We thank the editors for the opportunity to revise and resubmit our manuscript. In this memo we have separated the editor's and reviewers' comments into separate criticisms and suggestions. Under each comment, we describe how we have revised the manuscript in response to the feedback provided. We generally agree with the criticisms offered, and think that our manuscript has improved substantially as a result of incorporating this feedback.

Editor

E1 *R1 and R2 see a lot of potential in this manuscript and point you in some ways to improve it. We think following their suggested course of action will also speak to the concern of R3 that there is not much of a contribution here. While we do not think there is a problem with replicating and extending existing work, we agree with R3's broader point that you should do more to make clear what your intellectual contribution is in this manuscript; what are scholars learning here that they could not already get from the Bowers article?*

Addressed: In addition to making several specific improvements in response to the reviewers' comments, which we document below, we have revised the introduction to clarify the overall contributions offered by this paper. First, we provide a useful theoretical typology that researchers can draw upon when formulating hypotheses regarding interference. Second, we provide a review of the types of networks through which scholars in legislative research might consider testing for interference. Third, we illustrate the use of this methodology through two empirical applications (i.e., via the replications).

Reviewer: 1

R1.1 *Overall, I like this paper, and I think it's a good fit for SPPQ. The literature on elite experiments is growing quickly, and it's useful to both combine recent innovations into a single paper, and to demonstrate how they can be applied to a number of well-known experiments. I'm not really convinced that this paper demonstrates any substantial spillover effects, but it does show how one could test for such effects, and I could see the methodology being applied to other field experiments where spillover was more likely (either previous studies, or future studies).*

Addressed: We appreciate R1's encouragement. To be completely transparent, we have revised the introduction to note that we do not find universal evidence of spillover/interference effects in our replications.

R1.2 *I'd like the author to make the unique contribution of this article more clear. It's sometimes difficult to tell how much of the method is taken from other work,*

and how much is new suggestions. The additional empirical tests are clearly new, but beyond that it gets a bit fuzzy at times. Much of the paper borrows methodology from other publications (with citations throughout). I'd like the original contributions of this paper to be made just as clear.

Addressed: We revised the introduction to clarify this point (see response to E1), and noted in Section 3, "Considerations in Testing for Interference," that the set of recommendations we provide represents a novel contribution. Our main conceptual/methodological contributions are all presented in Section 3. The legislative networks we suggest are drawn largely from the literature, and the modeling choices we present are also drawn largely from the literature on network dynamics, but, to our knowledge, this information has not previously been consolidated into a guide on structuring field experiments to study interference effects in legislatures.

R1.3 *2A. I like the discussion of potential networks that might commonly lead to interference and spillover among state legislators. This is my favorite part of the paper, as it has the potential to guide future research. Again – I'd like the author to make clear how much of this is original suggestion versus collecting what's been done elsewhere. Either is worthwhile, in my opinion.*

Addressed: We have revised the paper to make clear that we draw this list of potential networks from the literature (i.e., we do not offer any new recommendations). We also organized this list into a table, and added a few new networks from recent studies.

R1.4 *2B. I'd like to see more discussion about when each type of network might be most likely to have spillover. For example, we might expect some information to be shared with anyone you contact (in which case committee co-members would be good); whereas in other cases, we might expect information to be shared most with those who are ideologically close to the legislator (in which case ideological proximity would be a more likely choice). Some guidance to authors in the future along these lines would set the piece up to make a bigger contribution to the future literature in this area.*

Addressed: We have added a paragraph to the discussion of legislative networks in which we provide guidance regarding which network(s) may be most relevant to spillover effects. We introduce the concepts of exposure and uptake. Explicit communication ties are important in cases where the network would control awareness of other legislators' beliefs or behaviors (i.e., exposure). Ideological similarity is important in cases where the network would control whether or not a legislator would adopt another's beliefs or behavior if aware of what the other

legislator had done (i.e., uptake).

R1.5 *How should one go about choosing which networks to test for spillover? Should we always test all of them?*

Addressed: We have added a paragraph to the manuscript in which we discuss this issue. We recommend (1) combining networks through a composite measure, (2) including more than one network in a single interference model, and/or (3) evaluate separate interference models for different networks. We note that researchers should be mindful of multiple testing bias in this process, and indicate that it would be valuable in future methodological work to develop a Bonferroni-style adjustment for the BFP methodology.

R1.6 *2C. The author suggests two additional network alternatives – based on party share, and percent black population. These should be integrated into the list of suggestions on pages 8-9.*

Addressed: We have removed the Broockman replication, in which these networks arise, to the appendix, but we added a reference to the Bratton and Rouse piece that studies demographic similarity between legislators' constituencies to our discussion of options for legislative networks.

