Reconsidering the Measurement of Political Knowledge *

Matthew S. Levendusky and Simon D. Jackman
Stanford University

December 15, 2003

*The authors thank Doug Rivers and Mo Fiorina for helpful comments.

1 Introduction

A basic feature of representative democracy is that citizens use elections as a means of controlling elected officials. A critical maintained hypotheis implicit in that idea is that citizens know enough about politics and public affairs to hold politicians accountable for their behavior in office. If citizens cannot pass even this basic test, then representative government seems an unsure proposition. In short, understanding how elections function in a democracy rests upon being able to evaluate what citizens know about politics and government. If we ever want to be able to address such questions, we need a high quality measure of citizen's political knowledge.

However, despite its importance, our measures of political sophistication are not as accurate as one would like. Traditionally, most scholars choose one of two routes: either relying on a single item, most typically the interviewer's subjective assessment of the respondent's level of political sophistication (Bartels 1996), or alternatively, constructing a knowledge scale built from several factual items (Zaller 1992, Mondak 1999, Mondak 2001). Yet as discussed below, both methods have associated drawbacks, so the question still remains: how we can accurately measure political sophistication? I outline here a methodology, item response modeling, that gives us a new measure of political knowledge. Here, because of the complexity of the model, I use a set of Bayesian tools and techniques to identify and estimate the model. Using this strategy, my analysis suggests several important revisions to how we measure political information. First, and perhaps most importantly, my results suggests that although many scholars rely on the interviewer's subjective assessment to operationalize political sophistication, there is considerable heterogeneity in how interviewer's use that scale. That is, what constitutes a "very high" level of political information for one judge may only be "average" for another. Analyses that rest on the assumption that all interviewers use the scale the same way risk drawing incorrect substantive conclusions about the role of political sophistication in explaining political behavior.

Further, not only do we measure each individual's level of political sophistication, but we also get to learn alot about the items themselves. The Bayesian techniques here serve to aid in the estimation of and inference from that model. From the output, we get estimates our uncertainty on all of the parameters in the model, which allows us to examine the quality of each item as a measure of political knowledge. That is, we can determine which items more fully tap political knowledge, which ones allow us to discriminate between well-informed and less well-informed respondents, and so forth. We not only get an accurate and valid measurement, we also learn quite a bit about our instrument in the process.

2 A Brief Summary of Political Sophistication

But what exactly is political sophistication? Before considering the measurement issues, we should discuss from a theoretical standpoint what is is we are trying to measure. Converse (1964) argues that political sophistication is a cognitive property, a belief system "where the elements are bound together by some form of contstraint or functional interdependence" (207). Political sophistication is the ability to think carefully and abstractly about politics; understanding how disparate issue positions (i.e., lower taxes and support for abortion rights) fit together in a coherent framework. Luskin (1987) calls sophistication "congnitive complexity about politics," a dense and rich set of beliefs and attitudes (861). While the unsophisticated understand politics is crude terms (i.e., Campbell et al.'s (1980) "nature of the times" voter), sophisticates rely on a more abstract and complex conceptualization of the political world.

And the way in which citizens conceptualize the world has serious consequences for politics and political science. A long tradition in democratic theory prescribes an informed citizenry as a crucial element to democratic politics (see, e.g., Dahl (1979)). Yet even this proposition is not controversy free, given that others (i.e., Neuman (1986)) argue that low levels of citizen information are acceptable. Defending the proposition that citizens need a basic amount of political information to be "good" citizens, Delli-Carpini and Keeter (1996) aruge:

> Better-informed citizens are significantly more likely to participate in politics, are better able to discern their "self-interest properly understood" (de Toqueville [1850]1969, 525), are better able to connect their enlightened self-interest to specific opinions about the political world, are more likely to hold opinions that are internally consistent and stable over time, and are more likely to connect their opinions to their political participation in meaningful, rational ways. More informed citizens are also more likely to demonstrate other requisites of good citizenship, such as political tolerance. In short, informed citizens are better citizens in a number of ways consistent with normative and pragmatic notions of what constitutes good citizenship (19).

If we take evaluating the premises citizen government seriously, then we need a way of assessing what citizens know.

A long lineage in political science, beginning with Converse (1964), demonstrates that the politically informed, the politically sophisticated, are simply different from the rest of the electorate. Zaller (1992) argues (and shows empirically) repeatedly how sophistication conditions political atttidues, behaviors, and opinions. Achen (1992), reviewing past efforts to understand the development of individual's party ID, argues that "voter information is theoretically critical, and without particular nonlinear controls for it, no statistical estimates of vote or party ID equations should be trusted" (198). In short, for understanding political behavior, political information has become one of the most essential variables in the political scientists arsenel. A critical concept requires a well-developed measure.

3 Measuring Political Information: Previous Methods and Item Response Modeling

Based on the definitions above that relate political sophistication to cognitive complexity, the ideal measures of political sophistication would be to examine people's thought processes. However, such processes are unobservable, so we fall back on measuring the *application* of this congnitive complexity. While this might seem like a stretch, Luskin (1987) argues that it is not such a stretch: “A politically sophisticated person may or may not believe that free enterprise is a good thing, that government spending is the root of inflation, or that effective control of handguns will lower the murder rate. These points are debatable. But he or she is unlikely to believe that Ronald Regan is now a Democrat or that the Democrats currently hold a majority of seats in the U.S. Senate” (881). In short, there is a strong correlation between someone’s underlying latent ability (their political sophistication) and their ability to apply that trait to factual items (their political knowledge). As long as we use items that actually tap relevant facts of the political landscape, the resulting measures should be valid.

Typically, most scholars measure this political knowledge in two principal ways.First, some studies simply use an individual item, most typically the NES item asking for the interviewer’s subjective assessment of the respondent’s level of knowledge about politics (see, for example, Bartels (1996)). The second common measurement strategy is to take a series of factual questions from the NES (i.e., which party controlled the Senate before the November elections?) and build them into a knowledge scale, where respondents are ranked by how many questions they correctly answer (Zaller 1992, Mondak 2001, Gomez & Wilson 2001). For example, Delli-Carpini and Keeter (1996, 304-5) defend a five-item scale built from the NES asking respondents questions about the structure of American government, as well as identifications of political figures and placement of the political parties on the ideological scale. While others have used alternative measures (Neuman 1986), these two methods are by far the most popular.

Despite their popularity, both methods have drawbacks. First, and most importantly, a single item cannot measure the complexities of political sophistication, and inevitably measures based solely on one item will fall prey to being extremely imprecise. And although the interviewer’s subjective assessment of the respondent’s level of political information (the most popular choice for a single-item measure) taps some aspect of political sophistication, it does not fully capture the complexities of such a nuanced concept. Further, there is an untested (but critical) assumption made every time this rating is used as a measure of political sophistication—that each interviewer uses the scale the same way. Scholars assume that "very high" from interviewer A is the same as "very high" from interviewer B. However, its quite likely that some interviewers have a much stricter definition of what constitutes well-informed than others, which can introduce significant bias into the model.

And when constructing a scale, assigning a weight to each individual item becomes problematic: should knowing the chief justice of the Supreme Court count more or less than knowing which party controlled the House of Representatives before the most recent election? Most people get around this problem by simply assigning each question equal weight. While this means the researcher does not need to assign arbitrary weights to the questions, it seems unrealistic that some questions don't tap an individual's political sophistication better than other items[1].

