Computerized Adaptive Testing for Public Opinion Surveys[*]

Jacob M. Montgomery
Department of Political Science
Washington University in St. Louis
Campus Box 1063, One Brookings Drive
St. Louis, MO, USA, 63130-4899

Josh Cutler
Department of Political Science
Duke University
Perkins Hall 326 Box 90204
Durham, NC, USA, 27707-4330

April 23, 2012

[*] Funding for this study was provided by the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University in Saint Louis. We are grateful for helpful comments provided by Martin Elff. A previous version of this paper was presented at the 2012 Annual Meeting of the Midwest Political Science Association.

ABSTRACT

Survey researchers avoid using large multi-item scales to measure latent traits due to both the financial costs and the risk of driving up non-response rates. Typically, investigators select a subset of available scale items rather than asking the full battery. Reduced batteries, however, can sharply reduce measurement precision and introduce bias. In this paper, we present computerized adaptive testing (CAT) as a method for minimizing the number of questions each respondent must answer while preserving measurement accuracy and precision. CAT algorithms respond to individuals' previous answers to select subsequent questions that most efficiently reveal respondents' position on a latent dimension. We introduce the basic stages of a CAT algorithm and present the details for one prominent approach to item-selection. We then demonstrate the advantages of CAT via simulation and by empirically comparing adaptive and static scales measuring political knowledge.

1. INTRODUCTION

Survey researchers often avoid using large multi-item scales to measure latent traits on surveys. In part, this reflects the high financial costs associated with long surveys. However, it is also a result of the higher rate of attrition and non-response associated with lengthy and repetitive surveys. To avoid lengthy batteries, researchers typically select only a subset of available scale items to include on a survey rather than asking the full battery. Reduced batteries, however, can sharply reduce measurement precision and introduce bias in estimates, especially for individuals on the extreme ends of a latent scale.

In this paper, we apply computerized adaptive testing (CAT) to the field of public opinion surveys and introduce software that will aid researchers to more accurately measure important latent traits using a minimum number of question items. As its name suggests, CAT algorithms adapt dynamically to measure latent constructs while minimizing the number of questions each respondent must answer – similar to the functioning of the modern Graduate Record Examinations (GRE). The method is an extension to item response theory (IRT) (c.f., Clinton, Jackman and Rivers 2004), and like IRT derives from the educational testing literature. Each question item is classified based both on its average level of difficulty (its position on the latent scale) and its capacity to discriminate between respondents along a given latent dimension.[1] In essence, CAT algorithms respond to individuals' prior answers by choosing subsequent question items that will place them on a latent dimension with maximum precision and a minimum number of questions.

In the rest of this paper, we briefly review the basic elements of CAT algorithms and explain in detail one approach to item selection and stopping rules appropriate for public opinion research. We then evaluate the advantages of CAT relative to traditional static reduced batteries both theoretically and empirically. First, we conduct a simulation study to show how CAT surveys can increase precision and reduce bias relative to static reduced scales. Second, using a convenience sample of Amazon Mechanical Turk respondents, we conduct a survey experiment and compare the precision and accuracy of CAT surveys to static scales of political knowledge. In both cases, when the two competing methods of measurement are compared on a common metric, CAT methods provide improved measurement via improved precision and reduced bias.

2. DYNAMIC SURVEYS IN POLITICAL SCIENCE

CAT algorithms assume that no additional information is learned by asking individuals whose prior responses indicate that they score highly on some latent trait questions designed to discriminate between individuals on the low end of the scale. For such individuals, better questions should be chosen that reflect prior responses. In the language of educational testing:

> The basic notion of an adaptive test is to mimic automatically what a wise examiner would do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier. This stems from the observation that we learn little about an individual's ability if we persist in asking questions that are far too difficult or far too easy for that individual. We learn the most when we accurately direct our questions at the same level as the examinee's proficiency (Wainer 1990, p 10).

In many ways, CAT is similar to branching survey techniques that have been used since the early days of public opinion surveys. For instance, the standard American National Elections Study measures of party identification asks whether

[1] Although Political Science has primarily focused on dichotomous IRT models, CAT can easily be applied to data with ordinal response categories.

respondents think of themselves as “a Republican, a Democrat, an Independent, or what?” Interviewers then ask Democratic identifiers whether they would call themselves a “strong Democrat.” However, there is no purpose in asking self-identified Democrats if they would call themselves a “strong Republican.” Based on prior beliefs about Democratic respondents and the nature of the “strong Republican” question, researchers know that little information will be gained by administering the “strong Republican” item to Democratic respondents.

CAT takes the logic behind these branching question formats and extends it to large survey batteries containing dozens (if not hundreds) of potential items – far more items than can easily be placed in a branching hierarchy. In essence, CAT algorithms use prior information about respondents and question items to more quickly and accurately place survey takers on some latent scale. Prior information about respondents derives from their initial responses to items in the battery.[2] Prior information about the items derives from pre-testing the questionnaire, which establishes how specific questions relate to the latent scale of interest and provides needed item-level parameters for the CAT algorithm.

Beyond simple branching formats, CAT is also clearly related to computer-assisted interviewing techniques developed over two decades ago (Piazza, Sniderman and Tetlock 1989; Sniderman et al. 1991). Like branching questionnaires, this approach allows for survey items to change based on a respondent’s answers, experimental assignment, or any other criterion. However, computer-assisted surveys can change in an intricate fashion that would likely lead to errors or confusion in a traditional pencil-and-paper survey. That is, computer-assisted interviews can change in complex ways during the interview process in a manner pre-specified by researchers.

While related to computer-assisted surveys, CAT algorithms derive from an entirely different branch of research – educational testing. Computerized adaptive testing is itself an extension of item response theory (IRT) (Lord and Novick 1968; Lord 1980; Embretson and Reise 2000; Baker and Kim 2004), which has received considerable attention in recent years in Political Science (e.g., Clinton and Meirowitz 2001; Jackman 2001; Clinton and Meirowitz 2003; Clinton, Jackman and Rivers 2004; Bafumi et al. 2005; Treier and Jackman 2008; Gillion 2010).

IRT is a general method for measuring latent traits using observed indicators, which are binary or ordinal in most political science applications. CAT takes the

[2] As we discuss below, it is possible to develop priors based on demographics or responses to related questions in the survey. However, we do not focus on this approach in the current paper, and leave that to future research.

IRT framework and extends it to allow tests or surveys to be tailored to each individual respondent (Weiss 1982; Kingsbury and Weiss 1983; Weiss and Kingsbury 1984; Dodd, De Ayala and Koch 1995). Items are selected based on respondents' answers to previously administered questions with the goal of choosing items that will be most revealing. Numerous studies have shown that CAT tests outperform traditional static scales of similar lengths and are theoretically equivalent to administering a full battery (e.g., Hol, Vorst and Mellenbergh 2007; Weiss 1982; Weiss and Kingsbury 1984).

Since its initial inception in the late 1970s, CAT has been extended in a number of directions to allow for refinements such as content balancing (van der Linden 2010), informative priors (van der Linden 1999), response times (van der Linden 2008), multidimensionality (Segall 2010), nonparametric assumptions (Xu and Douglas 2006), and more (van der Linden and Pashley 2010). CAT is widely used in the fields of educational testing, psychology (e.g., Waller and Reise 1989; Forbey and Ben-Porath 2007), and (to a much lesser extent) marketing (Jagdip Singh 2007). However, CAT methods have rarely been applied in the measurement of public opinion. Moreover, we are (currently) aware of no instances of its use in published political science research.

CAT differs from more familiar methodologies, such as branching questionnaires, in two important ways. First, CAT is able to easily deal with much longer dynamic batteries than is feasible using traditional methods. For example, a dynamic battery that asked each respondent only 11 dichotomous items would require the pre-specification of over $2^{10} = 1,024$ possible branchings.

Second, CAT methods assume that the branching procedure is not something determined in advance by researchers. Rather, questions are chosen "online" as the survey is completed to maximize a pre-specified objective function. Item selection is therefore extremely formalized and directly embedded in the mathematics of the scaling procedure that translates survey responses into the latent trait space of interest. Indeed, the method assumes researchers are not interested in answers to the questions *per se*, but only in accurately estimating respondents' position on the latent scale. One distinct advantage of CAT, therefore, is that choices about the ordering of questions need not be justified by researchers as item selection is explicitly determined by available data in a theoretically motivated manner.

