A Unifying Model for Hypothesis Testing Using Legislative Voting Data: A Multilevel Item-Response-Theory Model

Gyung-Ho Jeong*
University of British Columbia

Abstract
This paper introduces a multilevel item-response-theory (IRT) model as a unifying model for hypothesis testing using legislative voting data. This paper shows that a probit or logit model is a special type of multilevel IRT model. In particular, it is demonstrated that, when a probit or logit model is applied to multiple votes, it makes unrealistic assumptions and produces incorrect coefficient estimates. The advantages of a multilevel IRT model over a probit or logit model are illustrated with a Monte Carlo experiment and an example from the U.S. House. Finally, this paper provides a practical guide to fitting this model to legislative voting data.

Keywords
Item-response-theory, legislative voting, logit, and probit

* Associate professor in the Department of Political Science at the University of British Columbia. Email: gyung-ho.jeong@ubc.ca.

Legislative voting data are widely used to examine the relationship between legislators and other political actors, such as constituents and political parties. The most popular model for analyzing legislative voting data is a probit/logit model. However, probit/logit models have several limitations. First, such models do not utilize available information efficiently. Oftentimes, there are multiple votes on an issue. In this case, the selection of one vote out of several relevant votes can be arbitrary. Even if one fits a probit/logit model to each of the votes, combining the results of several models will be problematic when the results vary across votes.

Also, a probit/logit model is not useful when party-line voting is prevalent. In this situation, using an indicator variable for party as a predictor creates a 'separation' problem. That is, when an indicator variable perfectly predicts binary outcomes, this leads to infinite coefficients and standard errors (Zorn, 2005). Since party-line voting is common in many legislatures, this is a serious weakness of a probit/logit model. Even when party-line voting is not perfect, a variable for political party can cause a problem because this variable is often highly correlated with other predictors, such as legislator ideology. For instance, in the 110th U.S. House of Representatives (2007-2008), the correlation between party identification and ideology (measured by DW-NOMINATE dimension 1 score) was 0.96 ($p < 0.01$). Since both variables are important in explaining a legislator's voting decisions, omitting one of the variables can cause omitted variable bias.

To overcome these problems, one can alternatively pool multiple votes and fit a probit/logit model. This will help alleviate the problems mentioned above. However, a pooled probit/logit model has its own problems. By pooling multiple votes, the model imposes an assumption that every vote has the same degree of weight and extremity. As will be demonstrated later, this leads to biased estimates. In addition, pooling legislative votes and using a probit/logit model ignores the fact that a legislator's multiple votes are not independent, producing incorrect standard errors (Bailey, 2001). Using vote-fixed effects and legislator-fixed effects provides limited solutions to these problems.

Another problem of a probit/logit model is that researchers often need to make arbitrary coding decisions. To use an example of voting on trade legislation, in order to use a probit/logit model, we need to determine whether a vote for a proposal is a vote for free trade or for protectionism, and then assign to the vote either a one or a zero. However, this decision is not always straightforward. Some votes, like votes on final passage or omnibus bills, have ambiguous meanings. Even votes on amendments often defy such categorization. In such cases, these votes are either excluded from the analysis—which amounts to not using all available information—or coded either way based on the coder's judgements, which can be arbitrary.

Finally, a probit/logit model does not provide policy positions or the ideal points of individual legislators. Although this is not a problem in itself, the inability of a probit/logit model to estimate legislators' ideal points can be regarded as a weakness of the model, given that legislative votes have been a primary source of ideal point estimation. At the very least, it will be better if we can estimate the ideal points of legislators as well as test hypotheses using one model.

In this paper, I introduce a multilevel item-response-theory (IRT) model as a model that overcomes all of the aforementioned limitations of the existing method of analyzing legislative voting data. Although a multilevel IRT model was first introduced to legislative voting analysis by Bailey (2001), this model has been rarely used. Thus, the goal of this paper is not simply to reintroduce the model. This paper seeks to highlight the limitations of a probit/logit model and to introduce a multilevel IRT model as a unifying model for analyzing legislative voting data. To do so, I first demonstrate that a probit/logit model is a special type of multilevel IRT model with strong and unrealistic assumptions, especially when votes are pooled. I then use a Monte Carlo experiment and an example from the U.S. House to demonstrate that these assumptions produce biased estimates and have serious consequences on our inferences. Finally, I briefly explain the estimation procedure of a multilevel IRT model and provide a practical guide to fitting this model.