R1.7 *2D. I think an additional network that could be identified would be a combination of committee sharing with shared partisanship. I imagine sharing a committee gives one opportunity to share, and sharing a party gives one incentive to share. The combination of the two would be even more powerful – people you agree with, who you interact with often.*

Addressed: This is an excellent suggestion, and we incorporated this network into the Butler and Nickerson replication.

R1.8 *3A. 5.1 – Butler, Nickerson, et al*
The description of this experiment is ambiguous to me, and I have trouble following the discussion in the new analyses of the data. The author describes that the treatment increased support for the rebate overall "since the vote was popular", but that the treatment also decreased support in districts with low support. The discussion on page 16 is hard to follow – I can't tell when the author is analyzing only districts where the treatment indicated support, only those where the treatment indicated opposition, or all districts together. It seems the original article did at least two of the three groups. I'm not sure which ones were replicated, so it's hard to follow the comparisons. Also, on page 14, the author says the treatment had a significant effect, but on page 17, says the original effects were not significant. Lastly, do we need all these figures? They look very

similar to each other, and take up a lot of space.

Addressed: We agree that the description of this study in our initial submission was confusing. We have revised the description of the original study, the description of what we do, and actually how we model interference in this application, to fit more closely with the heterogeneous effects that were hypothesized and tested in the original study.

R1.9 *3C. 5.3 – Broockman*
The original experiment is explained very clearly. The author provides additional suggestions on networks to test for interference (beyond those originally proposed on pages 8-9). I don't fully follow the result described on page 26. The words are clear, but I can't grasp the intuition of why we would expect that kind of spillover to happen, or why that kind of spillover follows the original study. More explanation and discussion from the author would be useful here. Why would one legislator's responsiveness to an out-of-district constituent go down if other legislators from a district with similar racial composition also received an e-mail from an out of state constituent? Is it assumed that they would share that they received the e-mail, and suspect there was an experiment? Or assume the other person would do it? I don't follow the logic.

Addressed: After seeing the reviewers' comments, we are convinced that this is the weakest replication study--both in terms of the likelihood of interference and the networks available to us to use in testing for interference. We have removed the Broockman replication study.

R1.10 *4. Overall, I find the theory and method to be more interesting than the actual empirical results – though the empirics are important to document as well. The conclusion seems to jump into a whole new territory (talking about how relaxing SUTVA would change methodology).*

Addressed: Re-reading the introduction in the original submission, we realize that we did not note that the BFP methodology amounts to a relaxation of SUTVA, which made it seem as if we were jumping into something new in the conclusion. We have corrected this omission in the introduction and revised the conclusion to assure that we are not jumping into new material.

R1.11 *5. The manuscript is well-written overall, but there are a number of typos scattered throughout. Page 5: I believe should say "has not, as of yet" rather than "has, as of yet"*
Page 6: "The testing framework proposed by BFP 2012 is randomization test" reads poorly
Page 9: "of of"

Page 10: "For example, a state legislature is [a] relatively small…"
Page 11: "effects" should be "affects"
Page 12: Last full paragraph ends with a non-sentence (looks like first half was accidentally deleted)
There's more. I recommend proof-reading the whole draft thoroughly.

Addressed: We have proof read this draft carefully to minimize typos.

Reviewer: 2

R2.1 *1) The first section of the paper is too long. The authors can motivate the paper, discuss SUTVA, and signal where the paper is going in fewer pages. They need to do this because the second section feels out of place/is "clunky," even as it is centrally important to the sections that follow. I was able to understand the discussion/logic of the Bowers et al. (2012) randomization test, but this needs to be improved. Readers need to understand the basic intuition of comparing an observed statistic to a distribution, and the idea behind the permutations of the treatment vector. If some pages are eliminated in the first section, the authors can give a little more attention to detailing this, perhaps including a figure that shows a test statistic against the null distribution (and gives the intuition for what p-values represent here). The authors cut at all this isn't bad, but if this piece is going to be instructive, it needs to be clearer.*

Addressed: Re-reading our original section, we see how it was overly condensed and difficult to follow. We have expanded this section and revised it to follow along a toy example---conducted on an artificial dataset that is small enough to present in a table. We use this example to illustrate all of the pieces and steps in using the BFP methodology.

R2.2 *2) I wonder if the "Considerations in Testing for Causal Inference" are communicated as effectively as possible. Would a table summarizing this information be useful? (There would be room for this if the first section of the paper was cut to 3 or so pages.)*

Addressed: Responding to this comment substantially improved the presentation/ communication of our conceptual contributions. At the beginning of Section 3.2, we have added an overview in which we summarize the two dimensions fo interference model specification---neighborhood and functional form---along which our discussion is organized. We also added a table in which we include a simple visual example that highlights these two dimensions of model specification.

R2.3 *3) Sections 4 and 5 could be combined, with 4 ("research design") retitled to be more informative ("Replication Analyses: Testing for Network Effects")*

Addressed: We agree that the previous Section 4 in the initial version was just an introduction to Section 5. We have combined these sections and re-titled with R2's recommended title.

