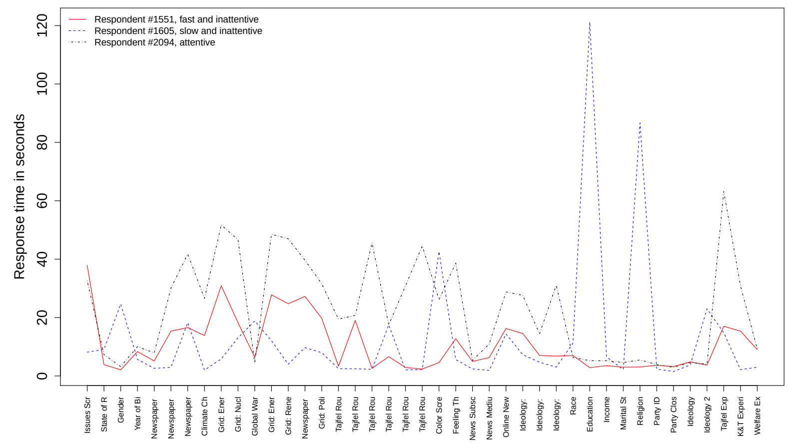



Figure 5: **Different response time patterns in each cluster:** This figure displays the response time for each question for three illustrative respondents, one from each cluster. The respondent assigned to the attentive cluster takes considerably longer on the more complex questions, including grid and ranking questions (`Climate Ch` and those starting with `Grid`), as well as experimental questions (`Tajfel` and `KT Experiment`). The respondent assigned to the fast and inattentive cluster exhibits a similar patterns but takes significantly less time on more complex questions. The respondent assigned to the slow and inattentive cluster takes a considerable amount of time to answer supposedly easy questions about their race, gender, and ideology, and is much faster to answer many of the more complex questions.

Given our theoretical motivation, we model a Gaussian mixture distribution with three distinct Gaussian distributions or groups, using the response time scores weighted by the first 19 principal components. We find that the algorithm clearly assigns the vast majority of respondents into one of the three categories. Graphic inspection of the EM sorting, available in the Appendix, shows that for each cluster, most respondents have a responsibility of either 1 or 0, meaning the algorithm assigns them probabilities of close to 1 or close to 0 of belonging to that group. From there, we hard assign each respondent to the cluster of which he or she is most likely to be a member (that is, the cluster with the highest posterior responsibility) (McLachlan, Lee and Rathnayake, 2019).[7]

We then inspect the group membership assignment more closely. Figure 5 shows the per-question response time from three illustrative respondents from our survey, one from each cluster. Recall Respondents #1551 and #1605 from Figure 1. Despite having similar global and average response times, these two respondents behave very differently. Respondent #1551, assigned to cluster 1, takes relatively longer on more complex ranking, grid, and experimental questions (e.g., those starting with `Grid:`) while respondent #1605, assigned to cluster 3, dwells on questions

[7] If researchers had a distribution where the algorithm was not able to decisively classify observations in any cluster (having posterior probabilities of .33, .33, and .33 for example), researchers could examine patterns among respondents whom the algorithm *was* able to classify with high probability. For one method for identifying observations that are statistically significantly consistent with one cluster or another, see Imai and Tingley (2012).

about their gender, education, and religious affiliation. Respondent #2094, assigned to cluster 2, takes the longest on the longer and most complex questions, including the experimental questions.

Consistent with this qualitative inspection, when we examine the mean global response time per cluster, respondents in the slowest cluster (n = 362) have much higher variance compared to respondents in the fastest cluster (n = 1,168) and the baseline – attentive – cluster (n = 987). Respondents assigned to cluster 1 are the fastest, cluster 2 the baseline respondents, and cluster 3 the slowest. We therefore label the respondents assigned to the first and fastest cluster *fast inattentive*, those in the third and slowest cluster as *slow inattentive*, and those in the middle cluster *"baseline duration"*, against whom we benchmark attentiveness. These frequencies correspond to 46% of our respondents being classified as fast inattentive, with 40% of respondents being classified as baseline.[8] Only 14% of the respondents are classified as slow inattentive, but as we show below, the algorithm fails to distinguish between the fast and attentive mixtures when we do not model and estimate a third cluster. The variance of global response time in each group also provides insight. While the global response time of the first and second cluster are just over a minute apart, their variances are quite different, pointing to very different distributions between the two groups.