Ⅰ. Models for Hypothesis Testing Using Legislative Votes

In this section, I demonstrate that a probit/logit model is a special type of multilevel IRT model with unrealistic assumptions. The key feature of a multilevel IRT model involves embedding ideal point estimation into a multilevel modeling framework. That is, we model the structural components of legislator *i*'s ideal point using some predictors or covariates, and then place this structural component on top of an IRT model, a commonly used ideal point estimation method (see Clinton et al. (2004)). In an IRT model, the probability of legislator *i* to vote 'yea' on vote *j* is modeled using three parameters: the ideal point of legislator *i* (Θ_i), the discrimination parameter of vote j (a_j), and the difficulty parameter of vote j (b_j). Thus, a multilevel IRT model is formally defined by

$$\Theta_i \sim N(X_i\beta,\ \alpha^2) \tag{1}$$

$$P(Y_{ij}=1|\Theta_i,\ a_j,\ b_j) = F(a_j(\Theta_i - b_j)) \tag{2}$$

where Y_{ij} is 1 if legislator i voted 'yea' on vote j and 0 otherwise. The symbols Θ_i, β, α^2, and X_i represent legislator i's ideal point, a vector of coefficients, the variance of ideal points, and a vector of covariates for legislator i, respectively. Finally, F represents a cumulative distribution function of either a standard Normal distribution (Φ) or a logistic distribution (Λ).

In comparison to a probit/logit model, which assumes $P(Y_{ij}=1|\beta) = F(X_i\beta)$, two more parameters are used to define the probability of voting yes in a multilevel IRT model: difficulty and discrimination parameters. In a multilevel IRT model, these parameters are used to capture some important differences between votes. First, the difficulty parameter (b_j) captures the extent to which a vote is extreme or moderate by measuring the location of the cutpoint for each vote. To illustrate the meaning of this parameter, suppose for a moment that a vote has a discrimination parameter of 1. Then, the probability of voting yes on this vote is defined by $P(Y_{ij}=1)=F(\Theta_i - b_j)$. This probability will be equal to 1/2 when the ideal point of legislator i is equal to b_j (i.e., $\Phi(0)=\Lambda(0)=\frac{1}{2}$). Since a probability of 1/2 means that the legislator is equally likely to vote yes or no, it indicates that a legislator whose ideal point is equal to b_j will be indifferent between yea and nay. On the other hand, the probability of voting yes will be greater than 1/2 when a legislator's ideal point is greater than b_j, making the legislator more likely to vote yes. Finally, the probability will be less than 1/2 when a legislator's ideal point is less than b_j, making the legislator less likely to vote for the proposal. As this example illustrates, the difficulty parameter estimates the location on a space that divides those who are likely to vote for and those who are likely to vote against a proposal. Thus, we can think of this parameter as measuring the degree to which a proposal is extreme. An extreme proposal will have a cutpoint on the far left or far right side of the space.

Table 1. An Example of the Role of Discrimination Parameter

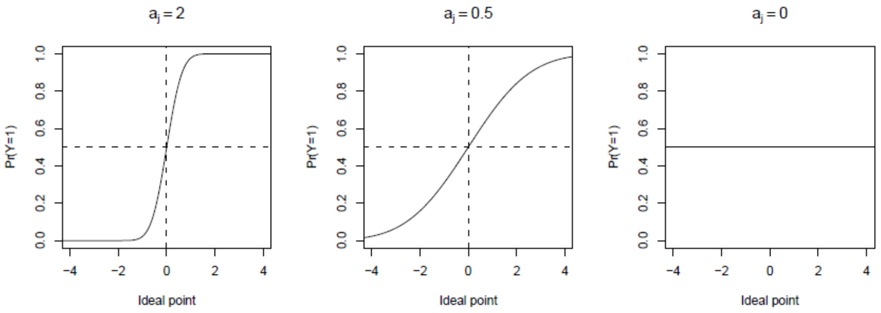
	Probability of Voting Yes=$\Phi(a_j(\Theta_i - b_j))$	
	Legislator A ($\Theta_A = 2$)	Legislator B ($\Theta_B = 0.1$)
Proposal 1 ($a_1 = 2$)	0.99=Φ(2×2)	0.58=Φ(2×0.1)
Proposal 2 ($a_2 = 0$)	0.50=Φ(0×2)	0.50=Φ(0×0.1)

Legislator A is assumed to be an extreme conservative, whereas Legislator B is assumed to be a moderate conservative. The two proposals are assumed to have different discrimination parameters (2 and 0) and the same difficulty parameter (0).

