

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

December 10, 2012

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 article, 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 approach to item-selection appropriate for public opinion research. We then demonstrate the advantages of CAT via simulation and empirically comparing dynamic and static measures of political knowledge.

*We are grateful for helpful comments provided by Martin Elff, Sunshine Hillygus, Walter Mebane, Brendan Nyhan, and two anonymous reviewers. A previous version of this article was presented at the 2012 Annual Meeting of the Midwest Political Science Association, the 2012 Saint Louis Area Methods Meeting, and the 2012 Summer Methods Meeting.

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 of long surveys. For most researchers, the primary cost associated with public opinion research is the per-question fee charged by survey firms. Thus, there are significant financial incentives for asking as few questions as possible when measuring any latent trait or attitude.

However, the desire to avoid large batteries also reflects the higher rate of attrition and non-response associated with lengthy and repetitive surveys. Numerous studies have shown that longer surveys are associated with higher rates of unit nonresponse (e.g., Heberlein and Baumgartner 1978; Yammarino et al. 1991; Burchell and Marsh 1992; Crawford et al. 2001; Galesic and Bosnjak 2009) a greater likelihood of halting an interview (e.g., Sheatsley 1983) and higher rates of item nonresponse (e.g, Anderson et al. 1983). Moreover, lengthy repetitive surveys increase the burden on respondents who compensate by beginning to satisfice when selecting answers (Krosnick 1991, 1999), increasing use of “don’t know” responses (Krosnick et al. 2002), and providing generally less informative responses (Herzog and Bachman 1981).

To avoid long batteries, researchers typically select a subset of available items to include on a survey. For instance, in their study of the role of personality traits in determining political attitudes, Gerber et al. (2010) rely on the Ten Item Personality Inventory (TIPI), a much reduced modification of the 44-item Big Five Inventory [BFI] (Gosling et al. 2003).¹ Likewise, the 2008-2009 American National Election Study panel included only two items adapted from the standard 18-item “need for cognition” scale (Cacioppo and Petty 1984). Indeed, developing reduced versions of larger scales constitutes a whole genre of research

¹The 44-item battery is itself an effort to develop a shorter (and publicly accessible) alternative to the 240-item NEO Personality Inventory-Revised.

in the fields of psychology, consumer research, public opinion, and more.² Yet, relying on reduced batteries can lower measurement precision and introduce bias – especially for individuals on the extreme ends of a latent scale.

In this article, we propose an alternative to the creation of static reduced scales. 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), 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 dimension) and its capacity to discriminate between respondents. CAT algorithms respond to individuals’ prior answers by choosing subsequent questions that will place them on the latent dimension with maximum precision and a minimum number of questions.

In the rest of this article, we 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 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 theory and practice, when the two competing methods of

²Just a few of the many recent examples of efforts to develop reduce scales in the literature include Podsakoff and MacKenzie (1994); Stanton et al. (2002); Russell et al. (2004); Richins (2004); Matthews et al. (2010); and Thompson (2012).

measurement are compared on a common metric, CAT provides improved precision and reduced bias.

2. CAT AND TRADITIONAL DYNAMIC SURVEY TECHNIQUES

Computerized adaptive testing: CAT algorithms are based on the notion that questions should be chosen for each respondent based, in part, on what we know about them from previous responses. Ignoring prior information leads us to waste valuable survey time asking questions to respondents that are not revealing. Just as we would not ask a simple algebra question to assess the mathematical aptitude of a theoretical physicist, we should not ask survey respondents who have already correctly explain the reconciliation process in the U.S. Senate whether they know what job or position is currently held by Joe Biden when measuring political knowledge. It is more informative to instead choose questions 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).

Relationship to traditional dynamic surveys: 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 measure of party identification asks whether 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. 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.

Beyond simple branching formats, CAT is also clearly related to computer-assisted interviewing techniques developed over two decades ago (Piazza et al. 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 criteria. Computer-assisted interviews can respond in complex ways during the interview process in a manner pre-specified by researchers.

Relationship to item response theory: 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; Martin and Quinn 2002; Clinton and Meirowitz 2003; Clinton et al.

2004; Bafumi et al. 2005; Poole 2005; Bailey 2007; Treier and Jackman 2008; Treier and Hillygus 2009; Gillion 2012).

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 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 et al. 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 tests of similar lengths (e.g., Hol et al. 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 (Jagdeep Singh 2007). However, CAT methods have rarely been applied in the measurement of public opinion. Moreover, we are aware of no instances of its use in published political science research. The purpose of this article, therefore, is to introduce the details of the CAT technique to the political science community, give guidance as to how it can be fruitfully incorporated into the study of public opinion, and provide a clear empirical demonstration of the significant advantages available from adopting and adapting this methodology.

Relative advantages of CAT surveys: 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. The disadvantage is that researchers must gather calibration data in advance in order to specify the item parameters.

A hypothetical example can demonstrate the basic advantages of 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 that there is space for five items and that all questions are dichotomous indicators (i.e., answers are either correct or incorrect).

One advantage of CAT, is that it allows researchers to include a larger number of question combinations. On a large national survey, scholars typically choose *just one* subset of these items. However, in our example there are actually $\binom{40}{5} = 658,008$ possible reduced batteries we could include. A five-item adaptive battery would allow us to include at least 16 question combinations, while a 10-item battery would include 512.³ Moreover,

³The response profiles can be modeled with a binary tree, where each node is a question and the branches correspond to correct/incorrect answers. The root node is the first question

using informative priors, response times, and other refinements, CAT would allow us to theoretically include the entire collection of potential question combinations in the latent question-item space.⁴

A second, and more important, advantage is that CAT allows us to ask *better* question combinations chosen to reveal the most information about each respondent. With both adaptive and static batteries, we will be able to partition respondents into $2^5 = 32$ categories based on their response profiles. 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 observe the full range of potential response profiles as it chooses questions that respondents are expected to answer either correctly or incorrectly with roughly equal probability.

Finally, this added precision comes with no expense in terms of survey time as each respondent is still asked only five questions. This advantage may be particularly important on national probability samples, where there are real and significant financial incentives to keep batteries short. It may also be important in situations such as interviews of political elites where response rates may be extremely sensitive to survey length.

3. COMPUTERIZED ADAPTIVE TESTING ALGORITHMS

Intuition: Having reviewed the basic motivation for applying CAT algorithms, in this section we provide additional details about one implementation of the method appropriate for

asked and the leaf nodes of this tree correspond to the number of response profiles we can obtain with the adaptive survey. In the five-item case we have a tree of height 5 and thus 2^4 possible response profiles. In the ten-item case we have 2^9 .

⁴We focus in this paper only on the simplest version of CAT. Implementation and analysis of adaptive algorithms that leverage prior information and response times for survey research remains a task for future research.

public opinion research. 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 et al. 2003), batteries on issue positions designed to place respondents into an ideological space (Bafumi and Herron 2010; Treier and Hillygus 2009), or a listing of forms of political participation respondents may have engaged in over the past year (Gillion 2012).

CAT may be particularly useful in instances where researchers care about individuals who are “extreme” along some dimension of interest. This might include political activism (e.g., Verba et al. 1995), racial attitudes (e.g., Feldman and Huddy 2005), political knowledge or sophistication (e.g., Zaller 1992), or political ideology (e.g., Bafumi and Herron 2010).

CAT is a method for taking a 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 most *precise* and *accurate* estimate of respondents’ position on a latent factor.

Ceteris paribus, CAT achieves this goal by choosing 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” to the current estimate of the respondent’s ability parameter. The algorithm prefers questions that it estimates the survey taker has a roughly equal chance of answering correctly and incorrectly. Additional intuition regarding how CAT chooses amongst available items is provided in our simulated example below.

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, p 4).

[Table 1 about here.]

First, estimates ($\hat{\theta}_j$) are generated for each respondents' position on the latent scale of interest (θ_j). Before the first item is administered, this estimate is based on our prior assumptions about 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, y_j , to specify an informative prior, $\theta_j \sim \pi(\theta_j|y_j)$ (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, y_j will include responses to items that have already been administered and answered.