### 3.3 Sensitivity to Researcher Decisions

As previously mentioned, there are no hard-and-fast rules for deciding how many components of the PCA to include, or how many clusters to model and estimate. Rather, researchers must decide the appropriate trade-off between variation ands parsimony for their particular needs. Yet we also want to ensure that our measure is not highly sensitive to the number of included components. Figure B.2 in Appendix B shows how our clustering of attentiveness changes according to the inclusion of different numbers of PCA components. In Figure B.2, we can see that once we include 80% of the variation or more, the number of observations assigned to each cluster begins to stabilize. Because the stabilizing point may change in other surveys, we recommend that researchers plot the respondent cluster assignment according to the number of PCA dimensions used and ensure that the number of dimensions where cluster assignment has stabilized by the point where researchers select the dimensions to include.

[8] This is roughly similar to the proportions of attentiveness as measured by screener questions. In the same survey, only 18% of respondents answered all four screener questions correctly, while 56% of respondents answered zero or only one screener question correctly.

# 4 Validation Strategy

RTAC clearly captures different speeds at which respondents take the survey, to what extent are we actually capturing attentiveness? Moreover, are those survey-takers assigned to the *slow* cluster actually slow *and* inattentive? Or are they simply slow because of potential cognitive restraints, yet still attentive and willing to spend time and effort to answer questions? We conduct three validation exercises to show that three-state classification by response time does appear to capture respondent attentiveness. The results of these three validation exercises are presented in the main text, while results from replications using different surveys are presented in Appendix C. By validating this approach across surveys, we show that our findings are generalizable across a variety of respondent pools, survey designs, and online survey firms.

First, we look at respondents' answer to an open-ended question in which they were asked to share as much or as little as they like with us about the most recent book they read or movie they watched. We hypothesized that attentive respondents are likely to provide longer answers, whereas inattentive and satisficing respondents would write much shorter answers. If those assigned to the *slow* cluster are not satisficing and instead attentive and willing to exert effort, we would expect them to write answers of similar length than those who are assigned to the *baseline* cluster. If they are indeed inattentive and trying to minimize the amount of effort needed to complete the survey, the word count of their answers should be closer to that of the *fast inattentive* cluster, whose members are rushing through the survey.

Figure 6 reports the results from that test, showing the average number of words provided in the answers of respondents to this open-ended question by assigned cluster. Not only do we find that those assigned to the *baseline* cluster write the longest answers (roughly 160 words), we also confirm that those assigned to the *slow* cluster provide on average the shortest answers (about 80 words), meaning they are not just slow in answering survey questions but also exert minimal effort. A qualitative inspection of the answers confirms this: respondents in either of the inattentive clusters tend to write just the title of their most recent book or film and give one-sentence summaries, whereas those assigned to the attentive cluster often describe the plot at length, and detail why they chose to read the book or see the film.

We then evaluated how our cluster assignment corresponds to two other survey patterns that provide hints as to respondent attentiveness: responding to a well-replicated and classic survey experiment, and noticing a flipped scale in a series of ideological questions. We first replicate a survey experiment to see how attentive and fast and slow inattentive survey-takers respond to a