3. COMPUTERIZED ADAPTIVE TESTING

A hypothetical example can demonstrate the basic logic the CAT framework. Consider a survey battery that contains 40 items. As a running example, assume that these items are factual questions about U.S. and international politics designed to measure political knowledge. Let us assume there is space for five items and that all questions are dichotomous indicators (i.e., answers are either correct or incorrect). On a large national survey, scholars typically choose *one* subset of these items. However, there are actually $\binom{40}{5} = 658{,}008$ possible reduced batteries we could include. CAT algorithms make it possible to expand the range of question combinations we can ask. A five-item adaptive battery would allow us to include at least 16 question combinations. Moreover, using informative priors, response times, and other refinements, CAT would allow us to ideally include the entire collection of potential question combinations in the latent question-item space.

More importantly, CAT makes it possible to ask "better" question combinations chosen to reveal the most about each respondent. For both an adaptive and non-adaptive survey, we will be able to partition respondents into $2^5 = 32$ response categories. For fixed batteries, however, many of these potential response profiles are rarely (if ever) observed. Individuals who recognize the name of the Chief Justice of the Supreme Court are extremely unlikely not to recognize the Vice President. CAT, however, makes it far more likely that we will actually observe the full range of response profiles – an advantage that grows exponentially with the battery length. Finally, this added precision comes with no expense in terms of survey time as each respondent is still asked only five questions.

3.1. *Overview of a CAT algorithm*

Intuition: CAT is designed for application in the context of a large survey battery or psychometric scale whose validity has already been established. That is, the method assumes that there actually is some latent trait to be measured and that each of the candidate items are appropriate indicators of that trait. Potential applications in political science include large psychometric batteries (e.g. Gosling, Rentfrow and Swann 2003) , batteries on issue positions designed to place respondents into an ideological space (Bafumi and Herron 2010), or a listing of forms of political participation respondents may have engaged in over the past year (Gillion 2010).

CAT is a method for taking this large population of potential items and selecting among them to efficiently place respondents on some latent scale such as ideology, political knowledge, or activism. Roughly speaking, the algorithm chooses items that are most likely to produce the *precise* and *accurate* estimate of respondents' position on some latent factor.

To better understand this, we return to our political knowledge example above. First, *ceteris paribus*, CAT prefers items with larger discrimination parameters. That is, it prefers items that are most revealing about respondents once they have answered. Second, *ceteris paribus*, CAT will choose items whose difficulty parameters are close – but not "too" close – to the current estimate of the respondent's ability parameter. The algorithm prefers questions that it estimates the survey taker has a chance of answering either correctly or incorrectly. That is, it avoids selecting items with difficulty parameters extremely "far" from the current estimate of respondent's ability parameter. On the other hand, the algorithm does *not* prefer items it estimates a survey taker has a 50% chance of answering correctly since little information will be revealed regardless of the actual response.

Algorithm essentials: A general overview of a basic CAT algorithm is fairly straightforward, although there are a wide array of increasingly complex implementations in the literature (van der Linden and Pashley 2010). The essential elements of computerized adaptive tests are shown in Table 1 (Segall 2005).

First, initial estimates, $\hat{\theta}_j$, are generated for each respondents' position on the latent scale of interest. Before the initial item is asked, this estimate is based on our prior assumptions about the value of θ_j. One option is to assume a common prior for all respondents, $\theta_j \sim \pi(\theta)$. An alternative is to use previously collected data points, $\mathbf{y}_j$, to specify an informative prior, $\theta_j \sim \pi(\theta_j|\mathbf{y}_j)$ (van der Linden 1999). For example, when administering a battery measuring political ideology it may be appropriate to assume that strong Democrats are more liberal than strong Republicans. In either case, after the initial item in the CAT battery is administered, $\mathbf{y}_j$ will include responses to prior items.

Second, the next question item is selected out of the available battery which optimizes some objective function. Multiple criteria appear in the literature, including maximum Fisher information (MFI), maximum likelihood weighted information (MLWI), maximum expected information (MEI), minimum expected posterior variance (MEPV), and maximum expected posterior weighted information (MEPWI) (Choi and Swartz 2009, p 421).[3] It is also possible to choose constrained

[3] An overview of the most common item selection criteria for dichotomous indicators are discussed in van der Linden (1998), and van der Linden and Pashley (2010). An excellent analysis of potential selection criteria for polytomous items is available in Choi and Swartz (2009).

Table 1: Basic elements of computerized adaptive testing batteries

Stage	Purpose	Description
1	Estimate respondent's positions	A provisional trait estimate, $\hat{\theta}_j$, is created based on first i responses. If no items have been given, the estimate is based on prior information.
2	Item selection	The item that optimizes some objective function is chosen. In our examples below, CAT chooses items that minimize *expected posterior variance*.
3	Administer item	
4	Check stopping rule	Pre-defined stopping rules may include reducing posterior variance, $Var(\hat{\theta}_j)$, below a certain threshold or reaching some maximum time allotment for the battery.
5a	Repeat steps 1-4	If the stopping rule has not been reached, new items are administered.
5b	Return final trait estimate	If the stopping rule has been reached, a final estimate for $\hat{\theta}_j$ is calculated.

optimization approaches to, for instance, ensure that scales balance positively and negatively worded items.

In general, these criteria aim to choose items that will result in accurate and precise measures. Moreover, all of these item selection criteria lead to similar results after a modestly large number of items (i.e., $n \geq 30$). However, there are significant differences in measurement quality when the number of items that can be asked is more limited (van der Linden 1998). In this paper, we have chosen to focus on the MEPV criterion as being the most intuitive and mathematically motivated approach of those that perform well with a small number of questions.

The third stage of the algorithm is to administer the chosen item and record the response. Fourth, the algorithm checks some stopping rule. In most survey settings, the stopping rule is likely to be that the number of items asked of the respondent has reached some maximum value. In these fixed-length CAT algorithms, all respondents will be asked the same number of questions. However, it is also possible that researchers wish to measure some trait up to a specific level of precision regardless of the number of items that are asked. In these variable-length CAT algorithms, items may be administered until this threshold is reached.

Finally, if the stopping rule has not been reached, the algorithm will repeat. Once the stopping criteria has been met, however, the algorithm produces final estimates of $\hat{\theta}_j$ and terminates.

3.2. *Mathematical foundations*

Outline of the general model for dichotomous indicators: As discussed in Section 3.1, there are numerous variants of CAT for both dichotomous and polytomous indicators. Rather than attempting to summarize all of these approaches here, we will focus on the particular specification we use in our empirical example below. We will also restrict ourselves to the dichotomous case, which is both more intuitive and more familiar to a political science audience due to its wide use in roll-call analyses (Clinton, Jackman and Rivers 2004; Bafumi et al. 2005).

We use a three parameter logistic model, where y_{ij} is the observed outcome (correct/incorrect or yes/no) for item $i \in [1, n]$ for person $j \in [1, J]$. The model assumes that the probability of a correct response for individual j is

$$p_i(\theta_j) \equiv Pr(y_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{exp\big(Da_i(\theta - b_i)\big)}{1 + exp\big(Da_i(\theta - b_i)\big)} \tag{1}$$

where $D = 1$ for a logistic model and $D = 1.702$ for an approximation of the normal ogive model.

For CAT, we assume that the item-level parameters (a_i, b_i, c_i) have already been estimated using some previously collected data, which we term the calibration sample below. These parameters are typically termed the discrimination, difficulty, and guessing parameters respectively. For CAT, we assume that these are known quantities and our interest is only in estimating the ability parameter, θ_j, for some new respondent.

The prior distribution for θ_j will be

$$\pi(\theta_j) \sim N(\mu_\theta, \frac{1}{\tau_\theta}), \tag{2}$$

where τ_θ denotes the precision of the distribution (the inverse of the variance). In our examples below, we set $\mu_\theta = 0$ and $\tau = 0.75$, a fairly diffuse (though proper) prior.[4]

Letting $q_i(\theta_j) = 1 - p_i(\theta_j)$, the likelihood function associated with the responses to the first $k-1$ items under a local independence assumption is

$$L(\theta_j|\mathbf{y}_i) = \prod_{i=1}^{k-1} p_i(\theta_j)^{y_{ij}} q_i(\theta_j)^{(1-y_{ij})} \tag{3}$$

Calculating skill parameter: We present one of the most prominent methods for calculating respondent-level positions on a latent scale.[5] The expected a posteriori

[4] In numerous simulation experiments, we found this setting to be ideal for the purposes of small batteries. Stronger priors (e.g., $\tau = 1$) result in item selection being dominated by the prior, which is only overcome after many items are asked. Weaker priors (e.g., $\tau = 0.01$) result in extreme and uninformative items being selected when n is small. Note that we only set this prior during item selection. For the final estimation of $\hat{\theta}_j$, we return to the same $\tau = 1$ prior used for parameter estimation on the calibration data. Future versions of this paper will more fully explore the issue of prior selection to provide guidance to applied researchers.

