July 28, 2018

Dear Professor Gibbs and the anonymous reviewers,

We thank the editors of *Communication Research* and the anonymous reviewers for their extremely thoughtful comments on our original submission. In response, we have reworked the manuscript substantially. We have incorporated the reviewers' and editor's suggestions and believe the revised submission improves greatly on the original as a result.

The most profound changes respond to concerns about conveying and interpreting the findings around the impact of requiring accounts on high and low quality contributions. We have built largely on R1's comments in clarifying the expectations from prior work as well as the implications of our results. We agree that these elements of the original manuscript were insufficiently clear or convincing and appreciate R1's insights, which have shaped our revisions deeply. We have also addressed each of the methodological questions and comments from both R2 and R3 and incorporated references to the literature suggested by all the reviewers.

The rest of this letter provides a detailed, point-by-point summary of all the changes we have made. We divide these into sections that correspond to the editor's summary of the reviews: (§1) addressing and citing literature brought up by reviewers; (§2) bolstering multiple aspects of our arguments; and (§3) clarifying various aspects of the methods. We also include a section (§4) listing several additional changes we have made. Within each category, we indicate which reviewer(s) raised concerns and describe how and where we have addressed the concerns in the manuscript.

As requested, we have provided a copy of our manuscript that highlights all the changes made in the text itself. We note that the software we used to do this highlighting (a tool called "latexdiff"[1]) identifies any new or altered content in the body of the paper, but apparently does not highlight additions or changes to the abstract, references, or to formatting or markup. These changes might not be captured perfectly by the highlighting as a result. We apologize for any inconvenience and hope it will not present an obstacle to reviewing the revised manuscript.

We are deeply grateful for the time and effort you have all invested in this manuscript. We have done our best to address the issues you each raised in your reviews and look forward to your responses.

Sincerely,

The Authors

[1] https://www.ctan.org/tex-archive/support/latexdiff

Summary of Changes

1. Adding literature recommended by reviewers

A number of reviewers pointed us to related literature. We have made the following additions to prior literature discussed by our paper:

- R1 suggested some additional references to Bennett and Segerberg (2012) and Margolin and Liao (2018) related to our finding that unregistered participation may stimulate and sustain participation by core (account-holding) community members.

  We thought these were excellent points of connection with prior research on connective action and online crowds. We have incorporated the references into the background and framing sections reviewing prior literature. We also included them alongside related references in the discussion.

- R2 suggested that we refine our references to Ostrom on boundary rules and group membership. We agree completely with the reviewer's suggestion and adopted this recommendation in the introduction, several points in the subsection on *Stable Identifiers as Catalysts of Cooperation*, and in some fine-tuning of the *Discussion* and *Conclusion*. In particular, we added sentences to the discussion connecting our findings with those described by Cox, Arnold, and Villamayor Tomás (2010) on the need for more precise specifications of community boundaries depending on the context of collective action.

- R2 also noted that the paragraph immediately preceding the *Stable Identifiers as Catalysts of Cooperation* subsection seemed to be missing citations. We have revised the text of this paragraph to clarify that it contains a very brief summary of the subsections that come immediately afterwards (and which includes many references). We appreciate the point and hope R2 finds the change consistent with their suggestion.

- R2 also suggested that a study by Frey, Bos, and Sumner (2017) that demonstrated a system design to resolve the tradeoff between privacy and safety in UGC moderation. We appreciate the suggestion and now point to it in the conclusion as an example of how future work might explore designs to manage the tradeoffs between community boundaries and collective action that we identify in our findings.

2. BOLSTER ARGUMENTS, IMPLICATIONS, AND TAKEAWAYS

The editor recommended that we bolster several aspects of our argument in response to reviewer concerns. In particular, she pointed to a question raised by R1 in regards to the relationship or ratio of the decreases we observe in high and low quality contributions:

> I wonder whether it is possible contextualize the aggregate effect on the ratio of individual reverted and PWR estimates within the expectations of theory. For example, using the figures reported on p. 20 it would seem that, prior to the intervention about 1% of contributions were reverted. The estimated "lost" contribution seem to be 4.48 reverts and 255 "good" ones, so about 1.8%. What does the existing literature say about these ratios? Presumably they are lower than catalysts would expect, but if not, this should be reported. Alternatively, do catalysts really argue that imposing a barrier would increase total contributions, or just contribution quality? If so, this should be re-emphasized in the discussion so it is clear that the authors' finding is non-trivial.