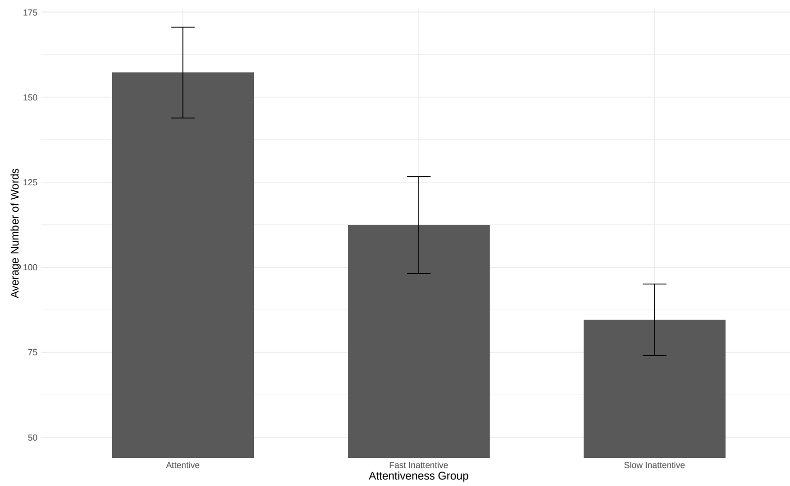



Figure 6: **Word Count of Open-Ended Question by Assigned Cluster:** This plot shows the average number of words provided in an open-ended question that asked about the respondents' most recently read book or watched movie or TV show. With roughly 160 words, respondents assigned to the *baseline* cluster provide the longest answers on average, while both those taking less and more time to complete the survey, i.e., respondents assigned to the *fast inattentive* and *slow inattentive* groups, exert much less effort and provide much shorter answers, with 112 and 80 words respectively.

classic experimental treatment. Previous work had noted that inattentive survey respondents may introduce enough noise in the data to attenuate or even nullify experimental results. Berinsky, Margolis and Sances (2014) in particular show that the average treatment effect of a well-known and widely replicated survey experiment first introduced by social psychologists Tversky and Kahneman (1981) is substantially lower for those respondents who fail to pass IMCs. The experiment asks survey respondents about their preference among two proposed policies to stop the outbreak of a contagious disease, where some respondents receive the proposed policies framed as potential gains and others as potential losses. We too replicate this survey experiment, finding that the change in framing of the disease is associated with a roughly 27% change in support.

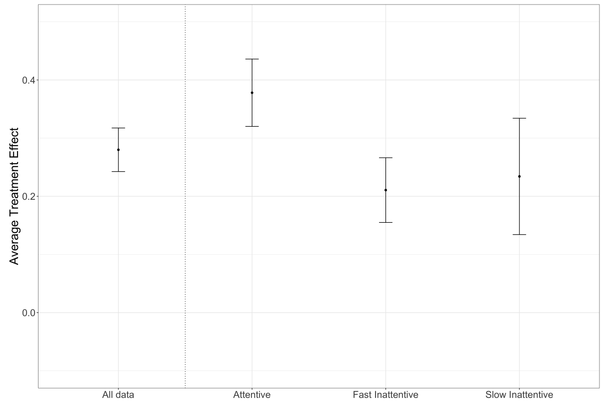
Yet when we stratify by attentiveness cluster, a different picture emerges. In addition to the global average treatment effect, shown as the dot-dash line, Figure 7a also displays the estimated treatment effect for each assigned cluster of respondents. Respondents in the attentive cluster display a high and significant treatment effect, whereas those assigned to one of the inattentive clusters have much lower and close to null effects. When examining the treatment effect only among those who complied with the treatment by paying attention to the survey, the average treatment effect jumps almost ten percentage points. Fast inattentive respondents have a significantly smaller treatment effect, and slow inattentive respondents have a smaller and much noisier treatment effect.

Of course, if we take seriously concerns about post-treatment bias, some researchers may be concerned that stratifying on response time for an entire survey, in which some items are measured post-treatment, would complicate interpretation. This issue can easily be addressed if researchers omit post-treatment timers from their response time dataset. In Appendix B, we show that omitting timers from either experimental items alone, or all post-treatment measures, does not meaningfully change the replication exercise. In other words, by using RTAC to identify attentive respondents, researchers can stratify their analysis on respondent attentiveness – regardless of whether they include or omit post-treatment timers.