[5] A second common technique is to estimate the maximum a posteriori (MAP), which is found by estimating the root of the first derivative of the log posterior (Equation 3).

$$\frac{\partial log L(\theta_j|\mathbf{y}_i)}{\partial\theta_j} = \sum_{i=1}^{k-1} \frac{p_i^*(\theta_j)(y_{ij} - p_i(\theta_j))}{p_i(\theta_j)q_i(\theta_j)},$$

where

$$p_i^*(\theta_j) \equiv \frac{\partial p_i(\theta_j)}{\partial\theta_j} = Da_i(d_i - c_i)\frac{exp\big(Da_i(\theta_j - b_i)\big)}{\Big(1 + exp\big(Da_i(\theta_j - b_i)\big)\Big)^2}.$$

(EAP) estimate of individual j's position on the latent scale is calculated as

$$\hat{\theta}_j^{(EAP)} \equiv E(\theta_j|\mathbf{y}_i) = \frac{\int \theta_j \pi(\theta) L(\theta|\mathbf{y}_i) d\theta_j}{\int \pi(\theta) L(\theta|\mathbf{y}_i) d\theta_j}. \tag{4}$$

Neither of these integrals can be analytically derived. However, using numerical methods we can approximate these quantities with precision.[6]

Item selection: We use the minimum expected posterior variance (MEPV) criterion to select items. This requires that we first estimate the posterior variance associated with a correct ($y^*_{ik} = 1$) and incorrect ($y^*_{ik} = 0$) response for all remaining items and multiply by the probability of observing these outcomes conditioned on the current estimate of θ_j.

We first estimate $P(y^*_{ik} = 1|\mathbf{y}_{k-1,j}) = 1 - P(y^*_{ik} = 0|\mathbf{y}_{k-1,j})$ where $\mathbf{y}_{k-1,j} = y_{1,j}, \ldots, y_{k-1,j}$. This is done by simply entering the current value of $\hat{\theta}_j$ into Equation (1). Assuming we are using the EAP estimator, we then calculate

$$\hat{\theta}^{(EAP)*} \equiv E(\theta_j|\mathbf{y}_{k-1,j}, y^*_{kj}), \tag{5}$$

which is the estimator conditioned on the potential response for the candiate item k, denoted denoted y^*_{kj}. The posterior variance for each possible response to each potential item is

$$Var(\theta_j|\mathbf{y}_{k-1,j}, y^*_{ik}) = E((\theta_j - \hat{\theta}_j^{(EAP)*})^2|\mathbf{y}_{k-1,j}, y^*_{kj}) \tag{6}$$

$$= \frac{\int (\theta_j - \hat{\theta}_j^{(EAP)*})^2 \pi(\theta) L(\theta_j|\mathbf{y}_{k-1}, y^*_{kj}) d\theta_j}{\int \pi(\theta) L(\theta_j|\mathbf{y}_{k-1}, y^*_{kj}) d\theta_j}, \tag{7}$$

which is estimated via numerical integration as above. Equation (7) represents the posterior variance we will observe if the algorithm administers item k to respondent j and the answer given is y^*_{kj}. According to the MEPV criterion, the item chosen will minimizes the function

$$\begin{aligned} &P(y^*_{ik} = 1|\mathbf{y}_{k-1,j}) Var(\theta_j|\mathbf{y}_{k-1,j}, y^*_{ik} = 1) + \\ &P(y^*_{ik} = 0|\mathbf{y}_{k-1,j}) Var(\theta_j|\mathbf{y}_{k-1,j}, y^*_{ik} = 0). \end{aligned} \tag{8}$$

[6] In most IRT models in the political science literature, these estimates are done using Markov chain Monte Carlo (MCMC) simulation. However, since these are both one-dimensional integrals, we deem such an approach unnecessary. In the current version of our software, we use an approximation through the `integrate.xy()` function in the `sfsmisc` package.

Stopping criteria: In our 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 precision, $1/V(\theta_j|\mathbf{y}_j)$, rises above some pre-specified level τ_{stop}.

4. SIMULATION AND ILLUSTRATION

To illustrate the potential advantages of CAT, we conducted a small simulation study. These simulations represent circumstances that are as close to ideal as can be expected in a survey setting. The item pool consists of 60 items and the response probabilities align exactly with the three parameter logistic model in Equation (1). The discrimination parameters are drawn from $a_i \sim Gamma(50, 25)$, the difficulty parameters (b_i) are spaced equidistantly on the interval [-3,3], and the guessing parameters are drawn from $c_i \sim U(0, 0.1)$. In essence, the simulation assumes that we have 60 items that load strongly on the underlying latent dimension ($\bar{a} = 2$) with items available across nearly the entire range of likely ability parameters.

We begin by comparing how fixed and dynamic batteries of identical length, estimate differently the position of a single individual. The focus here is to illustrate why, under some circumstances, CAT can provide less biased and more precise estimates of θ_j than a fixed battery.

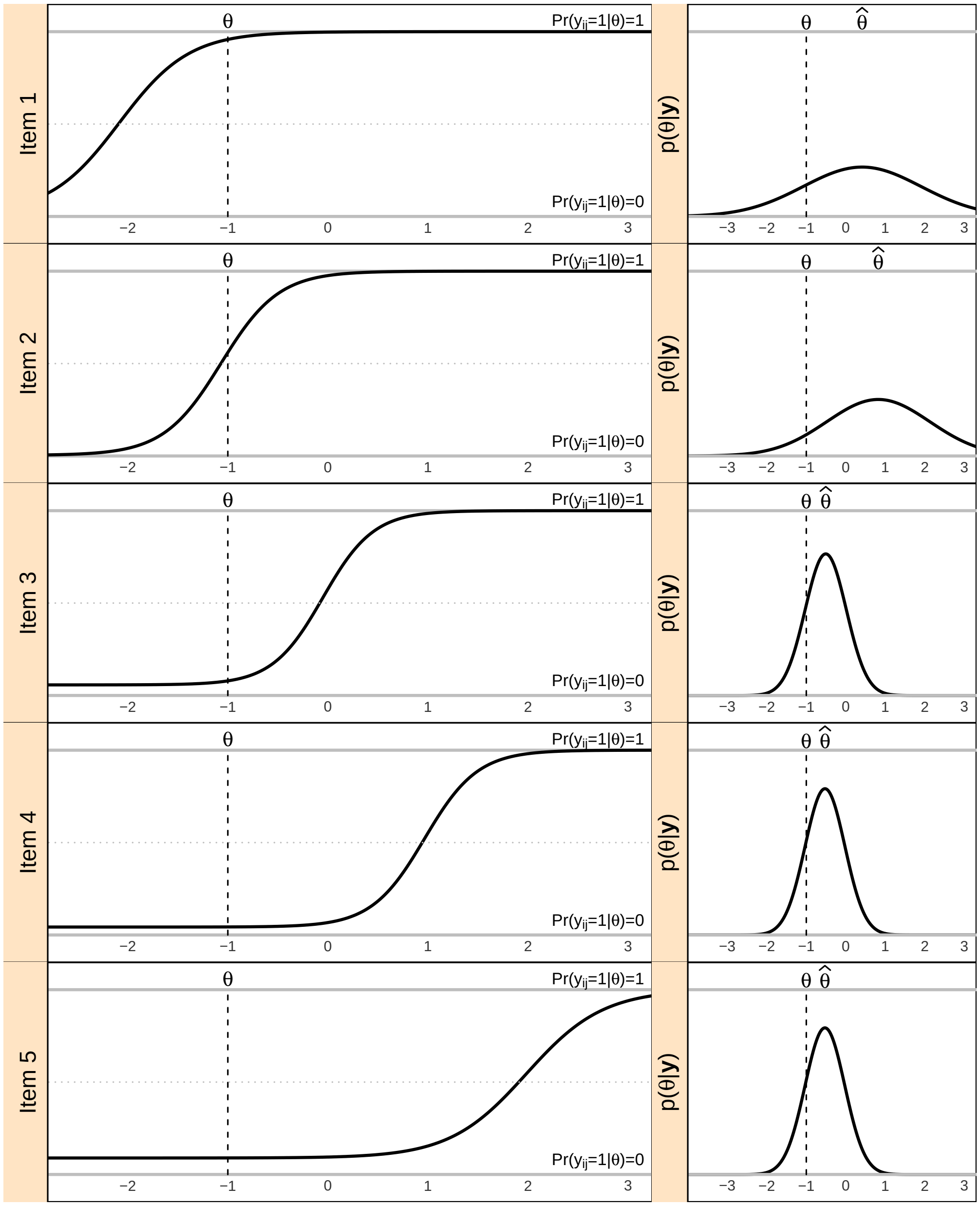
Reduced scales, both in our simulations and in the real world, are typically chosen to optimize measurement precision for respondents near the center of the latent distribution. For instance, in developing a reduced scale measuring aspects of personality, Gosling, Rentfrow and Swann (2003, pp 508) state that, “where possible we selected items that were not evaluatively extreme.” This is because it is items which reveal the most information about individuals at the *center* of the distribution that will minimize total MSE. Quite simply, most respondents are located in the middle of the distribution. However, this strategy often results in imprecise and even biased estimates of respondents towards the extreme ends of the latent scale.

