So many questions, so little time:
Integrating adaptive inventories into public opinion research*

Jacob M. Montgomery and Erin Rossiter

Department of Political Science

Washington University in St. Louis

ABSTRACT

One of the most difficult tasks facing survey researchers is balancing the imperative to keep surveys short with the need to measure important concepts accurately. Not only are long batteries prohibitively expensive, but lengthy surveys can also lead to less informative answers from respondents. Yet, scholars often wish to measure traits that require a multi-item battery. To resolve these contradicting constraints, we propose the use of adaptive inventories. This approach uses computerized adaptive testing methods to minimize the number of questions each respondent must answer while maximizing the accuracy of the resulting measurement. We provide evidence supporting the utility of adaptive inventories through an empirically informed simulation study, an experimental study, and a detailed case study using data from the 2016 ANES Pilot. The simulation and experiment illustrate the superior performance of adaptive inventories relative to fixed-reduced batteries in terms of precision and accuracy. The ANES analysis serves as an illustration of how adaptive inventories can be developed and fielded and also validates an adaptive inventory with a nationally representative sample. Critically, we provide extensive software tools that allow researchers with minimal technical expertise to easily incorporate adaptive inventories into their own surveys.

Abstract word count: 193
Manuscript word count: 6434

*Funding for this project was provided by the Weidenbaum Center on the Economy, Government, and Public Policy and the National Science Foundation (SES-1558907). We are grateful to Josh Cutler, Tom Wilkinson, Haley Acevedo, Alex Weil, Ryden Butler, Matt Malis, and Min Hee Seo for their programming assistance. Valuable feedback for this project was provided by Harold Clarke, Brendan Nyhan, and audience members at Washington University in St. Louis, the University of Chicago, Dartmouth College, New York University, and Princeton University. Previous versions of this paper were presented at the 2015 meeting of the Asian Political Methods meeting in Tapei, Taiwan, as well as the 2015 Annual Summer Meeting of the Society for Political Methodology in Rochester, New York.

1 INTRODUCTION

One of the most difficult tasks facing survey researchers is balancing the imperative to keep surveys short with the need to measure important concepts accurately. Surveys with nationally representative samples are expensive; long surveys are *extremely* expensive. Worse, lengthy surveys increase the burden on respondents and drive up attrition, item-nonresponse, unit non-response, and satisficing. Yet, scholars often wish to study latent traits or attitudes that can only be measured accurately using large multi-item batteries.

Facing these countervailing pressures, researchers almost universally choose a subset of questions from large batteries to administer to all respondents. However, this approach is inefficient since these reduced batteries inevitably include items that provide little additional information about respondents' true positions on the latent scale. In other words, while any given reduced battery might perform well *on average*, it will be poorly chosen for many specific respondents.

In this article, we provide an alternative approach that we refer to as adaptive inventories (AIs) that allows researchers to sidestep this balancing act. The advantage of AIs is that they adjust dynamically by using individuals' answers to items already administered in the battery to optimally choose subsequent questions that best measure respondents' positions. In short, using an AI rather than a fixed subset of questions comes at no additional cost in terms of survey time yet provides survey researchers more accurate estimates of respondents' positions on a latent trait by asking a reduced battery customized to each respondent.

Intuitively, AIs are founded on the premise that we should not ignore what we have already learned about a respondent when choosing questions. For example, if we are measuring political knowledge, we should not ask a respondent who correctly defined the Byrd Rule whether she knows what position is held by Mike Pence. Any respondent with sufficient political sophistication to answer the former almost certainly knows the answer to the latter. Instead, we should update our beliefs about respondents' level of knowledge as the survey progresses and choose questions calibrated to our current beliefs about their positions on the trait. For instance, we might ask that respondent to name the Secretary of Homeland Security.

On a more technical level, AIs are an application of computerized adaptive testing (CAT), an extension of item response theory (IRT) originating in educational testing (Weiss 1982). CAT builds on basic IRT models to allow tests to change dynamically for each respondent (Kingsbury and Weiss 1983). CAT is widely used in the fields of educational testing and psychology. Despite the vintage of the approach, however, it has rarely been applied in public opinion research and, to the best of our knowledge, has never before been used on nationally representative sample in social science research.

Superficially the lack of attention to AIs in public opinion research is surprising. After all, they are conceptually similar to branching schemes developed near the advent of scientific polling. The only differences are that in AIs: (a) the branching schemes are designed algorithmically to maximize some pre-defined criteria; and, (b) the resulting hierarchy can contain hundreds or thousands of branchings.

In reality the paucity of AIs on public opinion surveys is easy to explain. Researchers and survey firms are dissuaded from implementing AIs due to the lack of freely available software that they can integrate into their own data-collection systems. Further, the literature lacks a comprehensive guide to CAT methods that explicates the technique in an intuitive manner, illustrates its advantages beyond simulations or single batteries, and provides guidance as to how they can be implemented in collaboration with real-world survey firms.

In this article, we build on the limited previous work applying AIs to survey research (e.g., Montgomery and Cutler 2013) in two ways to address these obstacles. To begin, previous presentations of AIs have illustrated its advantages using only one or two latent traits tested in either simulations or convenience samples. Here, we provide simulation as well as experimental evidence demonstrating the benefits of AIs using ten well-established personality batteries. This includes an application of AIs on the 2016 American National Election Study (ANES) Pilot Study. To our knowledge, this is the first time an adaptive inventory has been administered to a nationally representative sample in the social sciences.[1]

[1]To be sure, CAT batteries have been administered to high-quality samples in other fields such

More critically we provide an extensive suite of software tools that resolve of the technical obstacles for large-scale adoption of APIs. This includes a freely available `R` package that can execute adaptive algorithms in near real time (item selection routines in all realistic settings execute in less than 0.01 seconds). We also provide a simplified approach to pre-calculating AIs for easy integration into online, interactive voice response (IVR), or computer assisted telephone interview (CATI) surveys. In addition, with support from the National Science Foundation, we provide a free webservice that allows any web-connected survey platform, such as Qualtrics, to include AIs.

In the next section, we discuss a particular valuable setting for APIs that we focus on below – measuring personality. In Section 3, we outline the general API approach and provide details on the specific implementation we use in our applications. We then provide evidence supporting the utility of AIs through an empirically informed simulation study, an experimental study using large convenience samples, and a case study of the 2016 ANES Pilot Study. Finally, in the Supporting Information (SI) Appendix we provide details about our freely available software and practical guidance to assist scholars implementing AIs.

2 MOTIVATING EXAMPLE: MEASURING PERSONALITY

When would researchers be interested in including an AI in a survey in the first place? The method is most appropriate when the following three criteria are met. First, it assumes that scholars are interested in measuring a latent trait rather than analyzing survey responses to specific questions *per se*. Second, it assume that the underlying concept is unidimensional. Third, the survey should have space for at least three question items drawn from a larger battery, although in principle it can be applied using only two (as we illustrate below).

Survey research tasks often meet these criteria. Potential applications include placing survey respondents into an ideological space using roll-call questions (Bafumi and Herron 2010), estimat-

as medicine (e.g., Hung et al. 2013; Gibbons et al. 2008). Yet, we find no applications of a CAT battery to a *nationally* representative sample, illustrating the degree to which the methodology has failed to penetrate public opinion research.

ing respondents' likelihood of voting (Erikson, Panagopoulos and Wlezien 2004), and measuring citizens' values (Schwartz 1992). For the sake of concreteness, however, we focus on a particularly valuable context for adaptive methods: measuring personality.