In a similar vein, R2 suggests that we strengthen the claims in our work:

> I do think the authors undersell the impact of their work. The question of requiring accounts in wikis seems somewhat narrow and applied, but the broader (or merely reframed) question of permitting anonymity is very relevant, within the communication community and beyond (as shown by the author's own excellent discussion of the broader anonymity literature).

We have made several changes in response to these concerns:

- Drawing from R1's suggestion to interpret our effects in terms of ratios of "good" and "bad" deterred contributions (this was a very useful suggestion!) we now emphasize the unanticipatedly severe nature of the tradeoff (i.e., 58 "good" edits deterred for every "bad" one). We make this point in the abstract and introduction as well as the *Discussion*, and *Conclusion* sections.
- We added text to the subsection reviewing the catalysts literature emphasizing the effects predicted by prior work and summarizing that, in the absence of the cost perspective, the perspective does indeed predict good-faith contribution rates that are stable and even increasing.
- We have also clarified a related point immediately following the statement of H4, underscoring that prior theories and empirical findings on the catalysts side would predict the opposite to our hypothesis and that most (but not all) theories adhering to the "costly barriers" perspective would fail to anticipate such outcomes as well.
- In the *Discussion* section, we have added material to further develop the impact of our findings and to specify how our results contradict the catalysts literature and diverge

from the costly barriers theories in important ways as well: We have reworked nearly our entire discussion section to do this. It now:

- very briefly summarizes our results in terms of prior work;
- discusses why the catalysts and costs perspective may have failed to anticipate our findings;
- walks through our findings in terms of ratios (as per R1's suggestion);
- discusses these ratios in terms of what previous theory would have predicted (as per R1);
- discusses the implications of the results more effectively in both substantive and theoretical terms.

3. IMPROVE/CLARIFY METHODOLOGY

3.1. Address sensitivity of inclusion criteria:

R1 raised a concern about the sensitivity of the results to the thresholds used in our inclusion criteria and described in the original manuscript's *Data* section. Specifically, R1 called out the fact that we required that there be a 90% decrease in edits before assuming that the policy had been implemented:

> In particular, if there is a 75% decrease in contributions from unregistered users, is that really insufficient evidence that the policy was implemented? Alternatively, the authors could show the typical variation in unregistered user contributions and choose their threshold (90% or whatever) based on that. In this case, given the authors' general rigor, I do not suspect cherrypicking. But in general, cherrypicking thresholds is a reasonable suspicion when only 1 threshold is chosen/presented. So as a work that introduces computational approaches to the field I think it is worthwhile showing sensitivity analyses here, too (in the supplement).

We have made several changes to address this threat:

- We have conducted robustness analysis in which we do not exclude *any* wikis (i.e., we do not require any decrease at all). The only exceptions are wikis for which we have no data at all and wikis that required accounts when they were created (neither group includes any variation we could use to estimate $\hat{\tau}$).
- We report results using this "full" sample in the *Sample Selection* section of our supplement. Our results are slightly moderated in magnitude but are not substantively different.

- We have added text to our manuscript in the section where we discuss our inclusion criteria to clearly point to these alternative results and to the fact that our inclusion criteria are not driving our findings.

We hope this demonstrates that our inclusion criteria are driven entirely by a desire to create reasonable estimates of our effects in communities that make the change. We appreciate the opportunity to make this clear.

3.2. More clearly justifying our use of RD

R2 wrote several paragraphs questioning our claims that RDD provides more credible estimates than other approaches—especially those that rely on case controls:

> The authors justify their choice of method (RDD) in detail in the supplement, where they mention alternatives that allow for control conditions, and dismiss them in favor of the method they used. ... I am actually content that RDD makes sense for this application, but not with the claim that it is better than using a method with controls. ... I don't think you need to argue that RDD is superior, or even ideal, in order to get readers to conclude that it is adequate to your purposes.