The left panels of Figure 1 show the item characteristic curves for five items chosen to span the space of likely ability parameters.[7] The right panels in the figure show the posterior distribution of $\hat{\theta}$ given responses to all previously administered items.[8]

[7] Specifically, the battery includes items 10, 20, 30, 40, and 50. As items are spread equidistantly, this represents a fairly high quality fixed battery.

[8] For illustrative purposes, in this example we assume that responses are deterministic and correct

Figure 1: Item characteristic curves and posterior densities for for a simulated *static* reduced battery administered to a single individual



The left panels show the item characteristic curves for five items in a static battery. The right panels show the posterior distribution of $\hat{\theta}$ after each item is answered. Note that there is little improvement in the estimate after the administration of items 1,4 and 5.

There are two aspects of Figure 1 we wish to emphasize. First, note that the final estimate $\hat{\theta} = -0.45$ is somewhat inaccurate as the actual value of θ for the simulated individual is -1. Second, neither the precision nor the accuracy of the estimate are improved after the administration of Item 1, Item 4, or Item 5. That is, almost no additional information about this respondent is gained by including these items in the battery.

In contrast, Figure 2 shows the same information for items as chosen by the CAT algorithm. As can be seen, the final estimate of θ_j is more accurate ($\hat{\theta} = -0.8$) and far more precise. Moreover, the items chosen are generally much more informative about this particular respondent. Indeed, only the fourth item fails to provide significant information about the respondent's true position on the latent scale.[9]

While the results in Figures 1 and 2 are illustrative, they do not provide systematic evidence in favor of CAT. We therefore seek to generalize these results across a broader range of values of θ. Figure 3 shows the mean squared error (MSE)[10] for the dynamic (blue) and static (red) batteries of various lengths.[11] The upper left panel shows the results for the case when the number of items is three, and the remaining panels show results when the battery-length is five, seven and nine items respectively. The lines shows the estimated loess curve for the distribution of MSE results across values of θ_j. These estimates were generated for 1,000 simulated respondents distributed equidistantly on the interval [-3,3].

There are three aspects of the results shown in Figure 3 that are helpful for understanding the advantages of CAT. First, for most of the range of observed values of θ_j, the dynamic survey results in a lower MSE. That is, for survey batteries of a similar length, a dynamic survey provides more accurate and more precise estimates

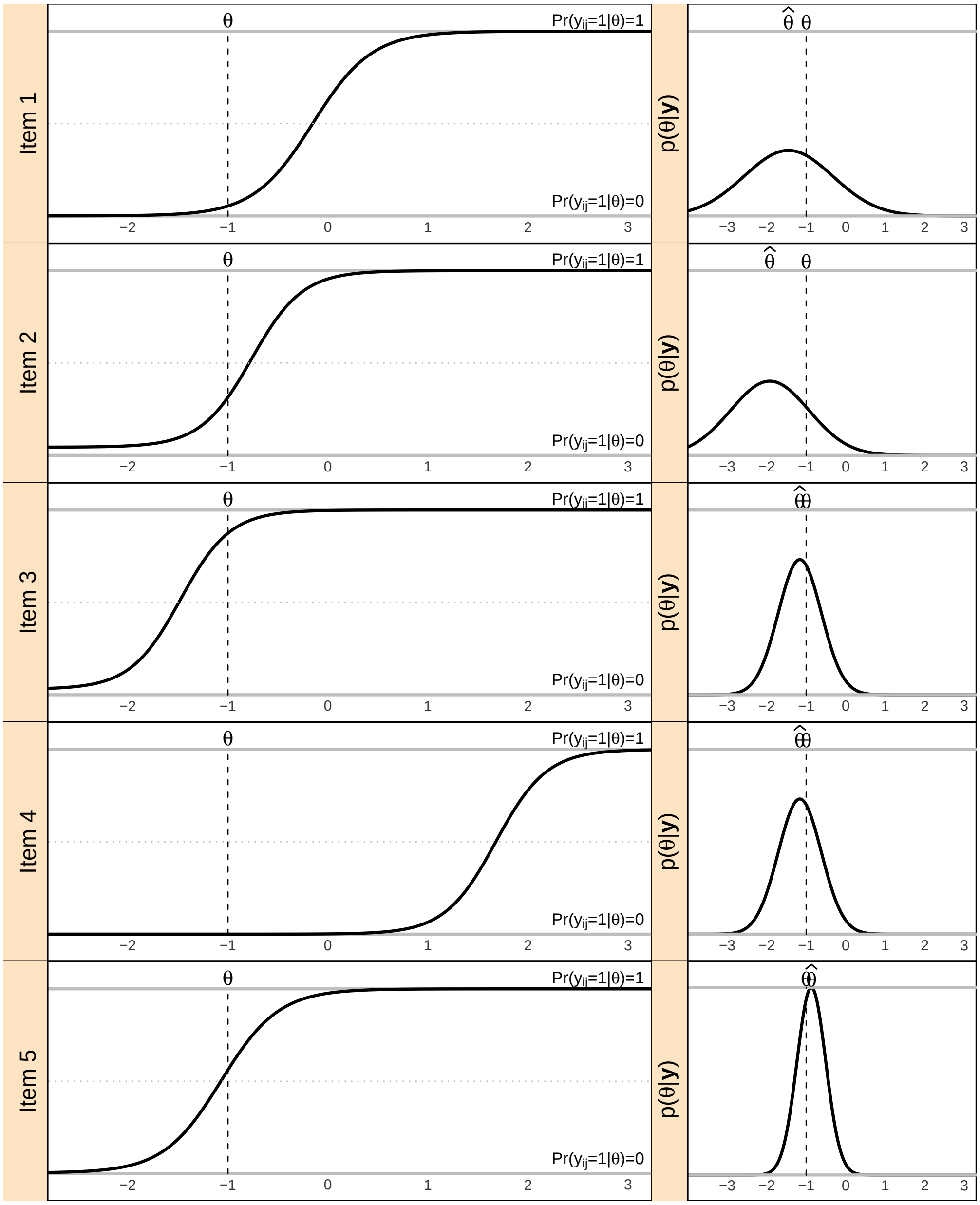
answers are always given if the predicted probability is greater than 0.5. In the larger simulations below, we assume they are probabilistic.

[9] In this example, the respondent answers the third item correctly leading the algorithm to ask a much more difficult question. Given this prior, the algorithm tends to "overshoot" a bit. Yet, the subsequent item selection indicates that the model continues to learn and the final estimates are quite accurate. We are still investigating alternative item selection criteria and priors that will reduce this type of over-compensation.

[10] MSE is defined as $Var(\hat{\theta}_j^{(EAP)}) + \left(\theta_j - \hat{\theta}_j^{(EAP)}\right)^2$.

