The Politics of Predicting the Future

Kenneth Lowande[*] Mark A. Weiss[†]

September 2, 2024[‡]

Abstract

Congress negotiates with vast uncertainty about the effects of proposed policies, and so relies on committees to report about the past and project the future. Little is known about their conclusions, their accuracy, and whether politics seeps into their judgements. We introduce a typology of informational statements contained in committee reports, which classifies them based on whether they are prospective or retrospective, and whether they contain an inference. We argue that bipartisan teams tend to focus on the past to avoid disagreements inherent in predicting out-of-sample. Using a new dataset of committee reports from the 106th-116th Congress, we identify and classify summary findings, evaluations, predictions, and recommendations. We show that bipartisan teams are more retrospective and fact-based, while partisan teams favor predicting the future—without making actionable recommendations. Our results demonstrate politics drives partisan teams to systematically communicate less accurate (and helpful) classes of information.

Keywords: congress, committees, information, prediction, oversight

[*]Associate Professor, Department of Political Science, University of Michigan: lowande@umich.edu

[†]Ph.D. Student, Department of Political Science, Ford School of Public Policy, University of Michigan: markweis@umich.edu

[‡]Prepared for the CORD Workshop at the Levin Center for Oversight and Democracy, July 2024. We thank Christian Fong and Jeff Jenkins for helpful comments and Alexis Castellanos for research support.

Picking the right policy is hard. Members of Congress negotiate with vast uncertainties about the effects of policies both enacted in the past and proposed for the future. To reduce that uncertainty, they rely mostly on committees and their permanent staff, who specialize and are charged with gathering information. That information, though, is a kind of power—in that committees might leverage what they know to get policies that advance their priorities over those of the median in Congress. Congress solves this problem, according to researchers, by making sure members of committees are broadly representative (Krehbiel, 1991). The diversity of opinion is supposed to lead to consensus, and the transfer of information between specialized committees and the floor.

But coming to consensus is a messy process. Committees might divide their voice, or form a bipartisan team. Moreover, the policy information committees communicate is more complex that the revelation of some discrete fact. Committees gather many facts, and much like social scientists, make causal inferences and out-of-sample predictions to ultimately come to actionable recommendations. Almost nothing is known about the variety of informational signals committees send, or who tends to send them. We do not know how often committees make inferences, how often they are wrong, or how politics influences their judgements.

In this study, we argue that the information sent by committees is conditioned by political incentives, as well as the inherent complexity of communicating the fruits of expertise. Like existing informational theories of legislative organization, we ask what would motivate expert committees to invest time and resources to learn on behalf of the rest of Congress. But we focus on the two critical features of the process mentioned above: committees ultimately decide whether to send a unified informative signal, and there is variation in the scope and complexity of what they communicate. We argue that as information becomes more forward-looking and complex, it is more costly to develop and more likely to breed disagreement among experts. Thus, though bipartisan teams leverage the most resources and expertise, they tend to stick to examining the past. In contrast, partisans teams, self-aware that their conclusions may be taken less seriously, over-emphasize the future to demonstrate their effort and increase their influence on the rest of Congress.

To evaluate this argument, we analyze a new dataset of House and Senate committee reports from the 106th to 116th Congress. Many studies examine the composition of committees, the frequency of hearings and members' conduct in them, the amendment process, and more. Yet, committee reports, which are supposed to transmit the summary views of committee members, have been largely ignored. These views reside in committee reports. We also propose and implement a typology of summary statements. Using a large language model, we first identify and then classify findings, evaluations, predictions, and recommendations. Findings of fact are simple conclusions about the past, whereas evaluations are inferences about the past. Predictions involve inferences about the future, whereas recommendations are simpler, normative statements about what should be done. This novel measure indexes the types of conclusions committee authors are willing to come to in public.

Our analysis shows that politics is an important determinant of the kind of information transmitted by Congress' committees. In general, committees ground their reports in findings of fact, and less often, make inferences or recommendations for the future. However, bipartisan teams tend to be more retrospective, containing, on average, three additional retrospective findings or evaluations, relative to partisan teams. The contrast is even starker for reports authored by the minority party, which contain about five fewer retrospective conclusions. Yet, despite their access to less staff resources, partisan teams working alone make 1.5 additional predictions, relative to their bipartisan counterparts. These findings persist after accounting for political conditions that mark inter-branch conflict, like divided government. They are also not an artifact of counting—in that, the conclusive statements in reports are not more or less concrete or specific, depending upon their author. Finally, we show that partisan teams do not make more (or more specific) recommendations than bipartisan ones.

Our results reveal an important dynamic in how Congress informs itself. Bipartisan teams are not just ideologically moderate or better-resourced than partisans. They are motivated to make statements that are, in general, less likely to be wrong. Predicting the future is harder than reporting the past. Because it is also more politically palatable to agree about the past, bipartisan teams systematically shy away from classes of information that are less reliable.

Therefore, as bipartisanship in contemporary Congresses declines, these incentives may further undermine Congress' "role as an information proccesor"(Lewallen, Theriault and Jones, 2016).

Information and Policy Evaluation in Congress

Uninformed policies are typically bad polices. Congress needs information to do its work. To supply this information, it relies on committees with specialized jurisdictions and scope. These committees exert costly effort to get informed and communicate their conclusions to the rest of Congress. A lot of scholarship investigates the strategic considerations that arise from this kind of organization.

Most importantly, knowing more than the rest of Congress might be tempting to exploit. Committees might misrepresent the truth to get the policies they want (Fenno, 1973; Lowi, 1969; Kiewiet and McCubbins, 1991). One solution is to make sure committees are made up of members with diverse preferences, and, to the extent that Congress tolerates extremists, it should do so only when it will benefit from their superior expertise (Gilligan and Krehbiel, 1987; Krehbiel, 1991). This emergent way of doing things, however, requires diverse committees to find consensus. Together, the committee must collect information and come to what amounts to a median position on the conclusions the rest of Congress should draw from its work. In practice, this involves a lot of conflict, which social scientists have carefully catalogued and studied.

That conflict is often on display, for example, in open committee hearings. It is the most obvious evidence that coming to consensus involves fundamental disagreements about the facts, how to frame them, and their implications. Eldes, Fong and Lowande (2024) show that members of the opposition party are more confrontational in their questioning of witnesses. Out-party legislators tend to make fewer falsifiable statements and engage in more rhetorical bluster (Park, 2021). The likely goal is press that will aide their re-election prospects (Park, 2023). By making a public display or "going viral," they might attract positive press that aids their own re-election. Indeed, members' staff training emphasizes the desire to cultivate good press for their member (Fong, Lowande and Rauh, 2024).

