Reconsidering the Measurement of Political Knowlege
Matt Levendusky [1]
1 April 04

A basic feature of representative democracy is that citizens use elections as a means of controlling elected officials. A critical maintained hypothesis 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. In order to answer such questions empirically, we need a high quality measure of citizen's political knowledge.

However, despite its importance, our measures of political sophistication [2] 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 both methods have associated drawbacks. First, if we use a single indicator, we never get any idea of the uncertainty with which we measure a given concept. If there is considerable error in our proxy variable, then we risk biasing our inferences of interest (the classic errors-in-variables result). If we use a multi-item scale, we don't know that every item should be given the same weight—some items might do a better job of tapping into the underlying quantity of interest. So the question still remains: how can we more accurately measure political sophistication? Here, I conceptualize political sophistication, think of it as a continuous (but unobservable) variable. Even though we do not observe political sophistication directly, we do observe various indicators of it, and we can use these indicators to measure this underlying unobservable quantity of political sophistication. Here, the observable indicators are factual information items from the

[1] This is part of a larger joint project with Simon Jackman. Any errors contained here are my own.

[2] Essentially, political sophistication is how well you conceptualize and understand politics—do you know the rules of the game, the way various pieces of an ideology coherently fit together, etc.

quadrennial National Election Study (NES) from 1980-2000. Using these indicators, we can develop a model of political sophistication.

What does this sort of model look like? To actually model political sophistication, I estimate a two-parameter item-response (IRT) model. The model takes the following form:

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

Where j indexes items and i indexes respondents. Here, we've modeled the probability that individual i correctly answers item j as a function of two model parameters a_j and b_j and a latent trait θ_i. So here θ_i is the individual-specific unobserved trait, and the ys are the observable items (the factual information items drawn from the NES surveys). Since the items used here are binary, the F in the above distribution is the logistic CDF. 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 respondents 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 underlying latent trait.

However, items that can be scored correct or incorrect are not the only sources of information about political sophistication in the NES. The NES also asks the person conducting each interview to rate the respondents level of political knowledge on a 5-point scale from "Very High" to "Very Low". Indeed, this is perhaps the most common

single-indicator item used to measure political sophistication. In all of the studies using this item, there is an untested assumption that every interviewer uses the scale in exactly the same way. That is, that for interviewer A a "Very High" is the same as a "Very High" from interviewer B. To test this hypothesis, I include a random effect term for every interviewer. That is, we can write out the ordinal logit link as $\mu_i^{(p)} = \theta_i + \eta_p$, where η_p is the random effects term given to interviewer p. Let Z_i be the ranking given by judge p to individual i. Then

$$\begin{aligned} Pr(Z_i = \text{"Very Low"}) &= F(\kappa_{1t} - \mu_i^{(p)}) \\ Pr(Z_i = \text{"Fairly Low"}) &= F(\kappa_{2t} - \mu_i^{(p)}) - F(\kappa_{1t} - \mu_i^{(p)}) \\ &\vdots \\ Pr(Z_i = \text{"Very High"}) &= 1 - F(\kappa_{4t} - \mu_i^{(p)}) \end{aligned}$$

Here, p indexes interviewers, i indexes individuals, t indexes surveys, and F is the logistic CDF. Here, I've restricted the discrimination parameter (the parameter on θ_i) to be 1. This is done to achieve identification of this part of the model by ensuring that respondents who receive a higher interviewer rating receive higher values of the latent trait *ceteris paribus*, thereby fixing the direction of the latent scale. Also, it gives us over-time comparability. That is, this model restriction allows us to compare respondents across time (since we impose this restriction in all years). Otherwise, we couldn't be sure that scores from 1980 respondents were directly comparable to those from 1996.

The interviewer effects terms (η) allow us to see if each judge uses the scale in the same manner. 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 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.