A discrimination parameter (a_j) measures the extent to which a legislator's policy position affects his or her voting decision. To illustrate the role of this parameter, an example is provided in Table 1. In Table 1, the probabilities of two hypothetical legislators to vote yes on two proposals with different discrimination parameters are computed. In the table, it is assumed that Legislator A is an extreme conservative with an ideal point of 2, whereas Legislator B is assumed to be a moderate conservative with an ideal point of 0.1. The discrimination parameter of the first proposal is assumed to be 2, whereas that of the second proposal is 0. For simplicity, assume the difficulty parameter of both proposals to be zero. Then, if we assume the probit link function, the probability of voting yes on the first proposal for Legislator A is 0.99, while the probability for Legislator B is 0.58. On the other hand, the probabilities of voting yes on the second proposal for Legislator A and Legislator B are both 0.5. That is, when the discrimination parameter of a vote is 0, the ideal point of legislators is discounted as a predictor of their voting decisions. Extreme legislators and moderate legislators have the same probability of voting yes on the vote. Thus, we can think of a discrimination parameter as capturing the weight or salience of a vote on the issue at hand. Ideological differences between legislators will matter more when a vote is salient.

The role of the discrimination parameter is graphically illustrated in Figure 1. In all three panels of Figure 1, the difficulty parameter is set at zero (that is, $b_j = 0$). Thus, the cutpoint is zero. In the figure, the curve in each panel represents the probability of voting yes as a function of the legislator's ideal point. For example, in the first panel ($a_j = 2$), a liberal legislator with an ideal point of -2 has a near zero probability of voting yes on this proposal, whereas the probability that a conservative legislator with an ideal point of 2 votes yea on the same proposal is very close to 1. Thus, we can say that this vote does an excellent job of discriminating between liberal and conservative legislators. Now, consider the vote in the second panel ($a_j = 0.5$). This vote has a lower discrimination parameter than the first vote. In this case, a liberal legislator with an ideal point of -2 has a probability of 0.16 to vote yes on this proposal, whereas a conservative legislator with an ideal point of 2 has a probability of 0.84 to vote yes on this proposal. As such, this vote is a less effective means of distinguishing between liberal legislators and conservative legislators than the first vote. Finally, the vote in the third panel is the case where the vote has a zero discrimination parameter. In this case, a legislator's ideal point contributes nothing towards an explanation of their voting behavior, because liberal and conservative legislators have the same probability of voting yes. In other words, this vote does not discriminate between liberal legislators and conservative legislators at all.

Figure 1. An Illustration of the Role of the Discrimination Parameter



* The Difficulty parameters in all three panels are fixed at 0.

A discrimination parameter is also useful for determining the nature of a vote. As I mentioned earlier, another important advantage of a multilevel IRT model is it does not force the user to make arbitrary coding decisions. When we use a probit or logit model, we code a vote for free trade as

1 and a vote for protectionism as 0 or vice versa. However, it is not always clear whether a proposal is for free trade or for protectionism. This problem can be solved when we use a multilevel IRT model because the sign of the discrimination parameter will tell us about the nature of the vote. For example, when the discrimination parameter of a vote has the same sign as other protectionist votes, it tells us that a 'yea' vote for the vote is for protectionism. It thus eliminates the need for a researcher to make arbitrary decisions.