In short, hearings themselves may be tactical engagements in larger conflicts between the president and the opposition party (Kriner and Schickler, 2014; Lowande and Peck, 2017). But committee hearings are only the most dramatic part of a larger information-gathering effort. Earlier in that process, committees and individual members request information from targets of investigations (Lowande, 2018, 2019; Ritchie, 2018), as well as pre-interview and decide on witness lists (Ban, Park and You, 2021).

All of the research referenced above involves observable behavior at some stage of the committee's information collection. But, oddly enough, it omits the last stage: transmitting the committee's conclusions. Committees produce reports, sometimes hundreds of pages long, presenting their top-line conclusions, along with supporting reasoning and evidence. Committee staff might spend years on such reports. The hearings, private inquiries, witness testimony, expert consultation, and more, ultimately end up in the written report. Yet, there are no systematic studies of these reports or their content. Thus, while the conflict-ridden process of holding hearings and investigating has been well-studied, the terminal process of coming to consensus has not.

This is what our study aims to accomplish. Reports are to hearings and oversight, as roll-call votes are to bill introductions, amendments, and mark-up. They are the informational equivalent of the committees position, and their central tendencies reveal help reveal how Congress works.

Of course, the basics of committee reports differ from advancing legislation in important ways. Unlike bills, nothing requires committees to compose a single a report, and they often choose not to. This is because committees are heterogeneous, made up of members of both the majority and minority party. The final work product(s) might not represent the median of the committee, but the median of each partisan team. In short, the information transmitted by the committee might come from the majority, minority, or both.

Less obviously, however, is that these reports show the *kind* of information their authors want to communicate. Information transfer between committees and the rest of Congress is usually modeled as committees sending a signal about private information they have obtained. This information is singular, discrete, and invertable (Callander, 2008). In benchmark principal-agent models, information is just a fact that reveals a secret. This is stylized and appropriate for the purposes of understanding the core agency problems in Congress' organization, and in many other settings. But it also abstracts away from empirical variation in the "signal" that might be revealing. In this study, we introduce a typology to describe this variation, and argue it can help reveal strategic considerations in committees.

More concretely, committees share information that varies both in its subject and complexity. Suppose, for example, a committee chooses to conduct an investigation of the Strategic Petroleum Reserve (SPR). It requests and obtains information from the Department of Energy (DOE), interviews witnesses, consults outside experts, sends its staff down to West Hackleberry, Louisiana. It has a mound of raw information, and must distil its work down to something usable for an audience who did not do the investigation. As anyone who has conducted extensive research on a single subject knows, summarizing and communicating that research is real work, and the right way to do it often depends on author(s) and the intended audience.

This hypothetical SPR report might contain headline conclusions that simply summarize facts obtained by the committee—for example, that the reserve contains in excess of 700 million barrels of crude oil. Alternatively, the committee might take a step further, and draw an inference about policy. It could say that current crude levels in the SPR are the result of the DOE's severe energy supply interruption policy. It could take another step further, and say that if such policies continue, the levels will dip below tolerable thresholds. Finally, it could recommend a policy change. Each of these hypothetical involves a choice about the kind of information to communicate to the rest of Congress. Looking at the past might be less politically fraught than prescribing the future. Moreover, each type of conclusion leaves more or less room for error. Sticking to the facts is easier than predicting out of sample.

In short, even for a toy example, it is clear that facts do not speak for themselves. Committees know this, and their reports reflect it by the way they are written. The writing might change depending on the author's incentives. To put it more generally, the kind of informative signal committees send is a choice. These decisions can teach us about how and if committees come to consensus, and how the quality of information might be implicated.

Why Committees Report the Way they Do

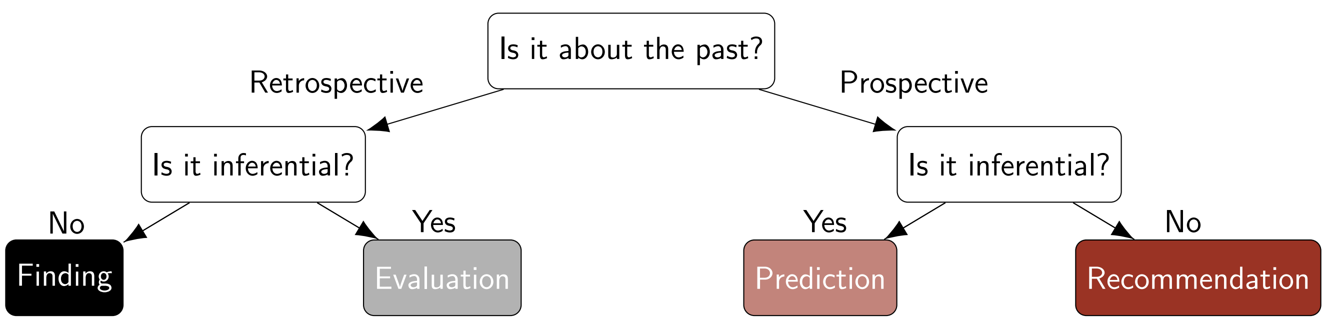
We argue committees report one of four types of conclusions: findings, evaluations, predictions, and recommendations. A "finding" is the most basic conclusion a committee could contain in a report.[1] This kind of information mimics findings of fact in the legal profession, due partly to the fact that many staff authors of committee reports are lawyers. Findings, by our definition, are the equivalent of revealing some discrete piece of private information the committee uncovered. We provide real-world examples in a later section, but for now, it is useful to think of this as the empirical analogue of the private information uncovered and communicated in principal-agent models.

As we lay out in Figure 1, other categories of conclusions come from the answers to two distinct questions. First, is the conclusion about the *past, or future*? Second, does the conclusion *make an inference*? Simple findings, by our definition, refer to the past, and do not make inferences. To return to our hypothetical, "the SPR contains 700 million barrels of oil" is a finding. But the committee could layer an inference by stating that this amount was observed because of a policy of the Biden administration. The *because* marks an inference about a particular policy, and implies a counterfactual that had this policy not been adopted, the level of oil would be different. We term this an "evaluation." It is a retrospective assessment of a policy, a causal inference about the past.

[1] Ours is a typology for research purposes. In practice, committees sometimes label *everything* a finding, regardless of whether it is a fact, projection, or implicit policy recommendation. We are careful in the following to distinguish between the empirical regularities and our conceptualization.