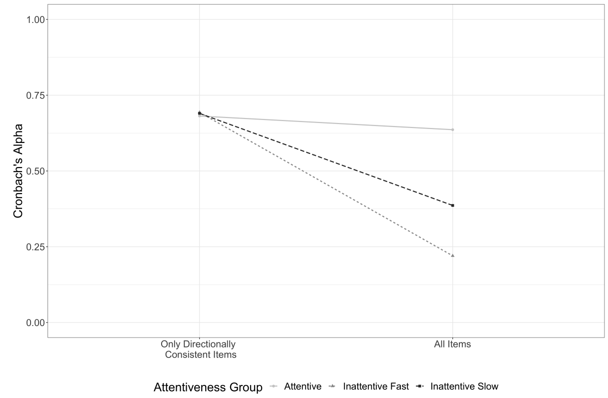
Finally, we examine attitude consistency. Previous work identified high correlations between questions measuring similar attitudes and whose wording requires close reading as a marker of attentive survey respondents (Berinsky, Margolis and Sances, 2014; Alvarez et al., 2019). We therefore expect those respondents labeled as attentive to be more consistent on a number of topically related questions with response scales coded in different directions than respondents classified as inattentive. These questions, which are designed to test ideologically consistent concepts, ask about attitudes towards government employment support, government involvement in health and education, and income redistribution. The first and third questions anchor more liberal values (a more interventionist government that provides public services) on the left-most side of the scale, and more conservative values on the right. In the second question, this scale is flipped. More interventionist beliefs are associated with *higher* values on the right-most side of the scale, while more conservative beliefs are on the left-hand side of the scale. If respondents are not taking the care to read the questions, they will likely pick similar points on the scale across all three questions, failing to catch the subtle directional change.

We use Cronbach's Alpha to measure consistency across those three ideological questions. A higher Cronbach's Alpha indicates that the responses to the three distinct questions are consistent with one another. Therefore, attentive respondents should have a high Cronbach's Alpha across all three items, whereas inattentive respondents should have a lower Cronbach's Alpha, as failing to notice the flipped scale would reduce internal consistency across the measures. We can extend this test by comparing Cronbach's Alpha when the reverse scaled item is included or excluded. If inattentive respondents are answering the items consistently but simply miss the reversal, then they should be highly consistent *without* the Item 2, but exhibit lower consistency when it is included.

Figure 7b shows our results graphically. The left-hand points show the Cronbach's Alpha for only the directional consistent items, whereas the right-hand points show the Cronbach's Alpha for all three items. The graph shows that when only directionally consistent measures are included,




(a) **Framing Experiment Results by Cluster**



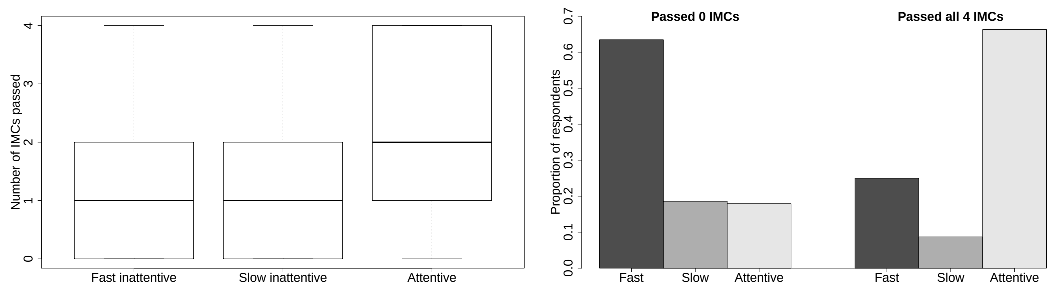

(b) **Cronbach's Alpha**

Figure 7: This left-hand figure shows the average treatment effect of a well-known and replicated framing experiment first introduced by Kahneman and Tversky (1981). Survey respondents in the attentive group display a high and significant treatment effect, whereas those assigned to one of the inattentive clusters display a much weaker and close to null effect. The right-hand figure shows Cronbach's alpha for three related ideology questions in which respondents had to position themselves on a scale ranging from liberal to conservative. Crucially, the scale was reversed for one of the questions, meaning that the correlation across the three questions will be lower for those respondents who simply click through the questions without carefully reading the instructions and thus noticing the change in the scale. While Cronbach's alpha is similar for all clusters when computed just over the two questions with the same scale, the coefficient dramatically decreases for all but the baseline cluster when computed over all three ideology questions when the reversed scale item is included.

all groups exhibit similar response consistency. Yet, when the flipped scale item is included, inattentive fast and slow respondents have substantially lower levels of internal consistency, while the level of consistency for attentive respondents remains largely unchanged. This result suggests that our measure of response time is able to discriminate between respondents who are attentive enough to notice and respond to the flipped scale, and those who are not.