Second, the next question item is selected out of the available battery. CAT chooses the item that optimizes some pre-specified 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).⁵ It is also possible to choose constrained optimization approaches to, for instance, ensure that scales balance positively and negatively worded items.

⁵An 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).

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 article, we have chosen to focus on the MEPV criterion because we (i) feel it is the most intuitive and mathematically motivated approach and (ii) it is among the criteria that previous research has shown performs 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, the algorithm produces final estimates of $\hat{\theta}_j$ and terminates.

Outline of the general model for dichotomous indicators: As discussed above, 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 examples 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 (c.f. Clinton et al. 2004; Bafumi et al. 2005).

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

of a correct response for individual j is

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

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

We assume that the item-level parameters (a_i, b_i) have already been estimated using some previously collected data, which we term the calibration sample below. These parameters are typically termed the discrimination and difficulty 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

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

where τ_θ denotes the precision of the distribution (the inverse of the variance). In our examples below, we set $\mu_\theta = 0$ and $\tau_\theta = 0.6$, a fairly diffuse (though proper) prior.⁶

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

$$L(\theta_j | \mathbf{y}_{k-1,j}) = \prod_{i=1}^{k-1} p_i(\theta_j)^{y_{ij}} q_i(\theta_j)^{(1-y_{ij})} \quad (3)$$

⁶In numerous simulation experiments, we found this setting to be ideal for the purposes of small batteries. Stronger priors (e.g., $\tau_\theta = 1$) result in item selection being dominated by the prior, which is only overcome after many items are asked. Weaker priors (e.g., $\tau_\theta = 0.01$) 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 $\tau_\theta = 1$ prior used for parameter estimation on the calibration data.

Calculating skill parameter: We present one of the most prominent methods for calculating respondent-level positions on a latent scale.⁷ The expected a posteriori (EAP) estimate of individual j 's position on the latent scale is calculated as

$$\hat{\theta}_j^{(EAP)} \equiv E(\theta_j | \mathbf{y}_{k-1,j}) = \frac{\int \theta_j \pi(\theta_j) L(\theta_j | \mathbf{y}_{k-1,j}) d\theta_j}{\int \pi(\theta_j) L(\theta_j | \mathbf{y}_{k-1,j}) d\theta_j}. \quad (4)$$

Neither of these integrals can be analytically derived. However, using numerical methods we can approximate these quantities with sufficient precision.⁸

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 ($y_{kj}^* = 1$) and incorrect ($y_{kj}^* = 0$) 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 $P(y_{kj}^* = 1 | \mathbf{y}_{k-1,j}) = 1 - P(y_{kj}^* = 0 | \mathbf{y}_{k-1,j})$ where $\mathbf{y}_{k-1,j} =$

⁷A 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).

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

where

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

⁸In 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. This is appropriate as the model is identified by fixing the distribution of ‘‘ability’’ parameters to cluster relatively tightly near zero making numerical integration less liable to significant error. Alternative parameterizations that allow for a broader possible range for θ_j may require alternative implementations.

$y_{1,j}, \dots, y_{k-1,j}$. This is done by simply entering the current value of $\hat{\theta}_j$ into Equation (1) for item k . We then calculate

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

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

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

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

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 minimize the value of

$$P(y_{kj}^* = 1 | \mathbf{y}_{k-1,j}) Var(\theta_j | \mathbf{y}_{k-1,j}, y_{kj}^* = 1) + P(y_{kj}^* = 0 | \mathbf{y}_{k-1,j}) Var(\theta_j | \mathbf{y}_{k-1,j}, y_{kj}^* = 0). \quad (8)$$

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_θ^{stop} . This option might be particularly useful when researchers seek homoscedastic measurement variance.⁹

⁹Note that one is not guaranteed the same level of variance in the estimate of $\hat{\theta}$ for each respondent when simply asking a fixed number of questions. Using this alternative stopping rule can help ensure that the estimates have equal variance and are more amenable to regression analysis.

4. SIMULATION AND ILLUSTRATION

In this section, we seek to demonstrate the potential advantages of CAT through a simulation study. These simulations represent circumstances that are as ideal as can be expected in a survey setting. The item pool consists of 60 items and the response probabilities align exactly with the two-parameter logistic model in Equation (1). The discrimination parameters are drawn from $a_i \sim \text{Gamma}(50, 25)$ and the difficulty parameters (b_i) are spaced equidistantly on the interval $[-3, 3]$. In essence, the simulation assumes that we have 60 items that load strongly on the underlying latent dimension ($\bar{a} = 2$) with item difficulty parameters spanning the range of likely ability parameters.

4.1. *Illustrative simulated example*

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

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 et al. (2003, p 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 absolute bias and error. Quite simply, most respondents are located in the middle of the distribution so it makes sense to choose items biased towards the center of the distribution. However, this strategy often results in imprecise and even biased estimates of respondents located towards the extreme ends of the latent scale.

More fundamentally, reduced scales almost inevitably include some questions that are either too “easy” or too “hard” for a given respondent. This results in inefficiencies since questions are administered to respondents which provide no additional insight as to their true position on the latent scale.

This is illustrated in Figure 1, which shows the item characteristic curves (ICC) for a fixed (left panels)¹⁰ and dynamic (right panels) battery administered to a single individual whose true position on the latent scale is indicated by the vertical dashed line.¹¹ Item characteristic curves show the predicted probability of answering a question affirmatively (i.e., getting the question “right”) for individuals of varying skill levels (θ_j). Thus, the horizontal axis shows the different potential values of θ_j , while the vertical axis shows the probability of answering affirmatively for each value of θ_j .

[Figure 1 about here.]

Note especially the ICCs shown in the four bottom-most panels of Figure 1. The ICC curve for the fixed battery indicates that the respondent, whose position is at -1 on the latent scale, will almost certainly not answer either Item 4 or Item 5 in the fixed battery affirmatively. The predicted probability of doing so is nearly zero. Thus, as we show below, no additional information about the respondent is gained by asking these questions. On the other hand, the items chosen by the CAT battery are all such that the respondent has a significant probability of answering in either directions. This suggests that we will learn

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

¹¹Recall that that MEPV selection criteria chooses the item which minimizes the function shown in Equation (8). The selected items will generally have (i) large discrimination parameters and (ii) difficulty parameters located near the algorithm’s current estimate of $\hat{\theta}$. If most items have fairly similar discrimination parameters, the latter criteria tends to dominate. Note that, even in a simulated example where all items perform well, there is not always a strictly smooth relationship between an item’s difficulty and its EPV due to heterogeneity in discrimination parameters.

more about the respondent as each additional question in the CAT battery that is asked and answered.

This intuition is confirmed by looking at the estimated posteriors for θ_j shown in Figure 2. The Figure shows the true value (θ_j) and the posterior estimates of ($\hat{\theta}_j$) given responses to all previously administered items for both the static (left panels) and dynamic (right panels) batteries.¹²

[Figure 2 about here.]

There are two aspects of Figure 2 we wish to emphasize. First, the final estimate for the static battery ($\hat{\theta}_j = -0.53$) is relatively inaccurate compared to the final estimate of the dynamic battery ($\hat{\theta}_j = -0.89$). Second, neither the precision nor the accuracy of the estimate are improved after the administration of Items 4 and 5 in the static scale. That is, over 40% of the battery provides almost no additional information about this respondent. In contrast, the right panels of Figure 2 show the posteriors as determined by items chosen by the CAT algorithm. As can be seen, the estimate of θ_j is more accurate and far more precise. Moreover, the posterior continues to converge towards the true value of θ_j after each question is administered and answered.

4.2. Systematic simulation

While the results in Figures 1-2 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 squared error¹³ for the dynamic (gray) and static (black) batteries

¹²For illustrative purposes, we assume that responses are deterministic. That is, respondents always answer affirmatively (correctly) when the predicted probability is greater than 0.5.