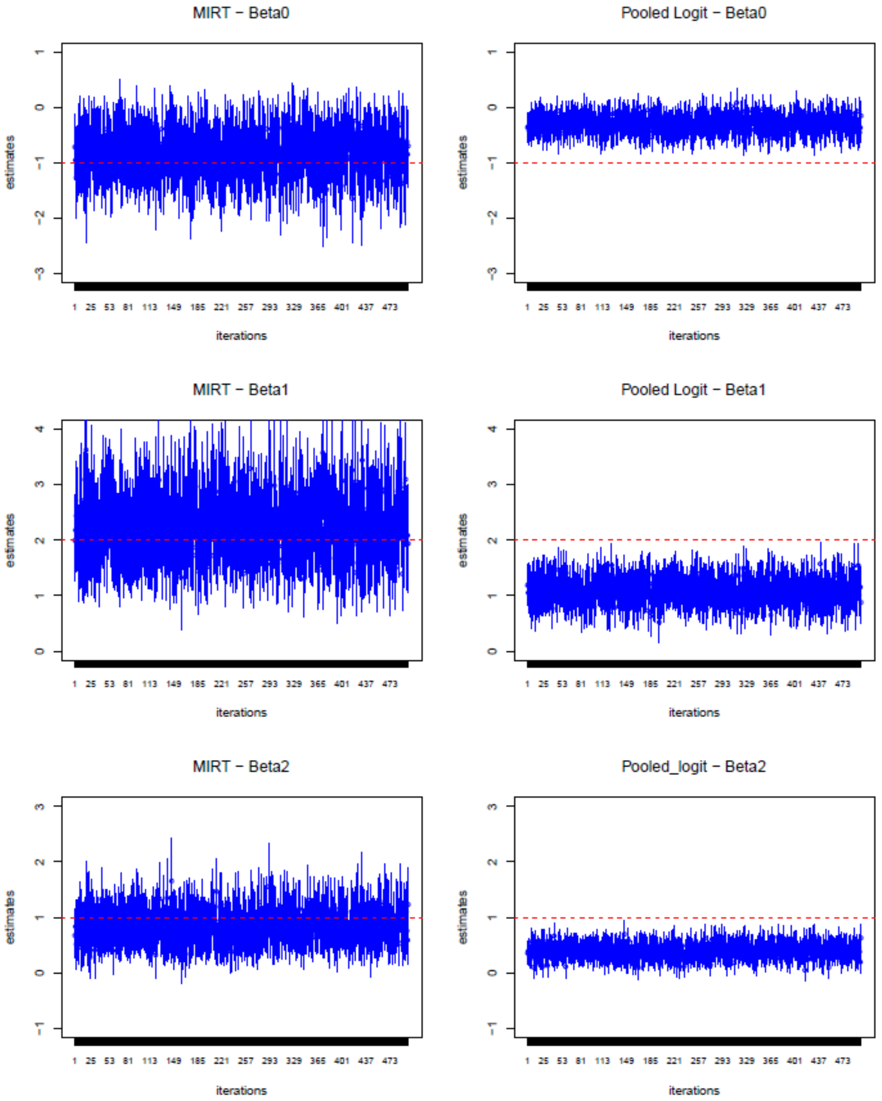
Next, I demonstrate that this multilevel IRT model provides a flexible and general approach to analyzing legislative votes by proving that a pooled logit or probit model is a special type of multilevel IRT model with strong and unrealistic assumptions. To transform a multilevel IRT model into a pooled probit/logit model, we need to make three highly unrealistic assumptions. First, we need to constrain a_j such that all votes are 1. Second, we need to assume that b_j for all votes are 0. Finally, we need to assume that the relationship between covariates and ideal points is deterministic, rather than being probabilistic. That is, $\Theta_i = X_i\beta$ rather than $\Theta_i \sim N(X_i\beta,\ \alpha^2)$. If we impose these three assumptions on a multilevel IRT model, a multilevel IRT model becomes a pooled probit/logit model. To see that, first, fix a_j and b_j in (2) to 1 and 0, respectively. Then, substitute $\Theta_i = X_i\beta$ into (2) and remove (1). We then have $P(Y_i = 1 | \beta) = F(X_i\beta)$, which features the typical structure of a probit model — $\Phi_i = X_i\beta$ — or a logit model — $\Lambda(X_i\beta)$. In other words, when we use a pooled logit or probit model to analyze multiple votes, we are making the unrealistic assumptions that all the votes have the same discriminatory power and the same cutpoint. In addition, we are assuming that constituent or party influences on legislators are deterministic rather than probabilistic. That is, we are assuming that changes in district interests or any other covariates will deterministically change legislators' ideal points. This proves that a pooled logit or probit model is a special type of multilevel IRT model with strong restrictions and unrealistic assumptions.[1] Since a multilevel IRT model allows us to estimate differences between votes instead of fixing them *a priori*, we can utilize multiple votes without making unrealistic assumptions. As will be demonstrated with a Monte Carlo experiment in the next section, making these unrealistic assumptions has serious consequences for our statistical inferences.