Estimation of this sort of a model would be taxing in a traditional framework. Since we're estimating a parameter for every respondent, plus parameters on the items and judges, we're estimating well over 12,000 total parameters. Trying to find the global maximum of a 12,000 dimension hyper-surface is a Herculean (if not Sisyphean) task. Further, even if you manage to find the MLEs in a frequentist framework, to actually conduct inference, you'll need to invert a $k \times k$ matrix, where k is the number of parameters (again, over 12,000), and that simply won't be possible even given modern computing power. To make estimation and inference possible, I move to a Bayesian framework. To identify the model now, I will specify prior distributions on the model terms (as well as impose two parametric restrictions discussed below). Further, the Markov-Chain Monte Carlo algorithm used simplifies greatly the computational issues associated with maximizing the likelihood (Jackman 2000). Further, as a tool for inference, a Bayesian setup allows us to avoid matrix inversion but rather rely on the fact that we can get standard errors directly from the posterior distribution of each model parameter.

For the prior distributions, I chose vague distributions to allow the data itself to inform us as to the values of the parameters. For the item discrimination and difficulty parameters, we let $a_j, b_j \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}\right)$. For the random effects for each judge, again a vague distribution is used: $\eta_p \sim N(0, \tau^2)$, where $\tau^{-2} \sim \Gamma(0.01, 0.01)$. By modeling the variance of η as a parameter, we have another test of whether or not the interviewers are all using the scale in the same way. If $\tau = 0$, this is evidence that every respondent is using the scale in the same manner (since there would be no dispersion across judges). However, if $\tau \neq 0$, then there is evidence of judges using the scale differently. Finally, consider the threshold parameters in the ordinal link function. Here, the κs all come from a uniform distribution: $\kappa_{1t} \sim U(-10, \kappa_{2t}), \kappa_{2t} \sim (\kappa_{1t}, \kappa_{3t}), \kappa_{3t} \sim U(\kappa_{2t}, \kappa_{4t}), \kappa_{4t} \sim U(\kappa_{3t}, \kappa_{5t}), \kappa_{5t} \sim U(\kappa_{4t}, 10)$ for all t. Here, the structure of these conditional distributions ensure proper ordering ($\kappa_{1t} < \kappa_{2t} < \ldots$), but also that the joint

distribution of the κ terms is proper [3].

To complete the identification of the model, I need to specify the final parametic restriction (the first restriction is above in the ordinal logit link function). The restriction above fixes the direction of the latent trait, but I still need to fix its scale. To do this, I take the 1980 respondents who correctly answer zero items and receive the lowest rating on the interviewer scale and fix them at -1 on the latent scale, and take the 1980 respondents who answer every factual item correct and receive the highest rating at +1. All other respondents' latent traits are given an extremely vague $N(0, 100)$ distribution. Note that due to their prior distributions, respondents who are not pinned down are not required to fall in the $[-1, 1]$ range, but most of them will, since those selected to pin down the scale are fairly extreme.

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 algorithm to estimate the model, which conditional 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 proportional to their posterior probability[6].

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

[3]Thanks to Doug Rivers for pointing this out to me.

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

[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]Sample traceplots are available from the author upon request. For more on the Gibbs sampler and the associated difficulties with assessing convergence, see Jackman (2000).

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. 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 high discrimination and positive difficulty (discrimination greater than 1.5 and difficulty greater than 0)[8], 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 outliers 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

[7] available from the author upon request.

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

have no hard evidence, but a similar 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 measuring 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 indices) 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.

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)[9]. 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 2 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

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

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.

[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.

In addition to these partial preliminary results (additional results are available from the author upon request), one further area is to give an actual application of the method being used. Following Zaller (1992), I examine what he calls a "polarizing issue," one where partisans of opposing stripes have different views. One such example is whether or not the United States should increase or decrease its cooperation with the USSR in 1984. Zaller (as well as others) suggest that political information plays an important conditional effect on behavior: the politically savvy will know which side of the argument they support (based on their partisanship and ideology), as opposed to the politically unsophisticated (who are generally uninformed about these sorts of matters). So a model here might look something like:

$$y_i = \beta_0 + \beta_1 \text{PID}_i + \beta_2 \text{ideo}_i + \beta_4 x_i$$

Where y_i is support/opposition to increased US-USSR cooperation, PID is the respondent's party ID, ideo is the respondent's ideology, and x_i is the respondent's level of political sophistication from the model outlined above. One obvious baseline comparison is just to use, say, the interviewer rating and see if the more complete measurement model makes additional gains on top of this simpler model–do the model estimates change? Do we get the same picture of political information's effect using the two different indicators? While I've yet to fully conduct this analysis, I'd expect to see results similar to those from the Monte Carlo simulation above.