¹³Squared error is defined as $Var(\hat{\theta}_j^{(EAP)}) + (\theta_j - \hat{\theta}_j^{(EAP)})^2$.

of various lengths.¹⁴ 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 ten items respectively.¹⁵ These estimates were generated for 1,000 simulated respondents distributed equidistantly on the interval [-3,3]. This provides a fairly precise understanding for how the CAT algorithms perform across the possible range of θ_j . The two sets of curves show the squared error (the vertical axis) that would result from the administration of a static and dynamic batteries for individuals with differing positions on the underlying latent scale (the horizontal axis).

[Figure 3 about here.]

There are three aspects of the results shown in Figure 3 that are helpful for understanding the advantages of CAT. First, across the entirety of the range of values of θ_j , the dynamic survey results in a lower squared error. That is, for survey batteries of a similar length, a dynamic survey provides more accurate and more precise estimates of the latent trait.¹⁶

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 more equally with the dynamic scale for values of θ_j near zero. This is an expected finding that replicates results from previous simulation studies (e.g., van der

¹⁴The static batteries are: Items 15, 30, 45 for $n = 3$; Items 10, 20, 30, 40, 50 for $n = 5$; Items 9, 16, 23, 30, 37, 44, 51 for $n = 7$; and Items 3, 9, 15, 21, 27, 33, 39, 45, 51, 57 for $n = 10$. Alternative methods for selecting fixed batteries make little difference in the substantive conclusions of these simulations.

¹⁵The stopping rule chosen for this and the empirical example is based on our belief that the primary constraint for public opinion researchers is time. While another rule, such as a measurement precision threshold, could have been used, this did not seem realistic given the budget and time constraints of most public opinion surveys.

¹⁶Indeed, examining the bias alone shows that CAT also provides lower bias for nearly all values of θ_j . Improvement in unbiasedness is especially large in the extreme range of θ_j (results not shown).

Linden 1998). It indicates the degree to which fixed batteries are optimized to accurately measure individuals near the center of the distribution.¹⁷

Finally, Figure 3 shows visually how CAT divides respondents into a larger number of categories or “bins” than traditional static scales. By asking questions each respondent is more likely to answer right or wrong, CAT increases the likelihood of observing a positive or negative response to *each* question. A clear example of this is visible in the upper-left panel of Figure 3. With three questions, we can potentially observe $2^3 = 8$ response profiles. Indeed, we can see that the CAT scale divides respondents into exactly eight categories as there are eight “U” shaped bins visible. However, with a static scale we observe only four.¹⁸ Increasing the number of observed response profiles improves both measurement accuracy and precision.

5. EMPIRICAL APPLICATION: POLITICAL KNOWLEDGE

The simulation results in Section 4 show the advantages of CAT relative to a static scale theoretically. In this section, we provide an empirical application of CAT to the domain of political knowledge (sometimes termed political sophistication). This example demonstrates the superiority of CAT relative to static scales for accurately and precisely measuring important concepts to political science.

Although scholars have developed a number of measures for knowledge and sophistication (e.g., Luskin 1987; Sniderman et al. 1991; Delli Carpini and Keeter 1993), one of the

¹⁷It makes sense that static scales are aimed at individuals in the “middle” of the latent space, as this is where most individuals will be located. However, the advantage of CAT is that the algorithm will tailor the battery to efficiently estimate a latent trait regardless of the respondents’ position in the latent space. As we show, this improves measurement for all individuals, but the comparative advantage is greatest for more extreme individuals.

¹⁸Recall that in these simulations answers are deterministic (see Footnote 12). In a static scale, therefore, there will always be $n + 1$ bins while there will be 2^n bins for the dynamic batteries.

most widely used methods for survey researchers is to ask questions measuring 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 four items allow open ended responses, other commonly used items 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?

While this method of measuring political knowledge is used widely in public opinion research (e.g., Barabas 2002; Brewer 2003; Delli Carpini and Keeter 1993, 1996; Gomez and Wilson 2001), it has also 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; Prior 2012). The coding of the open ended responses is of questionable reliability, and on occasion entirely incorrect.¹⁹ Some have argued that the heavy emphasis on identifying prominent individuals does not seem a valid indicator of “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).²⁰

¹⁹DeBell (2012) notes that in 2004 identifying Tony Blair as the “Prime Minister of the United Kingdom” was coded as an incorrect response.

²⁰These percentages change depending on how “correct” responses are coded. In addi-

In this section, therefore, we apply a much larger collection of closed form (i.e., multiple choice) questions designed to measure much broader areas of political knowledge necessary for successfully engaging in the political system. In addition, the questions were designed to supply sufficient variation in difficulty.

5.1. *Model calibration*

We developed a battery of 70 multiple choice knowledge questions, six of which were dropped due to poor performance. These items were largely drawn from questions used previously in national samples (e.g., Luskin and Bullock 2011). The remaining 64 items, which measure knowledge in areas including the legislative process, interest groups, foreign affairs, and constitutional rights, are listed in our online supplemental materials.

We administered the battery to 810 respondents based in the United States and over the age of 18. Respondents were recruited through Amazon Mechanical Turk.²¹ This sample primarily serves to calibrate the model and allow us to estimate appropriate difficulty tion, 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)

²¹Berinsky et al. (2012) provide a detailed analysis of Mechanical Turk participants. Among their findings, which is reflected in our results below, is that Turk participants have much higher levels of political knowledge than the general U.S. population. Although this limits the degree to which the item calibrations from this sample can be used for CAT algorithms administered to other populations, it does not alter the nature of our findings about the relative advantages of the CAT approach itself.

Since the calibration and test samples are drawn from the same population, the underlying distribution of political knowledge can be assumed to be similar. It is similarity in the distribution of political knowledge in the two samples that allows the CAT algorithm to outperform static batteries in the analyses below. To the degree that we might be concerned that this assumption is not correct as a result of our non-random sampling procedure, this fact only makes the significant improvement in measurement quality we show below more remarkable. In general, the more accurate the calibration of the item parameters, the better the performance of CAT will be relative to any given static scale.

and discrimination 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.²² However, this convenience sample serves the more limited purpose of illustrating the usefulness of the CAT method.

[Table 2 about here.]

Table 2 presents the item-level parameters associated with each of the questions in our battery. Items are ordered according to their difficulty parameters. The easiest question (the item with the lowest difficulty parameter) identified by this sample is Question 1, “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 Question 64: “On which of the following does the U.S. federal government spend the *most* money each year? (a) Education, (b) Medicare, (c) Interest on the national debt, (d) National defense.”²³ Broadly speaking, respondents ordered questions in difficulty as we would expect. While there are available items for all levels of difficulty, there is a skew towards the lower end of the difficulty spectrum. This is unsurprising as we relied on items originally designed for national probability samples.

²²Using 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. Researchers should avoid using convenience samples to calibrate the CAT algorithm unless they are comfortable that the distribution of the calibration sample is “representative” of the overall population of interest on the given latent trait.

²³Response options were randomized for all respondents except where responses had a clear numerical ordering. In addition, respondents were always allowed to answer that they did not know the answer or to simply skip the question (after a five-second delay). All “Don’t know” and skipped questions were coded as incorrect responses.

5.2. Empirical comparison with reduced scales

To assess the efficacy of CAT techniques outside of the calibration sample, we conducted a survey experiment on a fresh sample of 820 respondents.²⁴ In this second survey, roughly half of the respondents (n=401) first answered a ten-item fixed battery.²⁵ The remaining respondents (n=419) answered ten items as selected by the CAT algorithm discussed above. Respondents in *both* groups then answered the remaining 54 items presented 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 64-item battery. Thus, we approximated respondent j ’s true latent trait value, θ_j , using her answers to the 64 questions. We then computed estimated values of $\hat{\theta}_j^{(EAP)}$ based on the first $n \in (3, 5, 7, 10)$ questions administered in either treatment condition.²⁶ Our purpose is to evaluate how well $\hat{\theta}_j^{(EAP)}$ approximates θ_j across treatment conditions.

Note that the data from the calibration sample is used *only for item selection*. That is, we re-estimated the entire measurement model using only the fresh sample. This provides a fair test of the method as we are not simply assuming that the model estimated on the calibration sample is true for the second (experimental) sample.