# 5 Consistency with Other Attentiveness Measures

The analyses thus far suggest that RTAC can indeed distinguish between fast, slow, and attentive respondents. Yet how does this approach compare to, and improve upon, other attentiveness measures? Although other approaches – particularly IMCs – are imperfect proxies of respondent attentiveness, we should nonetheless expect that those who are categorized as attentive using our approach to be more likely to pass screener questions. However, as we show below, our approach still outperforms competing strategies for estimating attentiveness. We compare RTAC to three alternative proxies of response time: instructional manipulation checks, outlined and used by – among others – Oppenheimer, Meyvis and Davidenko (2009); Berinsky, Margolis and Sances



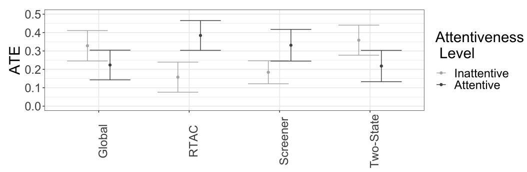

(a) (b)

Figure 8: **Respondents classified as attentive pass more IMCs:** The upper left hand plot (a) shows boxplots for the number of IMCs passed by respondents, disaggregated by the group to which they were assigned. The median number of IMCs passed by a respondent identified as fast and inattentive and slow and inattentive is one, with the 75th percentile correctly answering only two IMCs. Respondents assigned to the attentive cluster, on the other hand, have a median number of correctly answered IMCs of two, with the 75th percentile answering all four IMCs correctly. The upper right hand side plot (b) shows that the vast majority of respondents who pass not a single IMC are assigned to the *fast and inattentive* cluster, whereas those who pass all four IMCs are predominantly assigned to the *attentive* cluster.

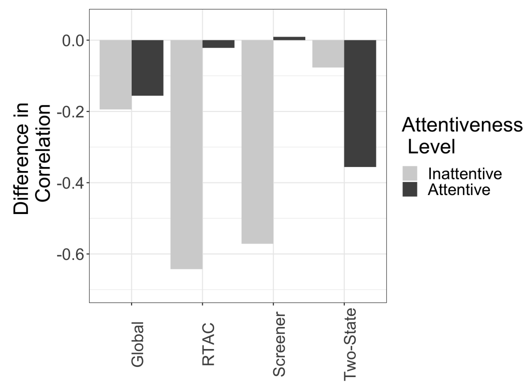
(2014); Berinsky (2017); global response time, as seen in Malhotra (2008), and two-state response time, which distinguishes between only fast and attentive respondents (Greszki, Meyer and Schoen, 2015).

Instructional manipulation checks are one of the most common approaches for identifying inattentive respondents; therefore, it is prudent that we benchmark our approach against this common alternative. To compare to screener questions, we included in our survey four instructional manipulation checks (IMCs) or screener questions, against which we can benchmark our method for estimating latent respondent attentiveness. In Figure 8, we show that screener passage rates are largely consistent with our response-time based measure of attentiveness. The left hand side figure shows boxplots for the number of IMCs passed by members of the three clusters we identified. While those classified as either fast or slow and inattentive pass a median number of one IMC and only two at the 75th percentile, those identified as attentive pass a median number of two IMCs and all four IMCs at the 75th percentile. Yet Figure 8 suggests that screener passage and response time categorization are distinct. The measures are most consistent at the extremes of correctly answering zero or four screener questions, perhaps because of fluctuating survey-taking behavior within the same respondent.