R2 goes on to support this argument further. We agree with R2's points in this portion of their review and have implemented several changes to address the issue:

- We have edited the the *Absence of Case Controls* subsection in the supplement. In particular, we try clarify the assumptions that underpin the panel RDD, the efforts we made to evaluate them empirically, as well as the strengths and limitations of this approach.
- We have revised the third paragraph of the *Analytic Strategy and Models* subsection. In this new paragraph, we discuss the assumptions and limitations underlying our approach in more depth; (2) we explain that alternative estimation techniques are available; and (3) we point readers to the supplement.
- We also made changes to address these points in the *Threats to Validity* section of paper.

In each of these edits, we attempt to underscore R2's point that alternative approaches are feasible and that the assumptions they require may be as reasonable as required for our approach. We also clarify that our reasons for preferring the identification strategy we pursue stem from our perception that some of the requisite assumptions can be empirically evaluated more directly.

3.3. Condense data section by moving material to supplement

R2 suggested that we condense the *Data* section by moving more material to the online supplement. As requested, we have moved details about the adjusted cut-off dates.

R2 also suggested that we also move details about our search of 10% of the wikis for evidence of announcements or discussion of the design change to the online supplement. We feel that this information is important for readers to evaluate the validity of the identification strategy and have retained it in the text. If R2, the editors, or other reviewers strongly disagree with this, we're happy to reconsider or discuss it further.

3.4. Discussion of fixed versus random effects for wiki

Both R2 and R3 raised questions about our use of fixed effects controls. In particular, they ask us to provide a more complete justification for this choice.

First we respond to R2 and R3's questions about random vs. fixed effects. The most widely used type of random effects model is a random intercepts model which divides residuals into within-group and between-group error. This is to address the concern that group-level residuals may be correlated with each other and lead to downward biased standard errors. The two key benefits of random effects are that they (a) allow a model to estimate parameters associated with variables that do not vary within group and (b) are more efficient, especially in models with many groups. Although (b) is typically an issue of statistical power, (a) means that random effects are essential for estimating the impact of a variable like gender in a longitudinal study of people (because gender does not typically vary within a single person).

In contrast, fixed effects are dummy variables added to a model—one per group. These dummy variables address both the concern about correlated error that random effects address as well as controlling for every feature (observed or unobserved) that has a consistent effect on the outcome across all observations within the group. Using fixed effects avoids an assumption (necessary to the use of random effects) that any omitted control variables are uncorrelated with any independent variables in the model. Because they estimate within-group differences and avoid this assumption, fixed effects—if they can be used—are seen as more conservative, powerful, and appropriate method for identifying causal effects in applied statistics and econometrics (Angrist & Pischke, 2008; Murnane & Willett, 2011).

The logic behind our modeling choices here stem from the overarching goal of identifying the effect of the change on our dependent variables (i.e., $\hat{\tau}$). What we lose by using fixed effects is efficiency (not a major concern the size of our dataset) and the ability to estimate and interpret parameters for specific control variables. What we gain is much more confidence in our estimate of $\hat{\tau}$.

In a related comment, R3 asked whether we could get additional data to control for wiki-level variation in variables such as audience size. As we hope is clear from the foregoing discussion, the fixed effects for wiki effectively control for the baseline audience size as well as changes in audience size that are reflected in a "secular" quadratic trend.

Although we do not have space in our paper to discuss this in as much depth as we do here,

we have made several edits to our *Analytic Strategy and Models* section in response to these concerns. In particular, we now mention:

- Fixed effects in repeated measure models are typically considered preferable to random effects in the context of causal inference because their inclusion ensures that estimates of $\hat{\tau}$ reflect only within group variation.
- We have added citations to Angrist and Pischke (2008) and Murnane and Willett (2011) to support this and to point interested readers to more detailed discussions.
- We explain that our three vectors of fixed effects capture all observed and unobserved characteristics of wikis that have a consistent effect across all weeks (such as each project's start date or initial audience size) as well each wiki's quadratic trend (such as growth or decline in the popularity of a topic covered by a wiki).
- We explain that, because our hypotheses focus exclusively on the estimate of $\hat{\tau}$, fixed effects serve as control variables and therefore we do not report the parameter estimates for them.

3.5. Unpack list of threats into plain English

R2 notes that the list of threats was not in plain English. Upon rereading, we completely agree! We have rephrased the opening paragraph of the *Threats to Validity* section to present this more gracefully and clearly.