Figure 4 shows the squared error of the estimated values of $\hat{\theta}_j^{(EAP)}$ for individuals in

²⁴Respondents were again based in the United States and over the age of 18 and were recruited using Amazon Mechanical Turk. Respondents who had participated in the first-round survey were excluded from this analysis.

²⁵We designed the fixed battery to provide a 3-item, 5-item, 7-item and 10-item batteries with good measurement properties. That is, we chose items that spanned the range of difficulty and had relatively large discrimination parameters in the calibration sample. Moreover, in choosing between similarly performing items, we selected questions most similar to the standard ANES measure of political knowledge (e.g., “Who is the Speaker of the House of Representatives?”). The items in the fixed battery are indicated in Table 2.

²⁶Observe that, because the CAT algorithm minimizes $\hat{\theta}_j^{(EAP)}$ at every step of the test, we can compare the results obtained by stopping at any given point without loss of generality.

the dynamic (squares) and static (triangles) treatment conditions. The lines represent loess curves for each population.²⁷ As in the simulations in Section 4, the dynamic algorithm outperforms the fixed battery for all values of θ_j and the difference is particularly noticeable for larger values of n and more extreme values of θ_j .²⁸

[Figure 4 about here.]

Note that the gains in measurement accuracy and precision are more significant than may be immediately obvious in Figure 4. Figure 5 compares the static and dynamic scales of various lengths at the population level. The upper-left panel shows the median squared error (MSE) while the upper-right panels shows the median absolute bias, measured as $|\theta_j - \hat{\theta}_j^{(EAP)}|$. Likewise, the lower-left panels shows the total absolute bias for each of the samples and the lower-right panel shows the median posterior variance for the estimates. The figure shows that CAT offers dramatic improvements in measurement accuracy relative to static scales on each of these metrics. The measures are *both* less biased and more precisely estimated. Indeed, by these metrics a CAT scale with only six items outperforms a 10-item fixed scale. Extrapolating from this example, Figure 5 suggests that CAT offers the potential for 40% reduction in battery length on surveys with no loss in measurement quality.

[Figure 5 about here.]

Finally, the increased accuracy and precision resulting from using the adaptive measure affects the inferences we draw from the data.²⁹ To begin with, the dynamic battery is able

²⁷We used a two-sided Wilcoxon test to determine whether the distributions were indeed statistically different. The results of the test confirm what we observed visually. That is, squared error is significantly lower in all cases with $p < 0.001$. The rank-sum statistics were $W = 69470, 43790, 28108, \text{ and } 13546$ for $n = 3, 5, 7, \text{ and } 10$ respectively.

²⁸Some caution is needed in comparing the performance of the two methods for the extreme values of θ due to the increasingly small sample size.

²⁹We thank an anonymous reviewer for suggesting this analysis.

to provide a more fine-grained measure that better reflects heterogeneity within the sample. For instance, the estimated sample variance on the political knowledge scale is only 0.65 when using a five-item static battery while it is 15% greater (0.75) for the sample that took the dynamic battery. Likewise, fully 21% of respondents taking the five-item static scale were placed in the most extreme category as either the most or least knowledgeable respondents in the sample. However, only 5% of respondents taking a dynamic scale of an identical length were so categorized due to the adaptive nature of the CAT battery.

The increased precision and reduced bias also improves the external validity of the knowledge measure. We can see this by examining the level of correlation between our political knowledge measure and other responses that we would expect to be correlated with political knowledge. Table 3 shows bivariate regressions for the five-item dynamic and static political knowledge measures. The “outcomes” in these regressions are answers to three additional survey items: the respondent’s level of interest in politics (a four-point scale), the frequency with which they discuss politics and current affairs (a seven point scale), and the degree to which they report paying attention to national and international issues (a seven point scale).³⁰

[Table 3 about here.]

To make the coefficients comparable, we re-scaled the knowledge batteries to range between 0 and 1. Table 3 shows that the coefficients for the dynamic political knowledge scale are always larger. Moreover, the differences are substantive. For example, our political interest questions asks respondent, “How interested would you say you are in politics and current affairs: Not at all interested; Not very interested; Somewhat interested; or, Very interested?” Moving from the minimum to the maximum of political knowledge as measured by the static five-item knowledge battery is associated with an average change of 1.31

³⁰All survey question wording and response options are shown in the online supplemental materials.

on this scale. However, because of the improved measurement properties of the five-item adaptive battery, moving from a minimum to a maximum level of political knowledge is associated with a 2.08 unit change for respondents randomly assigned to answer the dynamic battery. Table 3 shows similar differences for all three outcomes.

6. CONCLUSION

While we believe the evidence presented above suggests that CAT offers a superior approach to traditional static batteries, the methodology comes with several caveats and limitations that are important to note. First, CAT is only appropriate when researchers are interested in placing respondents onto some latent scale rather than examining responses to specific questions. Second, CAT should not be used for batteries where there is evidence of strong question order effects.

Third, CAT requires 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 costs of reducing the length of batteries and the costs of pre-testing batteries. The more accurate the pre-test (i.e., the larger the sample size and the more representative the sample), the greater will be the potential for reducing battery length while preserving measurement accuracy.

In part, pre-testing costs may be ameliorated by making survey data and item calibrations widely available to other researchers, thus sharing the costs of pre-testing. In any case, we note that even the selection of static reduced batteries is based on some kind of pre-test data or prior understanding. We strongly believe that, *ceteris paribus*, CAT algorithms will always outperform static batteries for any fixed level of pre-test information or prior beliefs about item performance.

Finally, for time-varying attitudes or traits, the calibration may not always remain cur-

rent 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 article implemented only uninformative priors, MEPV item selection, EAP ability estimation, and fixed-length batteries. Future research could explore which algorithms of the many available in the literature 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 article 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. Though the intuition behind such models is similar to that described above, 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 ordered-categorical survey questions.

Finally, we note that there are several extensions to CAT algorithms that may significantly improve performance beyond what we show here. These include the development of informative priors based on earlier survey response (van der Linden 1999) and accounting for response times (van der Linden 2008). In addition, we believe that additional research is warranted on the development of priors more appropriate for survey research because the battery size is likely to be quite short and responses may include more error relative to educational testing applications.

Although there is room for continued improvement and extension, we have shown in this article that CAT techniques 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. Adaptive testing allows for the administration of fewer questions while achieving superior levels of statistical precision and accuracy relative to any static reduced scale. 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 examples. Using a battery of political knowledge items, we administered a set of 64 questions to 810 respondents and calibrated the CAT algorithm on their responses. When compared to a fixed battery, CAT provided both improved measurement precision and accuracy for a fresh sample of 820 respondents. This was particularly true for larger numbers of questions and more “extreme” respondents. Finally, we have developed software to administer such dynamic surveys. This software will be made available to researchers who wish to adopt CAT techniques using a variety of survey platforms.³¹

7. FUNDING

This work was supported by grants from the Weidenbaum Center on the Economy, Government, and Public Policy at Washington University in Saint Louis and the National Science Foundation [SES-1023762 to J.M.M.].

³¹All data used to generate the results in this article will be made available to the public in the journal’s dataverse upon publication at <http://hdl.handle.net/1902.1/19381> (Montgomery and Cutler 2012).

