## 1 Introducing Adaptive Inventories

Technology has revolutionized how social scientists administer surveys, recruit participants, and design survey instruments. Even 20 years ago, collecting survey data was an onerous and expensive task requiring either the assistance of professional firms or hours of work by expansive research teams. Today, easy-to-use survey administration software can be paired with readily available pools of online respondents to implement and field a survey in a matter of hours.

However, in many ways, little has changed. Most surveys are entirely static. And to the extent surveys vary from person to person, the survey adapts not in *response* to inputs, but rather the instrument is pre-programmed (e.g., randomization). The textbook process remains that questions are written, evaluated, and placed on surveys. Once the instrument is designed, however, it changes little based on the input from respondents.

This is not to say that researchers have not innovated. Salganik and Levy (2015), for instance, propose a wiki survey battery where respondents can alter response options. Groves and Heeringa (2006) propose an influential responsive sampling framework where survey *mode* adapts systemically to optimize response rates while controlling costs. Moore and Moore (2013) propose a method for dynamic assignment in sequential experiments to maximize balance. However, none of these approaches is adaptive in the sense of active learning, where the goal is to alter the content of the survey interactively to maximize learning about some quantity of interest. Conceptually, the most similar work to our own is Offer-Westort, Coppock, and Green (2019), who apply a bandit approach to adapt treatment assignment propensities.

This is unfortunate for two reasons. First, the use of online surveys makes it possible to rethink how surveys work so they are responsive, less burdensome, and better suited to researcher needs. After all, wouldn't it be better if surveys could "think for themselves" a bit and adapt as the interview proceeds? Second, the active learning framework has been widely applied in computer science, engineering, and more. This research shows that algorithms designed to optimize data collection can dramatically improve estimates and reduce overhead in terms of time or money (see, e.g., Miller, Linder, and Mebane (2019) and Enamorado (2018) for recent examples).

The goal of this Element is to provide a detailed introduction to one such approach for making surveys "smarter." Specifically, we introduce adaptive inventories (AIs), a method that can help survey researchers measure important latent traits or attitudes accurately while minimizing the number of questions respondents answer.

In the following sections, we provide both a theoretical overview of the method and a suite of tools and tricks for integrating AIs into the survey process. But first, let's clarify the "problem" that needs solving. What exactly is wrong with traditional batteries?

## 1.1 Motivating Example: Measuring Political Knowledge

Imagine a researcher wants to measure respondents' levels of political knowledge. Tasks like this are common but actually represent a thorny problem. To begin, the concept of interest is latent; we cannot assess it directly. A respondent's "knowledge" is too abstract to be measured using one single question. Knowledge is not like, say, one's vote choice that respondents can plausibly report directly.

When measuring latent constructs, scholars must instead rely on survey items that relate to the concept indirectly and imperfectly. The consequence is that in order to measure the concept accurately, researchers need to ask each respondent multiple questions and aggregate their responses. Indeed, roughly speaking, the more questions we ask, the better our final measure will be. Following this logic, there is always an incentive to ask more questions about important latent constructs.

In the specific case of political knowledge, these items might ask about important political leaders, the rules of the constitutional system, and the basic contours of political debates and party competition. For example, Figure 1 shows questions adapted from the 1991 American National Election Study (ANES) Pilot, which included 20 political knowledge questions.

However, there are clear costs to batteries like this. In most situations, lengthy batteries are too expensive to administer as they take up too much valuable survey space. Moreover, answering long (and often repetitive) survey batteries is tedious for respondents. Attrition can increase (e.g., Sheatsley 1983) and answers become less informative (e.g., Herzog and Bachman 1981).

Faced with this dilemma, the standard approach for measuring concepts like political knowledge is as follows:

- First, administer a large set of items to a sample of respondents.
- Second, use this pilot data to evaluate the items and select *one subset* to include on the survey. (This second step is typically done with some measurement model, such as factor analysis.)