Interest is increasing across disciplines regarding the role personality plays in affecting behavior. In public opinion research, the most prominent example is research on the "big five" personality traits, which have been linked to policy attitudes (e.g., Gerber et al. 2010), turnout decisions (e.g., Mondak et al. 2010; Gerber et al. 2011), and more (e.g., Mondak 2010). However, the big five are only one broad form of the "multifaceted, enduring, internal psychological structure[s]" that constitute personality traits (Mondak et al. 2010, p 86). Other traits affect how individuals process information, including the need for cognition (e.g., Druckman 2004), the need to evaluate (e.g., Chong and Druckman 2013; Bizer et al. 2004), and the need for affect (Arceneaux and Vander Wielen 2013). Still more traits measure individuals' orientation towards specific social constructs such as the acceptable degree of inequality in society or the appropriate scope of state action,including social dominance orientation (Sidanius et al. 2004) and right wing authoritarianism (Altemeyer 1988).

However, public opinions scholars have only scratched the surface of how personality traits shape attitudes and behavior. Many widely used—and extensively validated—measures in the psychology literature have potential implications for public opinion research yet appear in the literature rarely or not at all. These include, for instance, narcissism (Raskin and Terry 1988), empathy-systematizing quotients (Baron-Cohen et al. 2003), and Machiavellianism (Christie, Geis and Berger 1970).

One reason many personality traits fail to filter into the public opinion literature is surely that most established inventories are too long. Standard practices in social and cognitive psychology result in batteries containing dozens or even hundreds of question items (see Table 1 below). Typically, survey researchers avoid these large scales because they are too time consuming for respondents and/or too expensive to administer.

Rather than including lengthy batteries, researchers typically develop a reduced version of a

battery by selecting a subset of items from the larger scale to administer to respondents. For example, one reason the Big Five personality traits became so prominent in recent years is the advent of the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow and Swann 2003). Prior to TIPI, the most common battery measuring the Big Five was the 44-item Big Five Inventory, itself a shorter alternative to the monstrous 240-item NEO Personality Inventory-Revised (Gosling, Rentfrow and Swann 2003; McCrae and John 1992).[2] In fact, developing reduced-form scales of larger batteries constitutes a considerable body of scholarship (see Table 1 for examples).

[Table 1 about here.]

Broadly speaking, researchers develop reduced scales in one of three ways. First, scholars may examine the properties of the scale to make theoretically motivated decisions about which items to preserve. Thus, Ames, Rose and Anderson (2006, p 441) developed the reduced-form narcissistic personality inventory (NPI-16) by choosing items with strong "face validity" that also ensured coverage of theorized subdomains.

Second, researchers may choose items based on factor loadings in the original publication. For instance, in designing a two-item battery measuring need for cognition for the American National Election Study (ANES), Bizer et al. (2000, p 13) chose, "the two items that loaded most strongly on the latent construct in Cacioppo and Petty's (1982) factor analysis." Thus, the need for cognition scale on the ANES is based on the responses of faculty and undergraduates at the University of Iowa, undergraduates at the University of Missouri, and workers on assembly lines in the Iowa City-Cedar Rapids in the early 1980s (Cacioppo and Petty 1982).

Finally, researchers may administer the original battery to one or more convenience samples and use these new responses to select a subset of items. For example, Muncer and Ling (2006) developed a 15-item reduced-form variant of the 40-item Empathy Quotient (Baron-Cohen et al. 2003) by analyzing responses from 362 students and parents at universities in North England.

[2]TIPI is unusual in that the items are not worded identically to the items in the larger scale. For this reason, we do not address the TIPI measure below.

In each approach, scholars developing reduced scales rely on parameter estimates from calibration samples. Once a reduced inventory is chosen, however, the same set of questions are administered to *all* respondents. AIs *also* rely on parameters estimated from calibration samples. However, they differ in that the goal is not to use this prior information to choose a single battery for all respondents, but rather to tailor a reduced battery to each respondent in a manner designed to maximize measurement precision. The result is improved measurement relative to any fixed battery of the same length.

3 ADAPTIVE INVENTORIES

In this section we briefly provide the details of one implementation of AIs that we use below. In this presentation, we follow Chen, Hou and Dodd (1998), van der Linden (1998), Segall (2005), and Choi and Swartz (2009).[3] AIs take a set of potential items and choose questions that best place each respondent on the latent scale. Broadly speaking, they choose (a) items that are highly discriminatory, and (b) items where the respondent has a high probability of answering in multiple categories.

3.1. *Overview*

Figure 1 shows the basic elements of an AI.

[Figure 1 about here.]

For some respondent $j \in [1, \ldots, J]$, our goal is to estimate her true position on the latent scale, denoted θ_j. The first stage of the algorithm estimates her position, $\hat{\theta}_j$. If no questions from the battery have been administered, we estimate θ_j using only a common prior, $\pi(\theta)$. After a respondent has answered at least one item, we estimate θ_j based on both the prior and observed responses to previous questions ($\mathbf{y}_j$).

Second, the algorithm selects the next question item based on a pre-determined criterion discussed further below. Third, we administer the chosen item and records the response. Fourth,

[3]We direct readers to these sources and the works cited therein for a more technical discussion.

the algorithm checks some stopping rule. In our examples below, this rule is that the number of items asked has reached a maximum value. If the stopping criterion is *not* met, the process repeats. Otherwise, the algorithm calculates final estimates for $\hat{\theta}_j$ and terminates.

3.2. *The general model for ordered categorical responses*

Personality inventories typically include questions with multiple ordered response options (e.g., Likert scales). To handle ordered categorical responses, we use a graded response model (GRM) (Samejima 1969; Baker and Kim 2004). For each item i we assume that there are C_i response options and a vector of *threshold parameters* defined as $\boldsymbol{\kappa}_i = (\kappa_{i0}, \kappa_{i1}, \ldots, \kappa_{iC_i})$, with $\kappa_{i0} < \kappa_{i1} \leq \ldots < \kappa_{iC_i}$, $\kappa_{i0} = -\infty$, and $\kappa_{iC_i} = \infty$. In addition, each item is associated with a *discrimination parameter* a_i, which indicates how well item i corresponds to the underlying trait. Note that these parameters must be pre-calculated based on a calibration sample.

To calculate the likelihood function, we estimate $P_{ijk} = \Pr(y_{ij} = k|\theta_j)$, which is the probability of answering in the k^{th} category for item i given the ability parameter for respondent j.[4]
The likelihood function is then,

$$L(\theta_j) = \prod_{i=1}^{n}\prod_{k=1}^{C_i} P_{ijk}^{I(y_{ij}=k)} = \exp\Big[\sum_{i=1}^{n}\sum_{k=1}^{C_i} \log\Big(P_{ijk}^{I(y_{ij}=k)}\Big)\Big] \tag{2}$$

where $I(\cdot)$ is an indicator function.

To complete the model, we specify the prior, $\pi(\theta_j)$. A natural choice is a conjugate normal prior $\pi(\theta_j) \sim N(\mu_\theta, \tau_\theta)$, where τ_θ denotes the standard deviation. We found that a standard normal prior works well in most settings. In our third application, however, we discuss selecting a prior based

[4]This quantity cannot be calculated directly. Instead, we define $P^*_{ijk} = \Pr(y_{ij} \leq k|\theta_j)$. Assuming a logistic response function, this is:

$$P^*_{ijk} = \frac{\exp(\kappa_{ik} - a_i\theta_j)}{1 + \exp(\kappa_{ik} - a_i\theta_j)}. \tag{1}$$

Note that this implies that $P^*_{ij0} = 0$ and $P^*_{ijC_i} = 1$ and $P_{ijk} = P^*_{ij,k} - P^*_{ij,k-1}$.

on a calibration sample.

3.3. *Details for one adaptive battery*

We can now provide the details of the adaptive algorithm applied below. We have implemented all of the most common approaches for estimating $\hat{\theta}_j$ in our freely available R package `catSurv`. Optional methods include maximum likelihood, weighted maximum likelihood, and maximum *a posteriori* methods. In our examples below, we use the expected *a posteriori* (EAP) approach—a standard choice for those adopting a Bayesian perspective.