References

- Anderson, A., A. Basilevsky, and D. Hum (1983). Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, and A. B. Anderson (Eds.), *Handbook of Survey Research*, pp. 415–481. New York: Academic Press.
- Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis* 13(2), 171–187.
- Bafumi, J. and M. 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.
- Bailey, M. A. (2007). Comparable preference estimates across time and institutions for the court, congress, and presidency. *American Journal of Political Science* 51(3), 433–448.
- Baker, F. B. and S.-H. Kim (2004). *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.
- Barabas, J. (2002). Another look at the measurement of political knowledge. *Political Analysis* 10(2), 209–222.
- Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research: Amazon.com’s Mechanical Turk. *Political Analysis* 20(3), 329–350.
- Brewer, P. R. (2003). Values, political knowledge, and public opinion about gay rights. *Public Opinion Quarterly* 67(3), 173–201.
- Burchell, B. and C. Marsh (1992). The effect of questionnaire length on survey response. *Quality & Quantity* 26(3), 233–244.
- Cacioppo, J. T. and R. E. Petty (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment* 48(3), 306–307.
- Choi, S. W. and R. J. Swartz (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement* 33(6), 419–440.
- Clinton, J., S. Jackman, and D. Rivers (2004). The statistical analysis of roll call voting: A unified approach. *American Political Science Review* 98(2), 355–370.
- Clinton, J. D. and A. Meirowitz (2001). Agenda constrained legislator ideal points and the spatial voting model. *Political Analysis* 9(3), 242–259.
- Clinton, J. D. and A. Meirowitz (2003). Integrating voting theory and roll call analysis: a framework. *Political Analysis* 11(4), 381–396.

- Crawford, S. D., M. P. Couper, and M. J. Lamias (2001). Web surveys : Perceptions of burden. *Social Science Computer Review* 19(2), 146–162.
- DeBell, M. (2012). Harder than it looks: Coding political knowledge on the ANES. Paper presented at the 2012 meeting of the Midwest Political Science Association in Chicago.
- Delli Carpini, M. X. and S. Keeter (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science* 37(4), 1179–1206.
- Delli Carpini, M. X. and S. Keeter (1996). *What Americans Know About Politics and Why it Matters*. New Haven: Yale University Press.
- Dodd, B. G., R. De Ayala, and W. 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*. Mahwah, New Jersey: Lawrence Erlbaum.
- Feldman, S. and L. Huddy (2005). Racial resentment and white opposition to race-conscious programs: Principles or prejudice? *American Journal of Political Science* 49(1), 168–183.
- Forbey, J. D. and Y. S. Ben-Porath (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment* 19(1), 14–24.
- Galesic, M. and M. Bosnjak (2009). Effects of questionnaire length on participation and indicators of response quality in web surveys. *Public Opinion Quarterly* 73(2), 349–360.
- Gerber, A. S., G. A. Huber, D. Doherty, C. M. Dowling, and S. E. Ha (2010). Personality and political attitudes: Relationships across issue domains and political contexts. *American Political Science Review* 104(01), 111–133.
- Gibson, J. L. and G. A. Caldeira (2009). Knowing the Supreme Court?: A reconsideration of public ignorance of the high court. *Journal of Politics* 71(2), 429–441.
- Gillion, D. Q. (2012). Re-defining political participation through item response theory. Unpublished Paper.
- Gomez, B. T. and J. M. Wilson (2001). Political sophistication and economic voting in the American electorate: A theory of heterogeneous attribution. *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.

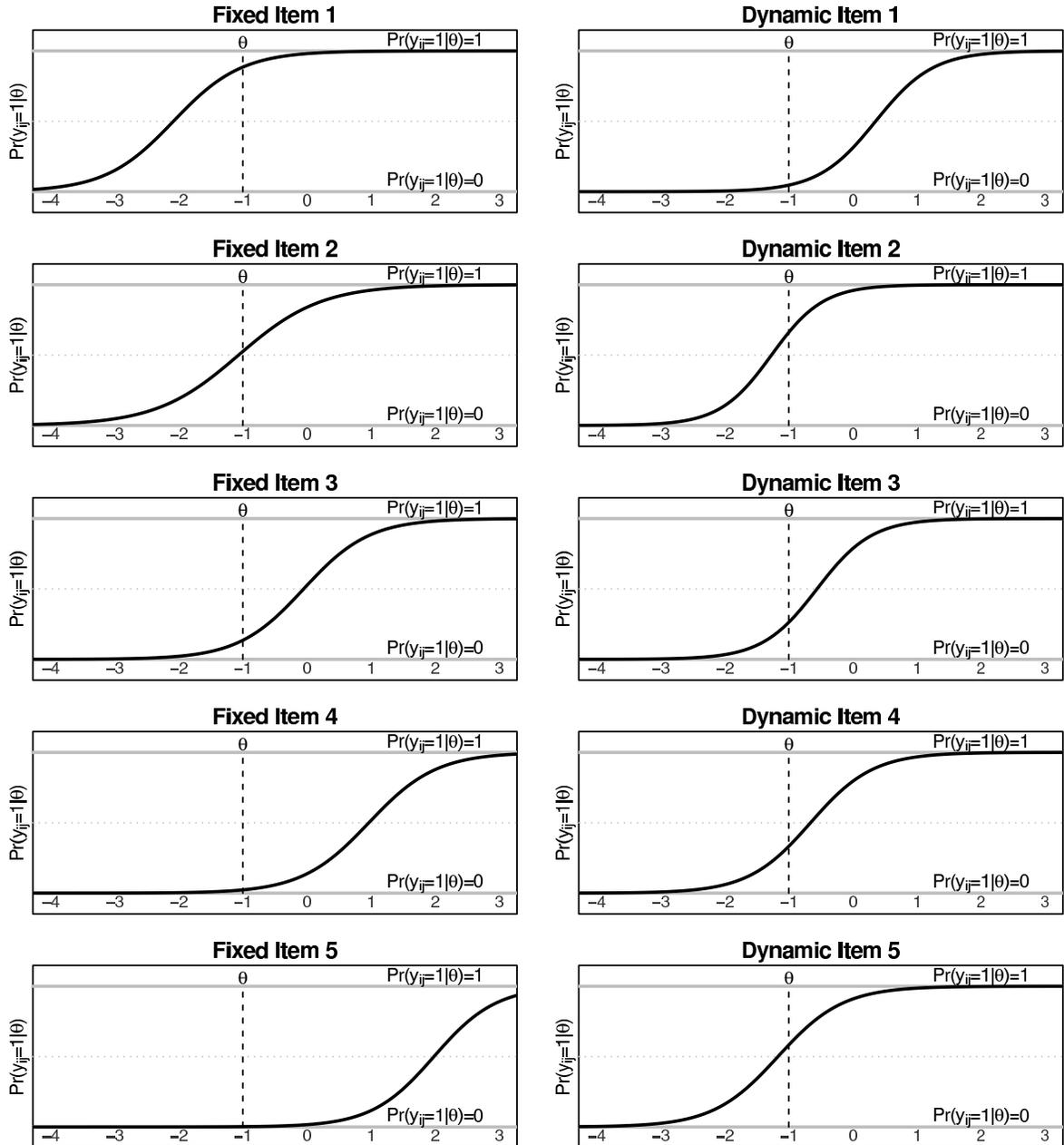
- Heberlein, T. A. and R. Baumgartner (1978). Factors affecting response rates to mailed questionnaires: A quantitative analysis of the published literature. *American Sociological Review* 43(4), 447–462.
- Herzog, A. R. and J. G. Bachman (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly* 45(4), 549–559.
- Hol, A. M., H. C. Vorst, and G. 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, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis* 9(3), 227–241.
- Jagdeep Singh, Roy D. Howell, G. K. R. (2007). Designs for likert-type data: An approach for implementing marketing surveys. *Journal of Marketing Research* 19(1), 12–24.
- Kingsbury, G. and D. J. Weiss (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. New York: Academic Press.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5(3), 213–236.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology* 50, 537–67.
- Krosnick, J. A., A. L. Holbrook, M. K. Berent, R. A. B. T. Carson, W. Hanemann, R. J. Kopp, C. Mitchell, Robert Cameron, S. Presser, P. A. Ruud, V. Smith, W. R. Moody, M. C. Green, and M. 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.
- Lord, F. and M. R. Novick (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: L. Erlbaum Associates.
- Lupia, A. (2006). How elitism undermines the study of voter competence. *Critical Review* 18(1-3), 217–232.
- Lupia, A. (2008). Procedural transparency and the credibility of election surveys. *Electoral Studies* 27(4), 732–739.