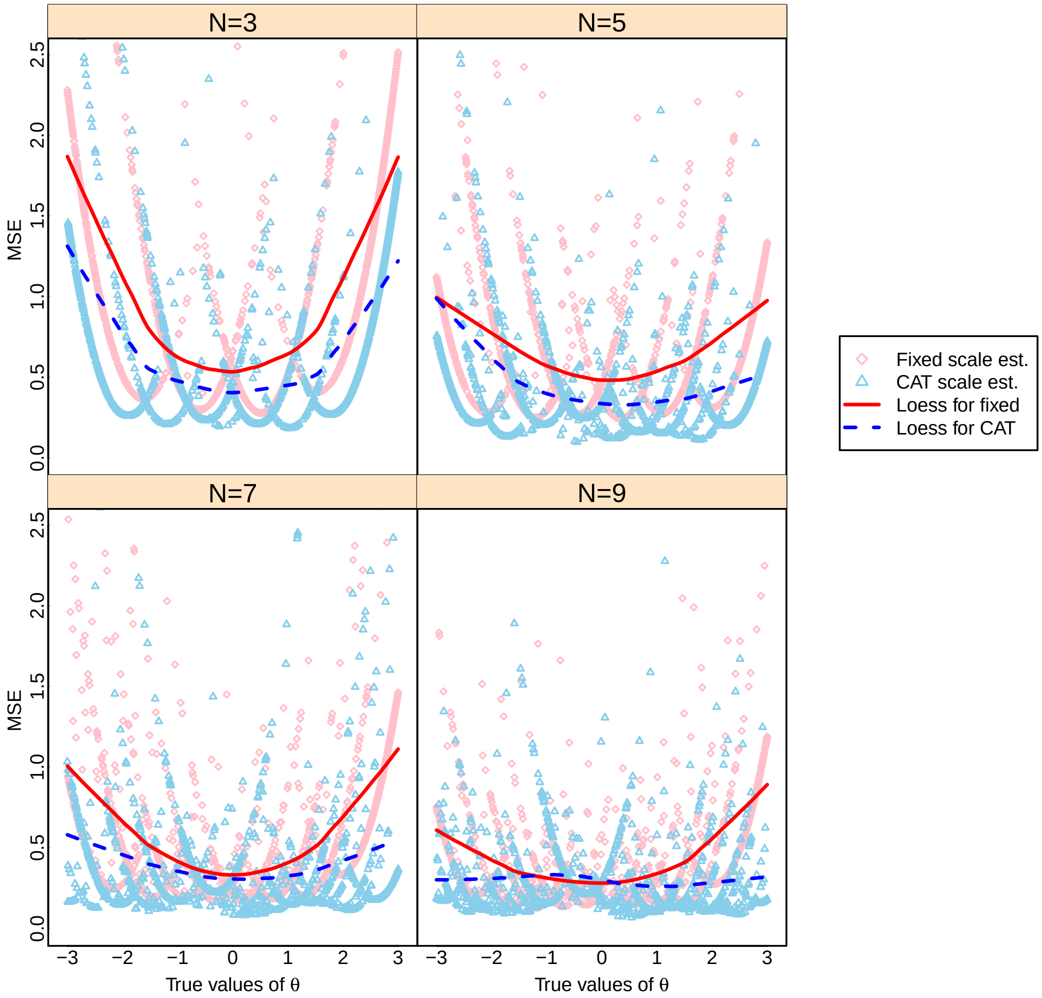
[11] The static batteries are: items 15, 30, 45 for $n = 3$; items 10,20,30,40,50 for $n = 5$; items 10, 17,23, 30, 37,43,50 for $n = 7$; and items 10,15,20,25,30,35,40,45,50 for $n = 9$. Alternative methods for selecting fixed batteries make little difference in the substantive conclusions of these simulations.

Figure 2: Item characteristic curves and posterior densities for for a simulated *dynamic* battery administered to a single individual



The left panels show the item characteristic curves for five items in a CAT battery. The right panels show the posterior distribution of $\hat{\theta}$ after each item is answered. Note that the estimate is both more precise and more accurate when measured by the items as selected by CAT.

Figure 3: Mean squared error for dynamic and static scales of four different lengths.



The points shows the MSE, defined as $Var(\hat{\theta}_j^{(EAP)}) + \left(\theta_j - \hat{\theta}_j^{(EAP)}\right)^2$, for each individual using both the CAT (red circles) and static (blue circles) batteries of length 3, 5, 7 and 9. The lines are loess estimates of the MSE for differing values of θ_j.

of a latent trait.[12] Second, the relative advantage of CAT diminishes somewhat towards the middle of the range of values for θ_j. For example, in the lower left panel of Figure 3 the static scale performs about equally with the dynamic scale for values of θ_j near zero. This actually an expected finding and replicates results from previous simulation studies (e.g., van der Linden 1998). It indicates the degree to which most fixed batteries are optimized to accurately measure individuals near the center of the distribution. Finally, the comparative advantages of CAT diminish noticeably as the number of items increases. Indeed, as the number of questions asked approaches the maximum of 60 (the total number of available items), the estimates will converge entirely.

5. EMPIRICAL APPLICATION: POLITICAL KNOWLEDGE

The simulation results in Section 5 show the advantages of CAT relative to a static scale theoretically. In this section, however, we provide an empirical application of CAT to the domain of political knowledge (sometimes termed political sophistication). This example demonstrate the superiority of CAT relative to static scales for accurately and precisely measuring important concepts to political science.[13]

Although scholars have developed a number of measures for knowledge and sophistication (e.g., Luskin 1987; Sniderman, Brody and Tetlock 1991; Delli Carpini and Keeter 1993), one of the most widely used methods for survey researchers is to ask respondents questions measuring their knowledge of basic facts about American politics, public officials, and current events. Since 1986 the American National Election Study (ANES) has asked respondents to identify the "job or political office" of officials such as the Vice President, the Speaker of the House, the Chief Justice of the United States Supreme Court, and the Prime Minister of the United Kingdom (DeBell 2012). While these items are open ended responses, others are similar to standard multiple choice questions used in educational testing. For instance, the 1992 ANES asked respondents:

> Who has the final responsibility to decide if a law is or is not constitutional . . . is it the President, the Congress, the Supreme Court, or don't you know?

[12] Indeed, examining the bias alone shows that CAT also provides slightly lower Bias, especially in the extreme range of θ_j. These results are available upon request.

[13] The data presented here is a pilot study. Future versions of this paper will include data from a larger number of respondents and a less terrible political knowledge battery.

> Do you happen to know which party had the most members in the House of Representatives in Washington before the election last month?

This method of measuring political knowledge has been used extensively in public opinion research (e.g., Barabas 2002; Brewer 2003; Delli Carpini and Keeter 1993, 1996; Gomez and Wilson 2001). Yet, this approach to measuring political knowledge has been extensively criticized (c.f., DeBell 2012; Gibson and Caldeira 2009; Lupia 2006, 2008; Luskin and Bullock 2011; Mondak 2001; Mondak and Davis 2001; Mondak and Anderson 2004; Prior and Lupia 2008). The coding of the open ended responses is of questionable reliability, and on occasion entirely incorrect.[14]

On its face, the heavy emphasis on identifying prominent individuals does not seem a valid indicator "political sophistication," or the ability to engage coherently in the political system (Lupia 2006). Finally, the items themselves do not seem appropriately chosen to acquire useful information about most respondents. In the 2008 ANES, 4% of respondents correctly identified the office held by John Roberts, 5% of respondents identified Gordon Brown, 37% identified Nancy Pelosi, and 73% correctly identified Dick Cheney (DeBell 2012).[15]

In this section, therefore, we apply a much larger collection of closed form (i.e., multiple choice) questions designed to measure a much broader swath of areas of political knowledge necessary for successfully engaging in the political system. In addition, the questions were designed to supply sufficient variation in difficulty to precisely estimate the latent position of respondents across the political knowledge spectrum.[16]

5.1. *Model calibration*

To demonstrate the usefulness of CAT methods, we developed a battery of 40 multiple choice knowledge questions, two of which were dropped due to poor performance. The remaining 38 items, which measure knowledge in areas including the

[14] DeBell (2012) notes that in 2004 identifying Tony Blair as the "Prime Minister of the United Kingdom" was coded as an incorrect response.

[15] These percentages change depending on how "correct" responses are coded. In addition, these numbers appear to fluctuate wildly from year to year. In 2004, 9.3% identified Dennis Hastert, 28% identified William Rehnquist, 62% identified Tony Blair, and 84% identified Dick Cheney (Gibson and Caldeira 2009)

[16] A second version of this study is now under review by the IRB, and will be in the field soon. This new version includes 71 total items and also allows respondents to more easily answer "don't know." We expect that this new battery will significantly improve the results below.

legislative process, interest groups, foreign affairs, and constitutional rights, are listed in Appendix A.