Assuming that person j has provided answers to at least one item ($\mathbf{y}_j$), we calculate EAP as,

$$\hat{\theta}_j^{(EAP)} = \mathbb{E}(\theta_j|\mathbf{y}_j) = \frac{\int \theta_j \pi(\theta_j) L(\theta_j) d\theta_j}{\int \pi(\theta_j) L(\theta_j) d\theta_j}. \tag{3}$$

Thus, $\hat{\theta}_j^{(EAP)}$ is the expected value of the posterior distribution. The posterior variance is,

$$\text{Var}(\hat{\theta}_j) = \mathbb{E}\big((\theta_j - \hat{\theta}_j^{(EAP)})^2|\mathbf{y}_j\big) = \frac{\int (\theta_j - \hat{\theta}_j^{(EAP)})^2 \pi(\theta_j) L(\theta_j) d\theta_j}{\int \pi(\theta_j) L(\theta_j) d\theta_j}. \tag{4}$$

As calculating these quantities involves solving only one-dimensional integrals, we can estimate both using numerical methods.[5]

The next step is to choose an item based on our current estimate of θ_j. Adaptive batteries choose items to optimize some pre-defined objective function. Popular options include maximum Fisher's information, maximum expected observed information, and maximum expected Kullback-Leibler divergence, among others (Choi and Swartz 2009). All of these options are available in our `catSurv` software. Here, we use the minimum expected posterior variance (MEPV) item-selection criterion, largely to stay within a simple Bayesian framework.[6]

[5]The `catSurv` software relies on the adaptive quadrature methods from the GNU Scientific Library (GSL) in `C++`, which can approximate single-dimensional integrals with high accuracy.

[6]Choi and Swartz (2009) note that this approach performs "equally well" to the more commonly used methods, but that "the MEPV method would be preferred from a Bayesian perspective" (p

First, we need to use the current estimate of $\hat{\theta}_j$ to estimate P_{mjk} for *each* possible response k to a candidate (unasked) item m. Second, we need to calculate the posterior variance we *would* have given each possible response to question m using Equations 3-4. Third, we combine these elements to estimate the expected posterior variance (EPV) for the candidate item,

$$\text{EPV}_m = \sum_k P_{mjk} \text{Var}(\hat{\theta}_j | \ldots, y^*_{im} = k). \tag{5}$$

In words, EPV_m is the posterior variance for $\hat{\theta}_j$ we *would* have given each possible response to item m weighted by the probability of observing that response—where P_{mjk} is conditioned on our *current* estimate $\hat{\theta}_j$. Finally, we select the item with the lowest EPV value.

After the item is chosen and administered to a respondent, the final step is to check a stopping rule. In the examples below, the algorithm stops offering items when the number of questions reaches a pre-specified threshold n_{max}. However, our software also allows researchers to use other criteria based on the precision of the current estimate of θ_j or the expected information to be gained from the remaining items in the battery.

4 APPLICATIONS

In this section, we demonstrate the advantages of AIs in an empirically informed simulation, an experiment conducted with convenience samples, and a case study using data from the 2016 ANES Pilot Study. The simulation and experiment illustrate the superior performance of AIs relative to fixed-reduced batteries in terms of precision and accuracy. The ANES analysis serves as an in-depth case study of how AIs can be developed and fielded and also validates and AI with a nationally representative sample.[7]

436). Thus, our use of this criterion is more a reflection of taste than an indication that MEPV is in some way superior.

[7]All question wordings and response rates are provided in the SI Appendix.

4.1. *Simulation: Narcissism, Machiavellianism, empathy, and systematizing*

We first demonstrate the benefits of AIs using a dataset of responses to four personality inventories. The goal is to show the advantages of AIs relative to fixed reduced-form batteries. By using simulations, we can compare latent estimates under hypothetical counterfactuals where respondents received either a fixed, an adaptive, or a random reduced battery.

In our simulation, we relied on data collected by `personality-testing.info`, maintained by Eric Jorgenson, which provides tens of thousands of individual responses to prominent personality inventories.[8] We selected four personality inventories for which there exists a validated reduced-form version (see Table 1). These reduced-form scales have been published in peer-reviewed journals, and several have been used extensively in academic research. For our analytical approach, it is essential that the reduced-form battery consists exclusively of a subset of items from the larger battery. The names of these batteries, the size of the reduced and full batteries, and the sample sizes are shown in Table 2. All question items are shown in the SI Appendix.

[Table 2 about here.]

To begin the analysis, we first needed item parameters from the GRM and a prior distribution for the position of respondents on the latent trait. We fit a GRM for each battery using a randomly selected training sample of five-sixths of the respondents using the `grm` function from the `ltm` R package (Rizopoulos 2006; R Core Team 2017). The `catSurv` software includes functionality to extract these item parameters from the fitted model. This `grm` function identifies the model by assuming that the first item included in the dataset loads positively on the latent trait and that the θ_j parameters are distributed according to the standard normal distribution.[9]

[8]Respondents to these surveys were recruited online. In essence, individuals were willing to take these batteries for "fun." We have no information about the representative nature of this pool.

[9]Since we anticipated that the distribution of latent traits in the training and test samples would be similar, we used the standard normal prior. We discuss these issue of choosing priors more fully below.

Next, we turned to the remaining sample (the test sample) and use individuals' *recorded answers* to estimate their scores under the assumption that we know only their responses to question as chosen by (a) the reduced-fixed scale, (b) the reduced-adaptive scale, and (c) a randomly selected reduced battery. Our goal is to determine whether the reduced-fixed scale or the reduced-adaptive scale results in more accurate (less biased) estimates. In order to put our estimates on a meaningful scale, we evaluate bias relative to a naïve approach of selecting question items at random.

So, for example, if the fixed battery calls for asking question-items 20, 14, and 3, we first calculated all respondents' scores using their real responses to *just* those three questions. Second, we let the adaptive algorithm choose the first item for all respondents, and we recorded each respondent's "answer" using the real responses in the dataset. The algorithm then customized the selection of the next item for each individual, and so on. Finally, we administered a randomly constructed short battery. That is, for *each* individual, we chose three items from the full battery at random to administer and calculated scores based on the observed responses to just those items.

To evaluate the performance of each reduced battery, we also estimated respondents' positions on the latent trait using their recorded responses to the *entire* battery. For our calculations below, we treat these scores as the respondents' "true" positions on the latent trait and benchmark the various reduced batteries in terms of how well they approximate these estimates. Note that we use a GRM fit to the full response profiles in the test sample to estimate scores on the latent traits for both the reduced and full batteries. This ensures all estimates are on the same scale while also avoiding an undue advantage for the adaptive battery by relying on parameters estimated from the training set.

We look first at the narcissistic personality inventory (NPI) (Raskin and Terry 1988), which measures one's "grandiose yet fragile sense of self and entitlement as well as a preoccupation with success and demands for admiration" (Ames, Rose and Anderson 2006, p 440-441). Although the original battery contained 40 items, Ames, Rose and Anderson (2006) developed a widely used 16-item version (NPI-16). In the first column of Table 3, we compare the performance of NPI-16

with an adaptive inventory in terms of root mean squared error (RMSE).[10] Since asking any item from a validated battery will reduce bias to some degree, we evaluate the bias of the fixed and adaptive batteries *relative* to a random battery of the same length. Table 3 shows that the RMSE of the adaptive NPI battery is 51% lower than that of the random battery, while the fixed battery provides only a 30% improvement. Thus, the difference in these improvements (the difference in differences) shows a 21% advantage for the adaptive battery.

[Table 3 about here.]

