The efficient measurement of personality:
Adaptive personality inventories for survey research

Jacob M. Montgomery[*]

March 6, 2016

ABSTRACT

Recent scholarship in political science has expanded our understanding of how personality affects political behaviors, attitudes, and learning. However, a major obstacle to expanding this research agenda is that many established personality inventories contain far too many questions for inclusion on surveys. In response, researchers typically select a subset of items to administer, a practice that can dramatically lower measurement precision and accuracy. In this paper, I outline an alternative method – adaptive personality inventories (APIs) – for including large personality batteries on surveys while minimizing the number of questions each respondent must answer. Building on results in Montgomery and Cutler (2013), I implement a computerized adaptive testing technique for categorical survey items appropriate for the measurement of personality traits. I provide simulation and experimental evidence demonstrating the benefits of the method in terms of measurement accuracy and precision and validate an API included on the 2016 American National Election Study Pilot Study.

[*] Funding for this project was provided by the Weidenbaum Center on the Economy, Government, and Public Policy. I am grateful to Josh Cutler, Tom Wilkinson, and Haley Acevedo for their programming assistance. Valuable feedback for this project was provided by Harold Clarke and audience members at Washington University in St. Louis, the University of Chicago, and Dartmouth College. 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.

2:15

1 INTRODUCTION

One of the core goals of research on public opinion and political behavior has always been to explain how environmental factors shape attitudes and actions. Since the advent of survey research, scholars have explored the effects of individuals' social class, geographic location, group membership, and other factors on opinion and behavior (e.g., Campbell, Gurin and Miller 1954). This includes the effects of the political context such as campaign ads, mobilization drives, presidential speeches, and news coverage.

Only in recent decades have scholars more fully explored the psychological determinates of behavior – factors independent of the environment that have an effect on opinions and actions. These include, for instance, in-group bias, altruism, tolerance, and more. In particular, one of the most active trends in contemporary scholarship focuses on the broad, stable psychological characteristics usually described as personality traits. Several recent studies emphasize the role of personality in shaping important outcomes ranging from political ideology to voter turnout.

However, a major obstacle to researching personality in the political realm is that many established personality inventories contain far too many questions for inclusion on nationally representative surveys – the workhorse of public opinion scholarship. In response, researchers typically select a subset of items from existing batteries to include. Unfortunately, this practice can dramatically lower measurement precision and accuracy, especially if the selection of items is poorly motivated.

In this paper, I outline an alternative method – adaptive personality inventories (APIs) – for including large personality batteries in survey research while minimizing the number of questions each respondent must answer. As its name suggests, APIs adjust dynamically to respondents to measure personality traits accurately while minimizing the number of questions each respondent must answer. Specifically, APIs adapt to individuals' prior answers to questions in a given battery by choosing subsequent questions that will measure their score on the underlying trait with maximum precision.

The project builds on previous work applying computerized adaptive testing (CAT) techniques to survey research (Montgomery and Cutler 2013) in three ways. First, I extend the method to accommodate categorical, as opposed to binary, response options. Second, I provide simulation as well as experimental evidence demonstrating the benefits of the method in terms of measurement accuracy and precision using a variety of well-established personality batteries originating in social psychology and political science. I further validate the method using data from the 2016 American National Election Study (ANES) Pilot Study, which included an API measuring the need for cognition personality trait. Third, I provide software tools that will allow researchers to more easily incorporate CAT techniques in their research using a wider array of CAT algorithm specifications (including alternative item selection routines and ability estimation methods) and a simplified approach to pre-calculating APIs for easy integration into online, interactive voice response (IVR), or computer assisted telephone interview (CATI) surveys.

In the next section, I discuss the importance of personality traits to contemporary political science and provide an intuition as to why APIs provide a superior method of measurement relative to traditional fixed scales. In Section 3, I outline the method and provide details of one implementation of CAT that I use in my applications below. I then provide evidence supporting the utility of APIs in an empirically informed simulation study, an experimental study using a large convenience samples, and data from the 2016 ANES Pilot Study. Finally, in the Supporting Information (SI) Appendix I provide a detailed case study to illustrate how to implement APIs using freely available software.

2 MEASURING PERSONALITY IN SURVEY RESEARCH

Recent scholarship in political science has increasingly emphasized the role of personality traits in explaining public opinion and political behavior. This research has greatly expanded our understanding of how personality fundamentally structures behaviors, attitudes, and learning. The most prominent personality trait in contemporary scholarship is 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 the broadest 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 and engage the world around them. These include 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. Examples in this category include social dominance orientation (Sidanius et al. 2004) and right wing authoritarianism (Stenner 2005).

Yet, the political science discipline has only scratched the surface in understanding the ways in which personality traits shape attitudes and behavior and how traits are in turn affected by political events. Many widely used – and extensively validated – personality measures with potential political implications have appeared in the political science 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).

2.1. *Long batteries and short surveys*

One explanation for the failure of many personality traits to filter into the political science literature is surely that most established personality inventories contain far too many questions for inclusion on surveys. Standard practices in social and cognitive psychology result in evaluative batteries containing dozens or even hundreds of question items. For example, Cacioppo and Petty (1982) originally proposed a 40-item battery to measure need for cognition. Later, they proposed an "efficient" battery containing 18 questions chosen from the original 40 (Cacioppo and Petty 1984).

In most cases, survey researchers avoid these large scales. In part this is because including large batteries on surveys is expensive. For most researchers, the main cost associated with surveys is the per-question fee charged by survey firms. However, the desire to avoid large batteries also reflects the higher rate of attrition and non-response associated with lengthy and repetitive instruments. Longer surveys lead to higher rates of unit nonresponse (e.g., Heberlein and Baumgartner 1978; Yammarino, Skinner and Childers 1991; Burchell and Marsh 1992; Crawford, Couper and Lamias

2001; Galesic and Bosnjak 2009) higher rates of halting an interview (e.g., Sheatsley 1983) and more item nonresponse (e.g, Anderson, Basilevsky and Hum 1983). Indeed, lengthy and repetitive surveys generally increase the burden on participants who compensate by providing lower quality responses. (Herzog and Bachman 1981; Krosnick 1999; Krosnick et al. 2002).

Faced with this dilemma, most researchers seek to develop a reduced-form version of large personality scales. This involves selecting a subset of items from the larger scale to administer to respondents. For example, it can be argued that the entrance of the big five personality traits into theories of public opinion and political behavior is in part the result of the validation of the ten item personality inventory (TIPI) (Gosling, Rentfrow and Swann 2003). Prior to the advent of TIPI, the most common battery measuring the big five was the 44-item Big Five Inventory, which was itself a shorter alternative to the monstrous 240-item NEO Personality Inventory-Revised (Gosling, Rentfrow and Swann 2003; McCrae and John 1992).[1] Indeed, developing reduced-form scales is nearly a cottage industry within the personality literature. Some recent examples include: Stanton et al. (2002), Russell et al. (2004), Richins (2004), Matthews, Kath and Barnes-Farrell (2010), and Thompson (2012). Further examples are presented in Table 1, which provides basic information about several personality inventories and their reduced-form alternative that I reference below.

These reduced scales are usually developed in one of three ways. First, scholars may examine the properties of the scale to make theoretically motivated decisions about which items to include. So, for instance, Ames, Rose and Anderson (2006, p 441-442) developed the reduced-form narcissistic personality inventory (NPI-16) by choosing items with strong "face validity" and to ensure coverage of theorized domains or facets uncovered in previous studies of the NPI. Thus, scholars in essence calibrate the model based on their own intuitions as well as the samples used in previous validation studies.

Another approach is to choose items based on factor loadings in the original publication to select items. For instance, in designing a two-item battery measuring need for cognition for the

[1] TIPI is unusual in that the items are not actually identical to the items in the larger scale it replaced. The Big Five Inventory asks respondents if single adjectives apply to themselves (e.g., "disorganized"), while TIPI combines two adjectives into each item (e.g., "disorganized, careless"). For this reasons, I do not address the TIPI measure directly.

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 (2002)	Muncer and Ling (2006)
Length	40	15
Systemizing quotient	Baron-Cohen (2002)	Wakabayashi et al. (2006)
Length	40	22
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	15	2
Social equality	Pratto et al. (1994)	
Length	15	2
Need for affect	Maio and Esses (2001)	
Length	26	4

For each scale, the reduced-form battery contains a strict subset of items in the larger battery.

American National Election Study (ANES), Bizer et al. (2004, p 13) chose “the two items that loaded most strongly on the latent construct in Cacioppo and Petty’s (1982) factor analysis.”[2] In this way, the need for cognition scale on the ANES is, in essence, based on the responses of 96 individuals drawn from the faculty at the University of Iowa or workers on assembly lines in the Iowa City-Cedar Rapids area in the early 1980s (Cacioppo and Petty 1982, p 118-119).

A final approach is to administer the battery to one or more convenience samples and to then use these new responses to choose a subset of existing 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] A nearly identical approach was taken in developing the three-item need to evaluate scale.

2.2. *Why adaptive batteries?*

In each approach, scholars developing reduced scales rely on estimates from calibration samples of some kind. Once a reduced inventory is chosen, it is administered *all* respondents. The adaptive personality inventories I propose below also rely on calibration studies of external samples to aid researchers choosing amongst potential survey items. However, APIs differ in that the goal is not to use this prior information to choose a single battery for all respondents, but rather to taylor the reduced battery to each respondent in a manner designed to maximize measurement precision. Specifically, I use estimates generated from the calibration sample to choose a reduced-form battery for each respondent in a mathematically informed way with the goal of minimizing measurement error.

APIs are an application of computerized adaptive testing (CAT) algorithms, which is itself an extension of item response theory (Lord and Novick 1968).[3] CAT is based on the notion that questions should be chosen for each respondent based on what we know about them. Ignoring prior information leads us to waste valuable survey time asking respondents questions that are not revealing. So, for instance, there is little purpose in asking an individual who self-identifies as "Extremely Conservative" questions designed to distinguish between moderates and liberals.

In essence, CAT algorithms use prior information about respondents and question items to quickly and accurately place survey takers on some latent scale. Prior information about respondents derives from their initial answers to items in the battery. Prior information about the items derives from pre-testing the questionnaire. This establishes how specific questions relate to the latent scale of interest and provides the needed item-level parameters for the CAT algorithm to operate.

A simple example can help illustrate the two major advantages of the API framework. Consider the Empathy Quotient battery which contains 40 items, where respondents can choose among four possible responses to each item. Assume further that there is space for six items on our survey.

[3] Applications of item response theoretic models are by now familiar in political science areas ranging from public opinion (e.g., Treier and Hillygus 2009), to the analysis of roll calls (e.g., Jackman 2001), to the measurement of democracy (Treier and Jackman 2008).

