October 1, 2017

Dear Professor Waisbord and the anonymous reviewers,

We thank the editor of *Journal of Communication* and the anonymous reviewers for extremely thoughtful comments on our original submission. In response, we have reworked much of the manuscript. We have also built off of many of the reviewers' and editor's suggestions with respect to both conceptual and analytical aspects of the paper. We believe the revised submission improves enormously on the original as a result.

The most profound changes respond to concerns about the theoretical framing, contribution, and discussion. We have re-focused these aspects of the paper around a much more robust synthetic account of the role of tradeoffs in the requirement to adopt persistent identities and the management of community boundaries. We develop this argument in greater depth than before and position it as a central contribution of our work. We have added a new fourth hypothesis that was previously buried in our discussion and online supplement. We believe our findings in regards to this new hypothesis highlights a theoretically important and surprising result.

We also address a number of concerns raised by the reviewers related to the empirical analysis, including our use of a regression discontinuity design, measurement, control variables, and the fit between the empirical analysis and theoretical claims. We respond to some of these concerns in the paper itself and some in our supplement, which we would post (together with all of the data and analysis code necessary to reproduce the paper) online in a digital open access repository as a companion to the paper.

Based on several comments in the reviews, it seemed that the reviewers may not have had access to the original supplement despite our efforts to include it with the submission. We apologize if it was unavailable. As the reviewers speculate, we had made attempts to address some of the concerns raised in the reviews in this supplement. We have concatenated the appendix to the end of this letter to ensure that it will be available to reviewers.

The rest of this cover letter provides a detailed summary of all the changes we have made. We divide these into three primary categories: (1) theoretical framing, contribution, and discussion; (2) methodology and measures; and (3) changes not requested by the reviewers. Within each category, we indicate which reviewer(s) raised concerns and describe how and where we have addressed the concerns in the 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

Summary of Changes

1. THEORETICAL FRAMING, CONTRIBUTION, AND DISCUSSION

A key theme across the reviewers' and editor's comments concerned the theoretical framing, contribution, and discussion. Based on this feedback, we have tried to address these issues by profoundly restructuring and refining several elements of our manuscript.

1.1. Clarifying theoretical contribution and central claims

We agree with R2 that our submission did an inadequate job theorizing an original synthesis of the tradeoffs perspective on requiring accounts and that this inadequacy undermined our central claims. The revised manuscript contains a new section in which we develop this synthesis and motivate our hypotheses in greater depth. The key point of this new section (which was implied, but not stated clearly in the original submission) is that we expected/hypothesized aspects of all the prior theories would hold true because we believed that none of the prior theories ("catalysts," "barriers," or the work on the importance of unregistered contributions) accounted fully for the threat of damaging contributions, the benefit of peripheral participants, and the impact of barriers to entry. Our perspective emphasizes these tradeoffs as central to design decisions around community boundaries and contribution infrastructure in the peer production of online public goods. We agree with R2 that this was a critical deficiency in our results and we have updated our title to reflect this central claim.

1.2. Refining, streamlining, and clarifying our discussion of prior research

The new section detailing our argument about tradeoffs motivates a related set of changes to the way we presented prior research and our hypotheses in the original submission. Before, the discussion of prior research and the hypotheses were too tightly coupled. We used existing theories to motivate separate hypotheses in a format that obscured how we synthesized these theories and deviated from them in important ways. As discussed above, our new framing section on tradeoffs develops our synthesis in a much clearer way. We have also revised the sections presenting prior theoretical perspectives to fit with this restructuring. We still derive empirical predictions from earlier perspectives, but we differentiate these predictions from our own.

1.3. Building up stronger intuitions for the predictions from prior research

As part of the revisions to our theoretical background sections, we have also tried to address several reviewer comments indicating that we failed to achieve our goals in the original

manuscript. For example, R2 notes that certain predictions seem obvious whereas others seem implausible or poorly developed. R3 also notes that the logic behind the "catalysts" argument produces a hypothesis that may be trivial and therefore uninteresting. In response, we have worked to communicate the rationales of prior research more effectively. We have also added material to the discussion to flesh out potential explanations of the findings as well as ways in which our data and methods cannot answer certain questions.