We administered the battery to 604 respondents based in the United States and over the age of 18. Respondents were recruited through Amazon Mechanical Turk. This sample primarily serves to calibrate the model and allow us to estimate appropriate difficulty, discrimination, and guessing parameters. Ideally, this calibration would be done on a nationally representative sample, which would give more meaningful estimates that could be used by researchers in future studies.[17] However, this convenience sample serves the more limited purpose of illustrating the usefulness of the CAT method.

Table 2 presents the item-level parameters associated with each of the questions in our battery. It shows that the easiest question, the item with the lowest difficulty parameter, identified by this sample is Item 10, "How long is one term for the President of the United States?: (a) Eight years, (b) Six years, (c) Four years, (d) Two years." The hardest question, the item with the largest difficulty parameter, is Item 15 "Most cases are considered by the Supreme Court: (a) at the request of congress, (b) upon order of the President, (c) with the approval of at least four Justices, (d) only with the approval of a majority of Justices, (e) in even-numbered years."

5.2. *Empirical comparison with reduced scales*

To assess the efficacy of CAT techniques outside of the calibration sample, we conducted a small survey experiment on a fresh sample of 204 respondents.[18] In this second survey, roughly half of the respondents (n=101) first answered the ten-item fixed battery chosen according to the procedure in Appendix B. The remainder (n=103) answered ten items as selected by the CAT algorithm discussed above. Respondents in both groups then answered the remaining 28 items in a random order.

Requiring all respondents to complete the entire battery allows us to evaluate the two measurement techniques using a common metric – respondents' score as assessed by the complete 38-item battery. Thus, we approximated respondent j's true latent trait value, θ_j, using her answers to the 38 questions. We then computed

[17] Using a national sample would allow us to say that difficulty parameters would, for instance, indicate the degree to which an average American can correctly answer a specific question.

[18] Respondents were again based in the United States and over the age of 18 and were recruited using Amazon Mechanical Turk.

Table 2: Item-level parameters estimated from calibration sample

Item	Guessing (c_i)	Difficulty (b_i)	Discrimination (a_i)	$P(x = 1 \| \theta_j = 0)$
1	0.50	0.19	3.11	0.68
2	0.00	-1.05	0.73	0.68
†3	0.25	0.49	3.74	0.35
4	0.23	-0.35	0.84	0.67
5	0.11	0.55	1.31	0.41
†6	0.06	-0.77	2.32	0.86
7	0.00	-2.26	2.61	1.00
†8	0.00	-2.08	2.72	1.00
†9	0.23	-0.91	2.19	0.91
10	0.00	-2.64	1.66	0.99
11	0.50	0.68	2.32	0.59
12	0.31	0.49	2.85	0.45
13	0.09	2.20	2.06	0.10
14	0.34	0.67	1.92	0.49
15	0.00	2.29	0.27	0.35
†16	0.11	0.56	2.56	0.28
17	0.14	-1.35	0.79	0.78
18	0.01	-1.46	0.66	0.72
†19	0.37	1.03	2.94	0.40
20	0.00	-1.63	0.49	0.69
†21	0.18	1.88	4.47	0.18
22	0.00	-2.56	0.69	0.85
23	0.17	0.68	2.43	0.30
24	0.24	0.39	2.39	0.45
25	0.50	-1.14	2.19	0.96
26	0.00	-1.31	1.15	0.82
27	0.33	-0.61	2.24	0.86
28	0.00	-2.43	1.06	0.93
†29	0.19	-1.54	2.27	0.98
30	0.23	0.25	1.11	0.56
31	0.22	-1.06	2.03	0.92
32	0.23	-0.41	1.78	0.75
33	0.50	-1.36	1.83	0.96
†34	0.34	-0.02	2.95	0.68
35	0.10	-1.90	2.04	0.98
†36	0.23	0.20	1.63	0.56
37	0.15	0.23	1.27	0.51
38	0.00	-2.58	1.21	0.96

$n = 604$. †Item included in 10-item fixed scale. The model was estimated using the `tpm()` command in the `ltm` package in `Rv2.12`. We constrained guessing parameters to take a maximum value of 0.5.

estimated values of $\hat{\theta}_j^{(EAP)}$ based on the first $n \in (3, 5, 7, 10)$ questions administered in either treatment condition. Our purpose is to evaluate how well $\hat{\theta}_j^{(EAP)}$ approximates θ_j across treatment conditions.

Figure 4 shows the MSE of the estimated values of $\hat{\theta}_j^{(EAP)}$ for individuals in the dynamic (red circles) and static (blue triangles) treatment conditions. The solid lines again represent loess curves (no confidence intervals are shown due to the small sample size). As in the simulations in Section 4, the dynamic algorithm noticeably outperforms the fixed battery for lower values of n. As n increases, the mean squared error for the two treatments start to converge.[19]

While the estimated loess curves of our sample seem to indicate that CAT techniques outperformed the static battery, especially when $n = 3$ or $n = 5$, it is difficult to visually tell the extent and significance of the statistical effect. To properly compare the treatment effect of the CAT algorithm on MSE, we used a two-sided Wilcoxon test to determine whether the distributions were indeed different. The results of the test confirm what we observed visually. That is, MSE is significantly lower when $n = 3$ $(W = 4065, p = 0.01)$ and $n = 5$ $(W = 4120, p = 0.02)$. There are no statistically significant differences when $n = 7$ or $n = 10$, although we believe that this may be the result of our small sample size.

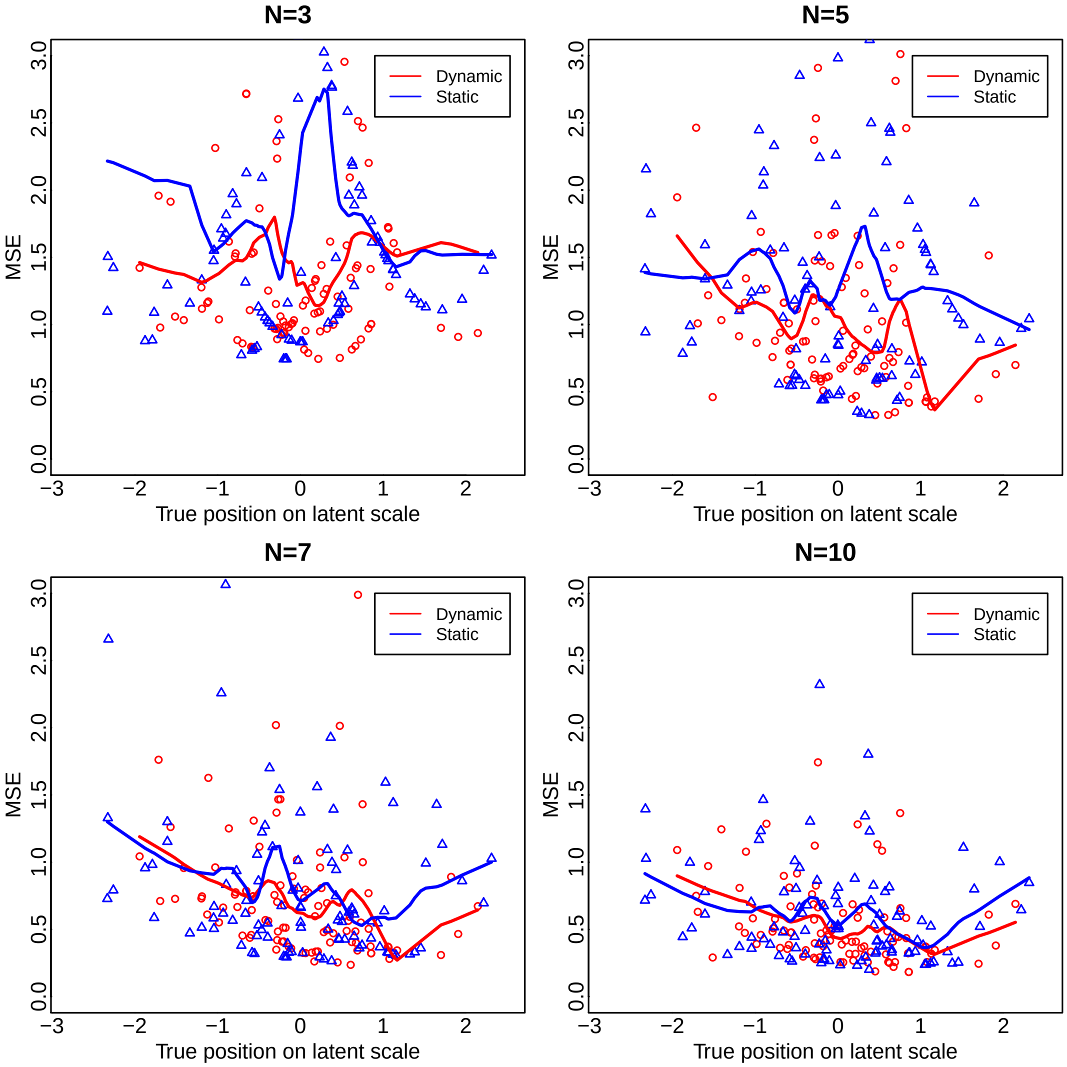
6. CONCLUSION

In this paper, we have shown that CAT techniques are capable of obviating the need for public opinion researchers to administer large multi-item scales to measure latent traits or to choose single reduced scales to administer to all respondents. Adaptive testing allows for the administration of many fewer questions while achieving superior levels of statistical precision and accuracy. We believe that CAT may provide substantial cost savings and efficiency gains for survey researchers while reducing attrition and non-response.