To avoid the problem of relying on a single measure, I combine the strength of the interviewer rating measure (arguably the strongest single-item indicator, see Zaller (1986)) and a series of factual knowledge and placement items to measure respondents level of political sophistication. In terms of actually selecting items, Zaller(1992, 333-344) gives an extensive discussion of various potential items, and concludes that general placement items (i.e., placing the Democratic and Republican Parties on a 7-point ideological scale), factual knowledge questions (i.e., "what office does Dan Quayle currently hold?") and the interviewer's rating of the respondent's level of political knowledge serve as superior measures of political knowledge. So in terms of the actual items selected, I follow Zaller's prescription and use the interviewer ratings, the placement items and the factual knowledge items to measure political information [2].

To actually estimate the model, and avoid simplying arbitrarily assigning weights to

[1] For a more detailed discussion about the problems associated with simply using an additive score of the number of items correct, see (Bullock 2002).

[2] As Zaller (1992, 338) argues, its not always unambiguous which candidate or party is to the left on some issue (i.e., women's rights in 1976) so I rely on just using the general party and candidate liberal/conservative placements, thereby avoiding these potentially thorny issues

items, I use a technique popular in the educantional testing literature, item response modeling. Essentially, item response theory (IRT) allows us to measure an unobserved latent trait of an individual (here, political sophistication) by taking multiple observed items (here, the individual's responses to the NES questions) and using them to identify the individual's level of the underlying trait. IRT modeling allows me to avoid the problems associated with other measures of political knowledge mentioned above. As it relies on more than one measure of political sophistication, I avoid the pitfalls of relying on only a single measure. Further, IRT allows me to estimate how fully each question taps political sophistication, meaning that I can empirically estimate which items are better/worse measures of political sophistication. In short, I can weight individual items differently yet do it in a systematic fashion as best suggested by the data.

Another alternative to IRT that would also seem to avoid the above pitfalls would be factor analysis, long used a measure of latent concepts (Jackman 1998). However, factor analysis contains two shortcomings relative to IRT. First, as traditionally analyzed, factor analysis cannot give us any measure of uncertainty for its estimates of the latent traits. To draw an analogy, it would be as if we ran a regression and only got point estimates and no standard errors. When we take results from a factor analytic model and then stick them into a regression equation, its as if we're saying we have measured the concept with no error, which clearly is not the case. In contrast to this, IRT gives us an explicit measure of the uncertainty of our estimates (a standard error), and we can take this uncertainty and directly account for it in our results or a second-stage model.

Second, to estimate categorical data (as we have in this case when people answer a factual question correctly or incorrectly), factor analysis must rely on asymptotic theory (as it was originally designed for continuous measures). In contrast, IRT does not rely on asymptotic theory, it can handle continuous, discrete or categorical data with equal ease. While this is a minor point, it is generally preferable to not depend upon asymptotic properties of estimators when possible. Now that we've discussed the data and measures, let us turn to the model itself.

3.1 The Model

To actually model a given individual's level of political sophistication, I estimate a two-parameter item response model. The data used to estimate the model comes from a series of factual questions asked of the NES respondents about American government and the interviewer's subjective assessment of the respondent's level of political information from the quadrennial Presidential election year surveys between 1980 and 2000, giving me a total of 6 surveys, 56 binary items, and over 11,000 total respondents [3]. In general, the model takes the form

$$Pr(y_{ij} = 1|\theta_i) = F(a_j\theta_i - b_j)$$

Where j indexes items, and i indexes respondents. Here, we're modeling the probability that a given individual i answers an item j correctly conditional on their level of the latent trait θ_i (political sophistication) as a function of the individual's latent trait θ_i and two parameters, a_j and b_j (Johnson & Albert 1999, 184). The function $F()$ in this case is the logistic CDF, as the items used are binary (i.e., the individual either gets the item correct or incorrect), hence a discrete choice method is appropriate.

In the above model, a_j (the slope) is termed the *item discrimination* parameter, and b_j (the intercept) the *item difficulty* parameter. The discrimination parameter measures how well the item distinguishes between individuals possessing differing amounts of the latent trait. An item with high discrimination distinguishes students with low amounts of the latent trait from those with middling and those with high levels. The difficulty parameter, on the other hand, measure the difficulty of a given item. The interpretation of this parameter is straightforward: a more difficult item is harder for individuals at all levels of political information than an easier item, and it is easier for people with higher levels of the latent trait to answer it than those with lower levels (Johnson & Albert 1999, 184-5). Together, these two parameters describe how well an item measures the latent

[3] The full list of questions used is available in the Appendix.

trait; how well it distinguishes between people with different levels of the latent trait, and how hard or easy it is to answer.

Note that estimation of this model will be complicated in a traditional frequentist framework. The model would require estimating 112 item parameter (a discrimination and a difficulty parameter for each of 56 binary items), the level of political knowledge for each of our 11,917 respondents and a number of other parameters: in all, a model with over 12,000 parameters. The maximum likelihood estimates would be the global maximum of the joint likelihood function, a 12,000 dimension hypersurface; finding such a region in a high-dimensional space is a near Herculian task. First, the parameters in the item-response model are not uniquely identifiable (that is, the parameters are only identifiable up to a linear transformation), and hence, it is not possible to derive unique numerical estimates (although parameters restrictions to help identification do exist). Further, as in many non-standard maximum likelihood estimation scenarios, there is the possibilty that the likelihood function has multiple nodes, and the MLEs found via numerical optimization may not correspond to the global maximum of the likelihood function. Often, to minimize these difficulities, scholars turn to an alternative method, such as marginal maximum likelihood, which circumvents the problem of maximum the full joint likelihood and instead maximizes the marginal likelihood function, see (Bock & Aitken 1981). In general, classical estimation of an IRT model, particularly one as complicated as this, is extremely onerous (Johnson & Albert 1999, 189-191).

However, a simpler, easier, and more direct approach is to adopt a Bayesian perspective for the purposes of estimation and inference. One major advantage of the Bayesian framework is the ease of identification. While this is a major hurdle to be overcome in the frequentist framework, a Bayesian setup circumvents this problem by specifying a prior distribution for each of the parameters of interest. These prior distributions help us identify the model, but we don't want them to do too much of the work for us (that is, we still want the data to speak to the true parameter values), so we specify vague prior distributions for the model parameters (Johnson & Albert 1999, 192-3). For the issues related to actually maximizing the joint likelihood, the MCMC algorithm used simplify the computational issues associated with estimating the model (see the next section on the Gibbs sampler). Further, as a tool for *inference*, MCMC has a tremendous advantage over classical methods. Since inference requires uncertainty estimates on the parameters (i.e., standard errors), we'd need to invert the $p \times p$ Hessian matrix in the classical setup, where p is the number of parameters. Here, that would require inverting a 12,000 by 12,000 matrix, which would require enormous computing power (and may not even by feasible in practice given the limits of conventional computing power). On the other hand, the Bayesian setup used here builds up posterior distributions on each of the parameters of interest, meaning that we don't need to invert a gargantuan matrix, but instead can get the standard errors from these distributions, greatly simplying the process of inference.