In particular, we have tried to underscore the means by which "looser" barriers would support cooperation by expanding our discussion of prior work on the importance of unregistered outsiders and newcomers for stimulating knowledge production communities. This integrates the suggestion from R2 to include a citation of Anthony et al.'s paper on unregistered editing in Wikipedia. We also follow R2's suggestion to provide a better explanation for potential mechanisms by which unregistered contributions could drive overall quality, pointing to the importance of group size (citing Olson as well as Zhang and Zhu), diversity, and heterogeneity in the production of collective intelligence (we cite Woolley et al. 2010 on this point although March's exploration/exploitation work and subsequent studies of that question are also relevant). We also explore the implications of our findings for these earlier theories in the Discussion section much more thoroughly. We appreciate R2's point that network analysis and other methods might support more direct identification of the mechanisms by which unregistered contributions drive overall activity. However, we feel that such questions lie beyond the scope of the empirical work we report here and are best pursued in future studies.

Building on the scholarship demonstrating that unregistered contributions can contribute to higher quality content overall (Anthony et al., Kane et al., and Gorbattai), we have also added a hypothesis (H4) to correspond to an aspect of our analysis that was obscured in the discussion section of the original manuscript. This tests whether editors with accounts registered before the intervention contributed at a different rate afterwards or not. We find that even these registered editors (who were not directly affected by the intervention) edit at a lower rate after the change, indicating support for the idea that unregistered contributions indirectly stimulate production and participation across the board. We believe that this far-from-obvious result also addresses several of the reviewers concerns that our results were not entirely surprising.

We think these changes have substantially improved on the earlier framing and appreciate that the reviewers pushed us in this direction. Ultimately, we believe the improbability and incompatibility of some earlier predictions helps to justify this study. Reading the literature we discuss in these domains and speaking to several of the authors of prior work, we maintain that scholars of cooperation and online public goods really do hold divergent expectations about the impact of the sort of intervention we analyze. As some of those divergent expectations were not convincing or clear in the original submission, we hope

that the revised manuscript better elaborates the different points of view in these debates while also distinguishing our own perspective more effectively.

1.4. Tightening the prose and flow of the framing sections

The changes to the theoretical framing described above also respond to R3's concern that the theoretical argumentation meandered without clear rationale. As per R3, we have removed the citation to legitimate peripheral participation. We agree it was not central to our argument. We hope the writing proves more compelling now and that the rationale framing the study emerges in a more convincing fashion.

R3 also indicated that they felt that the framing was not theoretically compelling because the analysis focuses primarily on the estimate of a single independent variable. We hope the reformulation of the framing helps address this concern. We also encourage consideration of two aspects of our approach. First, the identification and estimation of causal effects from observational data leads us to emphasize the effects of a single, concrete causal factor (see the Murnane and Willett book cited in our manuscript on this point). This contrasts with studies where the goal is to model an entire chain of factors that might (or might not) influence a given outcome through more or less direct means. Given the theoretical weight attributed to stable identifiers as well as the divergent findings in prior work, we also believe that understanding the effects of this particular cause merits focused scholarly attention. Our analysis gains theoretical depth by identifying and evaluating mechanisms through which an account requirement might work corresponding to prior theories.

Second, we should have communicated more effectively in the original manuscript that our models actually contain over 400 additional independent measures that we use to control for all wiki and time related factors. These factors are not the theoretical focus of the study, the prior theories, or the intent behind the intervention, and so we do not report these coefficients in our tables. This was unclear before and we apologize for any unintended confusion. We explain and justify this further later on in this letter (see the section on "Additional control variables and independent variables").