References

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

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

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.

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.

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

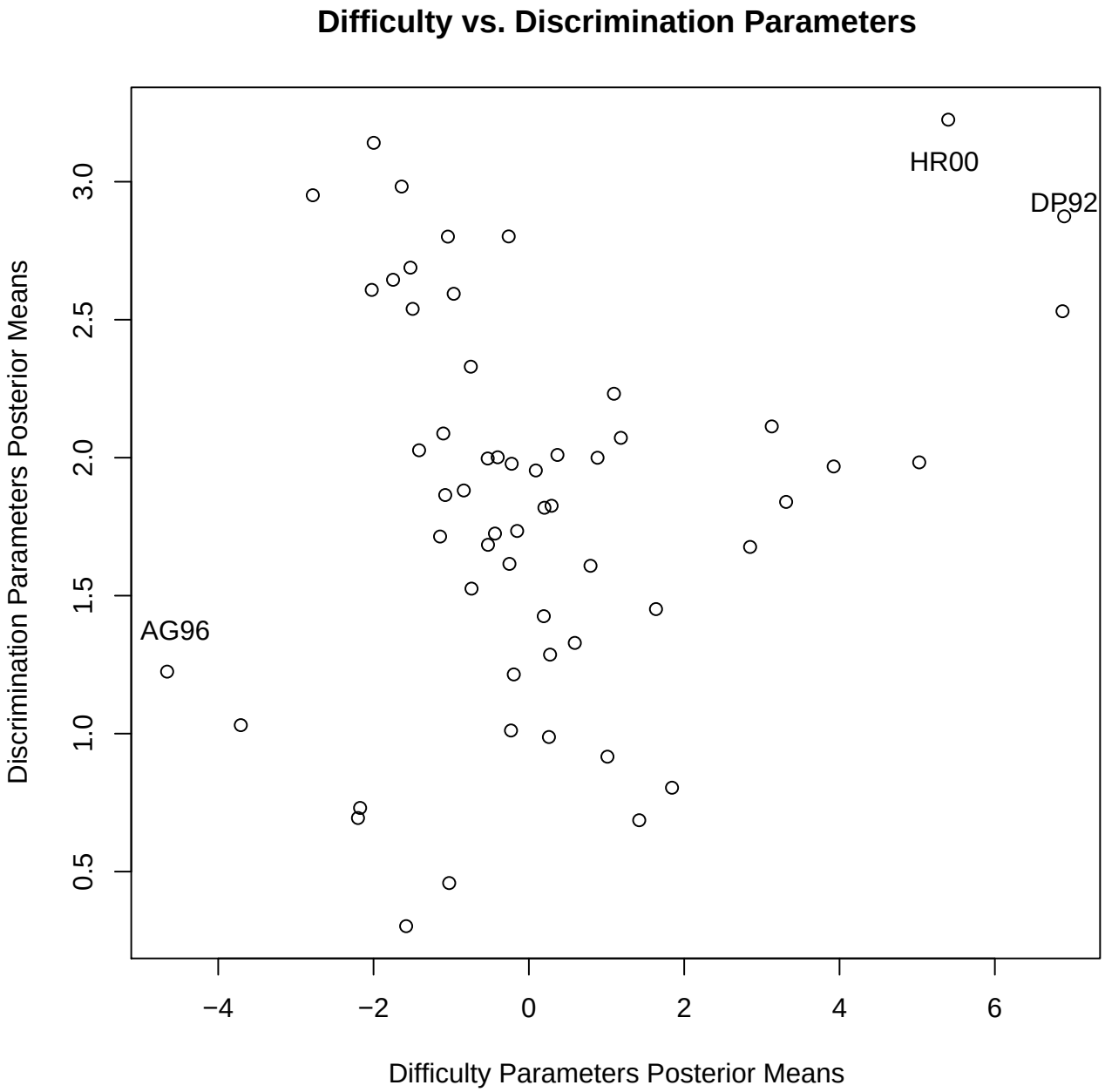

Figure 1: A plot of the difficulty vs. discrimination parameters. Several potential outliers are marked. "HR00" is the identification of the majority party in the House prior to the 2000 elections, "DP92" is the liberal/conservative placement of the 1992 Democratic Party, and "AG96" is the identification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