Next we turn to Machiavellianism, a measure inspired by the depiction of the manipulative, immoral, and power-hungry ruler in Niccolo Machiavelli's *The Prince*. The most widely used scale in the literature is the 20-item MACH-IV scale proposed by Christie, Geis and Berger (1970). We compare the adaptive inventory method to the 5-item Trimmed MACH proposed by Rauthmann (2013). The second column of Table 3 shows the proportional bias of the adaptive inventory versus the Trimmed MACH battery. Clearly, the adaptive battery is significantly better in reducing errors. Indeed, the Trimmed MACH scale performs worse than simply choosing survey items at random (a 15% *increase* in RMSE), while the adaptive method provides roughly a 8.5% decrease.

Third, we examine the empathizing and systematizing batteries developed by Baron-Cohen et al. (2003). Empathizing is defined as, "the way in which we understand the social world, the emotions and thoughts of others and how we respond to these social cues." By contrast, "Systemizing is concerned with understanding rules, how things work and how systems are organized" (Ling et al. 2009, p 539). Each scale originally contained 40 items.[11] We compare the adaptive inventory method with the 15-item reduced empathizing scale proposed by Muncer and Ling (2006). For systemizing, We compare an adaptive inventory with the 25-item reduced battery proposed by Wakabayashi et al. (2006).

[10]Recall estimates from responses to the entire battery are respondents' "true" positions (θ_j). RMSE is $\sqrt{\frac{\sum_j^J (\hat{\theta}_j - \theta_j)^2}{J}}$.

[11]Each scale also includes 20 "buffer" questions that we exclude.

The results are shown in the third and fourth columns of Table 3. Clearly the items in the reduced-form empathizing scale were not well selected. The third column shows that the fixed scale is nearly 54% worse relative to randomly selecting items. On the other hand, the adaptive inventory does much better, reducing RMSE by about 60%. The fourth column does not reveal such a stark contrast for the systematizing scale, but here again it is clear that the dynamic battery does well against both a random battery and the standard fixed-reduced battery in the literature. Further, the RMSE rate is quite low, showing that the adaptive inventory produces 40% less bias than random selection.

4.2. *Experimental study*

One feature of the simulations above is that even many of the reduced batteries include too many questions for a standard survey. We therefore turn to a more realistic setting where a researcher has space for only a handful of survey items. Specifically, we present results from an experiment conducted using convenience samples recruited via Amazon's Mechanical Turk (AMT) service to compare the performance of fixed-reduced batteries with an adaptive inventory of the same length. In the fall of 2014, we administered full-length versions of five personality inventories that have been included in reduced forms on the American National Election Study (ANES) to 1,204 subjects. The batteries were need for cognition (NFC), need to evaluate (NTE), need for affect (NFA), social dominance orientation (SDO),[12] and right wing authoritarianism (RWA) (see Table 1). Using these responses, we calibrated an adaptive inventory for each battery using the `ltm` package in `R` as described in the previous section.

In the spring of 2015, we then recruited 1,335 new respondents who were randomly assigned to receive either a fixed-reduced battery as used by the ANES or an AI of the same length.[13]

[12]We used only items measuring dominance attitudes (Peña and Sidanius 2002).

[13]Random assignment occurred once before each battery was administered. For RWA, 639 respondents answered the adaptive battery while 684 answered the fixed battery. The corresponding numbers the other scales are as follows: SDO (adaptive=682, fixed=652), NFC (adaptive=667, fixed=666), NTE (adaptive=682, fixed=649), NFA (adaptive=650, fixed=661).

After completing the reduced battery, all subjects then answered all remaining questions in the *full battery* in a random order. We then estimated scores using only questions selected by the fixed batteries and the AIs and compared them, as in Application 1, using respondents' "true" positions (estimated using responses to the complete battery) as a common benchmark. Finally, to put these numbers on a meaningful scale, we estimated respondents' positions on the trait using a random subset of responses.

Before turning to the results, it is worth noting that this experiment represents a far more difficult test for the adaptive batteries than the simulations above. To begin, these batteries are short (between two and five questions), giving the AI little opportunity to learn about respondents and choose items. Further, relative to the examples above, the underlying batteries themselves are small (between eight and 30 items), meaning there are fewer items for the algorithm to choose from. Further, since we estimate respondents' "true" positions using the full battery, error rates should be considerably smaller for almost any method of item selection.[14]

[Table 4 about here.]

Table 4 shows the root mean squared error (RMSE) for respondents answering either the fixed or adaptive scales.[15] The AIs provide more accuracy than the fixed batteries with improvements over random selection for the adaptive versus fixed batteries ranging from a modest 0.2% for the NFC battery to a substantial 13.3% improvement for the NTE battery. In all, these results show that AIs provide more accurate estimates than widely used fixed batteries, even when there is only space for a few items.[16]

[14]For example, if we ask seven out of a total of eight items, the resulting estimates will be similar no matter how the reduced inventory was selected.

[15]See Footnote 10. To ensure all estimates are on the same latent scale, we generate estimates using a GRM fit only with full response profiles from the second sample.

[16]Note, however, that the mean absolute bias for the adaptive NFC battery is actually higher than for the fixed battery, suggesting that some caution is needed in interpreting these results for the NFC battery.

We can demonstrate that this improved accuracy has important consequences beyond mere measurement. To do this, we focus on the RWA measure, which originally had 30 items but was reduced to five on the ANES 2013 Internet Followup Study. Figure 2 shows the distributions estimated for individuals assigned to fixed and adaptive battery conditions. The shaded distributions show the density estimated using only the reduced battery, while the unshaded distributions show the density estimated after these *same respondents* complete the entire 30-item inventory. The figure shows that the fixed battery does a particularly poor job estimating positions on the low end of the spectrum, shown by the difference in the shaded and unshaded densities in the left panel.

[Figure 2 about here.]

Our aim is to show that by inaccurately measuring RWA with fixed-reduced scales, we can inadvertently bias our understanding for how RWA relates to other important factors.[17] However, this bias is ameliorated by using an adaptive battery. To illustrate this, we measured several constructs theoretically related to RWA including: presidential approval, ideology, defense spending attitudes, civil liberties attitudes, symbolic racism, modern racism, and prejudice towards Arabs and Muslims (Sidanius et al. 2004).

We estimated separate regressions by treatment condition (adaptive or fixed battery) using RWA as an explanatory variable and these related constructs as dependent variables.[18] We then estimated the "true" value for these regression coefficients using respondents' scores as estimated from the full battery. We calculated bias as the difference between the regression coefficients (and 95% CIs) estimated using the *reduced battery measures* and the *full battery measures* of RWA. The results, shown in Figure 3, illustrate that the measure of RWA from the fixed battery upwardly biases these regression coefficients due to the censoring, which leads us to conclude that RWA is a stronger predictor than is actually the case. However, there is much less bias in these coefficient estimates when using a five-item AI.

[17]This bias can be in any direction depending on the resulting distortion in the measurement of the underlying trait.

[18]In these regressions, we control for race, gender, and level of education.

[Figure 3 about here.]

4.3. *Case study: 2016 ANES Pilot Study*

In our third application, we present a detailed case study of an AI measuring the need for cognition that was included on the 2016 ANES Pilot Study. (A more technical guide for calibrating and administering an AI with `catSurv` is shown in the SI Appendix.) In addition to providing an illustrative example, the purpose of this section is to test the validity of AI measures on a nationally representative survey conducted by a professional polling firm (YouGov).

Cacioppo and Petty (1982) originally proposed the need for cognition scale as a method for measuring, "the tendency for an individual to engage in and enjoy thinking" (p 116). While originating in social psychology, this trait has been used extensively in political science. Druckman (2004), for instance, shows that NFC moderates the degree to which individuals are susceptible to issue framing from elites, with individuals who score highly on NFC being less likely to respond to issue-framing attempts.