1.5. Enhancing the fit between the theoretical framing and empirical analysis

Another aspect of R2's comments related to the fit between the theoretical framing and the empirical analysis. Overall, we agree with R2 that the original manuscript did not adequately connect the measures and empirics to some aspects of the theories. In particular, we adopt R2's suggestions to better justify the linkage between the measures of quality and damage we use and the broader outcomes of community well-being advanced by our theories. We implement these suggestions with changes throughout the paper. In the framing, we go to greater lengths to clarify the empirical predictions of prior theories more fully (as already discussed). We also explain how our measures and models of short-term impacts connect to long term community well-being more clearly in the subsection on "Measures"

within the Methods section. In addition, we add material to the results section that (1) decomposes the effects across the wikis in the sample and (2) projects the impact of the effects we observe over a longer time horizon (we discuss this more below). Finally, we discuss the theoretical and empirical implications of these effects more fully in the latter parts of the paper.

1.6. Theorizing implications of the findings more deeply in discussion

We also adopt R1's suggestion to theorize the implications of the findings more deeply in the discussion. The perspective on tradeoffs we now synthesize in the revised framing lays the groundwork for this and enables us to return to the background in a more decisive manner. We conclude that the tradeoffs perspective characterizes the results more effectively than any of the prior approaches alone. We also discuss potential mechanisms driving the results, some of which we can support/refute based on our findings and some of which we cannot. Finally, we also add material that speculates on the reasons why our results might support/contradict specific claims from prior literature. As part of this, we explore why the outcomes we observe would have deviated so profoundly from the predictions and prior work arguing that requiring accounts would have catalyzed cooperation in various ways. We also identify additional questions raised by the present study.

2. Methodology and Measures

The reviewers also raised concerns about specific elements of our methods and measurement. We discuss these issues here.

2.1. Validating and further explaining the cutoff

R1 raised a series of questions about the degree to which we corroborated information about the date of the configuration change against the trace data from the wikis. As R1 surmises, we attempted to address this concerns in the online supplement and details of how we did so are provided there. The text of the manuscript now makes clear that we visually inspected every wiki in our sample and that the supplement contains additional information about wikis for which we apply alternate cutoff dates. If R1 feels that it is necessary to include data on anonymous editing for *all* the communities in our sample, we could share these or add them to the supplement as well. We have chosen not to do so because we feel it does not add much additional value or insight.

R1 also asked for more detail on why the change would not be publicly announced. We have added detail about how these decisions were made as well as why we believe they were unannounced. We removed one instance where we claimed they were "always … unannounced." We hope that these changes convey why we are confident that most were completely unannounced and that any announcement would not have reached unregistered editors (i.e., the population targeted by the intervention).

2.2. Providing more detail on inclusion criteria

Both R1 and R2 raise questions about our inclusion criteria. Although some of the this was described in detail in our supplement, we have added to this material. We have provided summary statistics for the excluded wikis. They are smaller and less active than the wikis we have included. We have also provided estimates of models that include these 37 excluded wikis. The estimates are slightly smaller and the standard errors are slightly bigger, but pattern of results is identical.

R2 asked about increasing the stringency of our inclusion criteria (e.g. requiring that wikis have more anonymous contributions to be included). Because this inclusion criteria will strengthen the effects of the findings, we prefer to use a looser (and more conservative) inclusion criteria that errs on the side of analyzing wikis that are likely to see more minimal effects of blocking unregistered contributors. We would prefer our results underestimate the effect than risk designing a sample that inflates the magnitude of the estimates.

On a related point, R3 questions whether and how the results may be systematically driven by the age of the wikis in the study. We note that the inclusion criteria do not specify anything about wiki age and that the wikis are actually quite diverse in terms of their age at the point of the intervention. We refer R3 to the first row of the table of summary statistics (now included as Table 1 in the online supplement). The age of wikis at the point of intervention has a range between 4 and 438 weeks (roughly zero to eight years) with a median of 142 weeks and a standard deviation of 93 weeks. Furthermore, any relationship between wiki age and the outcome measures is controlled for through the use of fixed effects for wiki and time (we discuss this further below), so our models have already accounted for this.