To identify the model, I specify prior distributions on the model parameters of interest and I make one further restriction. Because I have data drawn from 6 studies through time, I need to normalize one item appearing in each study to have a discrimination parameter to 1. Because it appears in every study, I chose the 5-category interviewer rating. This serves two functions. First, this normalization makes the results scale across years. That is, we know the scale in each year is the same, which means that the results are comparable across years. Otherwise, when we saw higher scores in some years than others, we would not know if the individuals had gotten smarter or if the scale had just been arbitrarily re-scaled with no change in the underlying latent trait. This normalization resolves this issue. Second, this constraint also gives the underlying latent scale a unique identification in each year. That is, in each year, higher ratings by the interviewers will generate higher scores on the latent trait, *ceteris paribus.*

To complete the identification of the model, let me know specify the prior distributions on the parameters of interest. To begin, consider the item discrimination and difficulty parameters, which are assumed to have multivariate normal prior densities such that $a_j, b_j \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}\right)$, where μ is simply a vector of mean values (all 0's) and σ is the variance-covariance matrix of the a_j and b_j. We don't asume a particularly informative prior distribution for the item discrimination and difficulty parameters, precisely because we want the data to speak to the true values of the parameters.

However, one of the items used to measure political information is the 5-category interviewer assessment of a respondent's level of political information. For this item, I fit a two-parameter item response model with a ordinal logit link function. Here, an ordinal link function is appropriate because the interviewers rank people into one of 5 categories based on their assessment of the respondent's level of political knowledge. The interviewer can rate the respondent's level of political knowledge as "Very High,", "Fairly High", "Average", "Fairly Low", and "Very Low", where the ordering is obviously high to low. As I mentioned above, one of the untested assumptions underlying much research using political information is that all the interviewers use the scale the same way; that is, that a rating of "very high" is the same for one judge as it is for another. To explicitly test this hypothesis, I include a random effects term for each judge. That is, the ordinal logit link is as follows: $\mu_i^{(p)} = \theta_i + \eta_p$, where $\theta_i \overset{iid}{\sim} N(0, 1)$. Let Z_i be the rating judge p gives to individual i. Then:

$$\begin{aligned}
Pr(Z_i = \text{``Very Low''}) &= F(\kappa_1 - \mu_i^{(p)}) \\
Pr(Z_i = \text{``Fairly Low''}) &= F(\kappa_2 - \mu_i^{(p)}) - F(\kappa_2 - \mu_i^{(p)}) \\
&\vdots \\
Pr(Z_i = \text{``Very High''}) &= 1 - F(\kappa_4 - \mu_i^{(p)})
\end{aligned}$$

Where p indexes interviewers, i indexes individuals, and F is the logistic CDF. Note here the implicit restriction of the discrimination parameter (the parameter on θ) to 1, for the reasons related to identification discussed above. Also note that I assume the distribution of the latent traits is a fairly vague $N(0,1)$ distribution, again because I want the data to speak to each individual's level of the latent trait.

Further, the η_p terms here are the random effects terms for each judge. The model estimates a different set of thresholds per year (the κ_t terms) and assumes that each judge uses these thresholds, but then allows for each judge to shift these cutpoints by differing amounts(the η_p terms). That is, I allow for the possibility that each each judge may have a higher/lower threshold for a given category than his/her peers. This allows me to test explicitly the idea that all judges are using the scale the same way. As an added test, I assume the distribution for η: $\eta_p \sim N(0, \tau^2)$, where $\tau^{-2} \sim \Gamma(0.01, 0.01)$.

That is, I assume the η terms have a 0 mean but have some precision τ^2, where τ^{-2} itself comes from a vague Gamma distribution. Note that if there is no dispersion, $\tau = 0$ and there is no variance in how raters use the scale. However, if $\tau \neq 0$, then there is evidence that judges use the scale differently. Examining η and τ will allow me a rich set of tests of the previously untested assumptions that all judges use the scale the same way.

Finally, to complete the specification of the model, the threshold parameters in our ordinal model, the κ terms, are uniformly such that $\kappa_1 \sim U(-10, \kappa_2), \kappa_2 \sim U(\kappa_1, \kappa_3), \kappa_3 \sim U(\kappa_2, \kappa_4), \kappa_4 \sim U(\kappa_3, 10)$. Here, the structure of these conditional distributions helps to ensure that the thresholds are both properly ordered ($\kappa_1 < \kappa_2 < \ldots$) but also that the joint distributions of the κ terms is proper. Now that we've discussed the various aspects of this model, let us turn to a discussion of the results.

4 Estimation

To actually estimate the model parameters, I use WinBUGS [4], a free program written for Bayesian estimation and inference via Markov-Chain Monte Carlo (MCMC) algorithms. More formally, I used the Gibbs sampling alogrithm to estimate the model, which condi-

[4] WinBUGS is freely distributed over the internet at `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`

tional sampling to learn about the parameter values based on the other parameters values (Jackman 1999).

After 1000 burn-in iterations to ensure that the MCMC algorithm has properly moved away from its starting values, I allowed the model to run for 10,000 iterations thinned by 20, resulting in 500 observations[5]. To assess convergence of the model, I visually inspected the trace and autocorrelation plots from the thinned MCMC output ; the model appeared to be visiting locations in the parameter space with probability porportional to their posterior probability. The interested reader is referred to the appendix, where sample trace plots are given for the reader's inspection[6].

5 Results

Given the output from the Gibbs sampler,we can summarize what we've learned about political information from the data. The first item of interest to consider are the item discrimination and difficulty parameters. Since the model output has 56 items, I have 112 total item parameters. Rather than simply presenting them as a table [7], I present several graphical summaries of the material. The first such summary is a scatterplot of the difficulty vs. the discrimination parameters, in figure 1 below.

[Figure 1 about here.]

The first thing to notice about the plot is that all of the items positive discriminate on the latent political knowledge scale. Further, none of the discrimination parameters has a 95% confidence interval that overlaps zero[8]. This indicates that all of the items contribute to our understanding of political knowledge, which is reassuring since all are frequently used as measures of political knowledge. Further, note that we have a cluster with

[5] that is, the program only saves every 20^{th} iteration, thereby reducing the autocorrelation between iterations in the data used for secondary analysis. The results that follow are based on this 500 iteration sample.

[6] For more on the Gibbs sampler generally, and some of the difficulties associated with convergence, see (Jackman 2000).

[7] available from the author upon request.

[8] for the actual parameter estimates and standard errors, see the appendix

high discrimination and positive difficulty (discrimination greater than 1.5 and difficulty greater than 0)[9], which is a desirable characteristic of an item—the item tells us something about where someone lies on the underlying latent scale. However, notice the empty block in the upper right-hand side of the graph (difficulty parameter between 1.5 and 4, and discrimination between 2.5 and 3). In terms of measuring political information across the full range of values (from low to high), we would like to see some items in this range with moderate difficulty and high discrimination.

Further, the items marked on the plot as potential outlyers are not particularly severe. These items all have large difficulty parameters: two in the positive direction (1992 Democratic Party placement and Control of the House in 2000) and one in the negative direction (1996 ID of Al Gore). For Democratic Party placement and Control of the House, I can only speculate why these items appeared to be so difficult: for Control of the House, its likely not low visibility but rather the fact that control was so close that people may have become confused as to who was currently in control; for the Democratic Party placement, Clinton's candidacy may have been the source of the issue. For Gore, I likewise have no hard evidence, but a similiar process could be at work: people simply heard so much about the Vice President that answering questions about him would be quite simple. The very low difficulty parameters for the Gore item (even given its moderate discrimination value)makes it a less useful item for measuering political knowledge.

Note also that we have *prima facie* evidence that constructing scales where we simply assign equal weight to each item (as people do when constructing simple additive indicies) is a flawed practice. Figure 1 clearly shows that the discrimination parameters vary considerably (even if many are quite large, suggest a sizeable ability to distinguish between better and worse informed respondents), and hence simply summing correct answers into an index introduces considerable measurement error (Bullock 2002). This approach allows us to systematically decide which items are better at discriminating between people of differing levels of political information.