The original battery was developed using a convenience sample of 96 individuals drawn from faculty at the University of Iowa and assembly line workers in the Iowa City-Cedar Rapids area engaged in the automotive parts industry and later validated using undergraduate students in Missouri and Iowa. The result was a 34-item inventory that was subsequently reduced to an 18-item "efficient" battery (Cacioppo and Petty 1984). It is from this 18-item inventory that Bizer et al. (2000) chose the items for inclusion on the ANES.

To calibrate the adaptive personality inventory, we combined data from three separate samples.[19] First, we used data from the December 2014 wave of The American Panel Survey (TAPS). TAPS is a monthly online panel survey for which panelists were recruited as a national probability sample with an address-based sampling frame in the fall of 2011 by GfK-Knowledge Networks. Individuals without Internet access were provided a laptop and Internet service at no cost. After removing respondents who completed less than 25% of the items we had 1,506 respondents.

[19]Recall that in our first applications, we calibrated using only a single training sample.

To supplement TAPS, we used responses to the 18-item NFC battery from the two convenience samples recruited via Amazon's Mechanical Turk (AMT) online workforce used in the experiment described above. While not a representative sample, AMT provides an easy way to administer the survey to a larger set of respondents (Berinsky, Huber and Lenz 2012). As Embretson (1996) notes, one of the "new rules" of item response models is, "Unbiased estimates of item properties may be obtained from unrepresentative samples" (p 342). That is, AIs do not need to be calibrated using a representative sample as long the sample is large, the sample is diverse along the dimension of interest, and there is not large heterogeneity in how the questions relate to the underlying trait.

In order to create the adaptive test, we fit a GRM with the combined sample and select a prior based on the TAPS sample.[20] Then, we pre-calculated a complete branching scheme. Figure 4 depicts portions of the complete branching scheme for the four-item NFC AI. The labels on the branches indicate possible answers. (The NA indicates item non-response.) For example, a respondent who answers "1" to NFC23 will be asked NFC32, and a respondent who then answers "5" will be asked NFC29. On the other hand, a respondent who answers "5" for NFC23 will be asked NFC40.

[Figure 4 about here.]

For longer batteries, a full enumeration of the scheme might be difficult. However, since this battery is only four items in length, the tree contains only $6^3 = 216$ complete branchings, and the entire tree can be represented as a simple lookup table with 259 rows. We calculated this tree using the `makeTree` command in the `catSurv` software, and provided the lookup table to YouGov in advance of the survey.[21] The 2016 ANES Pilot Study administered the NFC AI to 1,200 respondents drawn from an opt-in online panel.

Since the ANES pilot did not include the fixed battery, it is **not** possible to compare the adaptive and fixed batteries as we did in the applications above. However, it is possible to evaluate predictive

[20]Futher details about this selection are in the SI Appendix.

[21]We provide further details on the lookup table in the SI Appendix.

validity. In particular, we test whether NFC (as measured by the AI) is a moderator for the effect of issue framing as has been argued in the existing literature (e.g., Druckman 2004).

We take advantage of a framing experiment on the ANES Pilot Study.[22] In the experiment, respondents were randomly assigned to answer the question, "Do you favor, oppose, or neither favor nor oppose allowing [*Syrian refugees* / *refugees fleeing the Syrian civil war*] to come to the United States?" (Emphasis added), where 587 received the "Syrian refugees" frame and 613 received the civil war frame. Respondents indicated their level of support on a seven-point scale. We test the hypothesis that the civil war frame will make respondents less opposed to allowing Syrian refugees to enter the United States, but that this effect will be moderated by respondents' level of NFC.

The main results are presented in Table 5, which shows the coefficients of interest from a weighted least squares regression where the dependent variable is the degree of opposition to Syrian refugees on a seven point scale.[23]

[Table 5 about here.]

The first column in Table 5 shows that the civil war framing does not by itself appear to have a statistically reliable effect on opposition to Syrian refugees being admitted to the United States. However, Model 2 shows that there is a significant interaction between this treatment and NFC as measured by the AI ($p = 0.046$). Figure 5 shows the estimated marginal effect of the civil war framing on opposition to Syrian refugees for differing levels of NFC. Consistent with expectations, the plot indicates that the framing experiment had little or no effect for respondents with high levels of NFC, but that it has a significant and negative effect for respondents lower on this trait.

[Figure 5 about here.]

[22] We provide a similar analysis of a second framing experiment in the SI Appendix.

[23] We also controlled for feeling thermometer towards Muslims, support for intervening in Syria to combat ISIS, racial resentment, party identification, ideology, gender, education, race, and ethnicity.

5 CONCLUSION

Survey researchers face a constant trade-off between the desire to better measure concepts and the need to reduce survey length. While these tensions will always exist, AIs are capable of obviating the need for public opinion researchers to choose between administering a large, costly multi-item scale or a single reduced scale that may drastically reduce measurement precision. Our results show that AIs allow for the administration of fewer questions while achieving superior levels of statistical precision and accuracy relative to any fixed-reduced scale. At a minimum, we believe that AIs can dramatically expand the ability for scholars to explore the role of various personality traits on public opinion and political behavior. However, we believe that AIs could be applied to many tasks beyond measuring personality.

Nonetheless, there are several potential limitations to AIs as well as areas for continued research. First, as noted at the outset, survey time is perhaps the greatest contraint for improving the measurement of latent traits. Yet the relative advantage of adaptive surveys to static batteries actually increases for *longer* batteries, which may lead some to question the usefulness of adopting the method. One answer to this concern is that adaptive surveys provide superior measurement of latent constructs even if space allows for three or four items, as we show above. However, an additional approach is to include informative priors based on earlier survey responses as part of the CAT algorithm (van der Linden 1999). This will allow the algorithm to begin tailoring question items for respondents at the outset, further improving performance.

A second concern is that random error will interfere with the performance of adaptive surveys, since noisy responses may lead to the "wrong" question being selected—especially in early stages of the battery.[24] One particularly promising approach to addressing this issue is using a stratified multi-stage adaptive algorithm, where less discriminating items are used early in the adaptive process and highly discriminating items are reserved for later stages when respondents' locations in

[24]While a widely studied issue in the adaptive testing literature, the focus is often not on accuracy but on reducing item exposure (e.g., Chang and Ying 1996, 1999; Chen, Ankenmann and Chang 2000).

the latent space are more accurately estimated (e.g., Chang and Ying 1999).

Third, the advantages of CAT depend heavily on the accuracy of the item-level parameters. Indeed, within the CAT framework, poorly estimated item parameters may have particularly pernicious effects on the quality of the final measure (van der Linden and Glas 2000). Survey researchers may therefore be particularly interested in uncovering parameter drift, wherein items are no longer functioning as expected based on the calibration sample. Fortunately, numerous solutions have been proposed in the adaptive testing literature for uncovering such changes, often termed "differential item functioning," (e.g., Kim, Cohen and Park 1995; Glas 2010; Wang, Tay and Drasgow 2013).

A final limitation of AIs is that they require pre-testing of battery items to calibrate the model. While this may seem burdensome, two factors make it a reasonable requirement. First, calibrating these models can be done using large convenience samples. AI performance will be improved if the models can be "normed" to national samples such that our prior beliefs are correctly calibrated towards the target population, however this is not strictly necessary. Ideally, researchers will work collaboratively to pair large convenience samples with nationally representative samples to calibrate and test AIs.

Second, pre-testing costs may be ameliorated by making survey data and item calibrations widely available to other researchers. The calibrations in this study, for instance, will be included in the replication archive for this article at the time of publication. Clearly, additional research is called for to develop, calibrate, and field-test specific AIs measuring other constructs. Our hope is that once these are developed, scholars will disseminate them to the wider academic community—facilitating adoption of this promising technology in public opinion research.

5 References

Altemeyer, Bob. 1988. *Enemies of Freedom: Understanding Right-Wing Authoritarianism*. San Francisco, CA: Jossey-Bass.