2.3. Clarifying the measure of reverts

R1 raised important questions about our measure of reverts. We have clarified that we are using a measure of reverts often called "identity reverts" which is based on comparing the text of revisions (e.g., through MD5 hashes). It is the same measure used by Aaron Halfaker in the articles in the papers cited by R1 in their review. In fact, we used the Wikimedia Foundation's revert detection library written by Aaron Halfaker (*python-mwreverts*) to compute this measure. We added several of the suggested citations, describe our revert detection procedure in more depth, and included a link to the software library we use.

2.4. Clarifying the measure of content persistence

We have also added detail on the method and software we used to measure persistence and included a related citation (Flöck et al. 2017) suggested by R1. Once again, we used the techniques and tools published by Aaron Halfaker and the research team at the Wikimedia Foundation to construct these measures.

2.5. Contextualizing our methodological approach in terms of similar works

Following R1, we have have added several references (Geiger & Halfaker 2013, Slivko et al. 2016, and Zhang & Zhu 2011) that have used or discussed the role of naturalistic experiments to understand social processes in peer production.

2.6. Justifying the validity of our causal claims

R3 raised a number of important concerns about whether the treatment was "exogenous" and discussed some of the ways that the decision was likely intentional or pre-meditated. We agree that our use of the terminology "exogenous" was sloppy and unclear. R3 is absolutely correct: persistent vandalism by unregistered users likely led some administrators to mandate accounts. That said, this is far from the complete story. More than half of the wikis in our sample experienced no reverts at all in the 8-week period before they made the change. Much more importantly, however, the fact the change was an intentional reaction to vandalism does not compromise our ability to draw inference about a causal effect at the time of the change.

With respect to exogeneity, we have refined the manuscript to emphasize that we claim that the exact timing of the change was independent of the outcomes. In other words, our causal identification rests of the degree to which the week immediately after the block and the week immediately before are equal in expectation in terms of the amount of new accounts, reverts, and high quality content (plus or minus the secular trend we control for).

R3 also notes several ways in which individuals who knew that the configuration change was coming might have shifted their editing behavior immediately before or after as a result. Methodologists sometimes describe this as the potential for "crossover" between the treatment and control groups. As R3 points out, it is an important threat to the validity of our estimates. We have made several changes to attempt to address this:

- We add text to the Methods section to clarify what needs to be true for our effects to be appropriate estimates of the causal effect.
- We remove most uses of the term "exogenous" from our paper. The only remaining use is when we refer specifically to the "exogeneity of the precise timing of the shift" in relation to our dependent variables.
- We add material to our Threats to Validity section describing the threats of non-compliance and crossover and how they could affect interpretation of our results. We also describe robustness checks we conduct to address these threats.

The test we conduct to address this concern appears in the section on "crossover" in the online supplement. Based on information from Wikia staff, administrators were most likely to know about the change ahead of time. We used the Wikia API to collect data on which users held administrator status. We then fit new models on a dataset that removed all edits from administrators. The results are similar but slightly stronger than in the main

analysis.

We think this new material acknowledges these concerns more explicitly than in the previous version. It also summarizes our empirical approach to assessing the sensitivity of the findings to this threat. In this respect, our approach aligns with those taken in classic RDD studies. For example, in a famous 1999 paper, Angrist and Lavy use the impact of a policy change to study the effect of class size on student learning outcomes. Of course, both the intention and the effect of the rule was to create smaller class sizes. Similarly, not all class sizes turned out to be exactly what the rule stated and there were some indications that some students and families had awareness of changes in advance. Angrist and Lavy were able to convincingly argue that the eventual class sizes were not a function of students shifting themselves between schools to achieve a smaller class size. The fact that they might have done so remains a threat to the validity of that work just as the possibility of administrator manipulation of editing activity remains a threat to the validity of our own. The best practice in such circumstances is to report the threat and assess the sensitivity of the findings. We hope our revised manuscript does this much more effectively.

2.7. Explaining regression discontinuity methods more fully

R2 encouraged us to add more explanation on regression discontinuity and a reference to introductory material for communication scholars. We have expanded a paragraph in our "Analytic Strategy and Models" section describing the technique in more detail and pointing readers toward an excellent introductory text.