[9] because of the scaling in our model, a *positive* difficulty parameter indicates a harder item, a negative difficulty parameter indicates an easier item.

However, its difficult to tell much from the plot with all 56 items included. To help get a sense of how individual items are performing over time, I made the same plot for each item that was asked in three or more NES surveys. This allows me to see if the items have similar results over time.

[Figure 2 about here.]

Looking at the top two plots, placement of the parties on the liberal-conservative scale, its interesting to note how similar they are. This tends to be an easy question, the difficulty parameter for either party rarely rises above 0, with the odd exception of 1992 for the Democrats. However, notice that for both parties 1996 and 1984 are years with high discrimination parameters, suggesting that in both years, it may not be particularly difficult for people to identify the ideological location of the parties, yet ability to do so in those years is quite informative about where a person sits on the latent scale. Taken as a whole, the variability in the discrimination parameters for these items suggest that they vary quite a bit from year to year in their ability to distinguish respondents on the latent scale, but in some of these years this knowledge gives us a lot of purchase as to someone's ideological location.

The plot for the control of the House is interesting as well. In most years this is an average difficulty item (i.e., a difficulty parameter of approximately 0, with the exception of 2000), and also has fairly good discrimination (with the exception of 1996, the first NES survey after the Republican takeover). This item seems to effectively gauge the respondents level of political information rather well, as does the corresponding item for control of the Senate.

The rest of the plots tell similar stories. Interestingly enough, in most years where its asked, its easier to identify the leader of Great Britain and the USSR/Russia than it is to identify the Speaker of the House or the Chief Justice of the Supreme Court for the average respondent, but perhaps this should not be so surprising given the additional coverage the international leaders (especially leaders of those two nations) are given relative most domestic politicians not in the White House.

Further, we can also look at the item response curves for several items. The item response curves trace out the probability that an individual with a given level of the latent trait will correctly answer a question. That is, how does the probability of a correct response vary as a function of the respondent's level of political knowledge? Figure 3 plots out the item response curve for the identification of Jim Wright in 1988, with the same curve for the 1980 liberal/conservative placement of Jimmy Carter overlayed.

[Figure 3 about here.]

The plot for Jim Wright(the solid curve in figure 3) shows a high discrimination/high difficulty item. Note that for this particular item, an individual with average levels of the latent trait ("0" on this graph) has essentially no chance of correctly identifying Wright. Indeed, only those with high levels of the latent trait have a great than 50% chance of correctly identifying Wright. This suggests that if someone correctly identified Wright, they were quite likely to have been from the top stratum of political knowledge in the sample. In contrast, for an item with low difficulty and low discrimination—the ideological placement of Jimmy Carter (the dashed curve in the graph)—even those with the lowest levels of political knowledge in the sample have a better-than-average chance of correctly placing Carter. Indeed, even those with the median level of knowledge can correctly place Carter on a left/right scale with a probability approaching 1. Knowing that an individual correctly placed Carter on the left/right scale tells us much less about their level of political information. This sort of question only allows us to distinguish those at the lower levels of the latent trait, because those are the only respondents who get that question incorrect.

5.1 The Interviewer Effects

But in the Bayesian setup described above, we not only learned about the items themselves, we also learned about the interviewers. In particular, we included a random effects term to the model (the η terms)to see if different interviewers were ranking individuals on the scale measuring respondent's levels of political information differently. Recall that we assumed $\eta_p \sim N(0, \tau^2)$, where $\tau^{-2} \sim \Gamma(0.01, 0.01)$. From our MCMC output, we get 2 quantities of interest, then: η_p for each interviewer and τ for the whole sample. First, consider τ^2, which measures how dispersed the interviewer effects are (indeed, it is the variance of the distribution of interviewer effects). τ (the standard deviation of the distribution giving rise to the interviewer effects) from our model is approximately 0.77, indicating that there is some dispersion in the interviewer effects, giving us some evidence that different interviewers indeed utilize the scale differently.

Now consider the η terms themselves. If the η terms are all 0, then each judge is using the scale the same way. However, non-zero η terms indicates that a given respondent would be classified differently by two different judges. The η parameters model the possibility that getting a given interviewer means a respondent is scored higher/lower on the scale than those with the same response profile (that is, the same level of the latent trait)[10]. That is, η_p measures whether or not interviewer p systematically rate respondents higher/lower than his/her colleagues. Figure 4 shows a histogram of the posterior means of the 751 η_p terms.

[Figure 4 about here.]

As the graph indicates, the η terms are definitely *not* all 0. In fact, there appears to quite a large number with a large random effect term, indicating that some judges are using the scale quite differently from other judges. Since we have a posterior distribution for each interviewer's η_p term, we can look at each interviewer's 95% confidence interval and see if it overlaps 0. If the individual's confidence interval overlaps 0, then we cannot reject the hypothesis that that individual is using the scale the same way as his/her colleagues. However, if an individual's confidence interval does *not* overlap 0, then we can say that getting such a interviewer means an individual is likely to be scored higher/lower than those with the same response profiles were scored. Table 1 shows the breakdown.

[10] Recalling that respondents are scored by the interviewer as to their overall level of political knowledge.

[Table 1 about here.]

As the table clearly illustrates, overall approximately 19.5% of judges have confidence intervals that do not overlap 0. And these interviewers are not all clustered into 1 or 2 years, but rather in each year, there are 18-26 interviewers who use the scale quite differently than their peers. Getting one of these interviewers means that a given individual will be systematically scored higher or lower than they would be scored by a different interviewer.

Of course, one could speculate that this doesn't have much of an impact—a few people appear to have higher/lower levels of political information than they otherwise would have, but this isn't particularly pernicious. However, this view would be mistaken, often with dramatic consequences. To show how dramatic these effects can be, consider interviewers rating how informed respondents are about politics on the 1-5 (low to high) scale used in the NES. Have a randomly selected interviewer rate a respondent with a given level of the latent trait under two scenarios. In one scenario, ignore the uncertainty associated with the interviewer-to-interviewer variation in mapping from the latent scale to the response categories, i.e., consider the "average interviewer." In the other, take those random effect terms into account. Given these two scenarios, I consider ranking people with latent scores of -3 to 3 (low to high levels of the latent trait) by increments of 1. I repeat this simulation for 500 iterations and report the results in table 2.

[Table 2 about here.]

As table 2 shows, when we ignore the random effects terms, rating respondents in fairly straightforward: using the average interviewer, the individual is scored the same way in nearly every trial. However, when we take the random effects terms into account, the results are much more muddled. While its fairly easy to rank those with high levels of political knowledge correctly, its much harder to rank those with low to middling levels of political information (and there are many more people in the population with low to middling levels of political information in the population than there are people with high levels of political information). When an interviewer ranks someone with an average level of political information ("0" in our simulation above), taking the cross-interviewer variation into account, that individual's ranking changes nearly 31% of the time (versus not taking the random effects terms into account). The results are similar for those with similar levels of political information. The scale for the interviewer ratings contains a good deal of pure randomness and hence is an inferior measure of political knowledge when used without additional information about the respondent's level of political information. Conclusions are jeopardized by this hefty measurement error. Although it is a convenient measure of political information, it is time for scholars of political behavior to take seriously the measurement of political knowledge and use IRT estimated using Bayesian methods that allow us to richly capture the full complexity of political knowledge.