3.6. Refine visualizations

R2 requested that we attempt to improve our visualizations saying:

> Figure 1 would be stronger (more immediately apprehensible) with a big black arrow point up the zero point. Another thing that would help it scan better (help readers quickly ignore the many visual contrasts they should ignore and attend to those that make the point) is if they communicated the before->after transition with grey->color (or less saturated -> more saturated) instead of green->purple.

We have added a thick black line to indicate the zero point in Figure 1 (we felt that this looked better than an arrow). We experimented with saturation and grey/black difference but found that this made it difficult to interpret differences in density which are currently represented by overplotting translucent points.

R2 also asked about the differences between Figure 1 and 2, saying:

> Am I right that Figures 1 and 2 tell the same story, one with the data, and one with a version of the final model? If so, I wouldn't mind Fig 2 being Fig 1, and Fig 1 being supplemental, but that may just be a matter of taste. Comparing

> Figs 1 and 2, I don't see the two top panels telling the same story: it scans as an increase in Fig 2 but a null result or decrease in Fig 1.

R2 understands our figures correctly. Although the figures represent the same set of relationships, we feel it's beneficial to show readers both a "raw" view of data (Figure 1) as well as model-generated estimates of effects (Figure 2). We believe the fact that the differences in terms of the way that the two visualizations of new editors scans usefully conveys some of the fragility of this result (discussed elsewhere in the paper).

We would prefer to keep both figures in the paper. If the editor or R2 have further considerations on this point, we would be willing to move Figure 1 to our supplemental material.

3.7. Robustness check for influential cases

R2 suggested that we interpret the results of the robustness check in which we drop the largest observations from the analysis:

> It would be helpful if you could do a bit more to interpret the results of the final robustness check, in which results on supporting measures, but not one key measure, may be driven by the top 1%, 5%, or 10% of wikis.

Working to address this issue caused us to clarify two decisions at the heart of this robustness check. First, we realized that the text of the online supplement was not clear about how we were dropping observations. Second, we became less comfortable with our decision to use log linear models in this robustness check rather than the negative binomial models used in the rest of the analysis. Although we explained our decision to use log-linear models at some length in the previous submission—and although none of the reviewers raised this as a concern—we became troubled by the fact that the results for *new editors* (M1) from the log linear model were somewhat different when run on the full dataset.

We have made changes to address both the issue raised by R2 and these other related issues:

- As requested by R1, we have added the following text interpreting the results of our final robustness check to our *Threats to Validity* section:

  > The estimates of *acc_req* in these models are consistently positive but smaller in magnitude and are not statistically significant in models that drop the largest 5% and 10% of wikis. Although we can imagine explanations the effect of requiring accounts on new editors might vary with project size or other factors, we know of no theoretical reasons to anticipate this. It is also possible that the fragility of the estimate is an artifact of fact that the effect on new editors is relatively small and noisy and that much of the estimated effect is concentrated in large wikis. As a result of these tests, the results of M1 (see Table 1) should be interpreted with care.

- We have almost completely rewritten the section of our supplement on this threat. We have revised our text to make it clear that we are taking the dependent variable values per wiki, calculating the within-wiki average over all the weeks, and then dropping the highest 1%, 5%, and 10% of wikis based on the sample distribution of these averages.

  This did not appear to be a point of confusion for reviewers but we felt it could be more clear.

- We now report robustness checks with the same negative binomial models used in the main analysis presented in the paper. The results are included in the revised *Influential Observations* section of the online supplement are substantively similar to our previous results.

3.8. Improved discussion of measures

R3 raised several questions about the measures of account creation, reverted edits, and non-reverted edits. We discuss each in turn.

3.8.1. Account creation

In our manuscript we explain that:

> Because the XML databases only include data on the contributions to each wiki, this measure will not reflect accounts that were created but that never went on to make a contribution. If a user had created an account previously, but never made a contribution to the wiki in question, the account will be marked as new during the week it makes its first contribution.

R3 asked us, "how is it going to affect the results?"

It is possible that accounts which edit for the first time after the intervention could have been registered earlier but never used to make an edit. While we think this is possible, we have no way of assessing this in our dataset which simply does not include this information.