In Figure 9, we show that RTAC outperforms the alternative measures when we stratify by attentiveness group to examine responsiveness to the Kahneman and Tversky survey experiment, and detecting the flipped ideological scale. The left-hand panel shows the ATE for inattentive




(a)




(b)

Figure 9: **RTAC measure outperforms other measures:** Panel (a) shows the average treatment effect for the well-replicated Kahneman and Tversky experiment by attentiveness group. Respondents classified as attentive by both RTAC and Screener questions behave as expected, with attentive respondents have higher treatment effects than inattentive respondents. However, for both global response time measures and the two-state measure, the opposite is true. Inattentive respondents are more reactive to the experimental treatment than attentive respondents. Panel (b) replicates the flipped ideological scale validation measure, but displays the *difference* in Cronbach's Alpha scores including and omitting the flipped scale. Negative differences indicate that the flipped scale resulted in lower consistency between the three items. In both panel (a) and (b), both fast and slow inattentive respondents in the global and RTAC measures are collapsed into one category of inattentive.

and attentive respondents separately (collapsing both fast and slow inattentive respondents into one category). The right-hand panel shows the difference in Cronbach's Alpha scores including and omitting the flipped scale. If a particular measure is a good indicator of attentiveness, we should expect there to be little difference between the Cronbach's Alpha when the flipped scale in included or excluded. For inattentive respondents, there should be a sizeable difference, as we would not expect inattentive respondents to notice the flipped scale, and therefore should have lower consistency among the three items. This is precisely what we see for both RTAC and Screeners. In particular, both global response time and a two-state GMM model identify "attentive" respondents that neither exhibit larger experimental treatment effect, nor do they detect the flipped scale in the battery of ideology questions.

Taken as a whole, screener question passage rates and response time-based attentiveness categorization suggests these measures are capturing similar underlying concepts of attentiveness. In comparison to these measures, however, RTAC allows for researchers to leverage information taken from the whole survey, rather than snippets, more clearly capturing variation in survey-taking behavior across the survey. At the same time, this measure is one that researchers can collect unobtrusively, preserving space in the survey instrument for substantively important questions, rather than data quality checks, without sacrificing information about how respondents are taking the survey.

# 6 Discussion: Applying RTAC

Now that researchers can easily estimate respondent attentiveness in surveys using response time, what should they do with that information? Of course, dropping inattentive respondents from the analysis would complicate external validity, given that the traits of inattentive respondents likely correlate with how they respond to treatment stimuli. Instead, stratifying treatment effects by attentiveness category can help researchers better identify the degree to which respondents complied with the experimental treatment by reading and processing the information, and the degree to which the results may be driven by noise. Researchers still have to determine how much inattentiveness may influence the external validity of a study; for example, in the case of media effects, researchers may *want* to see how experimental stimuli are received by inattentive respondents, mirroring how many individuals may interact with media in the real world. This method instead allows researchers to better evaluate the internal validity of their survey experiments, and understand the degree to which experimental results – both null and significant – are robust to the various ways in which respondents interact with survey instruments in an unsupervised setting. Researchers must understand the internal validity of their experiments before proceeding to understanding how their findings relate to the population of interest.

The new method we propose in this paper, RTAC, is intended to provide researchers with a framework for better understanding the nuances and limitations of their own online data collection. Theoretically, we provide researchers with a framework of three different types of respondents, turning an emphasis to understanding respondent survey-taking behavior on the slow-end of the distribution. Empirically, we advance dimension reduction *and* clustering as important steps in parsimoniously detecting inattentive respondents. For researchers not wedded to the idea of the clustering approach, simply examining response time patterns in the PCA framework could provide insight into the data as well as their limitations. Other researchers may prefer to use other dimension-reduction or clustering techniques, but we encourage them to incorporate the presence of slow and inattentive respondents, and to take advantage of fluctuations in survey-taking behavior across all questions when adopting their own approaches. But regardless of the approach, RTAC provides researchers with a proxy for attentiveness in self-administered surveys that is easy to implement, unobtrusive, and as effective at detecting inattentive respondents as IMCs.