Ames, Daniel R., Paul Rose and Cameron P. Anderson. 2006. "The NPI-16 as a Short Measure of Narcissism." *Journal of Research in Personality* 40(4):440–450.

Arceneaux, Kevin and Ryan J. Vander Wielen. 2013. "The Effects of Need for Cognition and Need for Affect on Partisan Evaluations." *Political Psychology* 34(1):23–42.

Bafumi, Joseph and Michael C. Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104(3):519–542.

Baker, Frank B. and Seock-Ho Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

Baron-Cohen, Simon, Jennifer Richler, Dheraj Bisarya, Nhishanth Gurunathan and Sally Wheelwright. 2003. "The Systemizing Quotient: An Investigation of Adults with Asperger Syndrome or High–Functioning Autism, and Normal Sex Differences." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358(1430):361–374.

Berinsky, Adam J., Gergory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Bizer, George Y., Jon A. Krosnick, Allyson L. Holbrook, S. Christian Wheeler, Derek D. Rucker and Richard E. Petty. 2004. "The Impact of Personality on Cognitive, Behavioral, and Affective Political Processes: The Effects of Need to Evaluate." *Journal of Personality* 72(5):995–1028.

Bizer, George Y., Jon A. Krosnick, Richard E. Petty, Derek D. Rucker and S. Christian Wheeler. 2000. "Need for Cognition and Need to Evaluate in the 1998 National Election Survey Pilot Study." National Election Studies Report.

Cacioppo, John T. and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality & Social Psychology* 42(1):116–131.

Cacioppo, John T. and Richard E. Petty. 1984. "The Efficient Assessment of Need for Cognition." *Journal of Personality Assessment* 48(3):306–307.

Chang, Hua-Hua and Zhiliang Ying. 1996. "A Global Information Approach to Computerized Adaptive Testing." *Applied Psychological Measurement* 20(3):213–229.

Chang, Hua-Hua and Zhiliang Ying. 1999. "a-Stratified Multistage Computerized Adaptive Testing." *Applied Psychological Measurement* 23(3):211–222.

Chen, Shu-Ying, Robert D. Ankenmann and Hua-Hua Chang. 2000. "A Comparison of Item Selection Rules at the Early Stages of Computerized Adaptive Testing." *Applied Psychological Measurement* 24(3):241–255.

Chen, Ssu-Kuang, Liling Hou and Barbara G. Dodd. 1998. "A Comparison of Maximum Likelihood Estimation and Expected a Posteriori Estimation in CAT Using the Partial Credit Model." *Educational and Psychological Measurement* 58(4):569–595.

Choi, Seung W. and Richard J. Swartz. 2009. "Comparison of CAT Item Selection Criteria for Polytomous Items." *Applied Psychological Measurement* 33(6):419–440.

Chong, Dennis and James N. Druckman. 2013. "Counterframing Effects." *Journal of Politics* 75(1):1–16.

Christie, Richard, Florence L. Geis and David Berger. 1970. *Studies in Machiavellianism*. New York: Academic Press.

Druckman, James N. 2004. "Political Preference Formation: Competition, Deliberation, and the (Ir)relevance of Framing Effects." *American Political Science Review* 98(4):671–686.

Embretson, Susan E. 1996. "The New Rules of Measurement." *Psychological Assessment* 8(4):341–349.

Erikson, Robert S., Costas Panagopoulos and Christopher Wlezien. 2004. "Likely (and Unlikely) Voters and the Assessment of Campaign Dynamics." *Public Opinion Quarterly* 68(4):588–601.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, Connor Raso and Shang E. Ha. 2011. "Personality Traits and Participation in Political Processes." *The Journal of Politics* 73(3):692–706.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling and Shang E. Ha. 2010. "Personality and Political Attitudes: Relationships Across Issue Domains and Political Contexts." *American Political Science Review* 104(1):111–133.

Gibbons, Robert D, David J Weiss, David J Kupfer, Ellen Frank, Andrea Fagiolini, Victoria J Grochocinski, Dulal K Bhaumik, Angela Stover, R Darrell Bock and Jason C Immekus. 2008. "Using computerized adaptive testing to reduce the burden of mental health assessment." *Psychiatric Services* 59(4):361–368.

Glas, Cees A. W. 2010. Item Parameter Estimationa and Item Fit Analysis. In *Elements of Adaptive Testing*, ed. Wim J. van der Linden and Cees A. W. Glas. New York: Springer pp. 269–288.

Gosling, Samuel D., Peter J. Rentfrow and William B. Swann. 2003. "A Very Brief Measure of the Big-Five Personality Domains." *Journal of Research in Personality* 37(6):504–528.

Hung, Man, Judith F Baumhauer, L Daniel Latt, Charles L Saltzman, Nelson F SooHoo, Kenneth J Hunt, National Orthopaedic Foot & Ankle Outcomes Research Network et al. 2013. "Validation of PROMIS® Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research." *Clinical Orthopaedics and Related Research®* 471(11):3466–3474.

Jarvis, W. Blair G. and Richard E. Petty. 1996. "The Need to Evaluate." *Journal of Personality and Social Psychology* 70(1):172–194.

Kim, Seock-Ho, Allan S. Cohen and Tae-Hak Park. 1995. "Detection of Differential Item Functioning in Multiple Groups." *Journal of Educational Measurement* 32(3):261–276.

Kingsbury, G. Gage and David J. Weiss. 1983. A Comparison of IRT-Based Adaptive Mastery Testing and a Sequential Mastery Testing Procedure. In *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, ed. David J. Weiss. New York: Academic Press.

Ling, Jonathan, Tanya C. Burton, Julia L. Salt and Steven J. Muncer. 2009. "Psychometric Analysis of the Systemizing Quotient (SQ) Scale." *British Journal of Psychology* 100(3):539–552.

Maio, Gregory R. and Victoria M. Esses. 2001. "The Need for Affect: Individual Differences in the Motivation to Approach or Avoid Emotions." *Journal of Personality* 69(4):583–614.

McCrae, Robert R. and Oliver P. John. 1992. "An Introduction to the Five-Factor Model and Its Applications." *Journal of personality* 60(2):175–215.

Mondak, Jeffery J. 2010. *Personality and the Foundations of Political Behavior*. New York: Cambridge University Press.

Mondak, Jeffery J., Matthew V. Hibbing, Damarys Canache, Mitchell A. Seligson and Mary R. Anderson. 2010. "Personality and Civic Engagement: An Integrative Framework for the Study of Trait Effects on Political Behavior." *American Political Science Review* 104(01):85–110.

Montgomery, Jacob M. and Josh Cutler. 2013. "Computerized Adaptive Testing for Public Opinion Surveys." *Political Analysis* 21(2):172–192.

Muncer, Steven J. and Jonathan Ling. 2006. "Psychometric Analysis of the Empathy Quotient (EQ) Scale." *Personality and Individual Differences* 40(6):1111–1119.

Peña, Yesilernis and Jim Sidanius. 2002. "US Patriotism and Ideologies of Group Dominance: A Tale of Asymmetry." *The Journal of social psychology* 142(6):782–790.

Pratto, Felicia, Jim Sidanius, Lisa M. Stallworth and Bertram F. Malle. 1994. "Social Dominance Orientation: A Personality Variable Predicting Social and Political Attitudes." *Journal of Personality and Social Psychology* 67(4):741–741.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Raskin, Robert and Howard Terry. 1988. "A Principal-Components Analysis of the Narcissistic Personality Inventory and Further Evidence of Its Construct Validity." *Journal of Personality and Social Psychology* 54(5):890–902.

Rauthmann, John F. 2013. "Investigating the MACH–IV with Item Response Theory and Proposing the Trimmed MACH*." *Journal of Personality Assessment* 95(4):388–397.

Rizopoulos, Dimitris. 2006. "ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software* 17(5):1–25.