In the case of the political knowledge battery, this standard procedure was followed almost exactly. Using the ANES pilot data, Delli Carpini and Keeter

1. **Do you happen to know what job or political office is now held by Kamala Harris?**
2. Do you happen to know what job or political office is now held by Chuck Schumer?
3. Do you happen to know what job or political office is now held by John Roberts?
4. Do you happen to know what job or political office is now held by Vladimir Putin?
5. Do you happen to know what job or political office is now held by Boris Johnson?
6. Do you happen to know what job or political office is now held by Nancy Pelosi?
7. Do you happen to know how many times an individual can be elected President?
8. How long is the term of office for a United States Senator?
9. **How much of a majority is required for the US Senate and House to override a Presidential veto?**
10. What are the first ten amendments to the US Constitution called?
11. **Whose responsibility is it to decide if a law is constitutional or not? Is it the President, the Congress, or the Supreme Court?**
12. Whose responsibility is it to nominate judges to the Federal Courts? Is it the President, the Congress, or the Supreme Court?
13. Which of the two major parties is considered to be more conservative?
14. Do you happen to know which party has the most members in the House of Representatives in Washington?
15. **Do you happen to know which party has the most members in the US Senate?**
16. **Would you say that one of the parties is more conservative than the other at the national level? Which party is more conservative?**
17. Would you say that one of the parties supports increasing defense spending more than the other at the national level? Which party most supports increasing defense spending?
18. Would you say that one of the parties is supports increasing defense spending more than the other at the national level? Which party most supports increasing defense spending?
19. Would you say that one of the parties supports having the government provide more services at the national level? Which party most supports the government providing many more services?
20. Would you say that one of the parties supports having the government make every effort to improve the social and economic position of blacks? Which party most supports the government making every effort to improve the social and economic position of blacks?

**Figure 1** Items measuring political knowledge (bold items selected for reduced measure)

chose the five items bolded in Figure 1. These became a canonical measure of political knowledge (Delli Carpini and Keeter 1993, 1996).

So, what's the problem? In reducing these larger batteries down to a manageable size, the traditional strategy requires researchers to choose a *single* subset of items to administer to *everyone*. That is, the most knowledgeable and the least knowledgeable respondents will answer the exact same set of questions. This strategy is inefficient. When we administer the same static battery to the entire sample, it inevitably includes some items that provide little additional information about specific respondents' true latent positions.

To make this clearer, consider Question 2 (Q2) asking respondents to identify Chuck Schumer as the Majority Leader of the US Senate. Would this be a good question to add to the canonical five-item battery? The answer is, *it depends on a respondent's answers to the other items*. For some respondents, their answer to this question would be useful. When interviewing a respondent who has

already answered five other questions correctly, Q2 might represent a good test to distinguish the modestly knowledgeable from the most knowledgeable. We know that the respondent is fairly knowledgeable, but we are not sure if she can recognize more obscure political figures. In other words, based on what we know so far, she could plausibly get the question right or wrong, and we will learn something from her answer.

But imagine a respondent who has failed to identify the Vice President and cannot identify the Republican Party as being conservative. Based on what we already know, asking him about Schumer serves little purpose. The respondent will almost surely answer incorrectly. Thus, when he gets it wrong (exactly what we expected based on his previous responses), we do not really learn anything. But what if we adjust the battery based on the respondents' previous answers? Then, we could ask a question designed to distinguish the somewhat unknowledgeable from the completely ignorant, and our estimates will be more precise.

These examples illustrate the root problem with the traditional approach. By choosing the same set of questions to administer to *all* respondents, researchers are neglecting valuable information that they have already collected. Namely, they are ignoring previous responses to questions in the battery. Instead, we should use what we learn from previous responses to *customize* batteries to each respondent. And by asking better, more informative questions, we will improve our final measure.

## 1.2 AIs and Computerized Adaptive Testing

The goal of AIs is to tailor the battery for each respondent based on what is learned during the course of the survey. Rather than ask the full battery or fixed subset of questions, an algorithm picks the next question for each respondent. This amounts to an alternative procedure for developing surveys:

- First, administer a large set of items to a sample of respondents.
- Second, allow *an algorithm* to use this pilot data to select a subset of questions *for each respondent*.

When implemented correctly, the resulting latent trait estimates will be less biased, more precise (lower variance), and more efficient. Moreover, the adaptive battery takes up no additional survey time. Better measurement; same low cost!