Both findings and evaluations deal with the past, but there are analogous classes for the future. The committee could project forward, arguing that if some policy is changed (or remains in place), we can expect to see the level of oil to reach some new, different level. This is a "prediction." Again, it involves an inference, because it implies a counterfactual projection in which the policy is not in effect. Finally, the committee can draw a conclusion about the future that does not involve an inference. It can make a "recommendation" or a normative statement about an action or policy some external body—in the executive branch, private sector, or the rest of Congress—should adopt. Because of these definitions, the typology is not mutually exclusive. It is nested, in that you cannot make an evaluation without connecting it to a finding, because a statement like "the severe energy supply interruption policy caused *something* to happen" is nonsensical. Most recommendations are supported by predictions and evaluations. Most importantly, even a cursory read of committee reports makes clear that each type of conclusion is frequently present.

Figure 1 – Classifying policy-related conclusions in congressional reports



But this typology is more than a descriptive enterprise, because it is clear that each kind of conclusion presents different costs and benefits. Most obviously, the conclusions involve different potential for error. A discrete piece of information like a finding is subject to simple measurement error. It may be that the actual amount of oil in the SPR is slightly slower or higher. Causal inference is harder. It involves both the measurement error of the finding it is based on, and modeling error—some mis-specification of the relationship between a policy and its supposed effects for numerous possible reasons, like omitted variable bias, the violation of temporal precedence, un-modeled non-linearity, and more. Maybe some other unmeasured policy or intervention caused the level of oil, or perhaps the policy came *after* the SPR reached the level measured. Even if there is no measurement error of the underlying oil volume, this kind of error can lead to a mistaken inference.

Predictions involve all the error of findings and evaluations, plus the error inherent in predicting out of sample. As hard as it is to determine the effects of the policy, it is harder to project forward because it often involves projecting other relevant circumstances. Naturally, making a definitive recommendation raises the greatest potential for error. Beneath every recommendation is some amount of measurement, modeling, and prediction error. A related point is that because evaluations, predictions, and recommendations are successively more complicated and difficult, they are require more effort and work on the part of the investigator. It requires far more skill and effort to make a definitive recommendation about a course of action than it does to describe a fact. In this way, we see this typology as a way of describing how congressional reports vary in both their costliness and informativeness, which may incentivize specialization in the first place (Diermeier and Fedderson, 2000).

Put differently, the cost of different kinds of information is linked fundamental problems of research design and the research process. Another important point follows, then, from these general results in research methodology. In expectation, the more complex the conclusion, the more likely it is to be wrong. When reports stick to basic findings, the list of potential mistakes is shorter than if they venture into evaluations, predictions, and recommendations. We return to this point in our discussion, as it has implications for the quality of the information produced by committees.

More complicated conclusions also leave more room for disagreement, especially the kind that is political. It is not just that the potential for error is more significant. The underlying logic and selection of relevant facts to support something like a prediction are more disputable than a simple finding. Two non-partisan researchers might disagree on whether it was, in fact, the Biden administration's policies that led to the amount of oil held in the SPR—even if they agree upon the level itself. They can disagree for reasons that have nothing to do with their individual incentives, like the list of relevant factors to include in a model of storage levels.

Again, causal inference is just harder. Thus, evaluations should be more difficult to agree upon than findings, predictions more difficult than evaluations, and recommendations more difficult than predictions. The more structurally complicated a conclusion, the more difficult it should be to come to consensus on it among the investigating parties.

This raises the obvious question, which is: why would anyone go out on a limb to evaluate, predict, or recommend? The short answer is: because these conclusions potentially more useful and influential than simply reporting findings. The audiences outside of the committee are, ultimately, looking for actionable advice about what to do next. This is the underlying legal justification for why committees conduct oversight at all, it must serve some legislative purpose, no matter how diffuse and (Levin and Bean, 2018). If investigators are unable to come to concrete recommendations, or even assess the effects of a past policy, they are leaving the rest of Congress, the private sector, or the executive branch to draw their own inferences and decide on the menu of potential future policies to undertake, if any. Thus, if a committee confines itself to the safest, lowest cost conclusions, it is also potentially limiting its influence. Moreover, the authors of a committee report may have to contend with other, competing reports issued from within their own committee.

In summary, as we move from findings to evaluations, to predictions and finally recommendations (left to right in Figure 1), the required effort to produce, potential error, and potential for influence should all increase. This brings us to the essential, strategic questions we posed in the previous section: how do the conclusions bipartisan and partisan teams come to differ?

Bipartisan teams, we argue, can be expected to focus on retrospective judgements, relative to their single-party reporters. The logic here is over-determined. The fact that a bipartisan coalition has come together to inform the floor is itself an indication of consensus. For the committee to form one, the conclusions it reports must be sufficiently uncontroversial. Moreover, as the report is authored, disagreements that arise will either rupture the team—rendering the final report no longer bipartisan—or be suppressed in service of maintaining the coalition. Because predicting the future is inherently more complex and disagreeable, we can expect bipartisan teams to stick to the past.

Partisan reports, on the other hand, will focus more on the future. These kinds of conclusions spark disagreements that generate one-party reports in the first place. But partisan teams may also be driven to make more complicated conclusions after considering how their audience will receive them. It is well-known that signals from extremists are less informative (Crawford and Sobel, 1982; Krehbiel, 1991). In this case, committees release reports to be read and considered by the rest of Congress. Their information may be ignored, contradicted, or acted upon. That is, it will be reviewed ex post. As Patty and Turner (2021) show, for example, this can create incentives for experts to "cry wolf" or exaggerate the extremity of their conclusions. The rest of Congress knows that partisan teams are less representative of broader interests. But it also knows that predicting the future is more costly, which lends more credibility to their conclusions. Put differently, partisan teams are motivated to venture into prognostication by the need to be taken seriously by outsiders. They know the fact that they were unable to work across the aisle risks their information being ignored, so they draw conclusions that, if true, would be consequential for Congress to ignore. This helps them make up for their credibility issue.

Before turning to our evidence for this argument, it is worth noting an important implication. All else equal, reports from bipartisan teams, should contain more accurate conclusions. Others have made this argument, but for different reasons. Some have argued, for example, that because bipartisan teams collect different skills, expertise, and worldviews, they simply produce higher quality work. Relatedly, we might expect bipartisan reports to have twice the resources, on average, devoted to them, which would lead to more reliable conclusions. We show here, however, that partisan reports would be less reliable, even if partisan teams brought to bear equally diverse staff and parity in material resources. Political disagreements mean that partisan teams are driven to draw conclusions that are more unreliable by their epistemological nature.