The first advantage of APIs is that it allows researchers to include a larger number of possible combinations of questions. As stated above, when confronting a large personality inventory, scholars typically choose *just one* subset of six items. However, in this example there are actually $\binom{40}{6} = 3{,}838{,}380$ possible reduced batteries. A six-item adaptive battery, on the other hand, would allow us to include as many as 1,024 unique question combinations.[4]

A second, and more important, advantage is that APIs allows us to ask *better* question combinations to reveal the most information about each respondent based on our expectations as to how they will score on the battery. With both adaptive and static batteries, we will be able to partition respondents into $4^6 = 4{,}096$ categories based on their response profiles.[5] For fixed batteries many of these potential response profiles are rarely (if ever) observed. Individuals who strongly disagree with the statement, “Seeing people cry does not really upset me” are unlikely to agree to any degree with the statement, “I really enjoy caring for other people.” By carefully selecting among potential questions, APIs makes it far more likely that we will observe the full range of potential response profiles as it chooses questions for which respondents have a higher probability of answering in many potential categories.

3 ADAPTIVE PERSONALITY INVENTORIES

Having motivated the need for APIs, in this section I briefly provide the details of *one* implementation that we use below.[6] In essence, APIs take a large population of potential items and selects among them to efficiently place respondents on some latent trait. Roughly speaking, the algorithm chooses items that match the respondent’s position on the trait (i.e., the respondent is likely to answer in one of several response categories) and items that are highly discriminatory.

[4] With four-category items, the total number of possible combinations is 4^{n-1} where n is the number of items.

[5] A response profile is the unique combination of values to the questions asked. So, for instance, responding “1” to the first question and “3” to the second question, the response profile would be (1,3).

[6] For more technical discussions of various CAT implementations see Chen, Hou and Dodd (1998), van der Linden (1998), Chen, Ankenmann and Chang (2000), Pastor, Dodd and Chang (2002), Van Rijn et al. (2002), Veldkamp (2003), and Penfield (2006).

Table 2: Basic elements of adaptive personality inventories

	Purpose	Description
1	Estimate positions	A provisional trait estimate, $\hat{\theta}_j$, is created based on first i responses.
2	Item selection	The item that optimizes some objective function is chosen.
3	Administer item	
4	Check stopping rule	Pre-defined stopping rules may include reaching the maximum number of questions for and battery.
5a	Repeat steps 1-4	If the stopping rule has not been reached, administer new items.
5b	Return estimate	If the stopping rule has been reached, calculate a final $\hat{\theta}_j$.

3.1. *Algorithm essentials*

The elements of an adaptive personality inventory are shown in Table 2 (Segall 2005, p 4). First, estimates ($\hat{\theta}_j$) are generated for each respondent's true position on the trait (θ_j). Before the first item is administered, this estimate is based on our prior assumptions about θ_j. I assume a common prior for all respondents, $\theta_j \sim \pi(\theta)$. After each item in the inventory is administered, these expectations will be calculated as $\mathbb{E}(\theta_j|\mathbf{y}_j)$, where $\mathbf{y}_j$ represents responses to previously administered items in the battery.

Second, the next question item is selected out of the available battery. CAT chooses the item that optimizes some pre-specified objective function. I use the minimum expected posterior variance (MEPV) criterion in the applications below. In a review of competing item selection rules for the graded response model, Choi and Swartz (2009) note that this approach performs "equally well" to the more commonly used methods such as maximum posterior weighted information (MPWI), but that "the MEPV method would be preferred from a Bayesian perspective" (p. 18). Thus, my use of this method is more a reflection of taste than an indication that MEPV is in some way superior. Indeed, despite the large number of potential estimation methods, Choi and Swartz (2009) note that "for item banks with a small number of polytomous items, any of the methods ... are appropriate" (p. 18), which indicates that this choice is not critical.

The third stage of the algorithm is to administer the chosen item and record the response. Fourth, the algorithm checks some stopping rule. In my examples below, the stopping rule is that the number of items asked of the respondent has reached some maximum value. Once this stopping criteria has been met, the algorithm produces final estimates of $\hat{\theta}_j$ and terminates.

3.2. *Details of the general model for categorical responses*

Most personality inventories include multiple response options (Likert scales). This requires that I alter the algorithm specified in Montgomery and Cutler (2013) to accomodate ordinal responses. In an IRT framework, this is commonly modeled using a graded response model (GRM) (Samejima 1969). For each item i we assume that there are C_i response options. There is therefore a vector of threshold parameters defined as $\kappa_i = (\kappa_{i,0}, \kappa_{i,1}, \dots, \kappa_{i,C_i})$, with $\kappa_{i,0} < \kappa_{i,1} \leq \kappa_{i,2} \leq \dots, < \kappa_{i,C_i}$, $\kappa_{i,0} = -\infty$, and $\kappa_{i,C} = \infty$. In addition, each item is associated with a *discrimination* parameter a_i, which indicates how well each item corresponds to the underlying trait in question.

To calculate the likelihood function, we need to estimate $P_{ijc} \equiv P(y_{ij} = c|\theta_j)$, which is the probability of answering in the c^{th} category for item i given the ability parameter for respondent j, denoted θ_j. This quantity, however, cannot be calculated directly. Instead, we need to define P^*_{ijc}, which is $\sum_{k=c+1}^{C_i} P_{ijk}$. Given these quantities, we calculate $P_{ijk} = P^*_{ij,k-1} - P^*_{ijk}$. Note that $P^*_{ij0} = 1$ and $P^*_{ijC_i} = 0$ in all cases. Under a logistic response assumption, therefore,

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

The likelihood function is then, $L(\theta_j) = \prod_{i=1}^{n} \prod_{k=1}^{C_i} P_{ijk}^{I(y_{ij}=k)}$, where $I(\cdot)$ is the indicator function.

Calculating skill parameter: Assuming that the respondent has answered one or more previous items, calculating the expected a posteriori (EAP) estimate for the position of individual j given previous responses ($\mathbf{y}_j$) and the prior distribution is,

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

The posterior variance is then

$$\text{Var}(\theta_j) = \mathbb{E}\big((\theta_j - \hat{\theta}_j^{(EAP)})^2|\mathbf{y}_j\big) = \frac{\int(\theta_j - \hat{\theta}_j^{(EAP)})^2\pi(\theta_j)L(\theta_j)d\theta_j}{\int\pi(\theta_j)L(\theta_j)d\theta_j}. \tag{3}$$

As this involves a single dimension, we can estimate both using numerical integration.[7]

Item selection: As stated above, I use the MEPV item-selection criteria. To choose an item under this scheme, we follow three steps. First, we need to use the current estimate of $\hat{\theta}_j$ to estimate P_{mjk} for *each* possible response to a candidate (unasked) item m using Equation 1. Second, we need to estimate $\hat{\theta}_j^{(EAP)*} = \mathbb{E}(\theta_j|\mathbf{y}_j, y^*_{mjk})$, which is the EAP for individual j for all possible responses to candidate item m. With this, we calculate $\mathbb{E}\big((\theta_j - \hat{\theta}_j^{(EAP)*})^2|\mathbf{y}_j, y^*_{mjk}\big)$, which is the posterior variance associated with each possible response to item m. Third, we combine these elements to estimate the expected posterior variance (EPV) for the candidate item, which is

$$\text{EPV}_m = \sum_k P_{mjk}\text{Var}(\theta_j|\ldots, y^*_{im} = k). \tag{4}$$

This is the posterior variance for $\hat{\theta}$ we *will* have given each possible response to item m weighted by the probability of observing that response all conditioned on our *current* estimate $\hat{\theta}_j$. We select the item that minimizes this quantity.

Stopping rule: In the examples below, the algorithm stops offering items when the number of questions reaches a pre-specified threshold n_{max}. An alternative, however, is to stop when the posterior variance, $\text{Var}(\theta_j|\mathbf{y}_j)$, falls below some pre-specified level.

Prior selection: The prior distribution for θ_j is denoted $\pi(\theta_j)$. A natural choice is a conjugate normal prior $\pi(\theta_j) \sim N(\mu_\theta, \tau_\theta)$, where τ_θ denotes the standard deviation. Generally, these are chosen to be diffuse, but not completely uninformative. In the examples below, I discuss further how to select these parameters.

[7] The `catSurv` software I have developed allows three different numerical integration routines so users can determine their optimal tradeoff in terms of speed and accuracy.

Table 3: 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	6	10,249	2,050
Empathy	40	15	4	10,145	2,029
Systemizing	40	22	4	10,145	2,029

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

4 APPLICATIONS

In this section, I demonstrate the advantages of APIs in an empirically informed simulation, an experiment conducted with convenience samples, and data from the 2016 ANES Pilot Study, which included an API batteries measuring need for cognition. The simulation and experiment illustrate API's superior performance relative to fixed-reduced batteries in terms of precision and accuracy. The ANES study serves as an in-depth case study of how APIs can be developed and fielded, and also validates API estimates using a nationally representative sample.

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

I first demonstrate the advantages of APIs in a simulation based on a dataset of responses to four personality inventories. The goal is to show the advantages of APIs relative to fixed reduced-form batteries. I use data collected by `personality-testing.info`, which provides tens of thousands of individual responses to prominent personality inventories.[8] I chose 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 my analytical approach, it is essential that the reduced-form battery consist 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 3.

To begin the analysis, I fit a GRM using five-sixths of the respondents (a randomly chosen training sample). I then complete the calibration of the API by choosing a standard normal prior

[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.

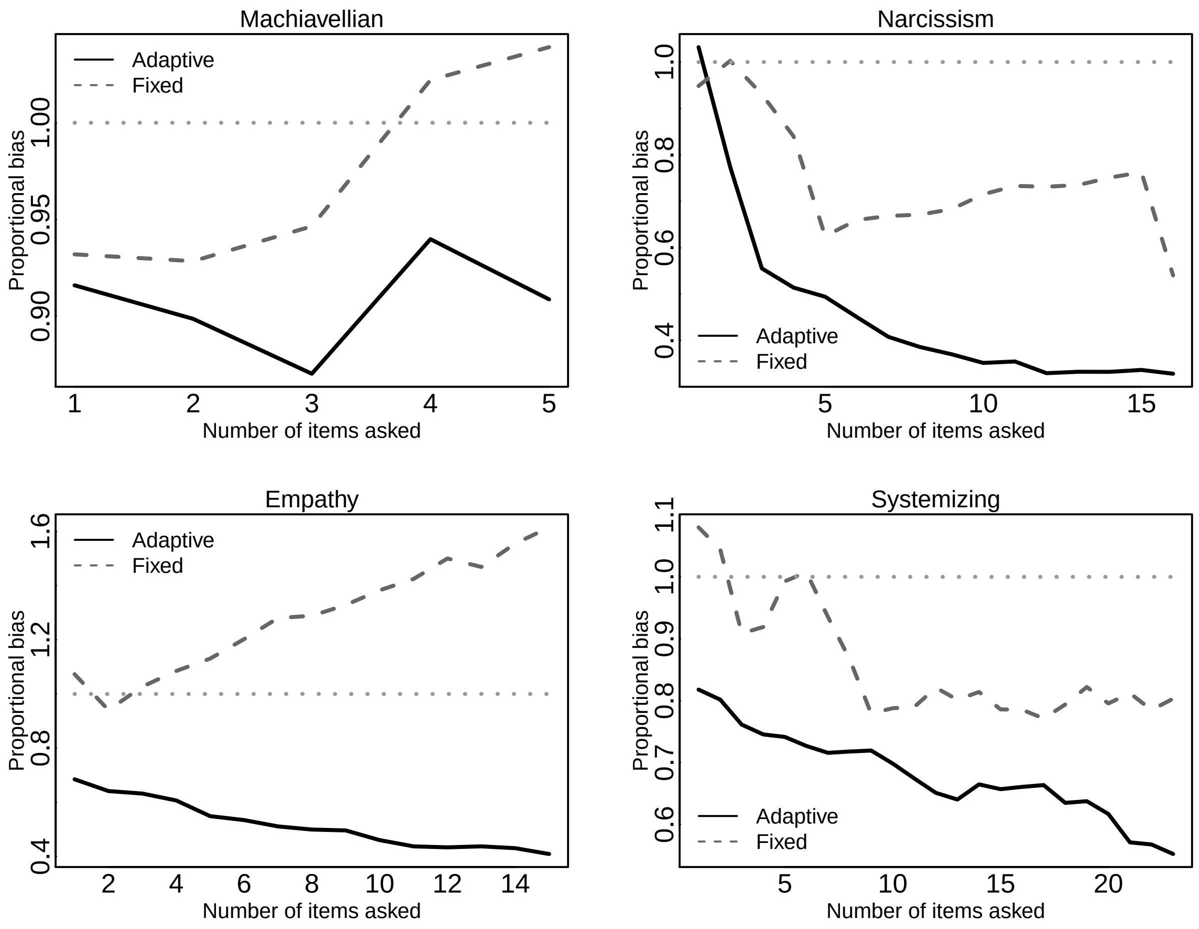
distribution.[9] I then turn to the remaining sample (the test sample) and use individuals' *recorded answers* to estimate their scores under the assumption that I know only their responses to question as chosen by (1) the reduced-fixed scale, (2) the reduced-adaptive scale, and (3) a randomly selected reduced battery. So, for example, if the fixed battery calls for asking question-items 20, 14, and 3, I calculate their score using their real responses to *just* those three questions. Likewise, I let the API choose one items and then recorded an "answer" according to the real responses in the dataset. The algorithm would then customize the selection of the next question as described above and so on. Finally, I administered a randomly selected reduced-form battery. That is, for *each* individual I chose items at random to administer and calculated scores based only on those responses.