To achieve this goal, we draw on the rich literature on computerized adaptive testing (CAT), which was originally developed in the field of educational testing (e.g., Kingsbury and Weiss 1983; Weiss 1982; Weiss and Kingsbury

1984). You are most likely to see CAT applied to educational testing programs (e.g., the Graduate Management Admissions Test), physical and mental health assessments (e.g., the National Institutes of Health Patient-Reported Outcomes Measurement Information System), or employee selection and placement (e.g., the Armed Services Vocational Aptitude Test Battery). Indeed, many readers may be intimately familiar with CAT methods as they have been used on the Graduate Record Exam (GRE).

## 1.3 A Brief Introduction to CAT

Figure 2 illustrates the basics of a CAT algorithm. Each of the four circles corresponds to a step in the procedure. These steps are as follows: estimate a respondent's position on the latent trait of interest; select the next item to administer that optimizes some objective function; administer that item and record the response; check the stopping rule(s) and either continue questioning or return the final estimate of the respondent's position (Segall 2005).

CAT needs two sources of information to complete the procedure: properties of the question items and respondents' answers as they advance through the survey. To explain, let's return to the researcher wishing to measure political knowledge. First, she needs information about the question items, such as how "difficult" the questions are. She gathers this information by pretesting

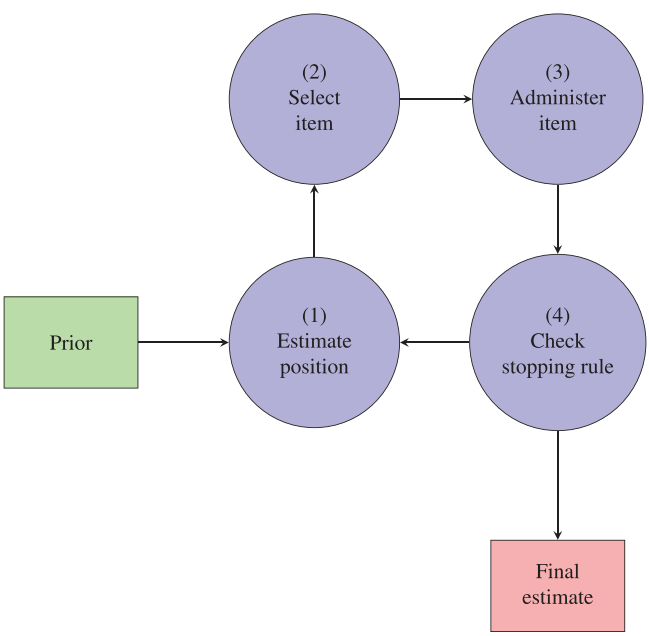


**Figure 2** Basic schematic of AIs

the battery with a calibration sample. That is, she must give the *full battery* of question items to some sample to understand which survey items are easier or harder and also which items are more reliable indicators of political knowledge. While the full battery needs to be evaluated to generate these item-level parameters, it is possible to do this effectively without giving the full battery to everyone with split sample designs.

Second, the researcher needs to generate guesses about the respondents' levels of political knowledge as they take the survey. For this, the algorithm simply needs the respondents' answers to questions in the battery so far.

Taking these pieces of information together, CAT tailors which questions are administered to each respondent. For a respondent who has already identified Nancy Pelosi (Q6 in Figure 1), it is not useful to ask her to identify Kamala Harris (Q1). Rather, the algorithm will ask the respondent to identify the Bill of Rights (Q10). How does it know to do this? CAT uses item-level information – Q1 is relatively easy and Q10 is relatively hard. Next, CAT will use the fact that the respondent answered Q6 correctly to inform our beliefs about her level of knowledge (that it is high). Combining our beliefs about the respondent and our beliefs about these questions, we can then reasonably infer that asking Q10 will be more informative than asking Q1.

## 1.4 Example: An Adaptive Measure of Political Knowledge

Of course, AIs are easier to describe than to implement. Intuitively, they are attractive because there are no additional costs in terms of survey time. But there are still a number of logistical hurdles to overcome and assumptions we must satisfy. How would this work in practice? Answering this question will be one of the primary goals of this Element. However, as a first pass, we provide a simple example focused again on measuring political knowledge.