Ⅱ. A Monte Carlo Experiment

What happens to estimates of coefficients when one fits a logit or probit model to multiple votes, ignoring differences between the votes? Bailey (2001) has already pointed out that pooling legislative votes and using a probit/logit model ignores the fact that a legislator's multiple votes are not independent, producing incorrect standard errors. However, the problem is not limited to standard errors. Ignoring differences between votes introduces bias in coefficient estimates as well. To demonstrate this, I conduct a Monte Carlo experiment.

[1] Note that a probit or logit model applied to a single vote is also a special case of multilevel IRT model. In this case, the only assumption needed to transform a multilevel IRT model into a probit or logit model is that $\Theta_i = X_i\beta$. a_1 and b_1 need to be assumed to be 1 and 0, respectively, for model identification in both models.

Figure 2. A Monte Carlo Experiment: Comparing the Performance of Multilevel IRT and Pooled Logit Models



* The red dotted line represents true value of each coefficient in each panel. The blue bars represent the 95% intervals of the estimated parameters.

In this experiment, five votes are generated from a set of fixed covariates, coefficients, and ideal points (100 legislators). Two covariates are used: one indicator variable—similar to a variable for political party—and one continuous variable. Five votes are generated by assigning different degrees of cutpoint and weight (i.e., a discrimination parameter) to each vote. Once a set of five votes is generated, I fit a pooled logit model and a multilevel IRT model to the data. Then, I determine whether the estimated coefficients and their confidence intervals capture the true value of the coefficients. This procedure is repeated 500 times.

The results of this experiment are reported in Figure 2. The left panels plot the 95% credible intervals from the multilevel IRT model, whereas the right panels plot the 95% confidence intervals from the pooled logit model. The true value of each coefficient is denoted by the dotted line in each panel (i.e., $\beta_0 = -1$, $\beta_1 = 2$, and $\beta_2 = 1$). The figure shows clearly that the pooled logit model failed to capture the true values. The coverage rate is less than 0.01 for all three coefficients. In contrast, the multilevel IRT model captured the true values with high coverage rates (0.97, 0.97, and 0.94). One might suspect that these high coverage rates are a natural outcome of the large credible intervals of the multilevel IRT model. This is only partially true. The credible intervals of a multilevel IRT model are generally larger than confidence intervals of a logit/probit model because a multilevel IRT model assumes that the relationship between covariates and ideal points is probabilistic (i.e. $\Theta_i \sim N(X_i\beta, \alpha^2)$), adding another layer of uncertainty. In contrast, a probit/logit model assumes a deterministic relationship (i.e., $\Theta_i = X_i\beta$). However, the size of credible intervals is not the only reason for the successful coverage of the multilevel IRT model. To verify this, I compute the root mean square error (RMSE) using the point estimates of the coefficients. The RMSE computes the average distance between the true value of the parameter and its estimate. Again, the multilevel IRT model performs better than the pooled logit model. The RMSEs for the multilevel model and the pooled logit model are: 0.33 vs. 0.72 for β_0; 0.44 vs. 0.93 for β_1; and 0.27 vs. 0.62 for β_2. Clearly, the multilevel model provides more precise point estimates than the pooled logit model.

Finally, the figure also reveals the tendency of pooled logit models to underestimate the size of the coefficients. That is, in comparison to the multilevel IRT model, the coefficients are estimated to be closer to zero in the pooled logit model estimate. This is the result of ignoring differences between votes and thus diluting the associations between covariates and a legislator's voting. Simply, a pooled logit/probit model is unable to tell whether a weak association between covariates and voting is due to the small size of the coefficients or the low discriminatory power of the vote. Given this inability of a pooled logit/probit model to take into account the possibility that a vote might have a low discriminatory power, the model will attribute any weak association between covariates and voting to smaller coefficients.