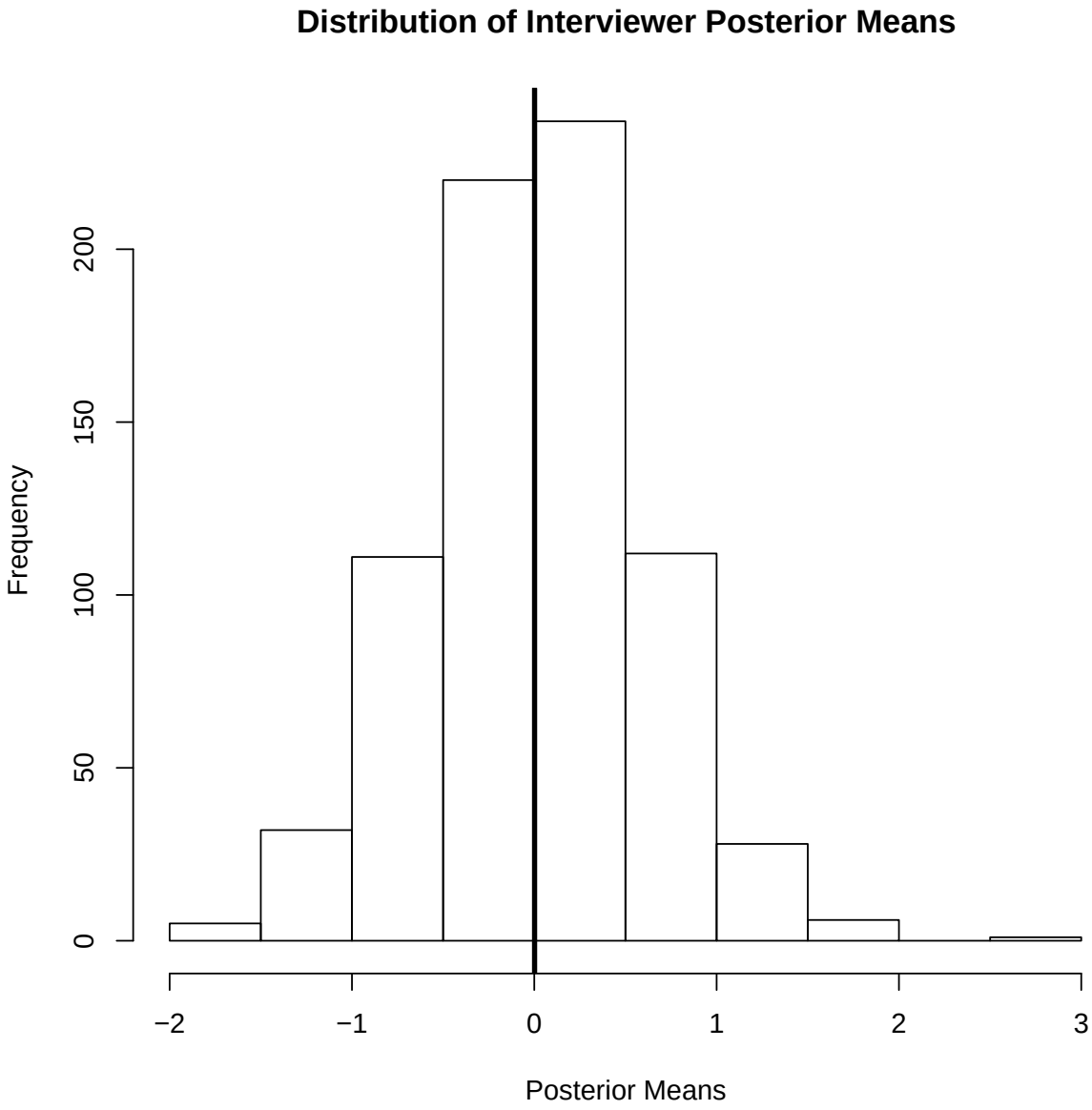

Figure 2: The posterior means of the interviewer random effects terms, with the overall mean given by the thick line.

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.