Satisficing is concerning to researchers conducting both observational and experimental research. For experimentalists, inattentive respondents indicate that not all respondents are receiving the same treatment, by virtue of some respondents' *actually not reading* the experimental stimulus. This behavior violates a key assumption of causal inference: that all units receive the same treatment. Moreover, it threatens the experiment's internal validity. Researchers cannot distinguish between their experiment having null effects and the case where a portion of their respondents simply failed to receive the treatment. Researchers can have confidence neither in the size of the treatment effect, nor in the presence of a null effect. Diagnosing noise is a crucial first step in understanding the limits of internal validity. Understanding the relationship between survey inattentiveness and *external* validity remains a fruitful research topic for future exploration, one that can leverage a measure of survey inattentiveness that travels across different surveys.

For observationalists, inattentive respondents generate missing data because their recorded attitudes do not reflect the attitudes of those who did not necessarily read the survey question carefully. This phenomenon is reflected in our analysis of ideological consistency in answering multiple items among different attentiveness groupings. If respondents do not read the question carefully enough to pay attention to directions, it is unlikely that they are taking the time to activate different beliefs and respond accordingly to survey questions. Moreover, Alvarez et al. (2019) show that inattentive respondents are quite different from attentive respondents, meaning that the data of inattentive respondents are not missing-at-random (MAR), and therefore contribute to biased estimates of key quantities of interest. When satisficers are non-random, researchers must make an effort to distinguish between such respondents and attentive survey-takers to reduce bias in estimates and increase the internal validity of surveys. Capturing this latent concept of respondent type allows researchers to stratify responses by attentiveness category, and control for it as a variable in quantitative models.

As public opinion scholars transition more to internet-based surveys due to cost and convenience, and researchers across a range of subfields turn to rich micro-level data to test existing theories, understanding how respondents interact with survey instruments in this setting will become increasingly important. Concerns over attentiveness will not go away, so researchers should continue to examine how attentiveness varies as a function of question structure, survey format and respondent behavior. Furthermore scholars should consider how to treat inattentive respondents when analyzing data. As this field of research progress, we hope that RTAC will be a useful starting point for understanding attentiveness throughout the survey using response time-based methods.

Not only is response time easily collected, it can also be consistently applied across surveys, relying on respondents' actual behavior, rather than IMCs that might vary from survey to survey. This consistency also allows researchers to compare between different survey dissemination strategies (e.g. mobile versus computer-based) and different survey pools, providing a measure of data quality that can be applied across these contexts. To this end, we see the development of a response-time based measure of respondent attentiveness as a first step towards more holistically evaluating data quality in internet-generated public opinion data.

# References

Alvarez, R Michael, Lonna Rae Atkeson, Ines Levin and Yimeng Li. 2019. "Paying Attention to Inattentive Survey Respondents." *Political Analysis* 27(2):145–162.

Anduiza, Eva and Carol Galais. 2016. "Answering without reading: IMCs and strong satisficing in online surveys." *International Journal of Public Opinion Research* 29(3):497–519.

Ansolabehere, Stephen and Brian F Schaffner. 2015. "Distractions: The incidence and consequences of interruptions for survey respondents." *Journal of Survey Statistics and Methodology* 3(2):216–239.

Baker, Reg, Stephen J Blumberg, J Michael Brick, Mick P Couper, Melanie Courtright, J Michael Dennis, Don Dillman, Martin R Frankel, Philip Garland et al. 2010. "American Association of Public Opinion Researchers' Report on Online Panels." *Public Opinion Quarterly* 74(4):711–781.

Barge, Scott and Hunter Gehlbach. 2012. "Using the theory of satisficing to evaluate the quality of survey data." *Research in Higher Education* 53(2):182–200.

Bassili, John N and B Stacey Scott. 1996. "Response latency as a signal to question problems in survey research." *Public opinion quarterly* 60(3):390–399.

Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241–262.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Berinsky, Adam J, Michele F Margolis, Michael W Sances and Christopher Warshaw. 2019. "Using screeners to measure respondent attention on self-administered surveys: Which items and how many?" *Political Science Research and Methods* pp. 1–8.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning.* springer.

Börger, Tobias. 2016. "Are fast responses more random? Testing the effect of response time on scale in an online choice experiment." *Environmental and Resource Economics* 65(2):389–413.