6 Application

INCLUDE AN APPLICATION OF THE METHOD.

7 Conclusion

The implications from this study are quite clear. First, scholars should stop using the interviewer ratings as the single measure of political information. Although is it a simple measure that appears to have a lot of face validity, the above analysis illustrates how it introduces measurement error into our studies of political knowledge. This is not to say the interviewer ratings have no place in our measurement strategy, however: they can contribute to a Bayesian approach that allows us to use multiple items to more specifically measure each individual's level of political information. However, they should not substitute for a richer framework.

Further, the Bayesian techniques used here also illustrated how we could directly examine a variety of quantities of interest directly from the output of the model. In contrast to other methods, even those that rely on multi-item indices, only the Bayesian framework allows us to get the rich picture of not only each individual's level of political information but also of the items themselves. This method allowed us to assess how well each individual item was measuring political knowledge (and how that measurement was changing over time). We get a better sense of the quantity of interest, and we further learn about the tools we use to measure it. The Bayesian methodology allows to gather a rich and nuanced portrait of our key concept. Given the centrality of political knowledge in our theories of political behavior, it is time we gave it the high-quality measure it deserves.

References

Achen, Christopher. 1992. “Social Psychology, Demographic Variables and Linear Regression: Breaking the Iron Triangle in Voting Research.” *Political Behavior* 14:195–211.

Bartels, Larry. 1996. “Uniformed Votes: Information Effects in Presidential Elections.” *American Journal of Political Science* 40:194–230.

Bock, R.D. & M. Aitken. 1981. “Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm.” *Psychometrika* 46:443–459.

Bullock, John. 2002. “Rasch Assumptions: Problems in the Measurement of Political Knowledge.” Stanford University. Manuscript.

Campbell, Angus, Phillip Converse, Warren Miller & Donald Stokes. 1980. *The American Voter*. Chicago: University of Chicago Press, Midway Reprints.

Converse, Phillip. 1964. ”The Nature of Belief Systems in Mass Publics”. In *Ideology and Discontent*, ed. David E. Apter. New York: The Free Press pp. 206–261.

Dahl, Robert A. 1979. Procedural Democracy. In *Philosophy, Politics, and Society*, ed. P. Laslett & J. Fishkin. New Haven: Yale University PRess.

de Toqueville, Alexis. [1850]1969. *Democracy in America*. Garden City, N.Y.: Doubleday.

Delli-Carpini, Michael X. & Scott Ketter. 1996. *What Americans Know about Politics and Why it Matters*. New Haven: Yale University Press.

Gomez, Brad T. & J. Matthew Wilson. 2001. “Political Sophistication and Economic Voting in the American Electorate: A Theory of Heterogeneous Attribution.” *American Journal of Political Science* 45:899–914.

Jackman, Simon. 1998. “Pauline Hanson, the Mainstream, and Political Elites: The Place of Race in Australian Political Ideology.” *Australian Journal of Political Science* 33:167–186.

Jackman, Simon D. 1999. "Bayesian Modeling in the Social Sciences: An Introduction to Markov-Chain Mote Carlo." available online at `http://jackman.stanford.edu/MCMC`.

Jackman, Simon D. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44:375–404.

Johnson, Valen E. & James H. Albert. 1999. *Ordinal Data Modeling.* New York: Springer.

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

Mondak, Jeffrey J. 1999. "Reconsidering the Measurement of Political Knowledge." *Political Analysis* 8:57–82.

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

Neuman, W. Russell. 1986. *The Paradox of Mass Politics: Knowledge and Opinion in the American Electorate.* Cambridge: Harvard University Press.

Zaller, John. 1986. "Analysis of Information Items in the 1985 Pilot Study." Report to the NES Board of Overseers. Center for Political Studies, University of Michigan.

Zaller, John. 1992. *The Nature and Origin of Mass Opinion.* New York: Cambridge University Press.

8 Appendix 1: Trace Plots of Parameter Values

This section contains trace plots for two representative parameters. The trace plots shows the value obtained at each iteration of the Gibbs sampler. One represents a parameter which converged quickly and settles down near the start of the iterations, another which is slightly more problematic and takes longer to feel confident that the model has converged. When the model has converged, we see movement in the parameter values, but only movement due to the underlying sochastic nature of the Gibbs sampler (that is, the movement that remains can be characterized as a random walk around some mean). Figure 7 presents these two trace plots.

[Figure 5 about here.]

By the end of the model's 10,000 iterations (thinned by 20 to result in the 1,000 iterations displayed here), it appears that all of the model parameters have converged.

9 Appendix 2: Questions Used

By year, here are the questions used from the NES in order to judge a respondent's political knowledge. After each item, the years for which the questions were asked are noted in square brackets.

Ideological Placements:
Q: Where would you place the Republican Party on the liberal/conservative scale[11]? [1980-2000]

Q: Where would you place the Democratic Party on the liberal/conservative scale? [1980-2000]

Q: Where would you place the Republican Party Presidential Candidate[12] on the liberal/conservative scale? [1980-2000]

Q: Where would you place the Democratic Party Presidential Candidate on the liberal/conservative scale? [1980-1988, 1996-2000]

Interviewer Ratings:
Q: Respondent's general level of information about politics and public affairs seemed: (Scored Very High to Very Low) [1980-2000]

Control of House and Senate:
Q: Do you happen to know which party had the most members in the House of Representatives before the election (this/last) month? [1980-2000]

Q: Do you happen to know which party had the most members in the Senate before the election (this/last) month? [1984-2000]

Identifications:
For the identification questions, the general format is as follows:
Q: The first (next) name is X. Do you happen to know what job or political office s/he now holds?

Respondents were asked to identify the following:
Ted Kennedy [1988]
George Schultz [1988]
William Rehnquist [1988-2000]
Mikhial Gorbachev [1988-1992]
Margaret Thatcher [1988-1992]
Yasser Arafat [1988]
Jim Wright [1988]
Dan Quayle [1992]
George Mitchell [1992]
Nelson Mandela [1992]
Tom Foley [1992]

[11] Here, a correct answer constitutes correctly identifying that the Republicans are the more conservative party and the Democrats are the more liberal party.

[12] in the NES studies, the name of the actual candidates are used.

Al Gore [1996]
Boris Yeltsin [1996]
Trent Lott [2000]
Tony Blair [2000]
Janet Reno [2000]

List of Figures

1 A plot of the difficulty vs. discrimination parameters. Several potential outlyers are marked. "HR00" is the identification of the majority party in the House prior to the 2000 elections, "DP92" is the liberal/conservative placement of the 1992 Democratic Party, and "AG96" is the identification of Al Gore in 1996. 27

2 Difficulty vs. discrimination plots for items asked 3 or more times. The year in which the question was asked is noted above each point. 28

3 Item Response curve for the Identification of Jim Wright in 1988 is the solid line, the dashed line is the same curve for the liberal/conservative identification of Jimmy Carter in 1980. The thick line at the bottom represents the actual values of the latent trait for the survey respondents, indicating the minimum, median, and maximum scores in the data. . . . 29

4 The posterior means of the interviewer random effects terms, with the overall mean given by the thick line. 30

5 The trace plots for two parameters. The top plot is for the discrimination parameter for the liberal-conservative placement of Jimmy Carter in 1980, the bottom plot is the discrimination parameter for 1996 Republican Party liberal-conservative placement. 31