Samejima, Fumiko. 1969. "Estimation of Latent Ability Using a Response Pattern of Graded

Scores." *Psychometrika monograph supplement* 34(4):100.

Schwartz, Shalom H. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In *Advances in Experimental Social Psychology*. Vol. 25 San Diego, California: Academic Press pp. 1–65.

Segall, Daniel O. 2005. Computerized Adaptive Testing. In *Encylopedia of Social Measurement*. Vol. 1 Oxford: Elsevier pp. 429–438.

Sidanius, Jim, Felicia Pratto, Colette Van Laar and Shana Levin. 2004. "Social Dominance Theory: Its Agenda and Method." *Political Psychology* 25(6):845–880.

van der Linden, Wim J. 1998. "Bayesian Item Selection Criteria for Adaptive Testing." *Psychometrika* 63(2):201–216.

van der Linden, Wim J. 1999. "Empirical Initialization of the Trait Estimator in Adaptive Testing." *Applied Psychological Measurement* 23(1):21–29.

van der Linden, Wim J. and Cees A. W. Glas. 2000. "Capitalization on Item Calibration Error in Adaptive Testing." *Applied Measurement in Education* 13(1):35–53.

Wakabayashi, Akio, Simon Baron-Cohen, Sally Wheelwright, Nigel Goldenfeld, Joe Delaney, Debra Fine, Richard Smith and Leonora Weil. 2006. "Development of Short Forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short)." *Personality and individual differences* 41(5):929–940.

Wang, Wei, Louis Tay and Fritz Drasgow. 2013. "Detecting Differential Item Functioning of Polytomous Items for an Ideal Point Response Process." *Applied Psychological Measurement* 37(4):316–335.

Weiss, David J. 1982. "Improving measurement quality and efficiency with adaptive testing." *Applied Psychological Measurement* 6(4):473–492.

5 List of Figures

Figure 1: Basic elements of adaptive inventories

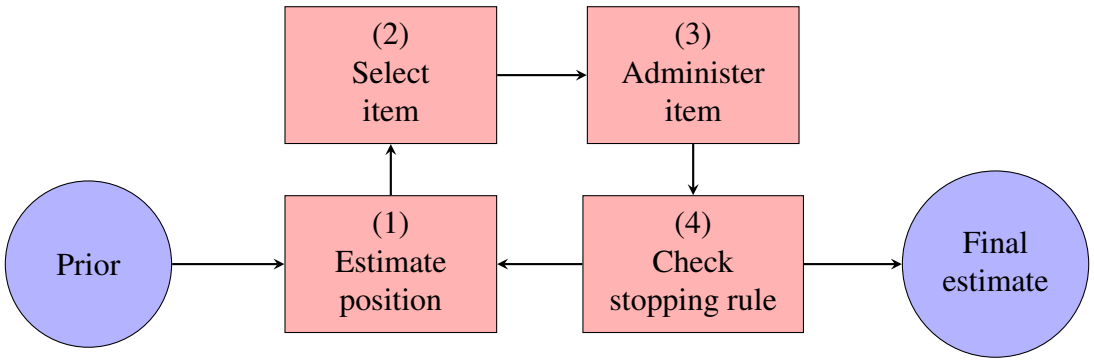
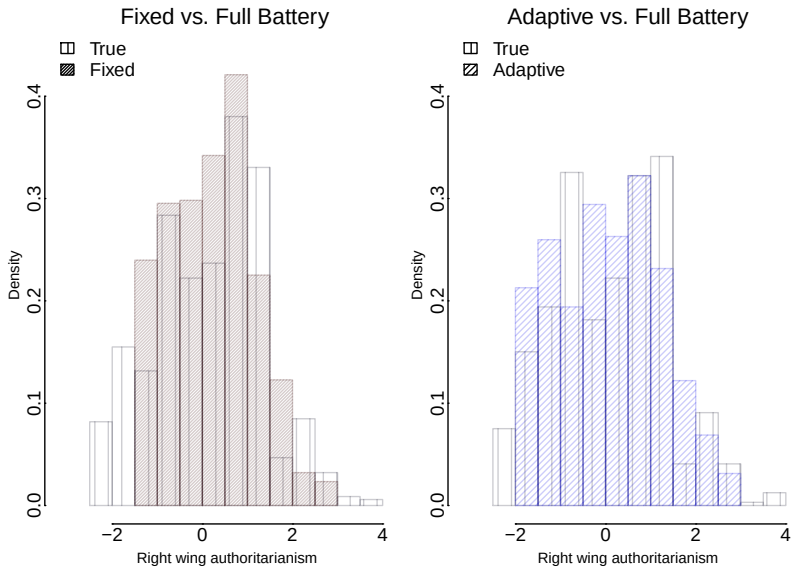
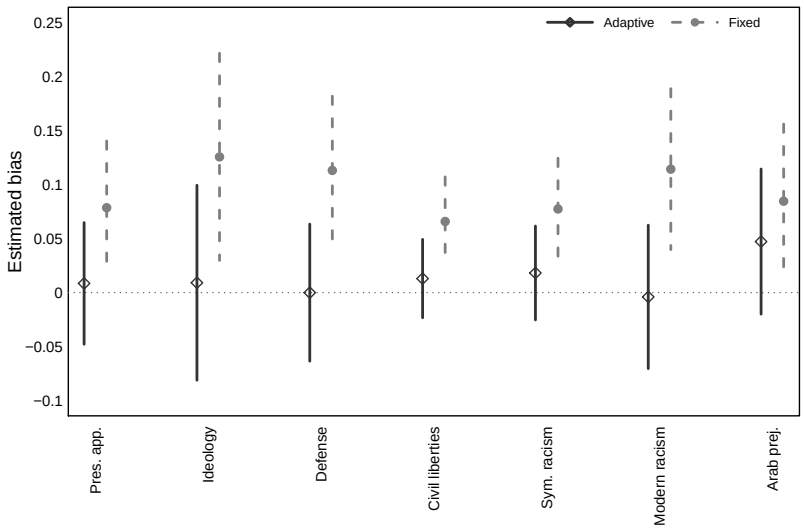


Figure 2: Revealed right wing authoritarian (RWA) estimates for adaptive and fixed measures



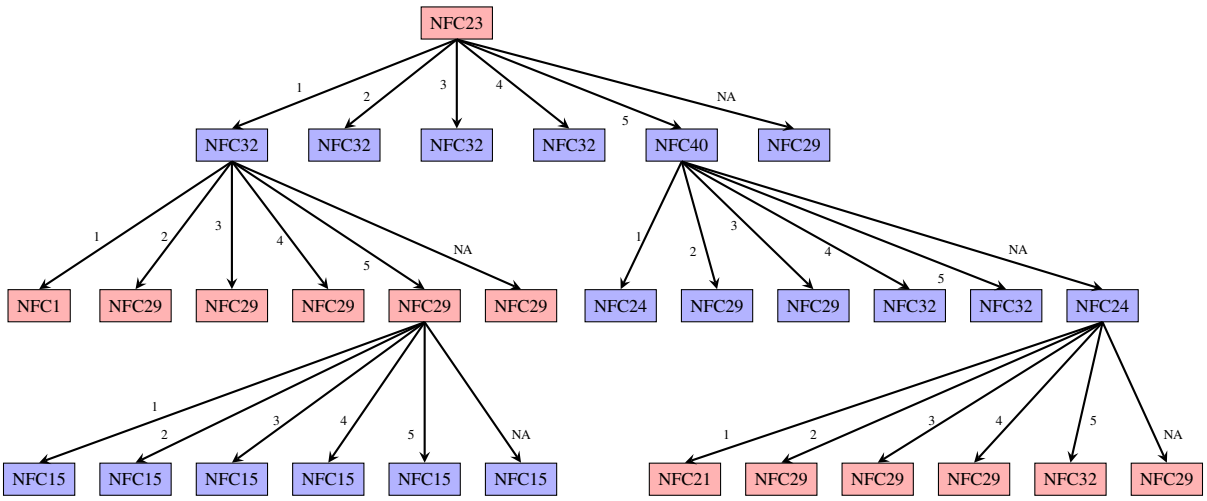
These figures show the distribution of RWA as estimated using the five-item reduced batteries (shaded histograms) and using the complete 30-item inventory (unshaded histograms). Estimates for respondents randomly assigned to answer a fixed battery (n=684) are on the left while estimates for respondents randomly assigned to answer the adaptive battery (n=639) are on the right. The adaptive battery does a superior job in recovering the positions of respondents with more extreme values on the latent scale.