To evaluate the performance of each reduced battery, I also estimated respondents' position on the latent trait using their recorded responses to the *entire* battery. That is, I used all of their responses to all of the questions to estimate their latent position. For my calculations below, I treat these scores as the "true" value and evaluate the various reduced batteries in terms of how well they approximate this benchmark. Note that I use a GRM fit to the entire 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.

I turn first to 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). For this inventory, respondents are asked to choose which of two statements best applies to themselves. For example, one item asks respondents to choose between: (a) "Sometimes I tell good stories" and (b) "Every-

[9] Specifically, I used the training sample to fit a GRM for each battery using the `grm()` command from the `ltm` R package (Rizopoulos 2006; R Core Team 2015). The package uses marginal maximum likelihood methods to estimate item-level parameters. It is important to note that the default settings for the `grm` function identify the model by assuming, first, that the first item included in the dataset loads positively on the latent trait and, second, that the ability parameters (θ_i) are distributed according to the standard normal distribution for these APIs. Since I anticipated that the distribution of latent traits in training sample would be similar to the test sample, I therefore used the standard normal prior. If a researcher has reason to believe that the calibration sample differs in some important respect from the target population where the API will be used, other decisions may be appropriate. I discuss these issues further in the ANES application below.

Figure 1: Proportional bias for adaptive and fixed batteries



Bias is calculated relative to a baseline of random selection. Solid black lines show the adaptive battery, and the gray lines show the fixed battery shown in Table 1.

body likes to hear my stories." Another item gives the options: (a) "Being an authority doesn't mean that much to me" and (b) "People always seem to recognize my authority."

Although the original battery contained 40 items, Ames, Rose and Anderson (2006) developed a widely used 16-item version (NPI-16). In the top-right panel of Figure 1, I compare the performance of NPI-16 with an API in terms of mean absolute bias.[10] To provide a common baseline, I estimate the bias of the fixed and adaptive batteries relative to a random battery. Let the mean absolute bias estimates for the adaptive, fixed, and random batteries be denoted MAB_{adapt}, MAB_{fixed}, and MAB_{random}. In Figure 1, I plot $\frac{\text{MAB}_{adapt}}{\text{MAB}_{random}}$ and $\frac{\text{MAB}_{fixed}}{\text{MAB}_{random}}$. This provides a meaningful

[10] Recall that we use estimates generated from using the entire 40-item battery as their "true" position (θ_j). For observation j, absolute bias is $|\hat{\theta}_j - \theta_j|$. Comparing scale performance using squared error ($(\hat{\theta}_j - \theta_j)^2$), posterior variance ($Var(\hat{\theta}_j)$), and squared loss ($(\hat{\theta}_j - \theta_j)^2 + Var(\hat{\theta}_j)$) produces substantively identical results.

baseline as any reduced-form scale should be able to outperform random selection (shown as the horizontal line at 1.0). The top-right panel of Figure 1 shows that the MAB of the adaptive NPI battery of length 16 is roughly 60% lower than the MAB for the random battery, while the fixed battery provides only a 40% improvement.

Next I 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). Items ask respondents to agree or disagree with statements such as, "The best way to handle people is to tell them what they want to hear" or "The biggest difference between most criminals and other people is that the criminals are stupid enough to get caught." I compare the API method to the 5-item Trimmed MACH recently proposed by Rauthmann (2013). The top-left panel of Figure 1 shows the proportional error of the API versus the Trimmed MACH battery. Clearly, the adaptive battery is significantly better in reducing errors. Indeed, by the fifth item, the MACH Trimmed scale is performing worse than simply choosing survey items at random, while the adaptive method provides a 10% improvement.

Third, I examine the empathizing-systematizing batteries developed by Baron-Cohen (2002) to explain sex differences in the mind. Empathizing is "the way in which we understand the social world, the emotions and thoughts of others and how we respond to these social cues." On the other hand, "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] Items on the empathizing scale asked respondents to agree or disagree on a four-point scale with statements such as, "I get upset if I see people suffering on news programmes" or "Seeing people cry doesn't really upset me." I compare our API, which draws from the full 40-item battery, with the 15-item reduced empathizing scale proposed by Muncer and Ling (2006). The systematizing scale asks respondents to agree or disagree with items such as "I can easily visualise how the motorways in my region link up," and, "I find it difficult to learn my way around a new city." I compare the API

[11] Each scale also included 20 "buffer" questions that I ignore in this analysis.

approach with the 22-item reduced battery proposed by Wakabayashi et al. (2006).

The results are shown in the bottom panels of Figure 1. The figure makes clear that the items in the reduced-form empathizing scale were not well selected for this sample. The bottom-left panel of Figure 1 shows that the fixed scale does uniformly worse relative to simply randomly choosing items out of the original 40-item battery. On the other hand, the dynamic battery does significantly better, having proportional error rates of about 45%. The bottom-right panel of Figure 1 does not reveal such a stark contrast for the systematizing scale, but here again it is clear that the dynamic battery does very well against both a random battery and the standard fixed-reduced battery in the literature. Moreover, the proportional error rate is quite low, showing that the API produces less than half of the level of error as the random selection method.

4.2. *Experimental study*

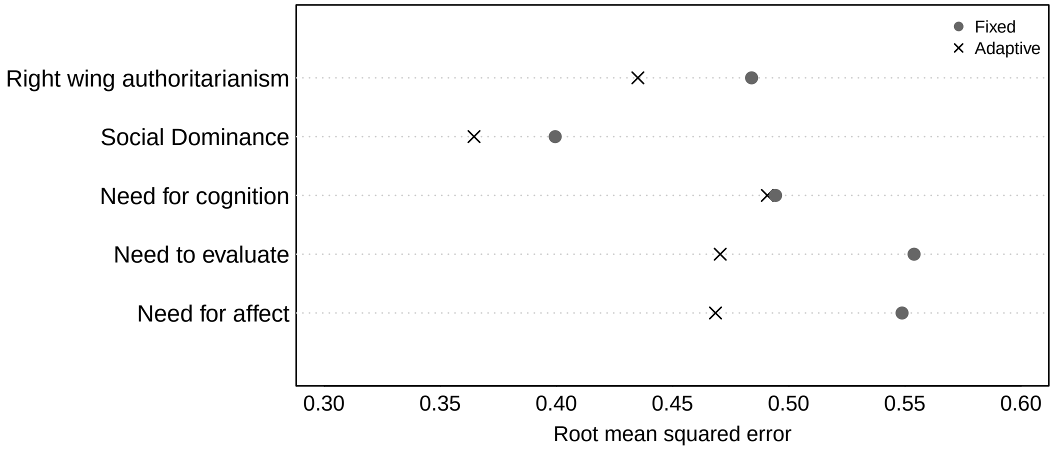
I next present results from an experiment conducted using convenience samples recruited via Amazon's Mechanical Turk (AMT) service to compare the performance of fixed-reduced batteries with an API of the same length. In the fall of 2014, I administered full-length versions of five personality inventories that have been included in reduced forms on the 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, I calibrated an API for each inventory as described above using a standard normal prior distribution.

In the spring of 2015, I then recruited 1,335 new respondents who were randomly assigned to receive either fixed-reduced battery as used by the ANES or an API of the same length.[13] After completing the reduced battery, all subjects then answered all remaining questions in the full battery in a random order. As before, I estimate respondents' scores using the full battery and treat them as respondents' "true" positions on the latent scale. I then compare the scores calcu-

[12] I 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).

Figure 2: Root mean squared error for adaptive vs. fixed batteries on the ANES



Points show the root mean squared error for observations assigned to answer the fixed batteries as they appeared on the ANES (see Table 1) and those randomly chosen to answer adaptive batteries of the same length. Randomization occurred before each battery. Point estimates were calculated relative to estimates generated for each respondent using the full inventory. In each case, the adaptive battery provided more accurate estimates.

lated using only questions selected by the fixed batteries with the scores calculated using questions selected by the API using these true positions as a common benchmark.