- Luskin, R. C. (1987). Measuring political sophistication. *American Journal of Political Science* 31(4), 856–899.
- Luskin, R. C. and J. 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.
- Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953-1999. *Political Analysis* 10(2), 134–153.
- Matthews, R. A., L. M. Kath, and J. 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.
- Mondak, J. J. (2001). Developing valid knowledge scales. *American Journal of Political Science* 45(1), 224–238.
- Mondak, J. J. and M. R. Anderson (2004). The knowledge gap: A reexamination of gender-based differences in political knowledge. *Journal of Politics* 66(2), 492–512.
- Mondak, J. J. and B. C. Davis (2001). Asked and answered: Knowledge levels when we will not take “don’t know” for an answer. *Political Behavior* 23(3), 199–224.
- Montgomery, J. M. and J. Cutler (2012). “Replication data for: Computerized Adaptive Testing for Public Opinion Surveys”. <http://hdl.handle.net/1902.1/19381> IQSS Data-verse Network.
- Piazza, T., P. M. Sniderman, and P. E. Tetlock (1989). Analysis of the dynamics of political reasoning: A general-purpose computer-assisted methodology. *Political Analysis* 1(1), 99–119.
- Podsakoff, P. M. and S. B. MacKenzie (1994). An examination of the psychometric properties and nomological validity of some revised and reduced substitutes for leadership scales. *Journal of Applied Psychology* 79(5), 702–713.
- Poole, K. T. (2005). *Spatial Models of Parliamentary Voting*. New York: Cambridge University Press.
- Prior, M. (2012). Visual political knowledge: A different road to competence. Unpublished paper.
- Prior, M. and A. Lupia (2008). Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science* 52(1), 19–183.
- Richins, M. L. (2004). The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research* 31(1), 209–219.

- Russell, S. S., C. Spitzmüller, L. F. Lin, J. M. Stanton, P. C. Smith, and G. H. Ironson (2004). Shorter can also be better: The abridged job in general scale. *Educational and Psychological Measurement* 64(5), 878–893.
- Segall, D. 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, D. O. (2005). Computerized adaptive testing. In *Encyclopedia of Social Measurement*, Volume 1, pp. 429–438. Oxford: Elsevier.
- Segall, D. O. (2010). Principles of multidimensional adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of Adaptive Testing*, pp. 57–76. New York: Springer.
- Sheatsley, P. (1983). Questionnaire construction and item writing. In P. H. Rossi, J. D. Wright, and A. B. Anderson (Eds.), *Handbook of Survey Research*, pp. 195–230. New York: Academic Press.
- Sniderman, P. M., R. A. Brody, and P. E. Tetlock (1991). *Reasoning and Choice: Explorations in Political Psychology*. New York: Cambridge University Press.
- Sniderman, P. M., T. Piazza, P. E. Tetlock, and A. Kendrick (1991). The new racism. *American Journal of Political Science* 35(2), 423–47.
- Stanton, J. M., E. F. Sinar, W. K. Balzer, and P. C. Smith (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology* 55(1), 167–194.
- Thompson, E. R. (2012). A brief index of affective job satisfaction. *Group & Organization Management* 37(3), 275–307.
- Treier, S. and D. S. Hillygus (2009). The nature of political ideology in the contemporary electorate. *Public Opinion Quarterly* 73(4), 679–703.
- Treier, S. and S. Jackman (2008). Democracy as a latent variable. *American Journal of Political Science* 52(1), 201–217.
- van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika* 63(2), 201–216.
- van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement* 23(1), 21–29.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics* 33(1), 5–20.

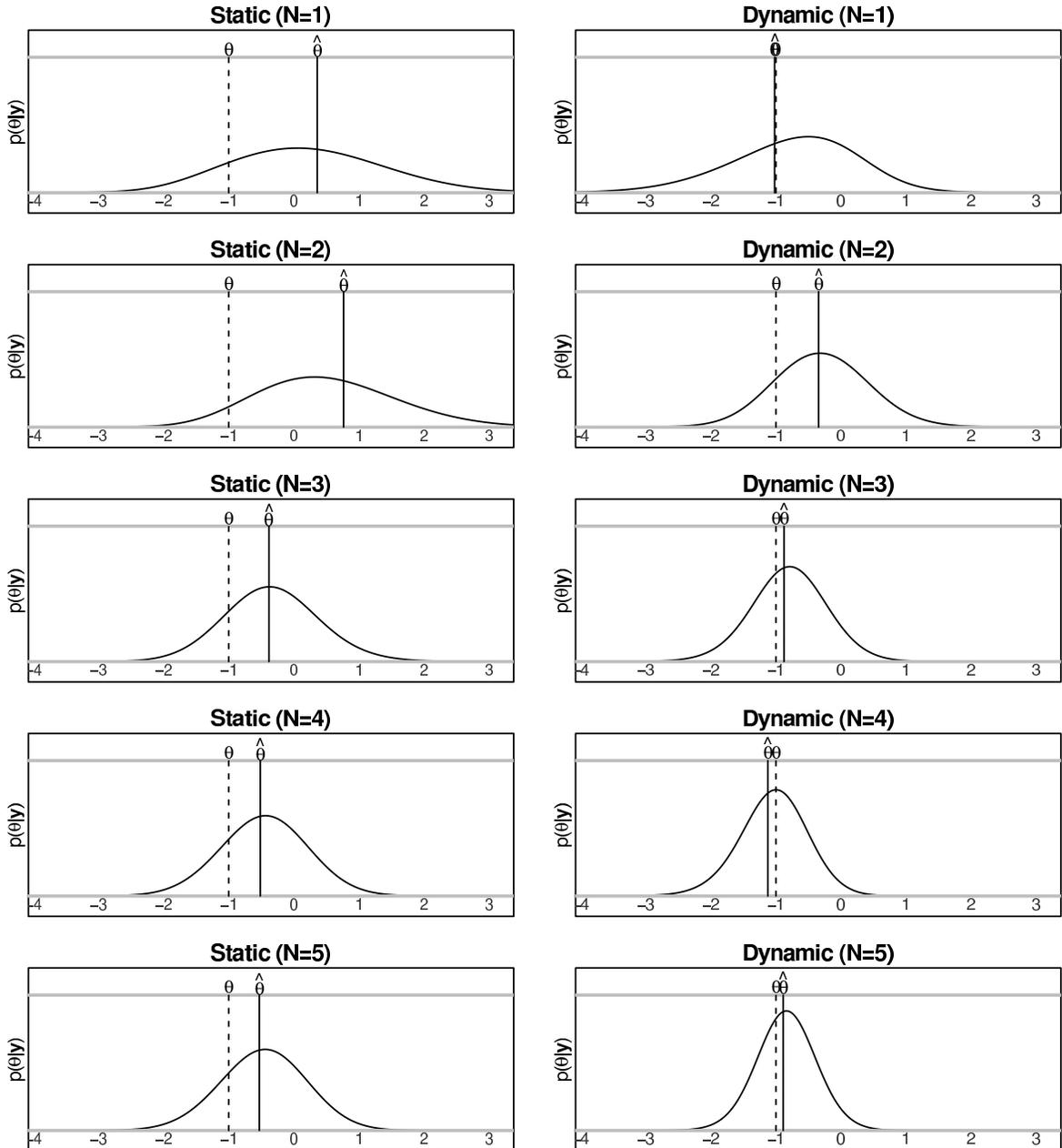
- van der Linden, W. J. (2010). Constrained adaptive testing with shadow tests. In W. J. van der Linden and C. A. W. Glas (Eds.), *Elements of Adaptive Testing*, pp. 31–56. New York: Springer.
- van der Linden, W. J. and P. J. Pashley (2010). *Elements of Adaptive Testing*. New York: Springer.
- Verba, S., K. L. Schlozman, and H. E. Brady (1995). *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.
- Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer*. 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, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement* 6(4), 473–492.
- Weiss, D. J. and G. G. Kingsbury (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement* 21(4), 361–375.
- Xu, X. and J. Douglas (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika* 71(1), 121–137.
- Yammarino, F. J., S. J. Skinner, and T. L. Childers (1991). Understanding mail survey response behavior: A meta-analysis. *Public Opinion Quarterly* 55(4), 613–639.
- Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.