Data and Measurement

To evaluate the nature of messages transmitted by committees, we mined a new dataset of Congressional oversight reports from the 106th to 116th Congress, compiled by the Levin Center for Oversight and Democracy.[2] These include 803 total reports produced by the two committees most responsible for legislative oversight: Senate Committee on Homeland Security and Governmental Affairs (HSGAC) and the House Committee on Oversight and Accountability (COA), along with their subcommittees and individual members. These reports involve official fact-finding activities, and are distinct from filings made in connection with specific legislation. Beyond the content of each report, the data also contain information on the report's author, publication date, and author partisanship.

These data offer several advantages. Most importantly, they allow us to examine the conclusions of investigators—the actual messages that personnel on the committee choose to deliver—rather than the investigative activity itself. There is reason to suspect that, ex ante, the kinds of conclusions contained in these reports are more measured and evidence-based than claims made in open hearings. The data span 1999-2021, meaning that there are reports published during presidential administrations of both major parties as well as during periods of unified and divided government.

One immediate question is whether data from two committees are indicative of information transfer in Congress, in general. We think they are indicative of a particular kind of information gathering and transfer. Since these reports concern oversight, the dataset excludes all reports tied to a specific piece of legislation. Most obviously, when committees author reports that concern future legislation, it's reasonable to assume that their tendency will be to focus on prospective, rather than retrospective information. Thus, the baseline descriptive figures we present later would likely differ significantly. It is also possible that consensus-building in the very public process of passing legislation is different from consensus-building in the very

[2]These can be found at: `https://cord-levin-center.org/home`. For a fuller description of their methodology for gathering and including reports, we refer you to their codebook.

private process of report-drafting. Our results are therefore only generalizable to oversight of policy, rather than the production of new policy. The data also pose a basic research challenge, which we address in the following section: Many reports are hundreds of pages long, which makes them difficult for humans to parse with speed and accuracy.

Identifying Conclusions

Our basic measurement challenge is to extract and classify informational conclusions contained in oversight reports. To accomplish this, we turn to a large language model (LLM). With text-as-data becoming increasingly common in computational social science (Egami et al., 2022), scholars have touted generative artificial intelligence, and LLMs in particular, as efficient tools for automating tasks often delegated to research assistants. Recent work has established that LLMs can parse corpora and summarize their contents with high accuracy (Bail, 2024; Ziems et al., 2024). If not properly trained, however, they may "hallucinate" (generate false information) or amplify harmful stereotypes (Chang et al., 2024).

Nonetheless, there are several advantages to using a LLM. The sheer quantity of data (over 73,000 pages of text and figures) would require assembling, training, and monitoring a team of research assistants (Goehring, 2024). The amount of time and resources involved would be substantially higher without the use of generative AI. Moreover, while both research assistants and LLMs are vulnerable to making coding errors, LLMs typically make the same class of mistakes repeatedly, whereas humans' mistakes vary more widely and are less predictable. It is therefore more efficient and reproducible to train a LLM to correct a systematic error than a single research assistant.

We used Open AI's Chat Generative Pre-Trained Transformer 4o (GPT4o) to parse the content of each report. We began by developing a prompt. Specifically, we attempted to follow best practices recommended by Lee et al (N.d.),by clearly articulating the task's context and objective as well as our desired output format. We included in our prompt clear definitions for each kind of statement as well as phrases commonly used to introduce them. We then randomly sampled 25 reports, and both authors hand-coded each report, which we then compared to results produced by GPT4o. The full prompt can be found in Appendix A of the Supplementary Information.

Ultimately, the model identified more than 12,000 unique statements from the reports. The model was relatively successful in extracting relevant statements from the reports. We encountered several classes of errors that occurred in our initial trial of 25 reports. The first, and arguably most concerning, were hallucinations. Many of the documents in our dataset were committee activity reports, which summarized a committee's actions over the course of their investigation. These reports typically do not contain any of the kinds of statements in which we are interested. Yet, GPTo falsely identified multiple statements of each kind for each of these reports. The second was a category error: GPTo often mislabeled findings as policy evaluations, likely because the two types of statements often involve similar language and syntax.

Both of these errors were attributable to limitations in the initial user-supplied prompt. We did not tell the model what to do when documents do not contain any relevant statements, so the model made them up. And findings, though similar in language to evaluations, differ substantially in context. In subsequent trials, we addressed these shortcomings by instructing the model not to parse activity reports and more clearly articulating the definition of a finding as opposed to an evaluation. We then repeated the process of hand-coding and comparing with GPT's output until we were able to minimize systematic errors.[3]

[3] This study remains a work in progress, and we plan on implementing additional changes to further validate the model in future versions of this work. Several parsed reports contained typographical errors, and others contained slight misclassifications. Future analyses may benefit from separating the trials by type of statement: the model will first identify all findings, *then* policy evaluations, etc. Using four separate prompts may reduce contamination across statement types.

Results

We find evidence in congressional oversight reports that reveal partisan conflict changes the kind of information transmitted to the rest of Congress. We firste examined whether simple empirical patterns are consistent with our argument and the typology it rests on. In particular, we argued that findings, evaluations, predictions, and recommendations should be successively more costly and error-prone. All else equal, we would thus expect findings to outnumber evaluations, evaluations to outnumber predictions, and predictions to outnumber recommendations.

This is mostly what we find. In particular, of our 12,619 conclusions contained in 731 parsible reports, 41.3% (5,216) are findings, while 29.6% (3,744) are evaluations.[4] However, recommendations did slightly outnumber predictions, at 14.7% (1,858) to 14.2% (1,801). Overall, this confirms the basic premises of our typology. More complex conclusions rest on a foundation of more basic, uncontroversial ones. Findings are easier to agree upon and obtain, therefore, they are the most prominent type of information.

Next, we examine whether there is evidence that report authorship impacts the kind of information communicated. As a first cut, we classify reports by the staff who wrote them, and the members that approved them. Some reports are co-authored by the majority and minority, some are solo authored. This gives us both the status and partisanship of authorship teams, which we report in Figure 2. Specifically, we report the raw frequency distributions for each type of conclusion, based report author.

A few stylized facts are worth noting. First, regardless of author, reports tend to be more retrospective than prospective. However, there is non-trivial variation in information type. Bipartisan reports appear to be driving the fact that recommendations are more prevalent than predictions. (In fact, if these reports were removed from the sample, predictions would

[4]Note, 72 reports were excluded either because they were committee activity reports, or their text contained too many errors related to optical character recognition, making them unintelligible to GPTo.

outnumber recommendations.) In addition, minority reports tend to focus on the future, relative to the majority and bipartsan teams. This is driven not by recommendations but their propensity to predict the future.