After presenting the details of one CAT algorithm, we demonstrated the method using both simulation and an empirical example. Using a battery of political knowledge items, we administered a set of 38 questions to over 604 respondents and calibrated the CAT algorithm on their responses. When compared to an optimally selected fixed battery, CAT can provided both improved measurement precision and accuracy for a fresh sample of 204 respondents. This was particularly true for

[19] When $n = 10$ we have administered over 25% of the questions used to determine the true value of θ_j. Thus, this relatively quick convergence is unsurprising.

Figure 4: Out-of-sample MSE for dynamic and static scales of four different lengths



The points shows the MSE, defined as $Var(\hat{\theta}_j^{(EAP)}) + \left(\theta_j - \hat{\theta}_j^{(EAP)}\right)^2$, for each individual using both the CAT (blue triangles) and static (red circles) batteries of length 3, 5, 7 and 10. The lines are loess estimates. The CAT algorithm outperforms or matches the Fixed battery for all n values but does particularly well for lower values.

smaller numbers of questions, which may be the more likely circumstance on large national surveys. Finally, we have developed our own software to administer such dynamic surveys. This software will be made available to researchers who wish to adopt CAT techniques.

While we believe the evidence presented above suggests that CAT may offer a superior approach to traditional static batteries, the methodology comes with several important caveats and limitations. First, CAT is only appropriate when researchers are interested primarily in placing respondents onto some latent scale rather than examining responses to specific questions. Second, CAT requires extensive 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. Third, CAT should not be used for batteries where there is evidence of strong question order effects. Finally, for time-varying attitudes or traits, the calibration may not always remain current or appropriate. This is not an issue for measurement of traits like personality, but could be problematic for less stable attitudes like presidential approval. Further research is needed to develop methods that can detect when specific item parameters have become obsolete (e.g., Segall 2002).

We will conclude by noting several promising paths forward for this research. While numerous variations in CAT algorithms are available, the examples in this paper implemented only uninformative priors, MEPV item selection, EAP ability estimation, and fixed-length batteries. Future research could explore which algorithms are most appropriate for various types of researcher constraints, whether they be time, cost, or measurement precision. Additional guidance as to the relative advantages and disadvantages of various CAT approaches may facilitate wider adoption of the methodology

Furthermore, this paper restricted itself to dichotomous data. While this is useful for many political science applications, there are also numerous latent traits that are more appropriately measured using polytomous models, e.g. "degree of conservatism" or "degree of authoritarianism". Though the intuition behind such models are similar to that described above, they have not yet been applied in the field of Political Science and implementation issues remain. Moreover, it may be that CAT offers limited advantages for Likert-type survey items relative to static batteries. Future studies should investigate the benefits of CAT surveys for multi-category survey questions.

References

Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.

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

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

Barabas, Jason. 2002. "Another Look at the Measurement of Political Knowledge." *Political Analysis* 10(2):209–222.

Brewer, Paul R. 2003. "Values, Political Knowledge, and Public Opinion About Gay Rights." *Public Opinion Quarterly* 67:173–201.

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

Clinton, Josh D. and Aadam Meirowitz. 2001. "Agenda Constrained Legislator Ideal Points and the Spatial Voting Model." *Political Analysis* 9(3):242–259.

Clinton, Josh D. and Aadam Meirowitz. 2003. "Integrating Voting Theory and Roll Call Analysis: a Framework." *Political Analysis* 11(4):381–396.

Clinton, Joshua, Simon Jackman and Doug Rivers. 2004. "The Statistical Analysis of Roll Call Voting: A Unified Approach." *American Political Science Review* 98(2):355–370.

DeBell, Matthew. 2012. "Harder Than it Looks: Coding Political Knowledge on the ANES." Paper presented at the 2012 meeeting of the Midwest Political Science Association in Chicago.

Delli Carpini, Michael X. and Scott Keeter. 1993. "Measuring Political Knowledge." *Measuring Political Knowledge: Putting First Things First* 37(4):1179–1206.

Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans Know About Politics and Why it Matters*. New Haven: YaleUniversity Press.

Dodd, Barbara G., R.J. De Ayala and William R. Koch. 1995. "Computerized Adaptive Testing with Polytomous Items." *Applied Psychological Measurement* 19(1):5–22.

Embretson, S.E. and S.P. Reise. 2000. *Item response theory for psychologists*. Lawrence Erlbaum.

Forbey, Jonathan D. and YOssef S. Ben-Porath. 2007. "Computerized Adaptive Personality Testing: A Review and Illustration with the MMPI-2 Computerized Adaptive Version." *Psychological Assessment* 19(1):14.

Gibson, James L. and Gregory A. Caldeira. 2009. "Knowing the Supreme Court? A Reconsideration of Public Ignorance of the High Court." *Journal of Politics* 71(2):429–441.

Gillion, Daniel Q. 2010. "Re-Defining Political Participation through Item Response Theory." Unpublished Paper.

Gomez, Brad T. and J. Matthew Wilson. 2001. "Political Sophistication and Economic Voting in the American Electorate: A Theory of Heterogeneous Attritbution." *American Journal of Political Science* 45(4):899–914.

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

Hol, A.Michiel, Harrie C.M. Vorst and Gideon J. Mellenbergh. 2007. "Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison with Short Forms." *Applied Psychological Measurement* 31(5):412–429.

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

Jagdip Singh, Roy D. Howell, Gary K. Rhoads. 2007. "Designs for Likert-Type Data: An Approach for Implementing Marketing Surveys." *Journal of Marketing Research* 19(1):12–24.

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

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.

Lupia, Arthur. 2006. "How Elitism Undermines the Study of Voter Competence." *Critical Review* 18(1-3):217–232.

Lupia, Arthur. 2008. "Procedural Transparency and the Credibility of Election Surveys." *Electoral Studies* 27(4):732–739.

Luskin, Robert C. 1987. "Measuring Political Sophistication." *American Journal of Political Science* 31:856–899.

Luskin, Robert C. and John G. Bullock. 2011. ""Don't Know" Means "Don't Know": DK Responses and the Public's Level of Political Knowledge." *Journal of Politics* 73(2):547–557.

Mondak, Jeffery J. 2001. "Developing Valid Knowledge Scales." *American Journal of Political Science* 45(1):224–238.

Mondak, Jeffery J. and Belinda Creel Davis. 2001. "Asked and Answered: Knowledge Levels When We Will Not Take "Don't Know" for an Answer." *Political Behavior* 23(3):199–224.

Mondak, Jeffery J. and Mary R. Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge." *Journal of Politics* 66(2):492–512.

Piazza, Thomas, Paul M. Sniderman and Philip E. Tetlock. 1989. "Analysis ofthe Dynamics of Political Reasoning: A General-Purpose Computer-Assisted Methodology." *Political Analysis* 1:99–119.

Prior, Markus and Arthur Lupia. 2008. "Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills." *American Journal of Political Science* 52(1):19–183.

Segall, Daniel O. 2002. “Confirmatory Item Factor Analysis using Markov Chain Monte Carlo Estimation with Applications to Online Calibration in CAT.” Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

Segall, Daniel O. 2010. Principles of Multidemensional Adaptive Testing. In *Elements of Adaptive Testing*, ed. Wim J. van der Linden and Cees A. W. Glas. New York: Springer pp. 57–76.