Figure 1: Item characteristic curves for a fixed and dynamic five-item battery



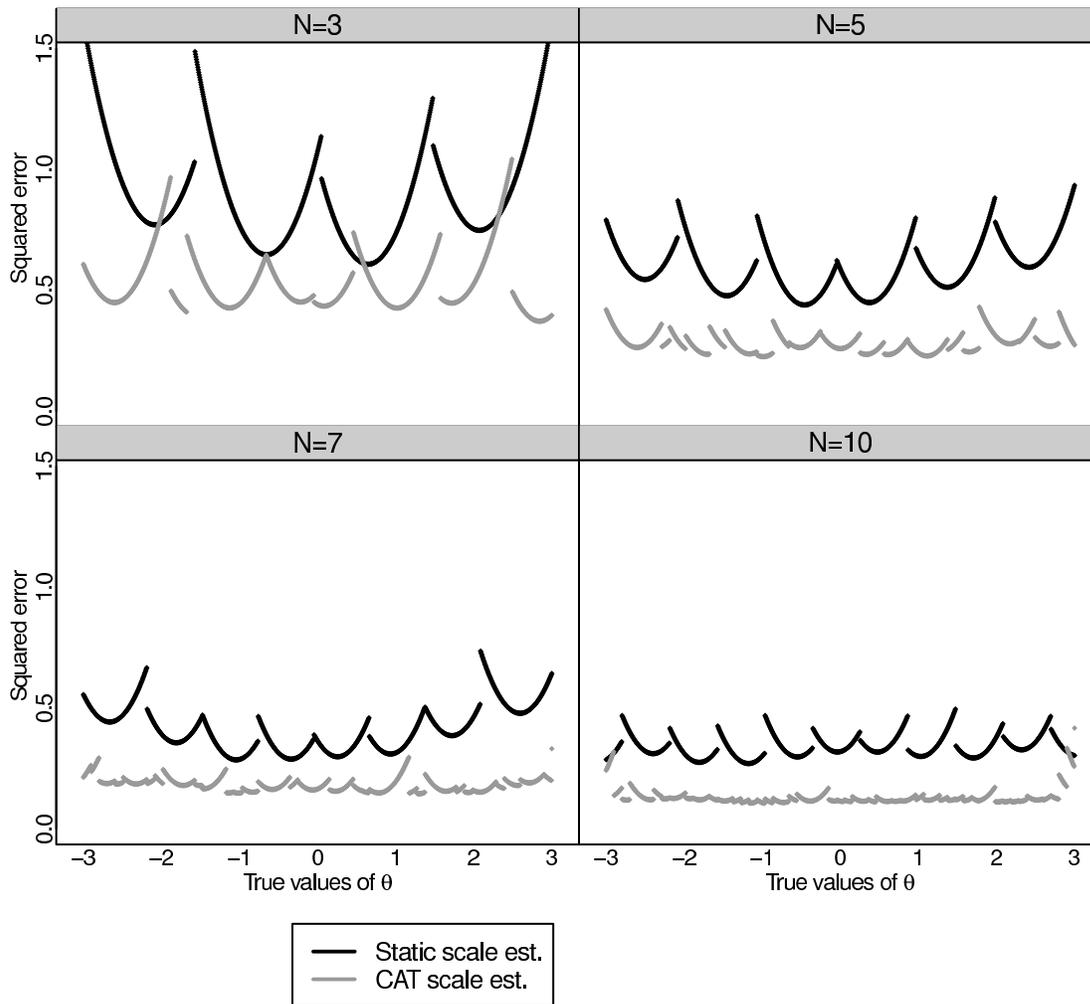
The left panels show the item characteristic curves for the five items in the static scale is administered. The right panel shows these same curves for the items as chosen by the CAT algorithm are administered. Note that the CAT algorithm chooses items that the respondent, whose position is indicated with a vertical line, has a significant probability of answering either correctly or incorrectly. For the static battery, the respondent is extremely unlikely to answer Items 4 and 5 correctly.

Figure 2: Exemplar posterior estimates for a five-item static and dynamic battery



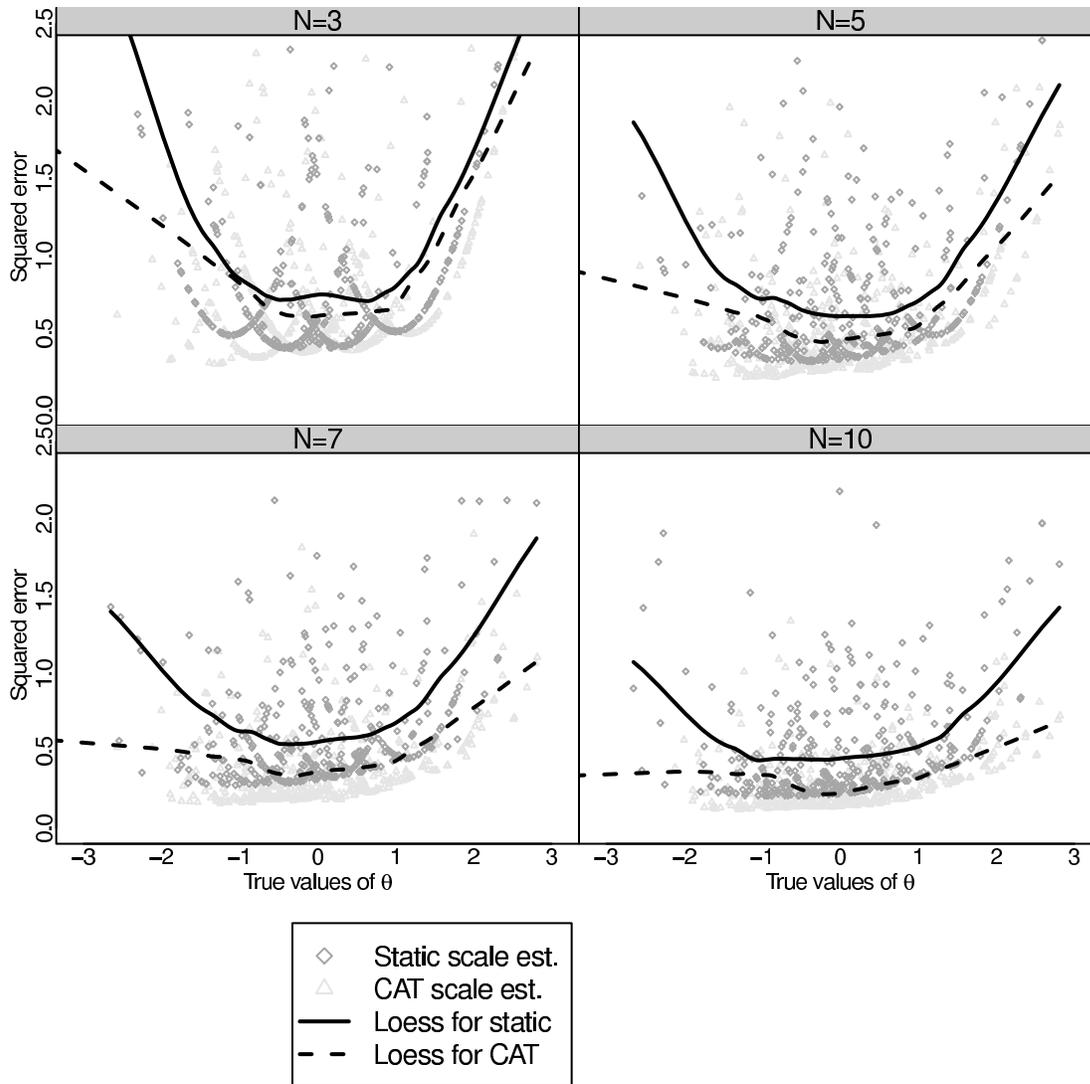
The left panels show the posterior estimates for the position of a single individual after each of five items in a fixed scale. The left panel shows the posterior estimate after items as chosen by a dynamic scale. The true value (θ) and estimate ($\hat{\theta}$) are indicated with the dashed and solid vertical lines respectively. Note that the posterior for the dynamic scale continues to converge to the true value of θ for all five items, while little additional information is garnered from the administration of the final two items in the fixed sale.