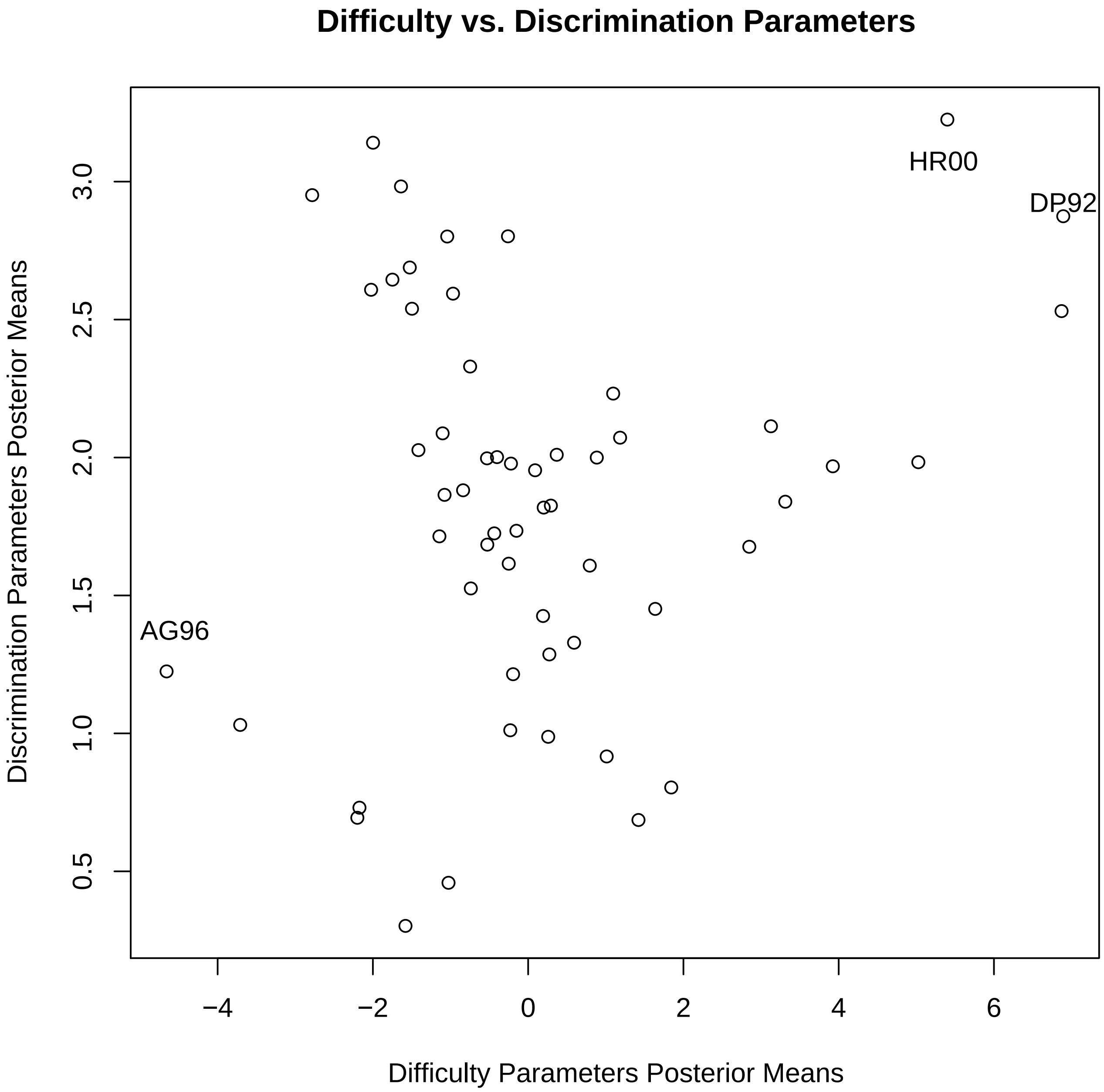
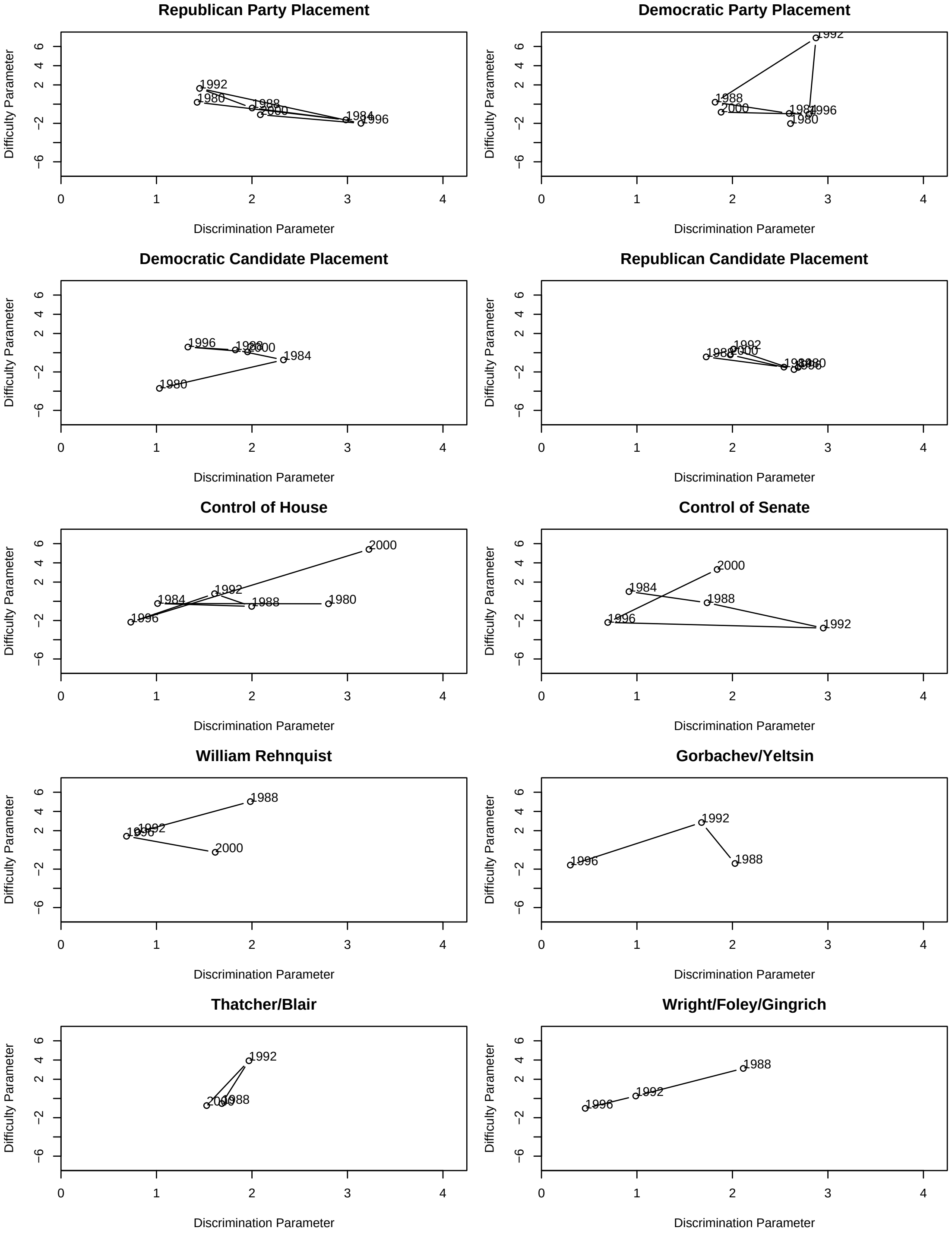


Figure 1: A plot of the difficulty vs. discrimination parameters. Several potential outlyers are marked. "HR00" is the identification of the majority party in the House prior to the 2000 elections, "DP92" is the liberal/conservative placement of the 1992 Democratic Party, and "AG96" is the identification of Al Gore in 1996.