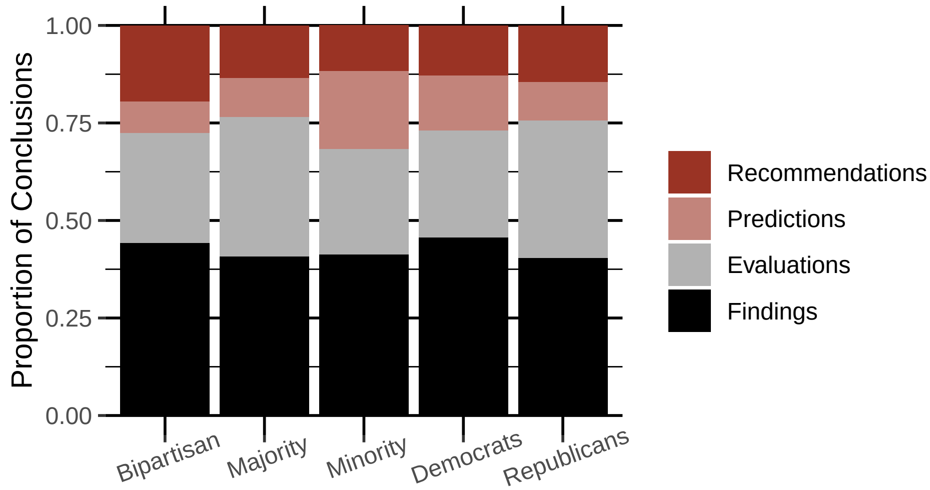


Figure 2 – Investigations tend to be more retrospective, but there is variation depending upon the author. Reports the proportion of conclusions classified as either a finding, evaluation, prediction, or recommendation, by the author of the report. Excludes 84 reports which are either committee activity reports or full committee reports without a designated author.

Raw frequencies, of course, cannot tell the whole story. They are partly a function of circumstances and events beyond the committees' control, as well as the partisan makeup of Congress at the time. They do not distinguish between oversight that targets an opposition or co-partisan administration. Most importantly, these authorship classifications obscure a lot of variation in disagreement between the floor and committee. Some ranking members of committees are more or less extreme, and the median in chamber routinely shifts.

To account for these factors, we model counts of each conclusion type separately as a function of political circumstances. Specifically, we predict the number of conclusions as a function of the ideological distance between the author and the chamber median, whether the report was authored by co-partisans of the president, the opposition party, or was bipartisan, the chamber of the author, and whether the report was authored under divided government. We use the count of conclusions for several reasons. First, in general, we found that the number of conclusions was not sensitive to apolitical features of reports, like page length, chamber, and committee. This makes a spurious relationship between our political covariates driven by these other factors unlikely.

Second, estimating separate count models imposes the fewest modeling assumptions on the relationship between the counts. If there is a substitution effect between categories of information, it will emerge from the data summary, not be assumed. Finally, we can estimate predicted counts, which tend to be more interpretable than models for unordered factors. There are downsides to this measure. In particular, it says nothing about the quality of individual conclusions, treating eat as equivalent within category. Thus, while our primary results rely on counts, we later move beyond these to investigate the specificity of conclusions.

Our primary covariate of interest is the ideological distance between the relevant committee leader(s) and the chamber floor. Chairs and Ranking members are closest politicians to report authorship, since they hire and direct the staff who conduct investigations and write reports. Our measure is the absolute distance in DW-NOMINATE score between the report author(s) and the chamber median (`voteview.com`). For each report, we first identify the authors. We assign either the chair or ranking member as majority and minority author, respectively. If the report is bipartisan, take the midpoint between the leaders as the author score. We then take the absolute distance between that the author score and the chamber median in that Congress.

Our commmittee leader data come from Stewart and Woon (2005), as updated by Eldes, Fong, and Lowande (2024). There were 115 reports authored by staff in more than one committee. For these reports, we again take the midpoint between any multi-author team, then the absolute distance between that hypothetical median and the floor median. Ultimately, this variable amounts to a continuous specification of the categorical authorship on display in Figure 2. Bipartisan reports tend to be close to the chamber median, with a mean NOMINATE distance of 0.21, whereas one-party reports are more than double that distance, at 0.5 on average. Minority reports, naturally, show the most disagreement, with 0.65 average NOMINATE distance to the floor.

Since these are congressional oversight reports, it is important to account for the relationship between the committee author and the executive branch. Though some of these reports target other entities in Congress or the private sector, often the target is an organization ostensible under the control of the president. For example, the opposition party might focus intently on evaluating the policies of the sitting president, hoping to score political points while ignoring constructive conclusions like recommendations. Each of these suggest the intended audience for reports is not the floor, as we have argued. Instead, in this reading, reports would be about the broader political combat between parties in government. If they predict the content of reports' conclusions, our informational theory of committee-floor relationships may simply miss the mark.

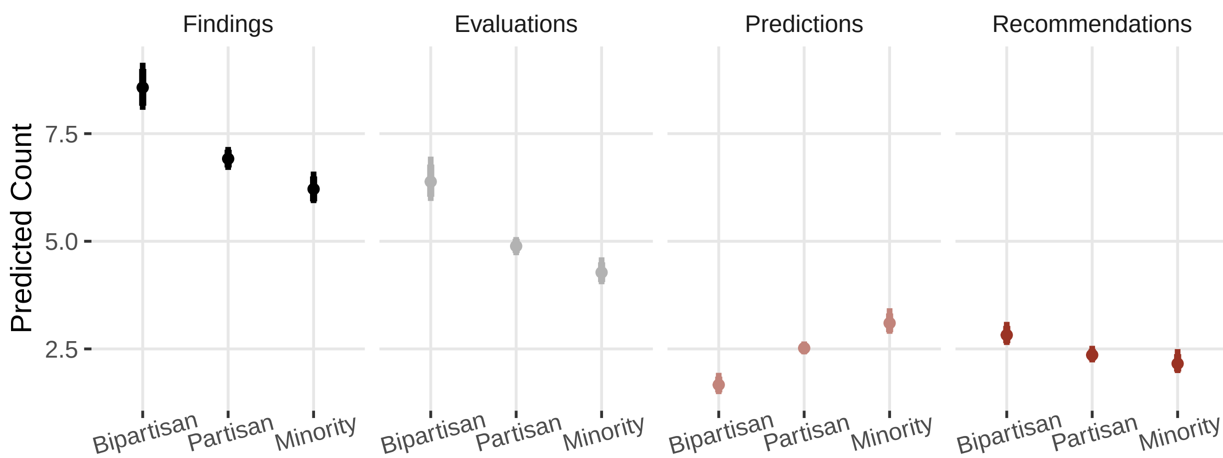


Figure 3 – Bipartisan teams tend to be more retrospective, while partisan teams are more prospective. Reports estimated counts of findings, evaluations, predictions, and recommendations, simulated from Poisson regressions that predict the count of each conclusion type with NOMINATE distance to the floor, along with binary indicators for opposition author, copartisan author, House committee, and divided government. Bipartisan, partisan, and minority estimates constructed by setting the value of NOMINATE distance to the chamber floor to the mean observed value for each type of staff team. Full model results reported in Table B.1.