This example will use the `catSurv` software, an open-source R package that we designed specifically for the needs of survey researchers (Montgomery and Rossiter 2017). This Element will provide detailed examples using `catSurv` to facilitate understanding of the CAT algorithm and how researchers might implement AIs in their own work. In addition to this Element, we provide extensive help files with examples in the package documentation at https://cran.r-project.org/web/packages/catSurv/catSurv.pdf.[1]

To follow along with the code using R or RStudio on your machine, you need to install and load the package using the following two lines:

[1] Users who wish to compile their own version of the package will need to take additional steps to configure their machine to correctly handle the `C++` code contained in `catSurv`. Visit https://catsurv.com for additional details.

```
install.packages("catSurv")
library(catSurv)
```

Alternatively, all code used throughout the Element is stored in a Code Ocean Capsule and can be executed online at `https://codeocean.com/capsule/2685679/tree`.

### 1.4.1 Setting Up an Adaptive Inventory

The first step is to write a set of question items that could potentially be administered to respondents. In this case, we will use the knowledge battery described by Montgomery and Cutler (2013). This consists of 64 multiple choice questions on topics similar to those in Figure 1.

The second step is to administer these questions to a calibration sample. We administered these items to a sample of about 800 respondents recruited from Amazon's Mechanical Turk service in 2012. This dataset codes answers as either correct (a value of "1") or incorrect ("0"). The data is included in the `catSurv` package and can be accessed using the following command.

```
data("polknowMT")
```

You can also view the help file for this dataset, or for any data or function in the package, by using the `?` function, (for instance, `?polknowMT`). This help file includes question wordings and response options.

With sample data in hand for the *full* battery, the next step is to calibrate the question item parameters using the `ltmCat()` function. This function creates an object of the class `Cat`. You may get a message that the measurement model is poorly estimated. This is largely a result of the quite small sample size used in this example. Section 6 provides additional guidance on appropriate sample sizes for calibration samples.

```
knowledgeCat <- ltmCat(polknowMT)
```

The `Cat-class` help file describes all of the elements of a `Cat` object. Importantly, a `Cat` object holds all the information about the question items needed for the battery to work. For instance, it contains the item parameters that indicate the difficulty of each question. You can see the first six difficulty parameters with the `@` symbol.

```
head(knowledgeCat@difficulty)
```

```
##        Q1       Q2       Q3       Q4       Q5       Q6
## 4.697249 3.887995 5.889785 1.191926 3.946480 2.535870
```

The design of `catSurv` means that all of the CAT options are chosen up front when you design the AI. For instance, we can indicate that our adaptive battery should only include two questions by setting the `LengthThreshold` slot. To do this, you alter the `CAT` object itself rather than including it as an argument when doing item selection.

```
setLengthThreshold(knowledgeCat)<-2
```

While this makes the initial setup more complex, actual administration of an AI requires little coding. Kaufman (2020), for instance, was able to incorporate an AI into an RShiny survey using only three lines of code.

### 1.4.2 AIs in Action

Now that the inventory is set up, choosing the next item to ask to a respondent is simple and fast using the `selectItem()` function.

```
selectItem(knowledgeCat)$next_item
```

```
## [1] 41
```

This function estimates the position of the respondent on the latent trait and chooses the optimal question to ask next. In this case, it indicates that item 41 should be asked.

As shown in Figure 2, the next step is to administer the question and record the response. In this case, we will imagine that the respondent answers question 41 correctly.

```
knowledgeCat<-storeAnswer(catObj=knowledgeCat, item=41, answer=1)
```

The next step is to check whether the stopping criteria was met. In this case, we specified that we only wish to ask two questions.

```
checkStopRules(knowledgeCat)
```

```
## [1] FALSE
```

Since the stopping point has not been reached (`checkStopRules()` returned a `FALSE`), the process repeats.

```
selectItem(knowledgeCat)$next_item
```

```
## [1] 49
```

```
knowledgeCat<-storeAnswer(catObj=knowledgeCat, item=49, answer=0)
checkStopRules(knowledgeCat)
```

```
## [1] TRUE
```

After administering a second item and storing the answer, the stopping rule has now been reached and the AI is finished. The final step is simply to estimate the respondent's position on the latent trait.

```
estimateTheta(catObj=knowledgeCat)
```

```
## [1] 0.0627642
```

This is the final estimate for this respondent's score on the political knowledge scale.