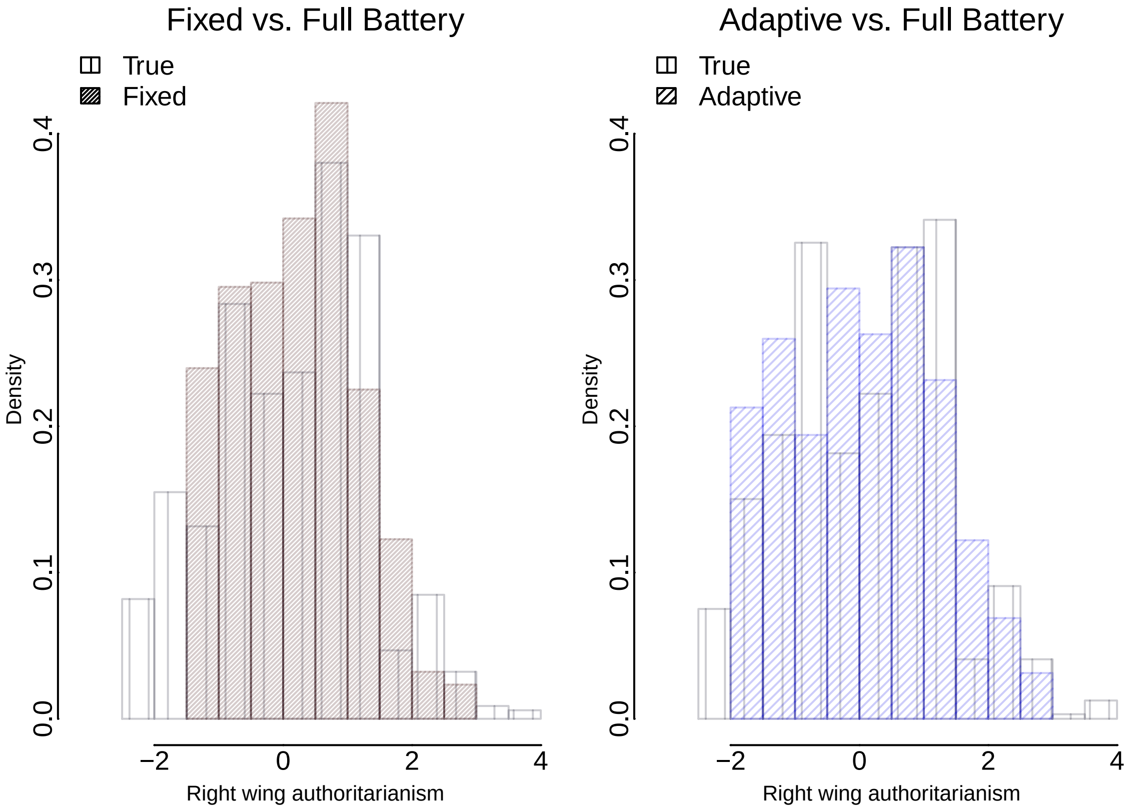
Figure 2 shows the root mean squared error (RMSE)[14] for respondents answering either the fixed or adaptive scales.[15] As can be seen, the APIs provided more accuracy than the fixed batteries except in the case of NFC. (The NFC battery on the ANES is only two items long, and is therefore not well suited for APIs.) These results show that APIs provide more accurate estimates than widely used fixed batteries, even when there is only space for a few items.

I can further demonstrate that this improved accuracy has important consequences beyond mere measurement. To do this, I focus on the RWA measure, which originally had 30 items but was reduced to five on the ANES 2013 Internet Followup Study. Figure 3 shows the estimated distributions of RWA for individuals assigned to complete the fixed and adaptive batteries. The shaded distributions show the density estimated using only the reduced battery, while the unshaded distributions show the density estimated for these same respondents using the complete 30-item inventory. The figure shows that the fixed battery does a relatively poor job estimating the positions at

[14] See Footnote 10.

[15] Estimates were generated using a GRM model fit with the complete responses from the second sample. This ensures that the estimates are all on a common scale and are thus comparable.

Figure 3: 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=649) 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.

the extreme ends of the spectrum, shown by the difference in the shaded and unshaded densities in the left panel.

By failing to accurately recover the position of more evaluatively extreme individuals, the fixed RWA battery is censoring the distribution of RWA values. This, in turn, biases our understanding for how RWA relates to other important factors. To illustrate this, I included measures of 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. I estimated separate regressions by treatment condition using RWA as an explanatory variable and these related constructs as dependent variables.[16] I then estimated the "true" value for these regression coefficients using respondents' scores as estimated from the full battery.[17] I calculate the bias as the difference between the regression coefficient estimated when RWA was measured using the reduced and full batteries. The results, shown in Figure 4, illustrate clearly that the censored measure of RWA from the fixed battery upwardly biases regression coefficients, leading us to conclude that RWA is a stronger predictor of these constructs than is actually the case. That is, the point estimates and 95% confidence intervals for the regression coefficients relating RWA to these outcomes are biased upwards when measured using a fixed battery (solid circles with dashed lines) while the upward bias is reduced when using an adaptive battery (open squares and solid lines).

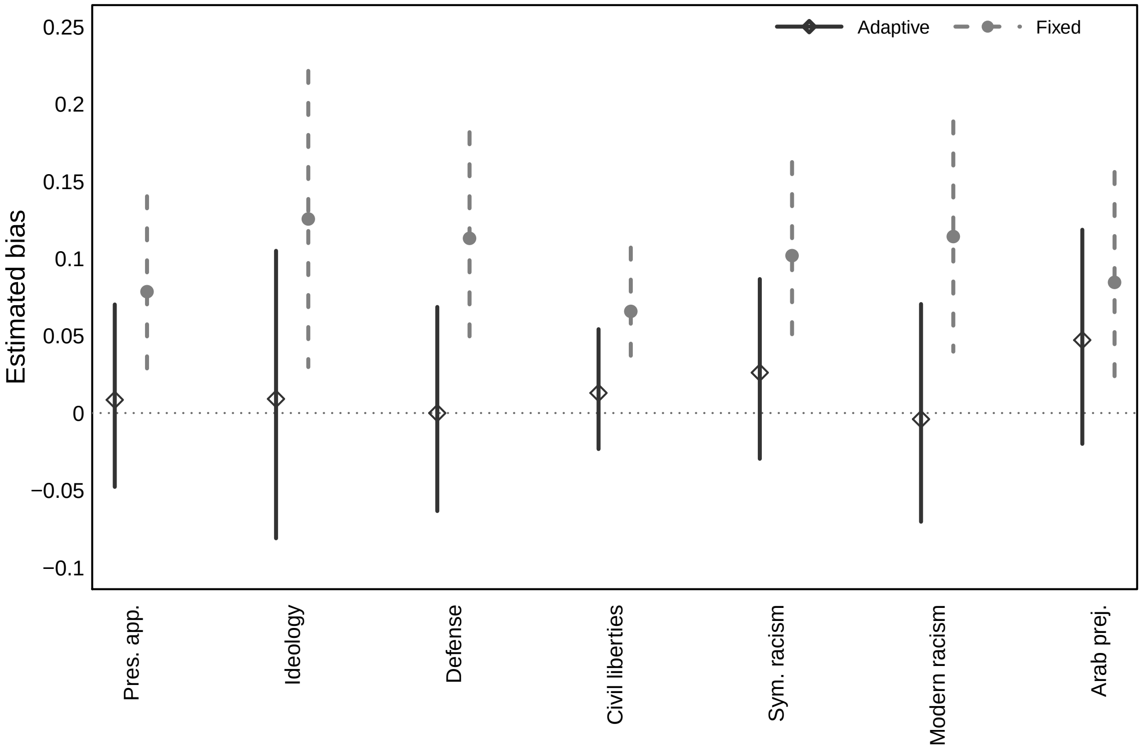
4.3. *2016 ANES Pilot Study*

In my third application, I present a detailed case study of an API measuring the need for cognition that was included on the 2016 ANES Pilot Study. (A more technical guide for calibrating and administering an API is provided in the SI Appendix.) In addition to providing an illustrative example, the purpose of this section is to test the external validity of API measures on a nationally representative survey conducted by a national polling firm (viz., YouGov).

[16] In these regressions, I control only for race, gender, and level of education.

[17] These baseline regressions were again fit within each treatment group.

Figure 4: 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 is biased upwards when estimates of respondents' latent position are censored 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 scales 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 these same answered an API of the same length (n=649). All regressions controlled for gender, race, and level of education.

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 attempts to frame issues.

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. The model was then validated using 419 introductory psychology students at the University of Missouri, a 103 person sample of psychology students at the University of Iowa, and further sample of 97 drawn from the same source. The result was a 34-item inventory that has been widely applied in the fields of psychology and political science.

However, Cacioppo and Petty (1984) subsequently recommended an 18-item “efficient” version of inventory that is shown in Figure 5. Two aspects of this reduced battery are noteworthy. First, it is from this 18-item inventory that Bizer et al. (2000) chose the items for inclusion of the ANES. Second, in order to retain balance for positively and negatively oriented items, Cacioppo and Petty (1984) slightly altered the wording of four items to reverse their orientation.[18]

To calibrate the adaptive personality inventory, I combined data from three separate samples. First, I used data from the December 2014 wave of The American Panel Survey (TAPS). TAPS is a monthly online panel survey of over 2,200 people. 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. In a typical month, about 1,700 of the panelists complete the online survey. After removing respondents who completed less than 25% of the items, I have 1,506 respondents to the NFC battery.

[18] Specifically, NFC15 was originally worded, “The idea of relying on thought to make my way to the top does not appeal to me.” NFC16 was originally, “The notion of thinking abstractly is not appealing to me.” NFC 24 was originally, “I find little satisfaction in deliberating hard and for long hours.” NFC29 was originally, “I don’t like to have the responsibility for handling a situation that requires a lot of thinking.”

Figure 5: Need for cognition (Reduced)

Do you agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly with the following statement? (*Response options*: Agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, disagree strongly)

NFC1. I really enjoy a task that involves coming up with new solutions to problems.

NFC4. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

NFC10. Learning new ways to think doesnt excite me very much.

NFC12. I usually end up deliberating about issues even when they do not affect me personally.

NFC15. The idea of relying on thought to make my way to the top appeals to me.

NFC16. The notion of thinking abstractly is appealing to me.

NFC19. I only think as hard as I have to.

NFC21. I like tasks that require little thought once I've learned them.

NFC22. I prefer to think about small, daily projects to long-term ones.

NFC23. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

NFC24. I find satisfaction in deliberating hard and for long hours.

NFC29. I like to have the responsibility of handling a situation that requires a lot of thinking.

NFC31. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.

NFC32. Thinking is not my idea of fun.

NFC33. I try to anticipate and avoid situations where there is likely a chance I will have to think in depth about something.

NFC39. I prefer my life to be filled with puzzles that I must solve.

NFC40. I would prefer complex to simple problems.

NFC43. Its enough for me that something gets the job done; I dont care how or why it works.

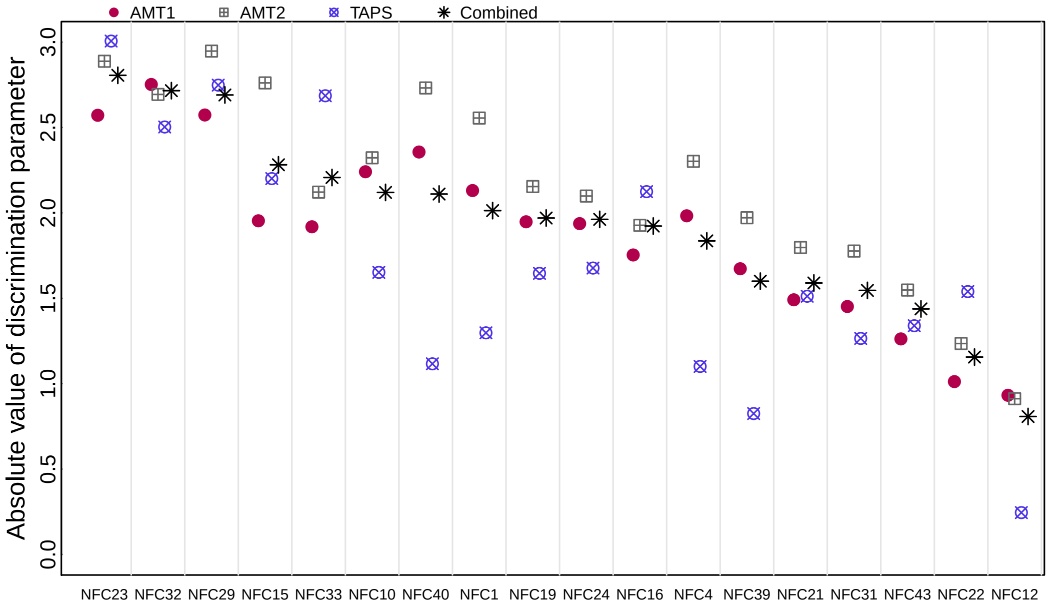
To supplement TAPS, I used responses to the 18-item NFC battery from the two convenience samples recruited via Amazon's Mechanical Turk online workforce used in the experiment described above. While not a representative sample, AMT provides an easy way to administer the survey to a diverse set of respondents (Berinsky, Huber and Lenz 2012). As Embretson (1996) notes, one of the 'new rules' of IRT models is, "Unbiased estimates of item properties may be obtained from unrepresentative samples" (p. 342). What is needed to calibrate APIs is not so much a representative sample, as a sample that is large and diverse along the dimension of interest.