Our variable *new accounts* captures the number of accounts making their first edit within each wiki week. The variable is measured consistently before and after the intervention. We think that there is some confusion from the variable name since accounts might be old even though the first edit might be new. This is true both before and after the change. We considered naming the variable "new editors" but we felt that this was also confusing because anonymous users are still referred to as editors and because an individual editor might edit both anonymously and with an account. Other options—like *accounts making first edits* struck us as also potentially confusing (in different ways) and infelicitously lengthy.

We are not sure whether this issue merits additional discussion in the paper and have not yet elaborated on it in the text. After talking the point over, we reached the conclusion that

we would leave the variable name as is for now. We would appreciate further input from the reviewers and the editor.

3.8.2. Low and high quality editing

R3 also raised questions about low and high quality editing:

> The measure for low quality edits is the number of reverted edits. Could you talk about whether the practice of reverting edits also requires an account? If so, then how would the treatment (account requirement) affect the practice of "reverting" and the measure of low-quality edits? In other words, could the treatment itself affect how the DV was measured?
>
> The same goes for your measure for high quality edits (the number of non-reverted edits). Could you talk about whether the treatment (account requirement) may have affected this particular measure?

For both reverts and non-reverted edits, the intervention does not impact the measurement in any way. A revert is simply another type of edit. As long as people do not require an account to edit, they do not require an account to revert. Because our measure of non-reverted edits is the total number of edits minus reverts, we believe that this addresses the issue with this measure as well.

We revised the *Measures* section to clarify that users do not need accounts or any special privileges to revert and that MediaWiki software includes affordances that makes it easy for all users to return pages to previous version in way that are recorded as identity reverts.

3.9. Logistic regression model for *reverts*

R3 raised a concern about the lack of variance in our measures of low quality edits saying:

> If half of the wikis experienced no reverts, then there is no variance whatsoever. Are negative binomial model necessarily the best options here? I wonder how this might have affected your results.

We agree that this is a potential threat to the results and appreciate the suggestion to evaluate it. To address this fully, we made two changes:

1. First, we've added a new section to our supplement, *Logistic Specification for Reverts*. In this new section we do several things:
   - Explain the threat raised by R3 in some depth.
   - Describe an approach to addressing this threat that involves using an alternative specification of model M2 where instead of a negative binomial regression model, we estimate a logistic regression model where the dependent variable is a dichotomous variable set to 1 if *reverted* > 0 and 0 otherwise.

- We include a new table in our online supplement (Table 18) that presents the results from this model.

  The parameter estimate for *acc_req* ($\beta = -1.77$) suggests that after the block, the odds of a wiki having at least one reverted edit were 17% the odds of experiencing at least one reverted edit before. We believe that these estimates are fully consistent with the pattern of effects reported in the negative binomial models.

2. We add a footnote to the sentence in our paper that explains that we use negative binomial regression to both briefly summarizes the threat raised by R3 and to point to the new section of the supplement.

We have opted to retain the negative binomial results as the primary results reported in the paper for three reasons. First, although much of the variation in *reverted* is between 0 and 1, there is also substantial variation in many wikis that see a major reduction in vandalism. We believe that this variation is important to capture. Second, by using the same model for all four of our dependent variables, we make interpretation of our tables and numbers much easier for readers. Third, the results are substantively similar and we feel that our takeaways would not be altered by the choice.

Although we believe that this is best solution, we would be willing to foreground the logistic specification instead of the negative binomial model if R3 or the editor feels strongly about change.

4. Other changes

We have also made the following other changes:

- We have fixed the typo pointed out by R3. We also carefully proofread the document and caught several other stylistic errors.
- We have introduced a new (darker) color palette for the visualizations. The color palette should make it easier to differentiate the colors in grey-scale copies of the paper.

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action. *Information, Communication & Society*, *15*(5), 739–768. doi:10.1080/1369118X.2012.670661

Cox, M., Arnold, G., & Villamayor Tomás, S. (2010). A review of design principles for community-based natural resource management. *Ecology and Society*, *15*(4). doi:10.5751/ES-03704-150438

Frey, S., Bos, M. W., & Sumner, R. W. (2017). Can you moderate an unreadable message? 'Blind' content moderation via human computation. *Human Computation*, *4*(1), 78–106. doi:10.15346/hc.v4i1.5

Margolin, D., & Liao, W. (2018). The emotional antecedents of solidarity in social media crowds. *New Media & Society*. doi:10.1177/1461444818758702

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.