Callegaro, Mario, Yongwei Yang, Dennison S Bhola, Don A Dillman and Tzu-Yun Chin. 2009. "Response latency as an indicator of optimizing in online questionnaires." *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 103(1):5–25.

Fazio, Russell H. 1990. "A practical guide to the use of response latency in social psychological research." *Research methods in personality and social psychology* 11:74–97.

Greszki, Robert, Marco Meyer and Harald Schoen. 2015. "Exploring the effects of removing "too fast" responses and respondents from web surveys." *Public Opinion Quarterly* 79(2):471–503.

Harden, Jeffrey J, Anand E Sokhey and Katherine L Runge. 2019. "Accounting for Noncompliance in Survey Experiments." *Journal of Experimental Political Science* 6(3):199–202.

Hillygus, D Sunshine, Natalie Jackson and M Young. 2014. "Professional respondents in non-probability online panels." *Online panel research: A data quality perspective* 1:219–237.

Höhne, Jan Karem, Stephan Schlosser, Mick P Couper and Annelies G Blom. 2020. "Switching away: Exploring on-device media multitasking in web surveys." *Computers in Human Behavior* p. 106417.

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.

Huckfeldt, Robert, Jeffrey Levine, William Morgan and John Sprague. 1999. "Accessibility and the political utility of partisan and ideological orientations." *American Journal of Political Science* pp. 888–911.

Imai, Kosuke and Dustin Tingley. 2012. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56(1):218–236.

Johnson, Martin. 2004. "Timepieces: Components of survey question response latencies." *Political Psychology* 25(5):679–702.

Kong, Xiaojing J, Steven L Wise and Dennison S Bhola. 2007. "Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior." *Educational and Psychological Measurement* 67(4):606–619.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3):213–236.

Malhotra, Neil. 2008. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72(5):914–934.

McLachlan, Geoffrey J, Sharon X Lee and Suren I Rathnayake. 2019. "Finite mixture models." *Annual Review of Statistics and its Application* 6:355–378.

Mulligan, Kenneth, J Tobin Grant, Stephen T Mockabee and Joseph Quin Monson. 2003. "Response latency methodology for survey research: Measurement and modeling strategies." *Political Analysis* pp. 289–301.

Olson, Kristen and Jolene D Smyth. 2015. "The effect of CATI questions, respondents, and interviewers on response time." *Journal of Survey Statistics and Methodology* 3(3):361–396.

Oppenheimer, Daniel M, Tom Meyvis and Nicolas Davidenko. 2009. "Instructional manipulation checks: Detecting satisficing to increase statistical power." *Journal of Experimental Social Psychology* 45(4):867–872.

Sendelbah, Anže, Vasja Vehovar, Ana Slavec and Andraž Petrovčič. 2016. "Investigating respondent multitasking in web surveys using paradata." *Computers in Human Behavior* 55:777–787.

Tourangeau, Roger, Lance J Rips and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Tversky, Amos and Daniel Kahneman. 1981. "The framing of decisions and the psychology of choice." *Science* 211(4481):453–458.

Vandenplas, Caroline, Koen Beullens and Geert Loosveldt. 2019. Linking interview speed and interviewer effects on target variables in face-to-face surveys. In *Survey Research Methods*. Vol. 13 pp. 249–265.

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):163–183.

Wood, Dustin, PD Harms, Graham H Lowman and Justin A DeSimone. 2017. "Response speed and response consistency as mutually validating indicators of data quality in online samples." *Social Psychological and Personality Science* 8(4):454–464.

Yan, Ting and Kristen Olson. 2013. “Analyzing paradata to investigate measurement error.”.

Yan, Ting, Lindsay Ryan, Sandra E Becker and Jacqui Smith. 2015. Assessing quality of answers to a global subjective well-being question through response times. In *Survey Research Methods*. Vol. 9 NIH Public Access p. 101.

Zandt, Trisha Van. 2002. “Analysis of response time distributions.” *Stevens' handbook of experimental psychology* .

Zhang, Chan and Frederick Conrad. 2014. Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In *Survey Research Methods*. Vol. 8 pp. 127–135.