The first AMT sample was collected in the fall of 2014 (n=1,204) and the second sample was recruited in the spring of 2015 (n=1,333). In order to fit the adaptive test, I fit separate GRM models for each of the three samples as well as for the combined sample. This history of the NFC scale is relevant for this exercise because the TAPS survey included the original full 34-item battery while only the 18-item battery was administered to the AMT samples. To make the data work similarly across samples, therefore, I calibrated the API using only responses to the 18-items included on the reduced battery. Further, I reverse-coded responses so that items were all oriented identically across samples.

The absolute values for the discrimination parameters for the GRM models fit to each sample as well as the estimates from the pooled sample are displayed in Figure 6. These parameters are typically the most influential in determining which survey items are administered most often to respondents. Specifically, items with larger discrimination parameters are considered to be more revealing about the underlying latent trait and, *ceteris paribus*, more likely to be selected by the adaptive battery. The figure shows that the importance of the items remains largely consistent across samples. The only noticeable pattern is that the TAPS estimates (blue boxes filled with an "x") are often lower in absolute terms, suggesting that this population overall gives somewhat "noisier" responses relative to AMT respondents. However, the discrimination parameters for the TAPS sample do diverge significantly in some instance—especially for the four items that were negatively oriented in the broader inventory.[19] However, note that the pooled estimates are very

[19] These are items NFC1, NFC4, NFC39, and NFC40.

Figure 6: Discrimination parameter loadings for the TAPS, AMT Sample 1 (Fall 2014), AMT Sample 2 (Spring 2015), and combined samples.



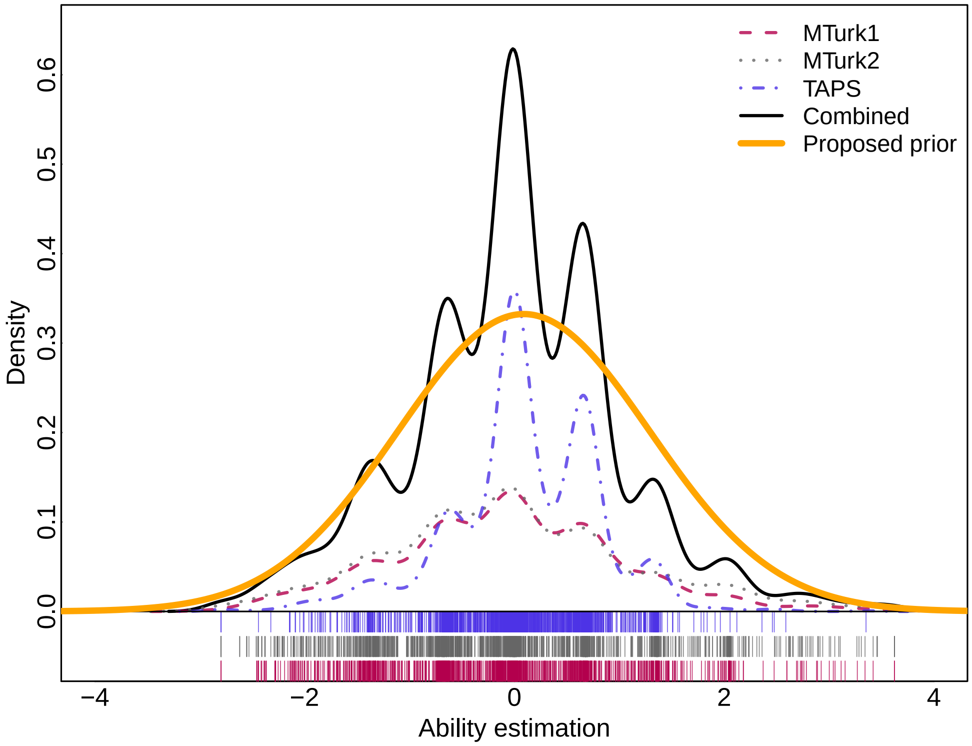
The figure shows the discrimination parameters estimated for three separate samples as well as the combined estimate. The estimates, which indicate the degree to which the item is informative about the underlying latent trait, roughly correspond across samples. I propose using the "combined" (black asterix) estimates for the ANES pilot study.

similar to those estimated for AMT Sample 1.

After fitting the GRM model, the next task is to select an appropriate prior for the adaptive battery. In particular, I aimed to choose a prior reflecting the true national distribution of NFC, which I assumed was best reflected in the TAPS sample. Figure 7 shows the distribution of NFC for the three samples as well as the pooled estimate. The thick orange line shows the prior I selected, which is a normal prior centered at the mean of the TAPS estimates and a standard deviation of 1.2. This is a proper prior that is sufficiently uninformative to encompass the NFC ability estimates in all three samples.

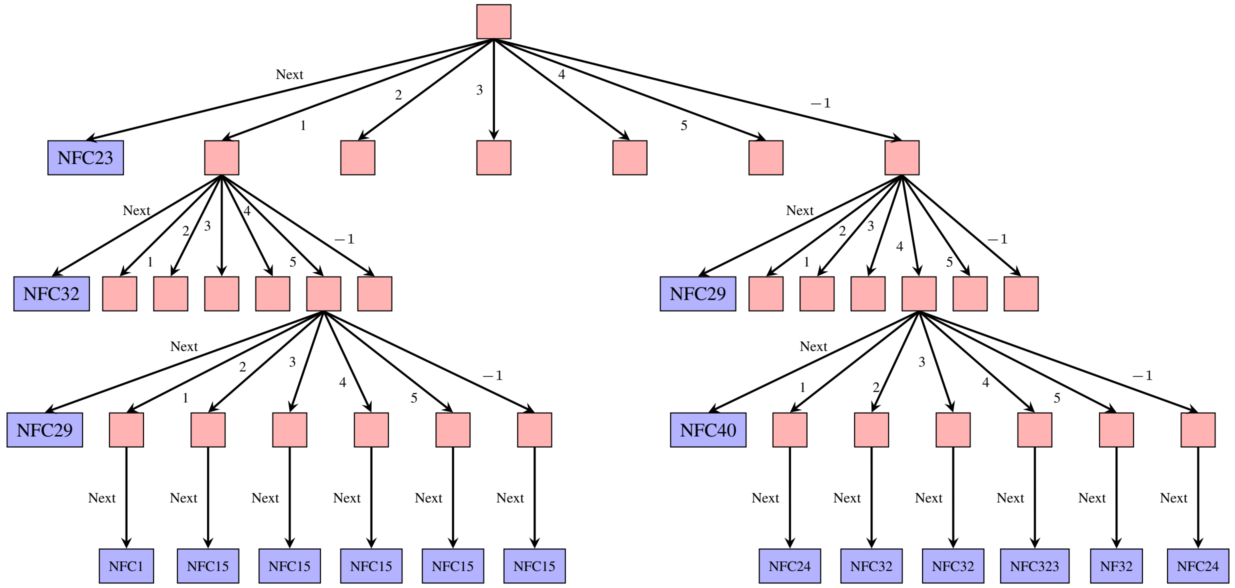
These two elements, a prior and the parameters from a GRM model, are all the ingredients needed for calibrating an API for the NFC personality inventory. However, it is worth briefly discussing some practicalities. In an educational testing setting, it is often assumed that item selection needs to occur in real time as test takers answer questions. This makes sense in a testing environment where the number of items a single test-taker complets will be relatively large and there are ancillary concerns (e.g., preventing overexposure of certain items, which can lead to

Figure 7: Densities of need for cognition ability parameters for the TAPS (December 2014), AMT Sample 1 (Fall 2014), AMT Sample 2 (Spring 2015), and combined samples.



The figure shows the density for the estimated ability parameters for three separate samples as well as the combined estimate. The estimates, which indicate the location of respondents' on the underlying latent scale, are similar across samples although the TAPS sample is more clustered towards the center of the distribution. The ANES used the prior density (shown as the thick orange line) which is centered at the mean position of the TAPS sample and is sufficiently uninformative to allow all observed values of NFC in all three samples.

Figure 8: 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 API included on the 2016 ANES Pilot Study. Blue boxes represent a terminal node, while red boxes are internal nodes in a tree. The labels on the branches indicate respondents' answers. A "-1" in this tree 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.

cheating by future test takers). In a survey setting, however, a more practical approach will often be to pre-calculate a complete branching scheme.

Figure 8 depicts portions of the complete branching scheme for the four-item NFC API. In this figure, blue boxes represent a terminal node, while red boxes are internal nodes in a tree. The labels on the branches indicate respondents' answers. A "-1" in this tree 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 refused to answer NFC23 will be asked NFC29. A respondent who then answers "4" will be asked NFC40.

For longer batteries, a full enumeration of the scheme might be difficult.[20] 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. This table was provided to

[20] In the SI Appendix I briefly discuss an online tool I have developed for implementing larger APIs that is currently in development for public release.

Table 4: Usage of NFC items in branching schemes and in observed response profiles for the need for cognition API on the 2016 ANES Pilot Study

	# of branchings including item	# of respondents answering item
NFC23	216	1,200
NFC29	211	1,158
NFC32	202	1,078
NFC15	159	796
NFC40	49	387
NFC24	15	122
NFC1	7	17
NFC21	3	21
NFC10	1	3

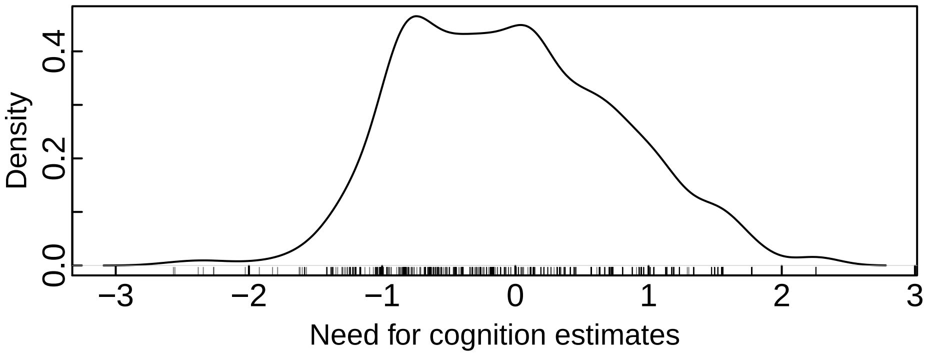
The first column shows the number of branching schemes (out of 216) in which each item appeared. The second column shows the number of respondents (out of 1,200) who answered each item.