To summarize, applying a logit/probit model to multiple votes is problematic in realistic situations where votes have different degrees of extremity and weight. The model produces biased point estimates. Also, the model tends to be overly confident about its estimates, providing narrow confidence intervals. Combined, these two tendencies can cause a serious problem as they will lead to incorrect inferences. In the next section, I illustrate this problem using a real-world example.

Ⅲ. An Application to U.S. House Votes on the Iraq War

The War in Iraq (2003) is one of the most significant foreign policy events in the US since the Vietnam War. Although the war on terror gained bipartisan support initially, the bipartisan consensus was broken up when the war on terror was applied to Iraq. As such, there were multiple votes in Congress dealing with the war. In particular, when the Democratic Party took control of the House after the 2006 elections, there were serious attempts to rein in President Bush's war policy. For instance, in the 110th Congress (2007-2008), there were 13 votes in the House that were directly related to the war, including votes on the 'surge', the setting of a deadline for troop withdrawal, and the prohibition of the permanent stationing of troops in Iraq (see Table 2 for the list of votes).

Table 2. House Votes on the War in Iraq in the 110th House

No.	Date	Yea	Nay	Bill No.	Description
98	2/16/2007	246	182	HCRes63	To disapprove the 'surge'
185	3/22/2007	218	212	HR1591	To pass supplemental appropriations. It set a 365 day deadline.
275	5/1/2007	222	203	HR1591	To override Presidential veto.
329	5/9/2007	171	255	HR2237	To call for troop withdrawals to start within 90 days of the enactment.
332	5/9/2007	221	205	HR2206	To require the President to direct the redeployment of troops from Iraq if there is a consensus with Iraq.
618	7/11/2007	223	201	HR2956	To direct the Defense Secretary to complete troop withdrawal by 4/1/2008.

No.	Date	Yea	Nay	Bill No.	Description
711	7/24/2007	399	24	HR2929	To limit the use of funds to establish the permanent stationing of United States Armed Forces in Iraq.
789	8/1/2007	229	194	HR3159	To provide troops with longer periods at home between tours.
1503	5/14/2008	141	149	HR2642	To provide $162.5 billion for the wars in Iraq and Afghanistan.
1504	5/14/2008	227	196	HR2642	To require a troop withdrawal by December 2009.
1534	5/21/2008	234	183	HR5658	To require congressional authorization for any agreement obligating the U.S. military to defend Iraq.
1536	5/21/2008	240	168	HR5658	To require that interrogations of detainees be videotaped.
1537	5/21/2008	218	192	HR5658	To bar the use of contractors as interrogators

* HR1591, HR2206, HR2642 are Supplemental Appropriations Bills. HR 5658 is Defense Authorizations Bill.

If one wants to understand why legislators supported or opposed the war, one would typically fit a logit/probit model to one of these votes. However, there are some problems in applying a probit or logit model here. First, selecting only one vote out of several relevant votes is problematic. Even if one fits a probit/logit model to each vote, combining the results of several models will be problematic when the results are different across votes. Second, increased partisan polarization in Congress resulted in some votes being highly partisan, making it impossible to include a party dummy variable. When a variable perfectly explains a binary outcome, coefficient estimates and standard errors will explode. Even when voting does not entirely fall along party lines, a high correlation between partisanship and ideology—again due to the increased party polarization in Congress—creates a problem. In the case of the

110th House, the correlation between a party dummy variable (Democrats) and ideology (DW-NOMINATE dimension 1 score) was 0.96 ($p < 0.01$). With such a high correlation between the two important variables, it is difficult to use a single vote to estimate the influence of partisanship while controlling for the role of ideology or vice versa. Thus, a better approach for dealing with such highly partisan votes is to utilize all available information by pooling the votes.