R2.4 *4) The authors should better signal why their replication improves upon the work of Coppock (2014). Also, I appreciate the authors demonstrating that not all the replications have indirect effects. Still, I think the authors could move the Bergan and Cole (2015) results to an online appendix without sacrificing much. They can still make the point in natural language that sometimes modeling for indirect effects will matter, and sometimes it will not.*

Addressed: We have revised the replication to both test different networks than those used in Coppock (2014) and use a different form of the interference model. We elected to remove the Broockman replication instead, since we could not gather much network data for that dataset, and needed the extra space in the manuscript.

R2.5 *5) By moving one of the examples to the appendix, this would free up space to talk more about some topics that are important if scholars are going to learn from the paper and put the advice into action. In the last paragraph of the manuscript the authors note that experiments really should collect data on networks of interdependence – linking this to previous discussion/elaboration here would be useful. Perhaps more importantly, the authors really should give more space to the issue of optimal randomization given interference effects (as this is pretty central to the main claims/contributions of the paper).*

Addressed: We have revamped and expanded the Butler/Nickerson and Bergan and Cole examples to incorporate more networks and more custom theoretical characterizations of the interference dynamics. We also removed the Broockman replication since we did not have much network data with which to test for interference, and it was less clear how interference could manifest in that application.

Reviewer: 3

R3.1 *In these replications, considering alleged potential spillovers, the author of this manuscript finds little differences in the results when accounting for the possibility of spillover effects.*

Addressed: We have revised the conclusion to discuss the fact that we do not find universal evidence of spillover in our replications. We do not think that null results degrade the merit of these replications, but do not dispute the fact that we

do not overturn previous findings in our replications.

R3.2 *One of the most serious problems in the manuscript is that it takes a good idea – reminding scholars of the possibility of SUTVA in experiments – and then mostly rehashes work presented by Bowers et al. in Political Analysis. I do not see what is particularly novel about the contribution, as potential SUTVA violations (whether studying state legislators or households or neighbors in GOTV studies) is well known. The manuscript argues that it would be best to model the network first, and then conduct treatments, but none of the three empirical analyses do so. Mainly, this article does not add a significant contribution to the study of state politics or state legislatures, nor is it methodologically novel, as it simply reminds the reader of SUTVA assumptions in experiments.*

Addressed: We realize that our contributions were not communicated effectively in the first submission. We have revised every section of the manuscript to sharpen, clarify, and communicate what we contribute beyond a review of existing work. This excerpt from our revised introduction lists our contributions, "Beyond the review of this methodology, we offer three contributions in this paper. First, we provide a typology of theoretical considerations that researchers can draw upon when formulating hypotheses regarding interference. Second, we provide a focused review of the networks through which scholars of legislative politics should consider in specifying tests for interference. Third, we apply this methodology by analyzing data from past studies that involved field experiments on state legislatures."

R3.3 *1. There is not a significant contribution in this article, and it is mostly a rehash of other work. The article's target audience appears to be scholars of experiments in states and state legislatures, and the authors assert that SUTVA violations are potentially severe in studies of small groups like legislatures. I am not convinced that SUTVA violations are likely in many of the types of state legislative experiments that the authors critique and replicate.*

Addressed: We have made three revisions to respond to this critique. First, we have clarified that our contributions are (1) to provide guidance in specifying the interference tests for field experiments on state legislatures, and (2) to provide a review of legislative networks that could be considered in tests of interference. Second, we added discussion of the literature that provides greater support for our claim that we should expect interdependence in legislative behavior (see, e.g., our added discussion of the Masket (2008) study of the effects of seating proximity on cue-taking in the CA Assembly). Third, we emphasize in the introduction that we do not yet know the extent of interference through legislative networks since the observational studies of legislative networks do not exhibit robust causal identification.

R3.4 *Audit studies with minimal email or constituent contact are unlikely to have spillover. For instance, in a professionalized legislature, in which there are many staff and legislators regularly receive 100s if not 1000s of emails per week, it strains credulity to think that a legislator is going to forward an email from a constituent to someone who is ideologically similar with them (ideological similarity is how the spillover network is modeled by the authors). I could be persuaded that legislators or their staff may speak to one another about particularly unusual interventions (or poorly designed interventions that are simply bad scholarship). But in the replications that the authors test, two of the three studies are fairly typical things legislators might receive. In these typical interventions, it is not clear that there would necessarily be spillover effects. Having worked in a legislative office as a staff member, I would not have thought to discuss with another office when a minor request was received by a constituent.*

Addressed: We have made two revisions to respond to this critique. First, since we agree that the likelihood of interference in the Broockman study is very low, and we have very little network data to work with in that replication, we have removed it from the paper. This has provided more space with which to develop the other two replications and expand the conceptual components of the paper. Second, in the new section on selecting legislative networks through which interference might occur, we clarify that interference can still be present if the treatment does not spread. In other words, if the treatment is presented in the form of information sent to a legislator, that legislator does not need to forward that information to other legislators in order for that legislator's treatment status to influence other legislators. It could be that the treatment influences the treated legislator's behavior, and that legislator's behavior influences the behavior of others (e.g., through cue-taking).