Figure 3: Bias in regression estimates for RWA and seven related constructs



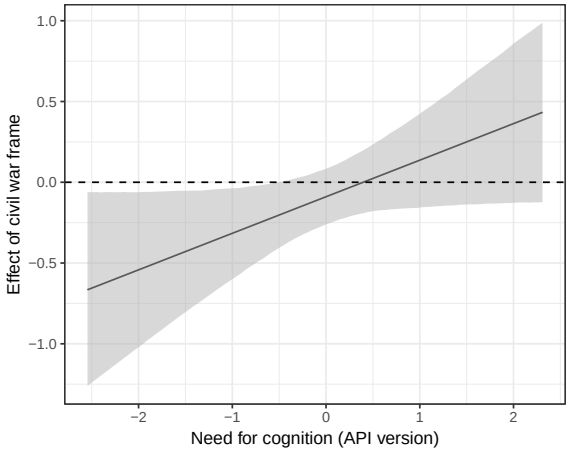
This figure shows that regression coefficients measuring the relationship between right wing authoritarianism (RWA) and related constructs are biased upwards when estimates of respondents' latent position are poorly estimated by fixed-reduced batteries. The vertical axis shows the degree to which regression coefficients between RWA and various outcomes *differs* when using a 5-item reduced scale relative to regression coefficients when RWA is estimated using the full 30-item inventory. The names of the various dependent variables are shown on the x-axis. The closed circles and dashed lines are point estimates and 95% confidence intervals for subjects randomly assigned to answer a fixed-reduced battery (n=684), while the open squares and solid lines show the same for subjects randomly assigned to answer an AI of the same length (n=649). All regressions controlled for gender, race, and level of education. All question wordings are provided in the SI Appendix.

Figure 4: Selected portions of a complete branching scheme for the four-item need for cognition adaptive personality inventory



The figure describes selected sub-trees of the compete branching scheme for the four-item need for cognition AI included on the 2016 ANES Pilot Study. The labels on the branches indicate possible respondent answers. An "NA" indicates item non-response.

Figure 5: Interaction plot estimating the effect of the civil war frame on opposition to Syrian refugees for differing levels of need for cognition



Lines represent point estimates and shaded region represents a 95% confidence interval. Parameter estimates for this model are shown in Table 5.

5 List of Tables

Table 1: Exemplar full and reduced-form measures of personality traits

	Original length	Reduced length
Example psychology scales with reduced-form scales		
Narcissistic personality	Raskin and Terry (1988)	Ames, Rose and Anderson (2006)
Length	40	16
Empathy quotient	Baron-Cohen et al. (2003)	Muncer and Ling (2006)
Length	40	15
Systemizing quotient	Baron-Cohen et al. (2003)	Wakabayashi et al. (2006)
Length	40	25
Machiavellian personality	Christie, Geis and Berger (1970)	Rauthmann (2013)
Length	20	5
American National Election Studies 2000-present		
Need for cognition	Cacioppo and Petty (1982)	Bizer et al. (2000)
Length	40	2
Need to evaluate	Jarvis and Petty (1996)	Bizer et al. (2000)
Length	16	3
American National Election Studies 2013 Internet followup		
Right wing authoritarianism	Altemeyer (1988)	
Length	30	5
Social dominance	Pratto et al. (1994)	
Length	8	2
Social equality	Pratto et al. (1994)	
Length	8	2
Need for affect	Maio and Esses (2001)	
Length	26	4

Note: The reduced-form batteries contain a strict subset of items in the original batteries. All question wordings are shown in the SI Appendix.

Table 2: Description of large personality inventories in simulation study

	Full battery length	Fixed battery length	Response categories	Training (n)	Test (n)
Narcissism	40	16	2	8,700	1,740
Machiavellianism	20	5	5	10,249	2,050
Empathy	40	15	4	10,145	2,029
Systemizing	40	25	4	10,145	2,029

See Table 1 for additional details. All data obtained from: http://personality-testing.info.

Table 3: Assessing fit of adaptive vs. fixed batteries in empirically informed simulation

	Inventory name			
	NPI	MACH	Empathy	Systemizing
Battery length	16	5	15	25
Random (RMSE)	0.38	0.45	0.38	0.21
Adaptive (RMSE)	0.18	0.42	0.15	0.13
% Improvement over random	51.41%	8.56%	59.81%	39.50%
Random (RMSE)	0.38	0.45	0.38	0.21
Fixed (RMSE)	0.27	0.52	0.58	0.17
% Improvement over random	30.26%	-15.09%	-53.86%	16.71%
Difference in improvement for adaptive vs. fixed	21.14%	23.65%	113.68%	22.79%

Values are the root mean squared error for respondents simulated to have answered fixed-reduced batteries (see Table 1) or adaptive batteries of the same length. Estimates were also calculated as if each respondent received a random battery of the same length by sampling from each response set. Point estimates were calculated relative to estimates generated for each respondent using the full inventory. In each case, a Wilcoxon Rank-Sum Test finds the adaptive battery provides less bias than the fixed battery ($p < 0.05$).

Table 4: Assessing fit of adaptive vs. fixed batteries in experimental study

	Inventory name				
	NFA	NTE	NFC	SDO	RWA
Battery length	4	3	2	2	5
Random (RMSE)	1.08	1.03	1.08	0.41	1.39
Adaptive (RMSE)	0.47	0.47	0.49	0.36	0.44
% Improvement over random	56.55%	54.53%	54.74%	10.27%	68.63%
Random (RMSE)	1.13	0.94	1.09	0.42	1.41
Fixed (RMSE)	0.55	0.55	0.49	0.40	0.48
% Improvement over random	51.52%	41.20%	54.51%	5.78%	65.75%
Difference in improvement for adaptive vs. fixed	4.96%	13.34%	0.23%	4.50%	2.88%

N=1,335

Values are the root mean squared error for respondents randomly assigned to either answer the fixed batteries as they appeared on the ANES (see Table 1) or adaptive batteries of the same length. Randomization occurred before each battery. Estimates were also calculated as if each respondent received a random battery of the same length by sampling from each response set. Point estimates were calculated relative to estimates generated for each respondent using the full inventory. In each case, a Wilcoxon Rank-Sum tests finds the adaptive battery provides less bias than the fixed battery with $p < 0.05$ for NFA, SDO, and RWA and $p < 0.10$ for NTE and NFC.

Table 5: Effect of civil war framing on opposition to Syrian refugees

	Model 1	Model 2
Intercept	4.563	4.553
	(0.319)	(0.319)
Civil war framing	−0.094	−0.092
	(0.088)	(0.088)
Need for Cognition	−0.083	−0.200
	(0.061)	(0.84)
Civil war × NFC		0.224
		(0.112)
N	1,064	1,064
R^2	0.530	0.532

Estimates from weighted least squares regression using survey weights. We also controlled for a feeling thermometer towards Muslims, support for intervening in Syria to combat ISIS, racial resentment, party identification, ideology, gender, education, race, and ethnicity. These coefficients are suppressed for clarity. All question wordings are shown in the SI Appendix.