However, pooling votes and fitting a pooled logit/probit model has its own problems. First, one has to code each vote as pro-war or anti-war. Although the coding decision is straightforward in many cases, it is not always clear-cut. For instance, Barbara Lee (D-CA) proposed an amendment to the Defense Authorizations bill that required congressional authorization for any agreement obligating the U.S. military to defend Iraq. Although we can infer that this was an anti-war proposal because it was put forward by a liberal Democrat, and because many Republicans and conservatives opposed this amendment, the text of the amendment itself is not clearly anti-war in nature. Another issue which arises with pooling votes is that a researcher has to treat equally votes with different degrees of extremity and weight. In the data, there are several votes on troop withdrawal. For instance, vote #185 requires troop withdrawal by a certain deadline, while vote #332 requires the President to order orderly redeployment of troops from Iraq if there is a consensus with the Iraqi government that directs a redeployment of US troops. Arguably, the former is a stronger anti-war measure than the latter. However, to fit a pooled logit/probit model, we need to assign 0 or 1 to both votes, which carries the assumption that the two votes have the same degree of extremity and weight. What then is the result of making these unrealistic assumptions?

Table 3. Determinants of Support for the Iraq War

	Pooled Logit	Multilevel IRT
Intercept	0.46^{**}	1.33^{***}
	(0.20)	(0.34)
Democrats	-1.93^{***}	-3.46^{***}
	(0.37)	(0.57)
Ideology	1.12^{***}	0.74^{**}
	(0.23)	(0.34)
South	0.16	0.26
	(0.11)	(0.17)
Presidential Vote	0.70^{***}	0.76^{***}
	(0.10)	(0.15)
HFAC	-0.06	-0.03
	(0.17)	(0.23)
HASC	0.30^{**}	0.17
	(0.14)	(0.22)
Seniority	-0.17^{***}	-0.22^{**}
	(0.05)	(0.08)
Military Employees (%)	-0.11^{**}	-0.05
	(0.05)	(0.08)
% Correctly Predicted	49.8	96.1

For the pooled logit model, standard errors are reported in parentheses.
*$P < 0.10$; **$P < 0.05$; ***$P < 0.01$ (two-tailed tests). For the multilevel IRT model, standard deviations of posterior densities are reported in parentheses. *,**, and *** indicates that 90%, 95%, and 99% credible intervals do not include 0, respectively.

For this, I fit a pooled logit model to the votes. As predictors, I include a party dummy variable (*Democrats*), a measure of ideology, using DW-NOMINATE dimension 1 score (*Ideology*), a dummy variable for the South (*South*), President Bush's vote share in each district in 2004 (*Presidential Vote*), memberships in the House Foreign Affairs (*HFAC*) and Armed Services (*HASC*) Committees, the number of terms a legislator served in the House (*Seniority*), and the percentage of district population employed in military installations (*Military Employees %*). These are standard variables used in analyzing votes on foreign policy matters (see Kriner (2010)). All continuous variables are standardized before estimation to allow for the comparison of coefficients.

Table 3 reports the estimate of the pooled logit model along with that of the multilevel IRT model. At first, the results seem similar. Both model estimates show that Democrats, liberals (with low DW-NOMINATE scores), and senior members were against the Iraq War, whereas President Bush's vote share is positively associated with support for the war. However, there are important differences. First, two variables are significant in the pooled logit model estimate but insignificant in the multilevel IRT model estimate: membership in the House Armed Services Committee (HASC) and the percentage of a district's population employed on military installations. Regarding the HASC, the model estimate suggests that the HASC members were more supportive of the war than non-HASC members. Although HASC members used to be more hawkish than non-members in the 1970s and early 1980s, the committee has lost its power in the 1990s due to increased control over committees by party leadership and party caucuses (see Deering (1993)). Thus, more recently, committee membership has lost its predictive power on military intervention votes (Kriner, 2010). Similarly, the influence of district interests on military intervention votes should have decreased in recent years due to the increased influence of party leadership. Yet the pooled logit model estimate suggests that representatives from districts that rely more on military installations were more likely to oppose the Iraq War. This result is not consistent with the result of a recent study that finds the number of military employees to be insignificantly related to a member's voting on military interventions (Kriner, 2010). Moreover, the idea that districts that rely more on military installations are more likely to oppose the continuation of the war seems counterintuitive, since opposition to the war would appear to go against their parochial interests. Thus, the significance of these two variables is likely to be the result of the pooled logit model producing overly confident estimates, as illustrated by the Monte Carlo experiment in the previous section.