R3.5 *If anything, given that the effects remain similar to the initially published studies, the small or null effects produced by the authors' replications suggest there are not spillover effects or SUTVA violations in these past studies (consistent with my expectations, but inconsistent with the authors' argument). Or, given that the results of the previous studies hold in the authors' replications, if there was spillover in the past studies, then the authors of this manuscript have not modeled it correctly.*

Addressed: See response to R3.1.

R3.6 *2. I am concerned that the network is modeled as ideological similarity between legislators. It is possible that a grassroots lobbying experiment (e.g., the Bergan and Cole study the authors replicate) has a treatment that is more likely to affect liberal legislators. This would mean not that there is necessarily a SUTVA*

violation, but instead that the treatment was conditionally more effective with liberal legislators. The authors do not justify their use of ideological similarity as the path of spillover in the manuscript. Do the authors theoretically expect that legislators will forward grassroots emails to other legislators of similar ideology? This does not seem likely, though some other network of spillover is definitely possible (friend networks, cohorts when elected, etc.) yet the authors do not model such a network.

Addressed: We have updated the Bergan and Cole analysis to test for different effects based on political party. We also gathered data necessary to create the cohort network, and have incorporated it into each analysis.

R3.7 *3. The empirical results suggest that there are not significant spillover effects in these past studies. This undercuts the authors' argument substantially. Or it suggests there is nothing specific to state legislative studies to be learned that does not apply more broadly to all field experiments. This then gets the reader back to the Bowers et al. piece, which is this more general contribution regarding experimental spillover.*

Addressed: We agree that the general contribution of BFP represents a useful methodology to several subfields. Our review and discussion of possible networks to consider in modeling interference is, however, specific to the study of legislatures. We have also added discussion to the ``Network Selection'' section of the paper in which we make the case that interference is particularly likely in legislatures since (1) legislators often need to influence their colleagues in order to accomplish their objectives effectively, and (2) legislatures are small enough groups to justify the efforts of legislators to influence their colleagues.

R3.8 *4. Finally, if there is spillover or SUTVA violations here, it is unlikely through the ideological similarity of legislators as the authors measure.*

Addressed: We have revised the conclusion to clearly acknowledge that the null findings regarding interference could result from using the wrong network(s) in the replications.

R3.9 *I would (1) rework the beginning of the manuscript significantly to make it less about a methodological contribution (as that contribution was the Bowers et al. article, which is important and well known);*

Addressed: We have revised the introduction to focus more on our contributions beyond reviewing the Bowers et al. methods, and to focus more on the context of legislative networks. We have removed much of the discussion of SUTVA and related concerns regarding experimental design.

R3.10 *(2) reframe the article as showing that SUTVA violations are perhaps not a threat to most field experiments of state legislators. Replications presented in the paper of simple audit studies or minor grassroots lobbying interventions of legislators do not yield major differences in results once spillover is modeled, so I would instead argue that you have checked whether spillover has negatively affected the interpretation of the initial studies' results and positively conclude that it has not.*

Addressed: We have clarified in the section on our replication analyses that our study is not intended as a meta analysis of interference in state legislatures. As such, we cannot draw conclusions from our results regarding whether SUTVA violations do or do not represent an empirically probable threat in field experiments on state legislatures. Rather, the goal of our replications is to illustrate the considerations and steps involved in testing for interference in state legislatures.

R3.11 *In addition, I would suggest the authors model the network, not via ideological similarity, but through other networks that are more likely (social media networks between legislators, proximity of offices) and then conduct or replicate an experiment with state legislators after modeling the network. The authors suggest this is preferred to the approach they use in the manuscript.*

Addressed: It is beyond the scope of the current paper to run an original experiment, but we have incorporated the cohort network in an attempt to measure more of an interpersonal network for our analysis.

R3.12 *Finally, I would like to hear what scholars of GOTV in state politics need to learn about SUTVA violations as well and how could scholars in other areas besides studies of state legislative field experiments learn from this paper?*

Addressed: We have added discussion to the section on the formulation of the interference model to note that our contribution presented in that section is applicable to research outside of the domain of legislative politics. We do not want to broaden our discussion of other domains too far, and risk repeating what past research has noted regarding interference and spillover effects in experiments. However, our discussion of interference model formulation represents a novel conceptual contribution, and it was worth noting that it is applicable to any context in which a researcher is testing for interference.