YouGov in advance of the survey. The first column of Table 4 shows how many branching schemes included each item from the broader NFC scale. Nine different items are included in the four-item API branching scheme.

As part of the 2016 ANES Pilot Study, the NFC API was administered to 1,200 respondents drawn from an opt-in online panel administered by YouGov. The second column of Table 4 shows how many respondents answered each question, which corresponds roughly with the number of branching schemes in which they appeared. In all, 101 complete branching schemes were traversed by respondents. Since there was no non-response in the sample and since 91 of the original branching schemes were programmed to handle item-nonresponse, this means that 80.5% of the relevant possible branching schemes were used.

Broadly speaking, the NFC API behaved much as expected. The unweighted mean estimated NFC value was -0.031 and the standard deviation was 0.801, while the survey weighted mean and standard deviation was -0.074 and 0.782 respectively. These numbers are smilar (though not identical) to the distribution in the unweighted TAPS sample used to specify the prior (M=0.089, SD=0.716). Indeed, the distribution of scores, shown in Figure 9, corresponds largely with the distribution revealed in the TAPS sample in Figure 7, although it is clear that the ANES sample

Figure 9: Distribution of need for cognition in 2016 ANES Pilot Study



The plot shows the distribution of estimates of the need for cognition trait, calculated using Equation (2) above. The distribution is similar to the TAPS sample used to calibrate the API.

scored somewhat lower than the TAPS sample.

While it is not possible to compare these estimates to those generated using the complete NFC battery as in the applications above, it is possible to evaluate their predictive validity. In particular, I test whether NFC (as measured by the API) is a moderator for the effect of issue framing as has been argued in the existing literature (Druckman 2004).

I take advantage of two framing experiments on the ANES Pilot Study. In the first, 587 respondents were randomly assigned answer the question "Do you favor, oppose, or neither favor nor oppose allowing *Syrian refugees* to come to the United States?" (Emphasis added.) Response options were: (1) Favor a great deal, (2) Favor moderately, (3) Favor a little, (4) Neither favor nor oppose, (5) Oppose a little, (6) Oppose moderately, and (7) Oppose a great deal. The remaining 613 respondents were randomly assigned to answer the question, "Do you favor, oppose, or neither favor nor oppose allowing *refugees fleeing the Syrian civil war* to come to the United States?" (Emphasis added.) I test the theory that the civil war frame will make respondents more favorable towards allowing Syran 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 variables is the seven point response to this

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

	Model 1	Model 2
Intercept	4.563	4.551
	(0.319)	(0.319)
Civil war framing	−0.094	−0.088
	(0.088)	(0.088)
Need for Cognition	−0.082	−0.200
	(0.061)	(0.85)
Civil war × NFC		0.227
		(0.113)
N	1,063	1,063
R^2	0.530	0.532

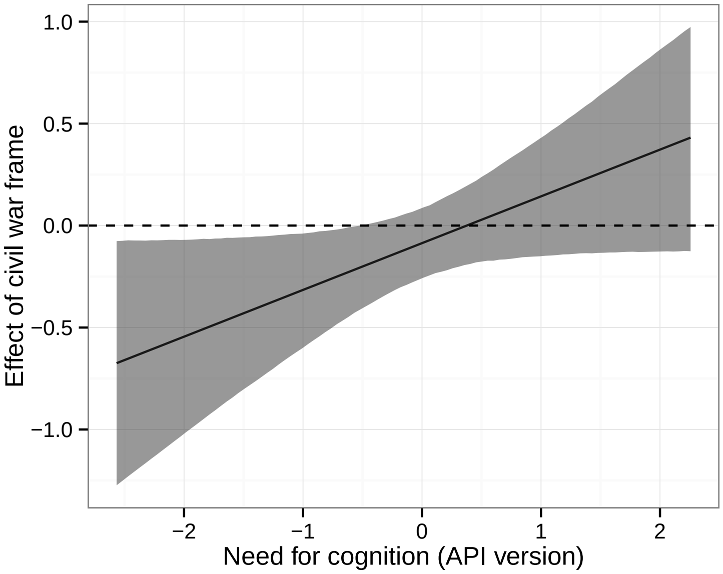
Estimates from weighted least squares regression using ANES weights. Additional controls for feeling thermometer towards muslims, support for intervening in Syria to combat ISIS, racial resentment, party identification, ideology, gender, education, race, and ethnicity are suppressed for clarity.

question.[21] 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 API ($p = 0.043$). To unpack this further, Figure 10 shows the estimated marginal effect of the civil war framing on opposition to Syrian refugees for differing levels of NFC. 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. These findings, that NFC is an effective moderator for attempts at framing issues, is consistent with expectations.

The second experiment relates to the concept of political correctness. A total of 574 respondents were asked, "*There's been a lot of talk lately about political correctness.* Some people think that the way people talk needs to change with the times to be more sensitive to people from different backgrounds. Others think that this has already gone too far and many people are just too easily offended. Which is closer to your opinion?" (Emphasis added) Response options were: (1) The way people talk needs to change a lot, (2) The way people talk needs to change a little, (3)

[21] I 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.

Figure 10: 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.

People are a little too easily offended, and (4) People are much too easily offended. Likewise, 626 respondents were asked a version of this same question without the framing in the first sentence. My expectation is that the political correctness frame will have a positive effect on responses to this item among white respondents,[22] but that this effect will be significantly moderated by respondents' level of NFC

The results from this analysis are shown in Table 6.[23] The first column (Model 1) shows that the political correctness frame had a significant and positive effect on responses as we would expect. However, Model 2 shows that there is again a significant interaction between the frame and NFC ($p = 0.007$). To understand this, I created an interaction plot for Model 2, which is shown in Figure 11. The plot shows that the effect of the framing experiment is again modified by respondents' level of NFC. For individuals who scored highly on the scale, the experiment had no

[22] Political correctness as a concept is highly racially loaded, and I therefore focus exclusively here on white respondents.

[23] I also controlled for ideology, party identification, education, gender, racial resentment, white racial identity, and a sense of white guilt.

Table 6: Effect of political correctness framing on belief that others are too easily offended among white respondents

	Model 1	Model 2
Intercept	1.07	1.09
	0.219	0.218
Political correctness framing	0.119	0.124
	(0.057)	(0.057)
Need for Cognition	0.065	0.158
	(0.041)	(0.054)
PC × NFC		-0.202
		(0.075)
N	791	791
R^2	0.399	0.404

Estimates from weighted least squares regression using ANES weights. Additional controls for ideology, party identification, education, gender, racial resentment, white racial identity, and white guilt are suppressed for clarity.

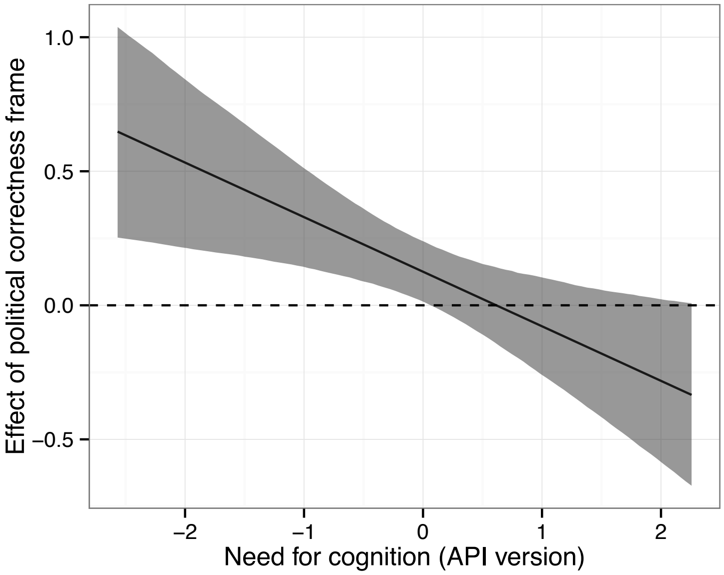
statistically reliable effect. However, for individuals low on this scale, the frame has a significant effect on responses consistent with past studies.

In all, the API measure appears to have provided a valid measure of NFC for the ANES sample. Of course, given the non-experimental nature of the data, it is difficult to asses whether the API version provided superior performance to a four-item fixed scale. However, in combination with the simulation and experimental results reported above, I believe that these results confirm that APIs are clearly a viable and valuable method for assessing latent personality traits on public opinion surveys.

5 CONCLUSION

Although there is room for continued improvement and extension, the results above show that APIs are capable of obviating the need for public opinion researchers to choose between administering large multi-item scales or selecting a single reduced scales to administer to all respondents that may reduce measurement precision. Above, I supported these claims using results from an empirically informed simulation, an experimental study, and a detailed case study of an API included on the 2016 ANES Pilot Study. In all, these results show that APIs allow for the administration of fewer questions while achieving superior levels of statistical precision and accuracy relative to

Figure 11: Interaction plot estimating the effect of the political correctness frame on the belief that others are too easily offended 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 6.

any fixed-reduced scale. I believe that APIs may greatly expand the potential for bringing modern personality research from psychology into the study of political behavior and public opinion.

Indeed, it is possible that the adaptive testing technology described above could, in principle, be applied for many tasks beyond measuring personality traits. When would researchers be interested in including an adaptive battery in a survey? CAT methods generally are most appropriate when the following three criteria are met. First, CAT assumes that scholars are interested in measuring a latent trait rather than analyzing survey responses to specific questions *per se*. Second, standard CAT methods are most appropriate when this underlying concept is itself unidimensional. Third, the survey should have space for at least three question items drawn from a larger battery, although in principle it could be applied using only two.

While many survey research tasks do not meet these criteria, they are nonetheless often satisfied – especially in academic research. For example, Montgomery and Cutler (2013) applied CAT in the measurement of political knowledge. Other potential applications include placing survey respondents into an ideological space using roll-call questions (Bafumi and Herron 2010), estimating respondents' likelihood of voting (Erikson, Panagopoulos and Wlezien 2004), and measuring citizens' values (Schwartz 1992).

Nonetheless, there are several potential limitations to this approach as well as areas for continued research. In particular, survey researchers face several data quality challenges including (a) strict limits on battery length, (b) relatively high levels of measurement error, and (c) concerns about the stability of item calibrations. Further research as to how these issues can be addressed within the CAT framework are clearly warranted.

First, as noted at the outset, survey time is perhaps the greatest contraint for improving the measurement of latent traits in a survey setting. Yet the relative advantages of CAT to static batteries is greatest for *longer* batteries, which may lead some to question the usefulness of adopting the method. One answer to this concern is to show, as I have above, that CAT provides superior measurement of latent constructs even if space only allows for three or four items. However, an additional approach is to include informative priors based on earlier survey response as part of the CAT algorithm (van der Linden 1999).[24] This will allow the algorithm to begin tailoring question items for respondents at the outset, further improving performance. So, for instance, it seems reasonable to expect that individuals who score high on a modern racism inventory to also be higher on the RWA scale and begin with more evaluatively extreme items.