We report the test of our main hypothesis in Figure 3, which plots estimated counts and standard errors for each conclusion type at three levels of distance to the floor: bipartisan author (0.21), partisan author (0.5), and minority party author (0.65). These estimates are simulated using an observed case approach from Poisson regressions which predict each dependent variable as a function of the five covariates we describe above, with standard errors clustered at the Congress level. Several important patterns emerge from these estimates.

First and foremost, as expected, bipartisan reports tend to be more retrospective than partisan reports. The typical bi-partisan report contains about 8 findings and 7 evaluative conclusions, whereas partisan reports contain around 1 fewer findings and 2 fewer evaluations. The difference is starker for reports authored by the minority party. The pattern is the reverse for predictions, with minority reports tending to make an additional 1.5 predictions, relative to bi-partisan reports. This pattern is not replicated for recommendations. Reports, regardless of author, tend to contain about 2.5 of them. It is worth noting that, if anything, these differences under-estimate the relative effort invested in each kind of informative signal. After all, partisan teams work with half the staff resources as their bipartisan counterparts. We therefore might expect them to have fewer conclusions as a function of their capacity. Yet, they issue *more* predictions and just as many recommendations, underscoring how much of their attention is focused on the future.

We take this as evidence consistent with our argument. Bipartisan teams, driven by the need to maintain consensus, tend to stick to less complex informational signals like findings and evaluations. Partisan ones, driven by the need to influence the decisions of the floor, focus on more complex informational signals like predictions.

There is also not strong evidence that the type of information is informed by inter-branch conflict. Authors made up of only copartisans of the president write reports that include about as many findings, evaluations, and predictions as their bipartisan, and even opposition party counterparts (Table B.1. Reports issued when the party controlling the chamber differs from the presidency do tend to contain fewer conclusions, in general. One potentially important difference, however, is worth noting and investigating further. We find that the opposition party, especially during divided government is less likely to include recommendations in their solo-authored reports. Specifically, their reports contain about 1.5 fewer recommendations, relative to bipartisan reports.[5]

[5]This estimate is not statistically distinguishable, by convention, from reports authored only

Conclusion Specificity

We have shown that partisan coalitions impact the focus of committee reports in terms of the raw frequency of conclusion types. This is consistent with the idea that some types of conclusions are more costly to transmit than others. However, it is still possible that some conclusions *within* type are more costly than others. Consider the following two conclusions: "policy X will have negative effects," and "policy X will increase inflation by 3.2% in the next fiscal year." Both are clearly policy predictions, but the former is less precise. This kind of imprecision indexes uncertainty, and may also be cheaper to develop. Without a measure of how specific a finding, evaluation, prediction, or recommendation is, it is difficult to determine precisely how costly it was for members of congress to transmit.

Moreover, it is easy to imagine how the specificity of conclusions could interact with their frequency in ways that cut against our theory. While bipartisan teams tend to be more retrospective than partisan ones in terms of frequency, suppose their retrospective conclusions are less specific, on average. The consensus-building of bipartisan teams might, in essence, produce watered-down conclusions. Likewise, their recommendations might omit clear, actionable steps, their evaluations might be difficult to falsify. Each of these interactions would change the implication for Congress' welfare, and for the process of information transfer. Bipartisan teams might write a lot without saying much, which would be contrary to our argument.

To investigate this, we first need to be specific about what it means to be specific. We apply a basic measure of conclusion specificity to help disentangle which is true. For each conclusion in our dataset, we compute a specificity score as follows:

$$\text{specificity} = \left(\frac{\text{conclusion word count}}{\text{conclusion sentence count}}\right) \cdot \text{conclusion average word length}$$

A manual check of the conclusions associated with the extreme values confirmed that conclusions with more concrete details and tangible claims scored higher on the specificity

by copartisans, however.

index than those without.[6]

We then tested whether accounting for conclusion-level specificity affected our main results. We computed new counts for the number of findings, evaluations, recommendations, and predictions within each report, weighting them by their specificity score. Figure 4 plots the results of our primary analysis, controlling for the specificity of statements contained within the reports.