Republican Party Placement

Democratic Party Placement

Democratic Candidate Placement

Republican Candidate Placement

Control of House

Control of Senate

William Rehnquist

Gorbachev/Yeltsin

Thatcher/Blair

Wright/Foley/Gingrich

Figure 2: Difficulty vs. discrimination plots for items asked 3 or more times. The year in which the question was asked is noted above each point.

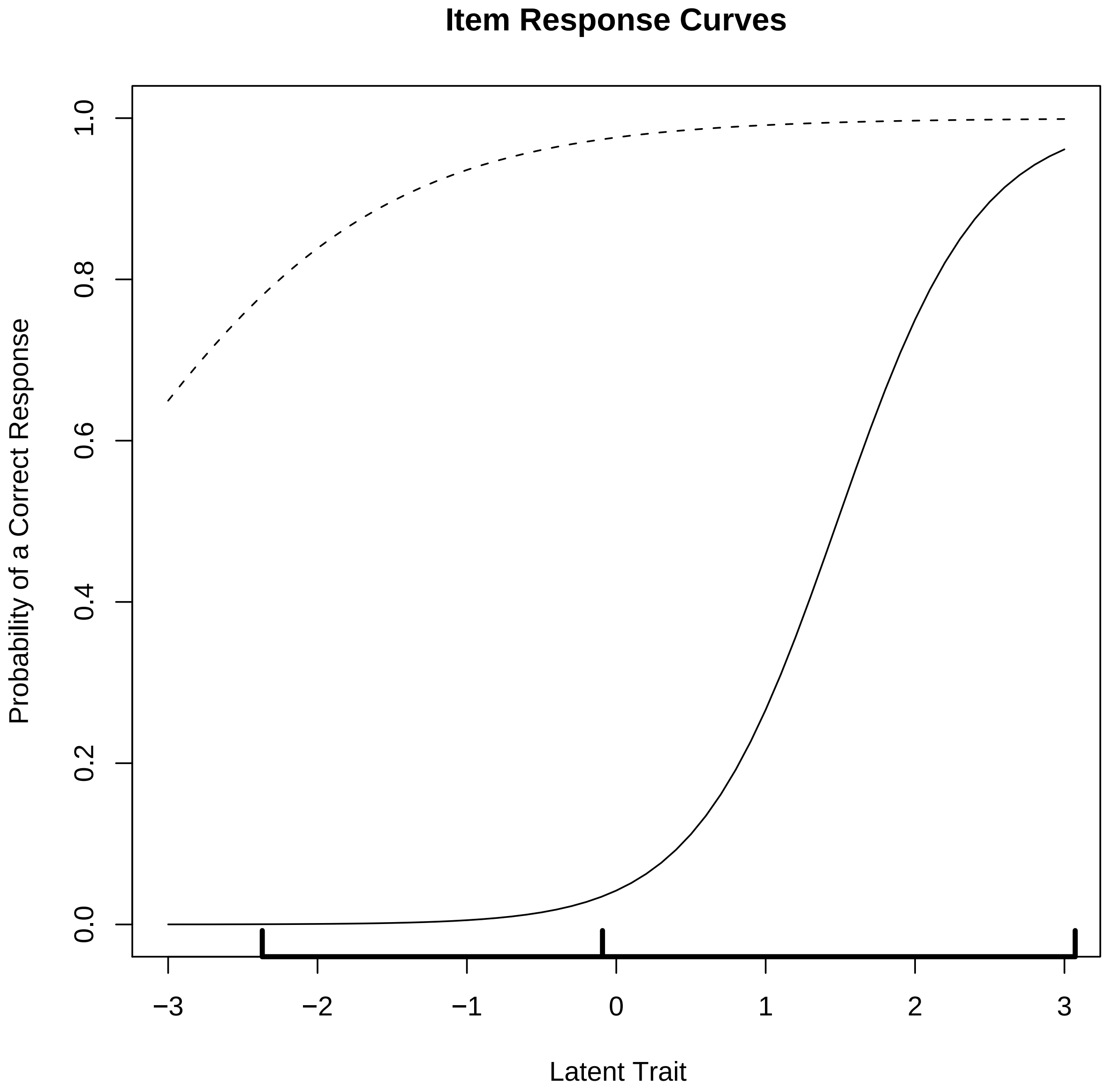


Figure 3: Item Response curve for the Identification of Jim Wright in 1988 is the solid line, the dashed line is the same curve for the liberal/conservative identification of Jimmy Carter in 1980. The thick line at the bottom represents the actual values of the latent trait for the survey respondents, indicating the minimum, median, and maximum scores in the data.