Figure 3: Squared error for dynamic and static scales of four different lengths for simulated respondents.



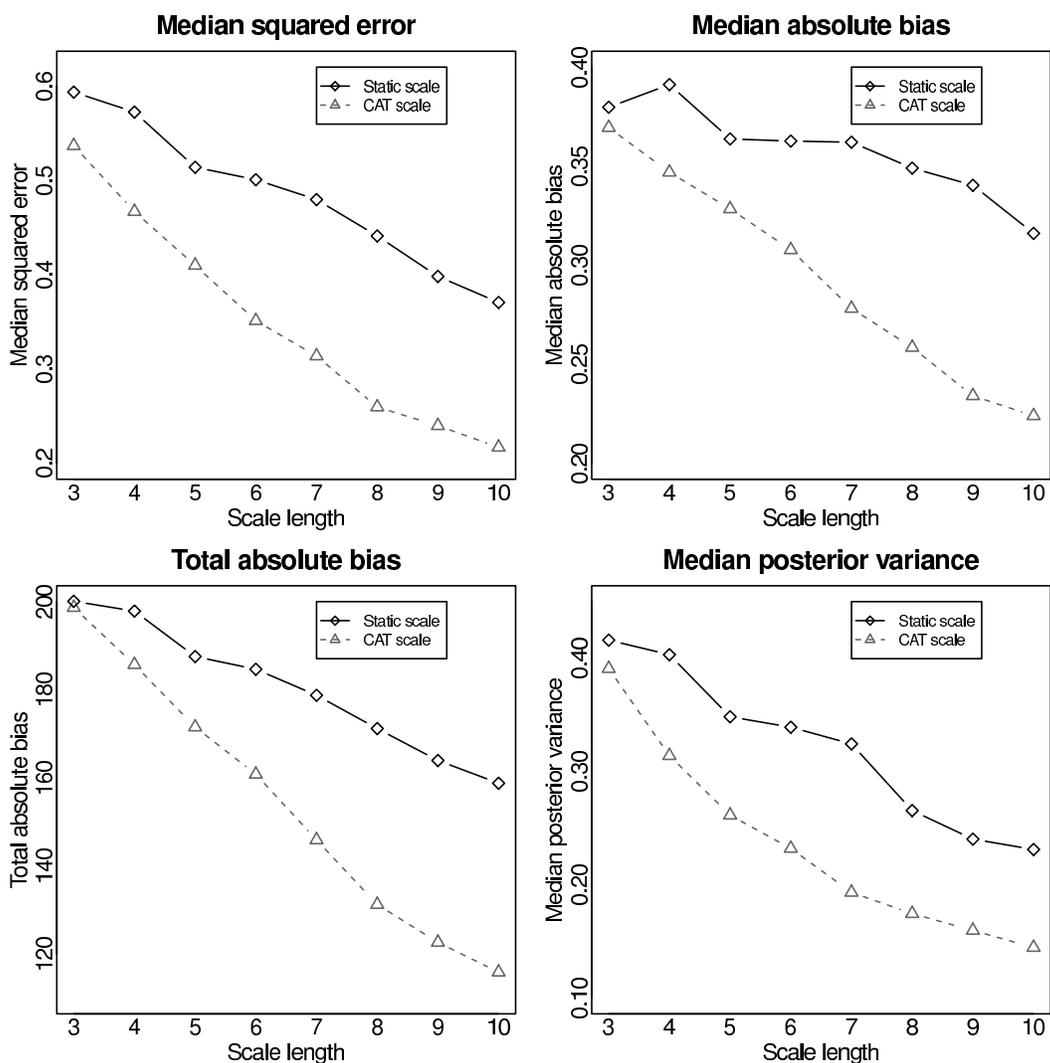
The curves show the squared error, defined as $Var(\hat{\theta}_j^{(EAP)}) + (\theta_j - \hat{\theta}_j^{(EAP)})^2$, for simulated individuals with differing values on the latent scale (θ). The point $\{0,1\}$ is associated with a squared error of 1 for an individual whose true position on the latent scale is $\theta_j = 0$. The dark curves show the estimated squared error that results administering a static scale, while the lighter curves show the squared error associated with a CAT scale of identical length ($n=3,5,7$, and 10).

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



The points show the squared error, defined as $Var(\hat{\theta}_j^{(EAP)}) + (\theta_j - \hat{\theta}_j^{(EAP)})^2$, for each individual using both the CAT (triangles) and static (circles) batteries of length 3, 5, 7 and 10. The lines are loess estimates for each method.

Figure 5: Comparing measurement quality by battery length



The points in the upper-left panel show the median squared error for the sample individual using both the CAT (light triangles) and static (dark circles) batteries of various lengths. The points on the upper-right panel show the median absolute bias, defined as $|\theta_j - \hat{\theta}_j^{(EAP)}|$. The points on the lower-left show the total absolute bias for the sample, while the lower-right panel shows the median posterior variance for the estimates.

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 <i>expected posterior variance</i> .
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.

Table 2: Item-level parameters estimated from calibration sample

Question	Difficulty	Discrimination	Question	Difficulty	Discrimination
1	-2.57	1.72	33	-0.61	0.97
2	-2.50	1.36	34	-0.54	0.71
†3	-2.34	2.12	35	-0.51	0.70
4	-2.17	0.51	36	-0.43	1.33
5	-2.17	1.77	37	-0.41	1.35
6	-2.03	1.18	38	-0.33	1.47
7	-2.00	1.00	39	-0.28	1.12
8	-1.97	1.55	40	-0.26	1.05
9	-1.89	1.23	41	-0.18	2.06
10	-1.81	1.57	42	-0.13	1.39
11	-1.62	1.84	†43	-0.07	1.65
12	-1.60	1.85	44	-0.04	1.60
†13	-1.59	1.63	45	0.12	0.92
14	-1.59	2.18	46	0.15	1.27
15	-1.57	1.22	47	0.17	1.57
16	-1.51	1.58	48	0.18	1.34
17	-1.46	1.38	†49	0.20	1.65
18	-1.42	2.03	50	0.31	1.51
19	-1.40	1.61	51	0.35	1.23
20	-1.30	1.78	52	0.39	1.44
21	-1.25	0.98	53	0.40	0.89
†22	-1.17	1.80	54	0.42	0.86
23	-1.16	0.67	55	0.43	1.05
24	-1.01	0.89	†56	0.53	1.45
25	-0.98	1.51	57	0.60	1.25
26	-0.95	0.66	58	0.60	0.92
†27	-0.85	1.71	†59	1.14	0.81
28	-0.84	1.00	60	1.34	0.66
29	-0.72	1.41	61	2.50	0.61
30	-0.71	1.95	†62	2.73	1.11
†31	-0.69	2.04	63	2.98	0.83
32	-0.64	1.63	64	4.25	0.53

†Item included in 10-item fixed scale. The model was estimated using the `ltm()` command in the `ltm` package in Rv2.15. Standard errors are suppressed for clarity. $n = 810$.

Table 3: Comparing external validity of five-item dynamic and static political knowledge batteries.

	Interest in politics		Frequently discuss		Attention to politics	
	Dyn.	Static	Dyn.	Static	Dyn.	Static
Constant	3.34 (0.11)	2.78 (0.08)	5.58 (0.23)	4.92 (0.16)	4.58 (0.21)	3.56 (0.14)
Knowledge	2.08 (0.17)	1.31 (0.13)	3.45 (0.35)	2.78 (0.25)	3.42 (0.32)	2.07 (0.23)
R^2	0.26	0.20	0.19	0.23	0.21	0.17
n	418	400	418	401	418	401

To make the coefficients comparable, the knowledge scores are re-scaled so that a one unit change represents moving from the minimum to the maximum observed value in each sample. Note that the coefficients are always substantially larger for the dynamic battery, indicating higher levels of correlation. All survey question wording and response options are shown in the online supplemental materials.