Sniderman, Paul M., Richard A. Brody and Philip E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. New York: Cambridge University Press.

Sniderman, Paul M., Thomas Piazza, Philip E. Tetlock and Ann Kendrick. 1991. “The New Racism.” *American Journal of Political Science* 35(2):423–47.

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

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. 2008. “Using Response Times for Item Selection in Adaptive Testing.” *Journal of Educational and Behavioral Statistics* 33(1):5–20.

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 Peter J. Pashley. 2010. *Elements of Adaptive Testing*. New York: Springer.

Wainer, Howard. 1990. Introduction and HIstory. In *Computerized Adaptive Testing: A Primer*, ed. Howard Wainer. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Waller, N.G. and S.P. Reise. 1989. "Computerized Adaptive Personality Assessment: An Illustration with the Absorption Scale." *Journal of Personality and Social Psychology* 57(6):1051.

Weiss, David J. 1982. "Improving Measurement Quality and Efficiency with Adaptive Testing." *Applied Psychological Measurement* 6(4):473–492.

Weiss, David J. and G. Gage Kingsbury. 1984. "Application of Computerized Adaptive Testing to Educational Problems." *Journal of Educational Measurement* 21(4):361–375.

Xu, Xueli and Jeff Douglas. 2006. "Computerized Adaptive Testing Under Nonparamteric IRT Models." *Psychometrika* 71(1):121–137.

A. QUESTION WORDING APPENDIX

1. Which party holds a majority of seats in the U.S. House of Representatives in Washington, D.C.?
   - Democrats
   - Republicans
   - Independents
2. How many votes are required in Congress to override a presidential veto?
   - A simple majority of one house of Congress
   - A simple majority of both houses of Congress
   - A two-third majority of one house of Congress
   - A two-thirds majority of both houses of Congress
3. How long is one term for a member of the U.S. Senate?
   - Eight years
   - Six years
   - Four years
   - Two years
4. The presiding officer in the House of Representatives is
   - the Majority Leader
   - the Sargeant at Arms
   - the Vice President of the United States
   - the Speaker
5. The President of the Senate is
   - the Majority Leader
   - the Sargeant at Arms
   - the Vice President of the United States
   - the senior senator of the majority party
6. The ability of a minority of senators to prevent a vote on a bill is known as
   - suspension of the rules
   - enrollment
   - a veto
   - a fillibuster
7. Who is the Vice President of the United States?
   - Harry Reid
   - Joseph Biden
   - John Boehner
   - Nancy Pelosi
8. A President may serve
   - any number of terms

- three terms
- two terms
- one term

9. The Secretary of State
    - serves a two-year term
    - serves the state governments
    - is nominated by the president
    - heads the armed services
10. How long is one term for the President of the United States?
    - Eight years
    - Six years
    - Four years
    - Two years
11. Who is the Chief Justice of the United States Supreme Court?
    - Hillary Clinton
    - Mitt Romney
    - Antoni Scalia
    - John Roberts
12. Who is the Speaker of the House of Representatives?
    - Harry Reid
    - Joe Biden
    - John Boehner
    - Nancy Pelosi
13. The Majority Leader of the House of Representative is
    - Nancy Pelosi
    - Kevin McCarthy
    - Eric Cantor
    - John Boehner
14. The head of the Department of Justice is
    - Kathleen Sebelius
    - Eric Holder
    - Timothy Geithner
    - Hillary Clinton
15. Most cases are considered by the Supreme Court
    - at the request of Congress
    - upon order of the President
    - with the approval of at least four Justices
    - only with the approval of a majority of the Justices
    - in even-numbered years

16. The President may NOT
    - declare war
    - pardon criminals without justification
    - appoint federal officials when Congress is in recess
    - refuse to sign legislation passed by Congress
17. Medicare is the health care program for
    - federal government employees
    - military veterans
    - low income families
    - senior citizens
18. Social Security is
    - funded by the personal income tax
    - operated by state government
    - the responsibility of the Department of Defense
    - the benefit program for senior citizens
19. On which of the following federal programs is the most money spent each year?
    - Education
    - Subsidies to farmers and agriculture
    - Medicare
    - Aid to foreign countries
20. On which of the following federal programs is the most money spent each year?
    - Welfare (Aid to Families with Dependent Children)
    - Interest on the debt
    - Environmental protection
    - Defense
21. The Byrd Rule is relevant
    - during the confirmation of cabinet members
    - for national party conventions
    - during appropriations hearings
    - for the reconciliation process
22. The federal debt is
    - much smaller than it was 20 years ago
    - the difference between imports and exports with foreign countries
    - the annual difference between spending and tax revenues
    - the accumulated borrowing of the federal government that has not been repaid
23. The Prime Minister of the United Kingdom is
    - David Cameron
    - Michael Howard
    - Tony Blair

- Gordon Brown

24. The President of Afghanistan is named
    - Bashar al-Assad
    - Hosni Mubarak
    - Hamid Karzai
    - Nouri al-Maliki
25. Who signs bills into law?
    - The President
    - The Vice President
    - The Chief Justice of the Supreme Court
    - The Secretary of State
26. The First Amendment to the United States Constitution guarantees all of these rights EXCEPT
    - the right to peaceably assemble
    - the right to remain silent
    - the right to the free exercise of religion
    - the right to free speech
27. Who is the Commander in Chief of the military?
    - The Attorney General
    - The President
    - The Secretary of Defense
    - The Vice President
28. What do we call the first ten amendments to the Constitution?
    - The Articles of Confederation
    - The Inalienable Rights
    - The Bill of Rights
    - The Declaration of Independence
29. The NRA is an organization that advocates for
    - election reform
    - a cleaner environment
    - the rights of gun owners
    - abortion rights
30. Common Cause is an organization that advocates for
    - election reform
    - a cleaner environment
    - the rights of gun owners
    - abortion rights
31. Roe v. Wade is a case decided by the Supreme Court that relates to
    - executive power

- campaign finance
- birth control
- abortion

32. Citizens United v. the FEC is a case decided by the Supreme Court that relates to
    - executive power
    - campaign finance
    - birth control
    - abortion
33. Which of these political parties is considered more conservative?
    - Green Party
    - Republican Party
    - Democratic Party
34. How many senators are elected from each state?
    - It depends on the population of the state
    - Four
    - Two
    - One
35. What are the two parts of the U.S. Congress?
    - The Senate and the Supreme Court
    - The House of Lords and the House of Commons
    - The House of Representative and the Supreme Court
    - The Senate and House of Representatives
36. The U.S. House of Representatives has how many voting members?
    - 441
    - 435
    - 200
    - 100
37. If both the President and the Vice President can no longer serve, who becomes the President of the United States?
    - The Secretary of the Treasury
    - The Secretary of State
    - The Speaker of the House
    - The President Pro Tempore of the Senate
38. In what month do we vote for President?
    - November
    - October
    - February
    - January

B. OPTIMAL REDUCED STATIC BATTERY

To make a fair comparison between CAT and traditional reduced static scales, we first developed a static scale to act as a baseline. From the 38-item pool of questions, our goal was to choose a three-item, five-item, seven-item and ten-item scale. To ensure that these scales represented an appropriate baseline for comparison, we used the following procedure:

1. We considered all $\binom{38}{3} = 8{,}346$ possible three-items batteries.
2. We constructed a population of 700 individuals with ability parameters drawn from the range of values implied by the empirical distribution of our sample.[20]
3. Using the parameters shown in Table 2, we simulated responses from our 700 hypothetical respondents to each of the 8,346 three-item batteries and generated posterior estimates of respondents' position on the latent scale, $\hat{\theta}_j^{(EAP)}$.
4. We calculated the mean squared error (MSE) for each respondent using the true ability level θ_j and the estimate $\hat{\theta}_j^{(EAP)}$.
5. For each possible combination of questions, we calculated a weighted average MSE, where weights were assigned each simulated agent based on the implied empirical distribution of the sample.[21]
6. We selected the battery that provided the lowest weighted average MSE.

After building our baseline three-item battery, we then added items using the same procedure. So, for instance, to build the five-item battery we considered all $\binom{35}{2}$ combination of questions to *add* to the three items we already identified. The items included in the ten-item reduced scale are indicated in Table 2.

[20] We generated 700 agents equally spaced in the range of the distribution of θ_j values implied by a kernel density of the $\hat{\theta}_j$'s estimated in the model shown in Table 2.

[21] We multiplied each agent's estimated MSE by the estimated density values.