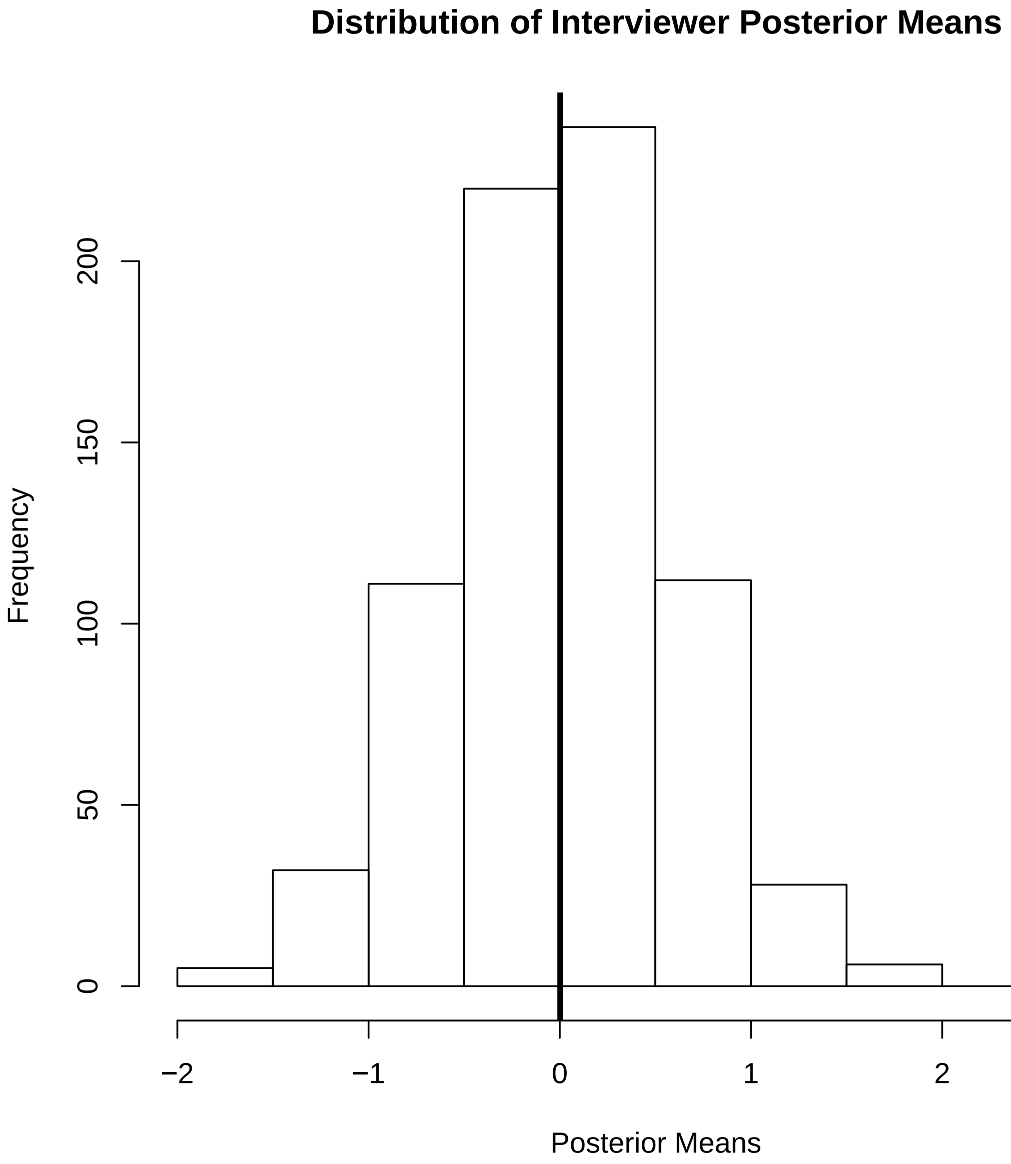


Figure 4: The posterior means of the interviewer random effects terms, with the overall mean given by the thick line.

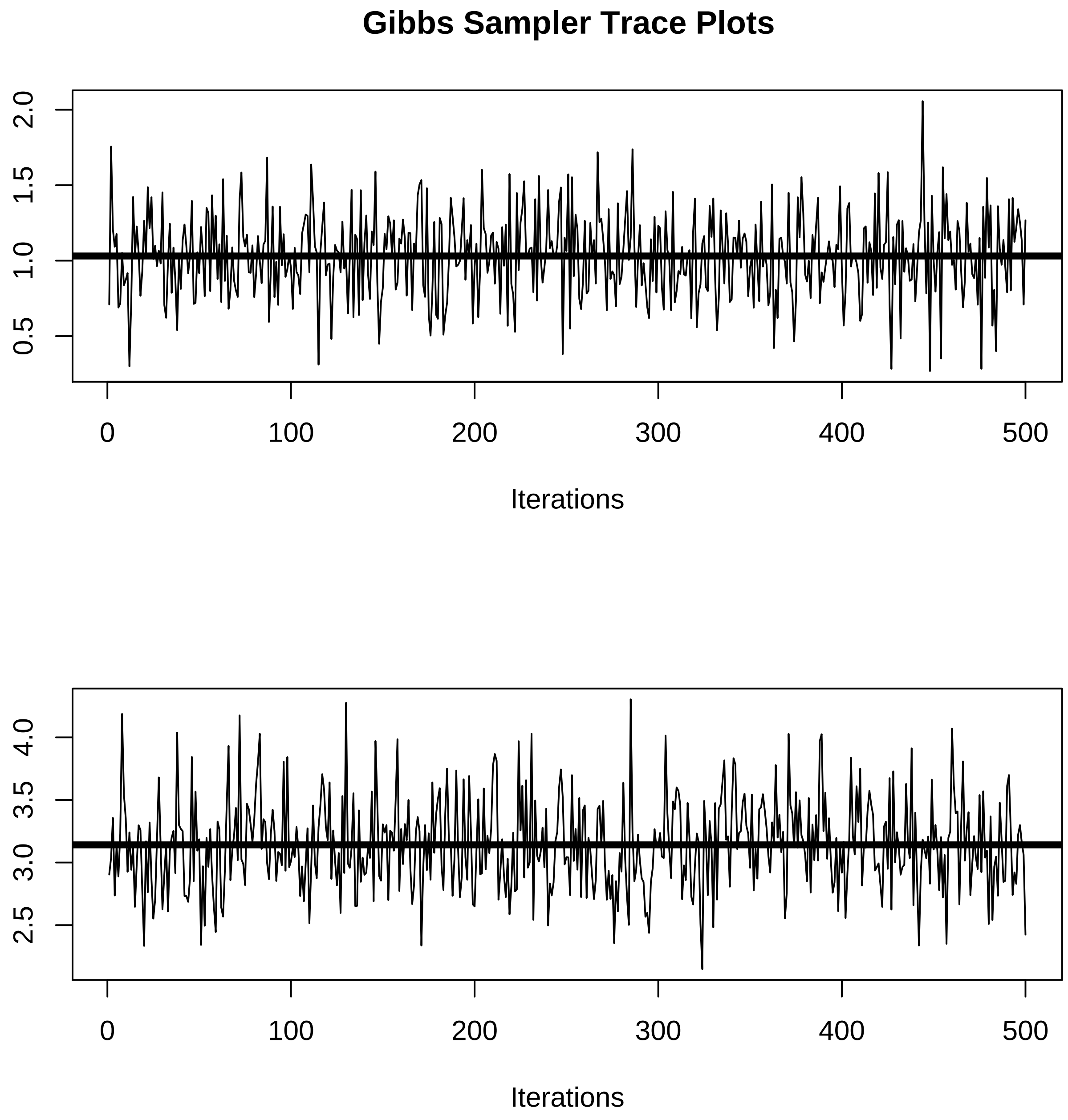


Figure 5: The trace plots for two parameters. The top plot is for the discrimination parameter for the liberal-conservative placement of Jimmy Carter in 1980, the bottom plot is the discrimination parameter for 1996 Republican Party liberal-conservative placement.

List of Tables

1 The breakdown, by year and overall, of how many of the interviewer effects terms overlap 0. If the interviewer effects terms do not overlap 0, then we can say that a given interviewer is using the scale differently than his/her peers. 33

2 Results of 500 simulations of a randomly selected interviewer assessing levels of political information of a respondent with a given level of the latent trait under two conditions: without the random effects terms and with the terms. The top half of the table reports the results without considering the random effects terms, the bottom half reports the simulation results considering the random effects terms. The cell entries are the number of iterations where an individual with that level of the latent trait is rated that score by the interviewer. 34

Year	Overlap Zero	Not Overlap Zero
1980	125	22
1984	91	18
1988	87	26
1992	119	25
1996	119	25
2000	99	19
Overall	629	123

Table 1: The breakdown, by year and overall, of how many of the interviewer effects terms overlap 0. If the interviewer effects terms do not overlap 0, then we can say that a given interviewer is using the scale differently than his/her peers.

Without the Random Effects Terms

Latent Score	1	2	3	4	5
-3	35	465	0	0	0
-2	0	500	0	0	0
-1	0	0	500	0	0
0	0	0	500	0	0
1	0	0	0	500	0
2	0	0	0	0	500
3	0	0	0	0	500

With the Random Effects Terms

Latent Score	1	2	3	4	5
-3	193	289	18	0	0
-2	16	346	138	0	0
-1	0	118	364	15	3
0	0	6	345	129	20
1	0	0	123	223	154
2	0	0	9	109	382
3	0	0	1	6	493

Table 2: Results of 500 simulations of a randomly selected interviewer assessing levels of political information of a respondent with a given level of the latent trait under two conditions: without the random effects terms and with the terms. The top half of the table reports the results without considering the random effects terms, the bottom half reports the simulation results considering the random effects terms. The cell entries are the number of iterations where an individual with that level of the latent trait is rated that score by the interviewer.