While AIs using `catSurv` require some work, the actual implementation requires only four functions: `selectItem`, `storeAnswer`, `checkStopRules`, and `estimateTheta`. And, as shown in Montgomery and Cutler (2013), AIs of political knowledge provide significantly improved estimates of the latent trait relative to fixed batteries of the same length.

## 1.5 Example Applications

Yes, AIs require extra work. Is it worth the trouble? Will implementing this technique actually change our substantive findings or help us to do better research? Our answer is that AIs offer a way to improve measurement of key latent variables, and this improved measurement can, in turn, improve our substantive findings. In essence, AIs can be used effectively whenever:

1. you want to measure a unidimensional[2] latent variable that will be an outcome, explanatory variable, or moderator;
2. you have space for at least three survey items for that trait;
3. the latent trait is associated with a large inventory where the number of potential items to include significantly exceeds the number of items you want to ask;
4. you are interested only in the underlying trait and not responses to individual items; and
5. you believe that the item calibration is stable from one sample to another.

These conditions apply in many settings across the social sciences whether one is trying to measure personality traits, ideology, cognitive skills, consumer sentiment, or other attributes. But some concrete examples will help set the stage. While the remainder of this Element focuses on mathematical foundations and practical considerations of AIs, we hope these examples serve as a

[2] As we discuss in Section 7, there are extensions to this framework that allow for multidimensional latent traits. However, all of the results in this Element assume a single underlying dimension.

backdrop to highlight that AIs can improve our ability to test theories and do better science. Our first example explicitly compares an adaptive and nonadaptive battery, while our second shows how AIs can be used on large national surveys. These examples are necessarily brief. For a fuller explanation of this data source and further results, see Montgomery and Rossiter (2020).

### *1.5.1 Adaptive Right-Wing Authoritarianism*

The right-wing authoritarianism (RWA) trait captures individuals' differences regarding submission to authorities, tolerance of outgroups, and conventionalism (Altemeyer 1988). Political scientists have shown RWA explains reactions to ethno-racial diversity (Velez and Lavine 2017), the support of war (Hetherington and Suhay 2011), and even partisanship and increasing polarization (Hetherington and Weiler 2009).

While a widely used battery for measuring RWA has 30 items (Altemeyer 1988), the American National Election Pilot Study has asked a fixed subset of five of these items to reduce the inventory's length. To assess the implications of measuring RWA with an adaptive inventory, we conducted a study where we randomly assigned 1,335 participants to take either a *fixed* or *adaptive* five-item inventory. In either case, after answering five items, respondents completed the remaining battery items in a random order.

Figure 3 shows the distribution of latent trait estimates when measuring RWA using each of these five-item inventories (dashed lines) compared against RWA as estimated using the same respondents' full 30-item answer profiles (solid lines). When comparing the dashed and solid lines of each plot, we see that the adaptive inventory does a better job than the fixed inventory of estimating extreme positions on the latent scale, especially for low values of RWA.

Does that alter our inferences? To answer this, Figure 4 shows the difference in our understanding of how RWA relates to modern racism (Sidanius et al. 2004). Specifically, we estimated separate regressions using RWA estimates for respondents randomly assigned to the fixed inventories and the AIs.[3] Further, we used respondents' RWA estimates from their full response profiles to estimate "true" values of the RWA regression coefficients in each model. The lines in Figure 4 shows the regression coefficients (and 95 percent confidence intervals) estimated using RWA estimates from each five-item inventory and from the full 30-item inventory. In all, the censoring that results from using the fixed five-item inventory, shown in Figure 3, leads to an upward bias in the

[3] We control for race, gender, and level of education in each model. We used the same measurement model fit to a separate wave to estimate respondents' positions and ensure that all measures are on the same latent scale.