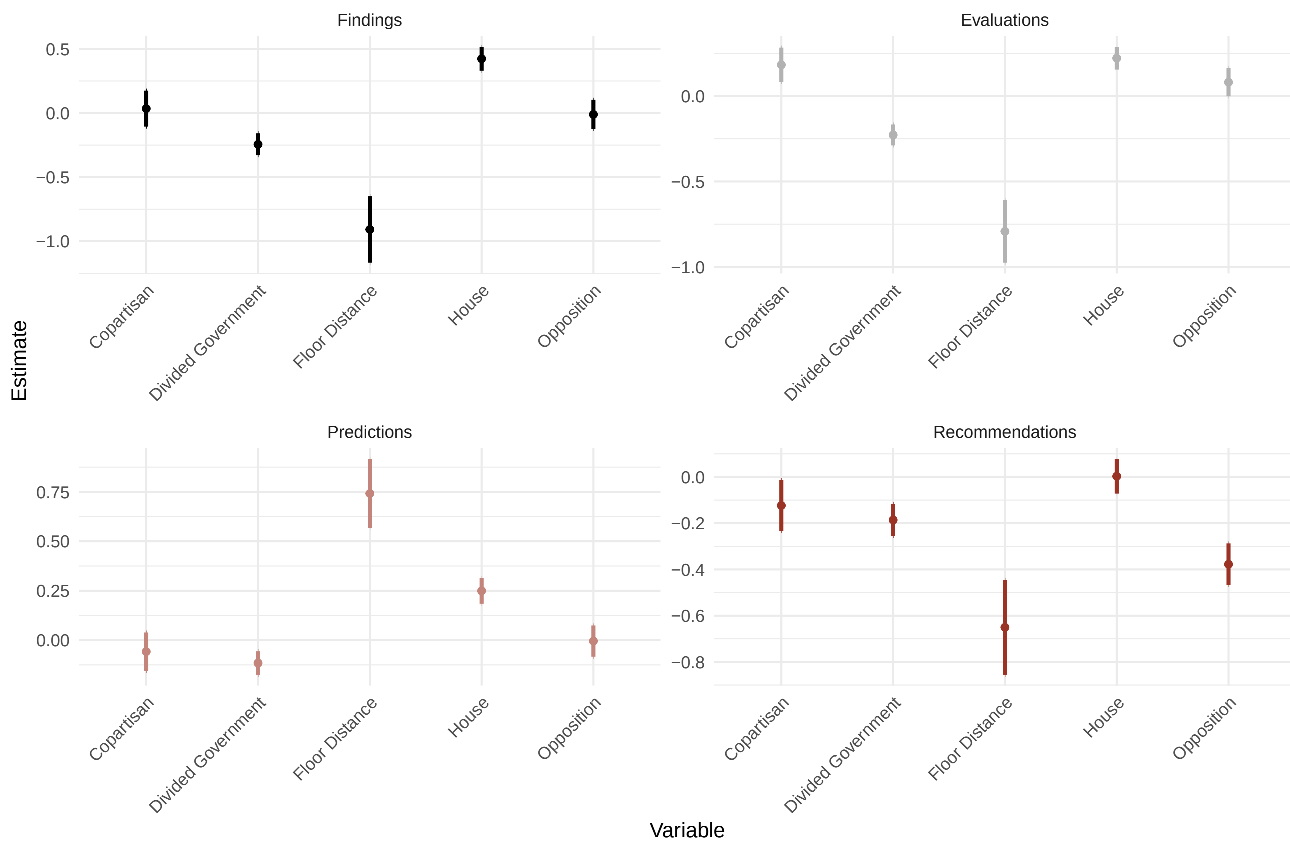


Figure 4 – Controlling for specificity does not change observed information transmission patterns. Reports the effect of political covariates on the weighted counts of each type of conclusion in the report, fit using a left-censored tobit regression model. Bipartisan, partisan, and minority estimates constructed by setting the value of NOMINATE distance to the chamber floor to the mean observed.

The empirical patterns observed before largely hold. Findings and evaluations become

[6]This score is an admittedly imperfect measure, as a conclusion's specificity score is highly correlated with its length in characters ($\rho = 0.62$).

less likely in cases of divided government and as the distance from the floor median increases. On the other hand, predictions become more likely and recommendations less likely as the distance from the floor median increases. Overall, we take this as evidence that no kind of team is systematically making more or less specific conclusions across their reports. Counts for each type of conclusion are therefore an appropriate measure of the amount of information transmitted by publishing a report.

Surprisingly, however, the coefficient estimates for the distance between the floor median and the presence of divided government are significant predictors of the number of recommendations present in a report, controlling for specificity.[7] As the distance from the floor median increases, the number of recommendations when accounting for specificity *decreases*. This suggests that recommendations are not tools for sincerely communicating actionable steps toward implementing policy.

Discussion

In this study, we provide evidence that partisan considerations affect information transmission in Congress. Oversight committees are aware of the political environment in which they exist and tailor their messages accordingly. Bipartisan teams, driven by the need for consensus, produce more findings and evaluations relative to minority reports, which tend to focus on prospective conclusions like predictions. These patterns are largely unaffected when controlling for the level of specificity of each conclusion.

We acknowledge several key limitations to address moving forward. At the time of writing, our LLM prompt is a work in progress. Several minor issues remain: the model occasionally selects fragments from reports that are difficult to interpret out of context, for example. Some conclusions contain typographical errors, and others are not conclusions at all, but rather titles of documents. Such irregularities are rare and therefore do not threaten our research design, but they do point to a need to further refine and validate our prompt in fu-

[7]The direction of the coefficients is unchanged from the previous analysis.

ture trials. The specificity score would also benefit from more careful construction. Its high correlation with a conclusion's length in characters raises questions about whether it is truly capturing the specificity of the conclusions' content, or if conclusion length is the main driver of those results.

Future work should also investigate the target or subject of each conclusion. While we provide evidence that a conclusion's specificity generally does not affect its likelihood of being transmitted by a committee, we are unable to determine if this changes when the report is directed at the conduct of a Congressional committee instead of the president, for instance. Incorporating the target into our work can help clarify the conditions under which politics does and not affect information transmission in Congress.

References

Bail, Christopher A. 2024. "Can Generative AI improve social science?" *Proceedings of the National Academy of Sciences* 121(21):e2314021121.

Ban, Pamela, Ju Yeon Park and Hye Young You. 2021. "How Are Politicians Informed? Witnesses and Information Provision in Congress." *American Political Science Review* 117(1):122–139.

Callander, Steven. 2008. "A Theory of Policy Expertise." *Quarterly Journal of Political Science* 3:123–140.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang and Xing Xie. 2024. "A Survey on Evaluation of Large Language Models." *ACM Transactions on Intelligent Sysems and Technology* 15(3):39:1–39:45.

Crawford, Vincent P. and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50(6):1431–1451.

Diermeier, Daniel and Timothy J. Fedderson. 2000. "Information and Congressoinal Hearings." *American Journal of Political Science* 44(1):51–65.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts and Brandon M. Stewart. 2022. "How to make causal inferences using texts." *Science Advances* 8(42).

Eldes, Ayse, Christian Fong and Kenneth Lowande. 2024. "Information and Confrontation in Legislative Oversight." *Legislative Studies Quarterly* 49:227–256.

Fenno, Richard F. 1973. *Congressmen in Committees*. Little, Brown and Co.

Fong, Christian, Kenneth Lowande and Adam Rauh. 2024. "Expertise Acquisition in Congress." *American Journal of Political Science* pp. 1–14.

Gilligan, Thomas W. and Keith Krehbiel. 1987. "Collective Decisionmaking and Standing Committees: An Informational Rationale for Restrictive Amendment Procedures." *The Journal of Law, Economics, and Organization* 3(2):287–335.

Goehring, Benjamin. 2024. "Improving Content Analysis: Tools for Working with Undergraduate Research Assistants." *PS: Political Science & Politics* 57(1):57–63.

Kiewiet, D. Roderick and Mathew D. McCubbins. 1991. *The Logic of Delegation*. American Politics and Political Economy Series Chicago, IL: University of Chicago Press.

Krehbiel, Keith. 1991. *Information and Legislative Organization*. University of Michigan Press.