Finally, my criticisms of the pooled logit model estimate can be justified by the fact that the pooled logit model poorly fits the data, as shown by the low classification success rate of 49.8%. That is, the model predicts only half of the votes successfully. In comparison, the multilevel IRT model has a very high classification success rate of 96.1%.

To summarize, the application of a pooled logit model to the Iraq War votes illustrates the problems demonstrated by the Monte Carlo experiment in the previous section. The overconfident expectations of the pooled logit model resulted in the production of small standard errors (resulting in the unlikely significance of the percentage of military employees and HASC membership) and fit the data poorly (by predicting the votes correctly less than half of the time).

Ⅳ. A Practical Guide on How to Fit a Multilevel IRT Model

In this section, I provide a guide to fitting a multlevel IRT model to legislative voting data. Although Bailey (2001) used an EM-algorithm, a multilevel IRT model can also be estimated via Markov Chain Monte Carlo (MCMC) methods. When we use the probit link function and assign conjugate priors for the parameters, we can use the Gibb-sampling algorithm to sample from the posterior density. The Gibbs algorithm is detailed in Appendix A.

One of the advantages of using the MCMC methods is that a multilevel IRT model can be fit using a ready-made package in R or using a relatively simple code in WinBUGS, a free Bayesian software. In R, the function MCMCirtHier1d in a package called MCMCpack (Martin et al., 2011) can be used to fit the model. To use this code, users simply need to provide voting data and covariates. If one wants more flexibility, one can use WinBUGS. A sample WinBUGS code is provided in the appendix. This code can be easily modified for use. The sample code assumes two predictors (X_1 and X_2). Users can add or remove predictors. Note that the discrimination and difficulty parameters of the first vote are fixed at 1 and 0, respectively, to identify the model (see the last two lines of the code). Users can choose a different vote to identify the model. This WinBUGS code can be implemented from R using the R2WinBUGS package.

V. Conclusion

Most legislatures make decisions by votes. These legislative votes are valuable data for legislative studies. Some researchers have used the data to test hypotheses regarding the relationship between legislators and other political actors, whereas others have used the data to estimate the ideal points of legislators. This paper has shown that we can do both using a hierarchical ideal point estimation or a multilevel IRT model.

Admittedly, this model was initially introduced more than a decade ago by Bailey (2001). Nevertheless, this model has been rarely used to test hypotheses using legislative voting data. One reason for this lack of attention or usage is that researchers have not realized the seriousness of the problems of a (pooled) logit/probit model. Another reason is the cost of computation. Bailey used the EM-algorithm. MCMC methods were not widely used then. This paper has addressed these two possible sources of hesitation or inattention. First, this paper has demonstrated that commonly used pooled logit or probit models make unrealistic assumptions and thus produce biased estimates. Regarding the cost of computation, although fitting a multilevel IRT model is still costlier than using a probit/logit model, the development of software and the availability of a ready-made package have significantly lowered the cost. In particular, the guide in this paper should help researchers in applying a multilevel IRT model to legislative voting data. Thus, this small cost of computation should not deter researchers from using this model, given the benefit of obtaining better estimates.

References

Albert, James. (1992). Bayesian Estimation of Normal Ogive Item Response Curves Using Gibbs Sampling. *Journal of Educational Statistics,* 17: 251–269.

Bailey, Michael. (2001). Ideal Point Estimation with a Small Number of Votes: A Random-Effects Approach. *Political Analysis,* 9: 192–210.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. (2004). The Statistical Analysis of Roll Call Data. *American Political Science Review,* 98(2): 355–370.

Deering, Christopher J. (1993). Decision Making in the Armed Services Committees. In *Congress Resurgent: Foreign and Defense Policy on Capitol Hill*, Ann Arbor, MI: The University of Michigan Press.

Kriner, Douglas L. (2010). *After the Rubicon: Congress, Presidents, and the Politics of Waging War*. Chicago: University of Chicago Press.

Martin, Andrew D., Kevin M. Quinn, and Jong Hee Park. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software,* 42: 1–21.

Zorn, Christopher. (2005). A Solution to Separation in Binary Response Models. *Political Analysis,* 13(2): 157–170.

Manuscript received: Jan 15, 2021; Review completed: Mar 2, 2021; Accepted: Mar 15, 2021