A second concern is that random error may interfere with performance of CAT, since noisy responses may lead to the "wrong" question being selected – especially in the early stages of the battery.[25] One particularly promising approach to addressing this issue is using a stratified multistage 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 the

[24] This has already been implemented in the current software. Specifically, we set $\pi(\theta_j|\mathbf{x}_j) \sim N(\mu_0, \frac{1}{\tau_0})$, where μ_0 and τ_0 are estimated using some calibration sample and $\mathbf{x}_i$ are relevant responses observed earlier in the survey. We also allow for the inclusion of scaled non-central t-distribution $\pi(\theta_j|\mathbf{x}_j) \sim \frac{t(\mu_0,\nu_0)}{\tau_0}$, where μ_0 is the non-centrality parameter, τ_0 is some measure of precision, and ν_0 is the degrees of freedom. For smaller values of ν_0, this prior has "fatter" tails that improves the responsiveness of the CAT algorithm when μ_0 is inaccurate.

[25] This is a widely studied issue in the area of adaptive testing, although the focus is often not on accuracy but on reducing item exposure (e.g., Chang and Ying 1996, 1999; Chen, Ankenmann and Chang 2000; Hau and Chang 2001; Chang, Qian and Ying 2001; Chang and van der Linden 2003; Chang and Ying 2008; Cheng et al. 2008; Barrada et al. 2008; Passos, Berger and Tan 2008; Rulison and Loken 2009).

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

Finally, 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.[26] On possible solution is to implement Lord's (1980) χ^2 test and the likelihood ratio test (Thissen, Steinberg and Wainer 1993) for detecting differential item functioning, which are both relatively straightforward to implement and which have performed adequately in past studies. A related point is that research is needed on allowing APIs to use constrained selection mechanisms to allow for content balancing (van der Linden 2010). APIs could, for instance, seek to balance positively and negatively loading items or facets of the overarching trait in question to improve performance.

A final limitation of APIs is that they require pre-testing of battery items to calibrate the model. Although pre-testing of items is generally considered ideal for public opinion research, it is not always feasible. This suggests that there may be a trade-off for the cost of reducing the length of batteries and the cost of pre-testing batteries. The more accurate the pre-test, the greater the potential for reducing battery length while preserving measurement accuracy.

However, pre-testing costs may be ameliorated by making survey data and item calibrations widely available to other researchers. (The calibrations in this study will be included in the replication archive for this article.) Thus, additional research is called for to develop, calibrate, and field test specific APIs measuring constructs of interest to the wider discipline.[27] In addition to the batteries specified in Table 1, particularly valuable research would be developing adaptive variantes of the 44-item Big Five Inventory and the widely used Schwartz values index (Schwartz 1992).

[26] Uncovering changes in item-level parameters across administrations (or groups of individuals) is often termed differential item functioning, and this issue has received considerable attention in the adaptive testing literature (e.g., Kim, Cohen and Park 1995; French and Miller 1996; Donoghue and Isham 1998; DeMars 2004; Glas 2010; Wang, Tay and Drasgow 2013; Woods, Cai and Wang 2013).

[27] For a discussion of issues surrounding building similar infrastructure in the context of health research, see Cella et al. (2007). For an excellent example of how a CAT measuring a specific latent trait can be validated and distributed, see Walter et al. (2007), which presents an adaptive measure of anxiety for psychiatric patients.

As noted above, calibrating these models can be done using large convenience sample. Nonetheless, the measurement properties of the APIs will be improved if the models can be "normed" to national samples such that our prior beliefs are correctly calibrated towards the target population. Ideally, researchers interested in adopting APIs in their own research will work collaboratively to pair large convenience samples with nationally representative samples to calibrate and test APIs measuring important constructs for general dissemination to the academic community. My hope is that this will significantly facilitate more widespread adoption of this promising technology within political science and beyond.

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.

Anderson, A.B., A. Basilevsky and D. Hum. 1983. Missing Data: A Review of the Literature. In *Handboook of Survey Research*, ed. Peter H. Rossi, James D. Wright and Andy B. Anderson. New York: Academic Press pp. 415–481.

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.

Baron-Cohen, Simon. 2002. "The extreme male brain theory of autism." *Trends in Cognitive Sciences* 6(6):248–254.

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.

Barrada, Juan Ramón, Julio Olea, Vicente Ponsoda and Francisco José Abad. 2008. "Incorporating randomness in the Fisher information for improving item-exposure control in CATs." *British Journal of Mathematical and Statistical Psychology* 61(2):493–513.

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):329–350.

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.

Burchell, Brendan and Catherine Marsh. 1992. "The effect of questionnaire length on survey response." *Quality & Quantity* 26(3):233–244.

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.

Campbell, Angus, G. Gurin and Warren E. Miller. 1954. *The Voter Decides*. Evanston, IL: Row, Peterson.

Cella, David, Richard Gershon, Jin-Shei Lai and Seung Choi. 2007. "The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment." *Quality of Life Research* 16(1):133–141.

Chang, Hua-Hua, Jiahe Qian and Zhiliang Ying. 2001. "a-Stratified multistage computerized adaptive testing with b blocking." *Applied Psychological Measurement* 25(4):333–341.

Chang, Hua-Hua and Wim J van der Linden. 2003. "Optimal stratification of item pools in α-stratified computerized adaptive testing." *Applied Psychological Measurement* 27(4):262–274.

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.

Chang, Hua-Hua and Zhiliang Ying. 2008. "To weight or not to weight? Balancing influence of initial items in adaptive testing." *Psychometrika* 73(3):441–450.

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.

Cheng, Ying, Hua-Hua Chang, Jeffrey Douglas and Fanmin Guo. 2008. "Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints: Balancing measurement efficiency and exposure control." *Educational and Psychological Measurement* 69(1):35–49.

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(01):1–16.

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

Crawford, Scott D., Mick P. Couper and Mark J. Lamias. 2001. "Web Surveys : Perceptions of burden." *Social Science Computer Review* 19(2):146–162.

DeMars, Christine E. 2004. "Detection of item parameter drift over multiple test administrations." *Applied Measurement in Education* 17(3):265–300.

Donoghue, John R and Steven P Isham. 1998. "A comparison of procedures to detect item parameter drift." *Applied Psychological Measurement* 22(1):33–51.

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.

French, Ann W and Timothy R Miller. 1996. "Logistic regression and its use in detecting differential item functioning in polytomous items." *Journal of Educational Measurement* 33(3):315–332.

Galesic, Mirta and Michael Bosnjak. 2009. "Effects of questionnaire length on participation and indicators of response quality in web surveys." *Public Opinion Quarterly* 73(2):349–360.

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(03):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(01):111–133.

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.

Hau, Kit-Tai and Hua-Hua Chang. 2001. "Item selection in computerized adaptive testing: Should more discriminating items be used first?" *Journal of Educational Measurement* 38(3):249–266.

Heberlein, Thomas A. and Robert Baumgartner. 1978. "Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature." *American Sociological Review* 43(4):447–462.

Herzog, A. Regula and Jerald G. Bachman. 1981. "Effects of questionnaire length on response quality." *Public Opinion Quarterly* 45(4):549–559.

Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9(3):227–241.

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.

Krosnick, Jon A. 1999. "Survey Research." *Annual Review of Psychology* 50:537–67.

Krosnick, Jon A., Allyson L. Holbrook, Matthew .K. Berent, Richard A. Brody T. Carson, W.Michael Hanemann, Raymond J. Kopp, C. Mitchell, Robert Cameron, Stanley Presser, Paul A. Ruud, V.Kerry Smith, Wendy R. Moody, Melanie C. Green and Michael Conaway. 2002. "The impact of 'no opinion' response options on data quality: Non-attitude reduction or an invitation to satisfice?" *Public Opinion Quarterly* 66(3):371–403.

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.

Lord, Frederick and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Lord, Fredric M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: L. Erlbaum Associates.

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.

Matthews, Russell A., Lisa M. Kath and Janet L. Barnes-Farrell. 2010. "A Short, Valid, Predictive Measure of Work-Family Conflict: Item Selection and Scale Validation." *Journal of Occupational Health Psychology* 15(1):75–90.

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):141–171.

Muncer, Steven J and Jonathan Ling. 2006. "Psychometric analysis of the empathy quotient (EQ) scale." *Personality and Individual Differences* 40(6):1111–1119.

Passos, Valeria Lima, Martijn PF Berger and Frans ES Tan. 2008. "The D-optimality item selection criterion in the early stage of CAT: A study with the graded response model." *Journal of Educational and Behavioral Statistics* 33(1):88–110.

Pastor, Dena A, Barbara G Dodd and Hua-Hua Chang. 2002. "A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model." *Applied Psychological Measurement* 26(2):147–163.

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.

Penfield, Randall D. 2006. "Applying Bayesian item selection approaches to adaptive tests using polytomous items." *Applied Measurement in Education* 19(1):1–20.

Pratto, F., J. Sidanius, L.M. Stallworth and B.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. 2015. "R: A Language and Environment for Statistical Computing.".
URL: *http://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.

Richins, Marsha L. 2004. "The material values scale: Measurement properties and development of a short form." *Journal of Consumer Research* 31(1):209–219.

Rizopoulos, Dimitris. 2006. "ltm: An R package for latent variable modeling and item response theory analyses." *Journal of Statistical Software* 17(5):1–25.

Rulison, Kelly L and Eric Loken. 2009. "I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT?" *Applied Psychological Measurement* 33(2):83–101.

Russell, Steven S., Christiane Spitzmüller, Lilly F. Lin, Jeffrey M. Stanton, Patricia C. Smith and Gail H. Ironson. 2004. "Shorter can also be better: The abridged job in general scale." *Educational and Psychological Measurement* 64(5):878–893.

Samejima, Fumiko. 1969. "Estimation of latent ability using a response pattern of graded scores." *Psychometrika monograph supplement* 34(4):100.

Schwartz, S.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. 1–62.

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

Sheatsley, Paul B. 1983. Questionnaire construction and item writing. In *Handboook of Survey Research*, ed. Peter H. Rossi, James D. Wright and Andy B. Anderson. New York: Academic Press pp. 195–230.

Sidanius, Jim, Felicia Pratto, Colette Van Laar and Shana Levin. 2004. "Social dominance theory: Its agenda and method." *Political Psychology* 25(6):845–880.

Stanton, Jeffrey M., Evan F. Sinar, William K. Balzer and Patricia C. Smith. 2002. "Issues and strategies for reducing the length of self-report scales." *Personnel Psychology* 55(1):167–194.

Stenner, Karen. 2005. *The Authoritarian Dynamic*. New York, NY: Cambridge University Press.

Thissen, David, Lynne Steinberg and Howard Wainer. 1993. Detection of Differential Item Functioning Using the Parameters of Item Response Models. In *Differential Item Functioning*, ed. Paul W. Howard Holland and Howard Wainer. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, pp. 67–113.

Thompson, Edmund R. 2012. "A brief index of affective job satisfaction." *Group & Organization Management* 37(3):275–307.

Treier, Shawn and D. Sunshine Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73(4):679–703.

Treier, Shawn and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1):201–217.