Kriner, Douglas L. and Eric Schickler. 2014. "Investigating the President: Committee Probes and Presidential Approval, 1953–2006." *The Journal of Politics* 76(2):521–534.

Lee, Kyuwon, Simone Paci, Jeongmin Park, Hye Young You and Sylvan Zheng. N.d. "Applications of GPT in Political Science Research." *Working Paper.*. Forthcoming.

Levin, Carl and Elise J Bean. 2018. "Defining Congressional Oversight and Measuring Its Effectiveness." *Wayne Law Review* 64:1–22.

Lewallen, Jonathan, Sean M. Theriault and Bryan D. Jones. 2016. "Congressional dysfunction: An information processing perspective." *Regulation and Governance* 10:179–190.

Lowande, Kenneth. 2018. "Who Polices the Administrative State?" *American Political Science Review* 112(4):874–890.

Lowande, Kenneth. 2019. "Politicization and Responsiveness in Executive Agencies." *The Journal of Politics* 81(1):33–48.

Lowande, Kenneth and Justin Peck. 2017. "Congressional Investigations and the Electoral Connection." *The Journal of Law, Economics, and Organization* 33(1):1–27.

Lowi, Theodore J. 1969. *The End of Liberalism: The Second Republic of the United States*. W.W. Norton Company.

Park, Ju Yeon. 2021. "When Do Politicians Grandstand? Measuring Message Politics in Committee Hearings." *The Journal of Politics* 83(1):214–228.

Park, Ju Yeon. 2023. "Electoral rewards for political grandstanding." *Proceedings of the National Academy of Sciences* 120(17):e2214697120.

Patty, John W. and Ian R. Turner. 2021. "Ex Post Review and Expert Policy Making: When Does Oversight Reduce Accountability?" *The Journal of Politics* 83(1):23–39.

Ritchie, Melinda N. 2018. "Back-Channel Representation: A Study of the Strategic Communication of Senators with the US Department of Labor." *The Journal of Politics* 80(1):240–253.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang and Diyi Yang. 2024. "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics* 50(1):237–291.

Supplementary Information

The Politics of Predicting the Future

Kenneth Lowande and Mark Weiss

Table of Contents

A Prompt

"Your task it is to extract statements from government reports. The user will provide the texts of reports. They will paste the text one at a time. You will extract four types of statements: policy predictions, policy evaluations, findings, and policy recommendations.

Define policy predictions as evaluations of proposed or potential policies. These statements try to estimate the future effects of implementing new policies, bills, laws, or regulations. Predictions must make concrete and testable claims about the effects of a policy. Look for phrases like "would," "will," "cause," "will lead to," "projected," "estimated."

Define "finding" as a statement of fact that does not make recommendations, predictions, or evaluate the effects or assess of a policy. It merely communicates determinations come to by the authors of the report. Note, many of the reports will contain enumerated lists of findings. Some of these qualify by our definition, but some will actually be what we have defined as "policy evaluations" or "policy predictions."

Define policy evaluations as retrospective assessments of existing policies, bills, laws, or regulations. These are statements of fact that assess the impacts of policies that have already passed. Look for phrases like "led to," "caused," "as a result," "impacted," "resulted in."

Define policy recommendations as those policies, bills, laws, and/or regulations that the authors of the report believe should be implemented. Look for phrases like "should," "ought to," "recommend" or "recommended." If the authors quote someone - an individual, a firm, or an interest group, for example - that expresses support for a policy, but the authors themselves do not endorse it, then the sentence does not count as a policy recommendation.

Some of these reports are summaries of certain committees' activity. These documents will not contain any of the types of statements in which you are interested. Do not parse reports of this kind. Extract all quotes that make policy predictions, policy evaluations, or policy recommendations. When isolating these quotes, please include the full statement in which they appear and make note of the name of the file from which it came.

Format your response as follows: Document title - file title - number of quote or information - type of statement. Store this information in memory; do not print it back to me. Please make sure you include all relevant statements from each document. After analyzing a document, reread it and extract any additional policy predictions, policy evaluations, or policy recommendations missed the first time. Only include additional statements from the second read if they are novel and meet the definitions provided.

Once you have received the text document, print the number of each kind of statement found in the document. Then, export the saved information to a csv file with the following information as columns: file title, quote or information, type of statement. Also generate a unique numeric value called report id for each file, such that if one file has two quotations in the data, those two rows should have the same value for report-id. Column names should be named as follows, in this order: file, report-id, quote, state-type-gpt. Do not print your output; just perform the task."

B Full Regression Results

Table B.1

	Dependent variable (counts):			
	Findings	Evaluations	Predictions	Recommendations
	(1)	(2)	(3)	(4)
Floor Distance	−0.752**	−0.935***	1.454***	−0.626
	(0.342)	(0.284)	(0.526)	(0.463)
Copartisan	0.046	0.274*	−0.037	−0.219
	(0.179)	(0.152)	(0.327)	(0.294)
Opposition	−0.010	0.107	0.005	−0.607***
	(0.129)	(0.146)	(0.256)	(0.205)
House	0.471***	0.309*	0.522***	−0.087
	(0.147)	(0.160)	(0.170)	(0.231)
Divided Government	−0.299**	−0.293***	−0.103	−0.306*
	(0.131)	(0.096)	(0.143)	(0.174)
Constant	2.094***	1.846***	−0.138	1.716***
	(0.140)	(0.102)	(0.256)	(0.170)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table B.2

	Dependent variable (weighted counts):			
	Findings	Evaluations	Predictions	Recommendations
	(1)	(2)	(3)	(4)
Floor Distance	-0.908^{***}	-0.792^{***}	0.742^{***}	-0.650^{***}
	(0.259)	(0.184)	(0.175)	(0.205)
Copartisan	0.035	0.183^{*}	-0.058	-0.124
	(0.140)	(0.101)	(0.097)	(0.110)
Opposition	-0.011	0.081	-0.005	-0.378^{***}
	(0.115)	(0.082)	(0.079)	(0.090)
House	0.424^{***}	0.222^{***}	0.250^{***}	0.003
	(0.093)	(0.067)	(0.065)	(0.075)
Divided Government	-0.244^{***}	-0.228^{***}	-0.116^{*}	-0.186^{***}
	(0.086)	(0.061)	(0.059)	(0.069)
Constant	0.952^{***}	0.813^{***}	-0.272^{***}	0.658^{***}
	(0.111)	(0.080)	(0.078)	(0.086)
Note:				*p<0.1; **p<0.05; ***p<0.01