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. 2010. Constrained Adaptive Testing with Shadow Tests. In *Elements of Adaptive Testing*, ed. Wim J. van der Linden and Cees A. W. Glas. New York: Springer pp. 31–56.

van der Linden, Wim J and Cees AW Glas. 2000. "Capitalization on item calibration error in adaptive testing." *Applied Measurement in Education* 13(1):35–53.

Van Rijn, P.W., T.J.H.M. Eggen, B.T. Hemker and P.F. Sanders. 2002. "Evaluation of selection procedures for computerized adaptive testing with polytomous items." *Applied Psychological Measurement* 26(4):393–411.

Veldkamp, Bernard P. 2003. Item Selection in Polytomous CAT. In *New Developments in Psychometrics*. New York: Springer pp. 207–214.

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.

Walter, Otto B, Janine Becker, Jakob B Bjorner, Herbert Fliege, Burghard F Klapp and Matthias Rose. 2007. "Development and evaluation of a computer adaptive test for 'Anxiety'(Anxiety-CAT)." *Quality of Life Research* 16(1):143–155.

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.

Woods, Carol M, Li Cai and Mian Wang. 2013. "The Langer-improved Wald test for DIF testing with multiple groups evaluation and comparison to two-group IRT." *Educational and Psychological Measurement* 73(3):532–547.

Yammarino, Francis J., Steven J. Skinner and Terry L. Childers. 1991. "Understanding mail survey response behavior: A meta-analysis." *Public Opinion Quarterly* 55(4):613–639.

SI-A APPLIED EXAMPLE: NEED FOR AFFECT

In this appendix, I demonstrate how APIs can be calibrated and calculated using my open source software package catSurv, which I intend to distribute on the CRAN network. I note here that this software is still in development, so some functions may change at a later date.

To illustrate this process, I focus on the need for affect scale. This personality trait measures individual's "motivation to approach or avoid emotion-inducing situations" (Maio and Esses 2001, p 583). The full question wording for the original 26-item scale is shown in Figure SI-1. In May of 2014, I administered this battery to 1,204 individuals recruited via Amazon's Mechanical Turk. Respondents were asked to read each statement and agree or disagree on a five-point scale

The first step to developing an API is to fit the graded response model to the data. After loading the package and specifying the working directory, this can be done using the following two commands (we assume that the data hase saved as a csv object called affectData.csv).

```
affectData<-read.csv("affectData.csv")
nfaAPI<-grmCat(affectData)
```

The grmCat function fits the GRM model using the ltm package in R. The parameters for this model are shown in Table SI-1. Using these estimates, the grmCat function then builds an object of the class Cat, which allows us to develop the API as described below.

Note that at this stage, we also can set up our API using a number of different options shown in Figure SI-2. The defaults for these options are the expected *a posteriori* (EAP) estimator and the minimum expected posterior variance (MEPV) item-selection criterion shown in the main text. However, there are many other options in the literature. For example, we might wish to use the maximum *a posteriori* (MAP) estimator of individual ability and the maximum expected information (MEI) item selection algorithm with an approximation of the probit response function. To do this, we use the following command.

```
nfaAPI<-grmCat(affectData, ability.estimator="MAP",
               item.selection="MEI", D=1.702)
```

Once the nfaAPI object has been set up, determining the next item that should be administered is simple.

```
nextItem(nfaAPI)
```

The nextItem function returns the name of the item that should be administered to the respondent. For this example, where the respondent has not answered any questions, the algorithm specifies item three (*Explore feelings*).

Next, the respondent answers the question and the Cat object must be altered to include this response. This is done using the storeAnswer function. For instance, if the respondent chose the second category (disagree somewhat) for item three, we could input this response as follows.

```
nfaAPI<-storeAnswer(nfaAPI, item=3, answer=2)
```

To determine what question to ask next, we then simply need to again use the nextItem function.

```
nextItem(nfaAPI)
```

Figure SI-1: Need for affect scale

Below are the kind of statements that people make when we interview them. Please read the sentences below carefully and indicate the degree to which you agree or disagree with each statement. (*Respons options*: Strongly disagree, somewhat disagree, neither agree nor disagree, somewhat agree, strongly agree)

1. I wish I could feel less emotion. (*Feel less*)
2. It is important for me to be in touch with my feelings. (*Touch feelings*)
3. I think that it is important to explore my feelings. (*Explore feelings*)
4. I am a very emotional person. (*Emotional person*)
5. It is important for me to know how others are feeling. (*Know others*)
6. Emotions help people get along in life. (*Emotions help*)
7. Strong emotions are generally beneficial. (*Emotions beneficial*)
8. I feel that I need to experience strong emotions regularly. (*Need emotions regularly*)
9. I approach situations in which I expect to experience strong emotions. (*Approach emotions*)
10. I feel like I need a good cry every now and then. (*Need cry*)
11. I like to dwell on my emotions. (*Dwell emotions*)
12. We should indulge our emotions. (*Indulge emotions*)
13. I like decorating my bedroom with a lot of pictures and posters of things emotionally significant to me. (*Decorating room*)
14. The experience of emotions promotes human survival. (*Promotes survival*)
15. I do not know how to handle my emotions, so I avoid them. (*Handle emotions*)
16. I find strong emotions overwhelming and therefore try to avoid them. (*Emotions overwhelming*)
17. Emotions are dangerous – they tend to get me into situations that I would rather avoid. (*Emotions dangerous*)
18. I would prefer not to experience either the lows or highs of emotion. (*Not experience*)
19. If I reflect on my past, I see that I tend to be afraid of feeling emotions. (*Reflect past*)
20. I would love to be like “Mr. Spock,” who is totally logical and experiences little emotion. (*Mr. Spock*)
21. I have trouble telling the people close to me that I love them. (*Trouble love*)
22. Displays of emotions are embarrassing. (*Emotions embarrassing*)
23. Acting on one’s emotions is always a mistake. (*Acting mistake*)
24. I am sometimes afraid of how I might act if I become too emotional. (*Sometimes afraid*)
25. Avoiding emotional events helps me sleep better at night. (*Avoiding helps sleep)*)
26. People can function most effectively when they are not experiencing strong emotions.

Table SI-1: Model parameters for need for affect API

	Cut1	Cut2	Cut3	Cut4	Discrimination
Feel less	-1.286	0.327	1.108	2.502	1.445
Touch feelings	2.649	1.562	0.816	-0.724	-1.844
Explore feelings	2.851	1.656	0.880	-0.667	-1.749
Emotional person	2.679	0.561	-0.446	-2.401	-0.852
Know others	3.937	2.285	1.214	-1.161	-1.054
Emotions help	3.313	1.878	0.965	-1.093	-1.550
Emotions beneficial	2.878	1.328	-0.077	-1.739	-1.464
Need emotions regularly	2.149	0.368	-0.880	-2.736	-1.060
Approach emotions	2.921	1.220	-0.426	-2.591	-1.111
Need cry	2.828	1.283	0.419	-1.917	-0.741
Dwell emotions	4.051	-0.861	-3.185	-8.162	-0.364
Indulge emotions	2.694	1.178	-0.156	-2.064	-1.305
Decorating bedroom	2.757	0.235	-0.966	-3.907	-0.538
Promotes survival	3.498	2.446	1.078	-0.909	-1.434
People tell	-3.105	-1.395	-0.175	1.832	1.098
Handle emotions	-0.651	0.770	1.405	2.799	1.580
Emotions overwhelming	-1.225	0.201	0.798	2.172	1.766
Emotions dangerous	-1.144	0.209	0.800	2.268	1.748
Not experience	-1.029	0.215	0.852	2.055	1.861
Reflect past	-1.308	0.234	0.892	2.645	1.271
Mr. Spock	-0.848	0.347	0.954	2.084	1.369
Trouble love	-0.589	0.615	1.096	2.711	1.076
Emotions embarrassing	-1.433	-0.065	0.561	2.151	1.498
Acting mistake	-1.534	0.611	1.693	3.555	1.120
Sometimes afraid	-2.188	-0.415	0.262	2.582	0.847
Avoiding helps sleep	-1.791	-0.349	0.548	2.095	1.344

BIC=82,762
N=1,204

Figure SI-2: Primary options in the `catSurv` package

Model	Complete?
3-parameter logistic	✓
Graded response model	✓
Generalized partial credit model	
Prior structures	
Normal	✓
T-distribution	✓
Uniform	
Ability estimation	
Maximum likelihood	✓
Weighted likelihood	
Maximum *a posteriori*	✓
Expected *a posteriori*	✓
Item selection criterion	
Maximum Fisher's information	✓
Kullback-Leibler	✓
Posterior KL	
a-stratification	
Maximum weighted Fisher's information	
Maximum posterior weighted information	✓
Maximum interval information	
Maximum expected information criterion	
Minimum expected posterior variance	✓
Random	✓
Stopping rule	
Length	✓
Precision	✓
Classification (interval)	

In many instances, researchers may not wish to conduct all of these operations in the midst of a survey. Therefore, we have developed additional functionality that will build a tree of all possible response profiles and the question that should be administered next. This can be viewed as a complex sequence of branching. Note that this branching can quickly become *very* complex. In general, the number of response profiles is equal to C^{n-1} where C is the number of possible response options and n is the number of questions that will be administered. So, for instance, if we wished to have an API with eight questions based on the need for affect scale, this would result in tree with 5^7 =78,125 possible branches.

We can build such a tree using our example where the maximum number of questions that will be administered is three using the following command.

```
nfaTree<-makeTree(nfaAPI, maxNum=3)
```

This `nfaTree` object is a "list of lists." The advantage of this approach, is that we can quickly lookup the next item that must be asked with very few calculations. For instance, to find the item that should be asked first to all respondents, we use this command.

```
nfaTree[["Next"]]
```

To find the item that should be asked of a respondent who chose the second category for the first item presented and the fourth category for the second item presented, we use the following command.

```
nfaTree[["2"]][["4"]][["Next"]]
```

In general, the computations needed to find the next item are O(log n), where n is the number of questions respondents answer. However, it is also possible to translate this tree into a flat lookup table (as described in the main text) using the `tree2table` function.

```
nfaTable<-tree2table(nfaTree)
```

The resulting table can be saved in whatever data format is convenient to the survey firm.

I have also developed prototype webservice based on the `catSurv` package, which is hosted on the Heroku cloud application platform. The webservice is designed to respond to queries from external servers and to execute the specified item selection routines. So, for instance, a survey run on the Qualtrics platform can call on the webservice to determine which question should be asked next given a specific response history. Testing thus far in Qualtrics shows that incorporating the CAT algorithm in this manner results in virtually no delays with batteries as large as 60 items.[1] Further, the Heroku platform is fully scalable, allowing me to set up multiple instances such that speed is maintained even when many surveys occur simultaneously.

With support from an NSF grant, I am working to make this webservice accessible to all academic researchers wishing to integrate CAT into their surveys.[2] Further, I will provide instructional materials to integrate the webservice with existing major survey platforms such as Qualtrics. Finally, this infrastructure will be open source such that technically advance researchers will be able host their own webservice if desired.

[1] This is facilitated by pre-calculating next steps while the survey taker reads and answers the current question.

[2] The CAT model – including options for ability estimation, priors, item selection routines, and the response history – is passed in as a JSON object. The server then calculates which questions should be administered next.