2.8. Clarifying and justify control variables and independent variables

As mentioned earlier, R3 raised several issues around the lack of independent variables and controls in our models. In fact, our models include an enormous number of control variables. Although the parameter estimates were omitted from our tables for reason of space and parsimony, every model we fit with 408 control variables (separate intercepts and two slope parameters for time and for every wiki in our sample). Our intercepts controls for everything—observed and unobserved—that has a consistent effect across wikis (Angrist and Pischke, 2008; Murnane and Willett, 2011). We also control for a curvilinear second order polynomial trend in our dependent variables over time *within* each wiki. Additional controls could only explain remaining within-wiki variation not correlated with time and we believe this variation is likely to be quite small.

We have made several changes to clarify and justify our approach in this regard. We have updated our analytic plan section to describe our strategy with control variables more clearly. We have updated the captions for our regression tables to make it clear that we have included 408 controls that are not reported. We explain that we have controlled for all observed and unobserved factors with systematic temporal or wiki-level effects across our outcomes. We also explain that this includes baseline activity as well as within-wiki

variation. We cite two econometric textbooks that introduce and describe why this approach advances identification through controls. As R3 suggests, adding controls in a fixed effects framework is possible in that the controls would still capture additional within wiki variation. Unfortunately, the variables R3 suggests (i.e. amount of high quality words, vandalism) are already dependent variables in our models and so we cannot add them without undermining the entirety of our analysis. We welcome additional suggestions on this point.

2.9. Decomposing effects across communities and over longer time horizons

R2 encouraged us to assess the "health impact" of the intervention on a community-by-community basis in greater depth and broader scope. Among several related comments, R2 asks us if there is a way to "use RDD to get at a counterfactual...path of content each wiki would have undergone?" Although this might be possible in a differences-in-differences design or with matching (other econometric techniques for identifying causal effects in observational data) we contend it is neither possible nor responsible using this RDD due to the absence of credible case controls. We have added some text to both our manuscript and supplement discussing these concerns.

Despite this limitation, and while a detailed analysis of the 136 communities in the study is not possible in a paper of this length, we have tried to address R2's comments in two ways within the RDD framework we apply. First, we have conducted new analyses to decompose the effects and characterize the diversity of effects more directly. Second, we integrate our observed effects across longer time periods to help convey the potential long term impact of the intervention. We thank R2 for this suggestion and hope the new material speaks to this set of concerns. We describe both changes below.

To decompose the effects, we run supplementary models that estimate interaction terms between ***wiki*** and *acct_req*. These estimates allow us to produce individual wiki-level estimates of the effect of the intervention which we describe in a new section and table in our supplement. These new results suggest that there is diversity in our estimates of the effects across wikis, but that there are about 2 communities with a negative estimate for each community with a positive estimate across M2, M3a, and M3b. In the case of M1 (new accounts), we find that more wikis experience a negative effect than a positive effect—a surprising finding given our overall positive estimate. We discuss each of these results in the manuscript.

To estimate a longer term impact, we integrate the instantaneous effects estimated at the week of the intervention over a two month period (note that these estimates already take the underlying secular trends within that two month period into account). We then report these projected effects as a proportion of total editing activity at the time of the intervention. Extrapolated in this manner, the instantaneous effects correspond to a 121% increase in the total new accounts created, a 38% decrease in reverted edits, and a decrease

in non-reverted edits and PWR of 62% and 59% respectively. While such projections rely on a naive assumption that the instantaneous effects would be stable week-over-week, they convey the large-scale impacts of the intervention. We believe this speaks to the impact on overall community outcomes in a powerful way.

2.10. Fixing layout problems

We have fixed the layout issues in our regression tables identified by R1. Thank you for catching this!

3. Changes not suggested by reviewers

We also made several changes that were not suggested by the reviewers but that we felt would improve the paper:

- We found a minor error with the way that administrators were being identified in our original analysis code. Fixing it did not change any of our results.
- To accommodate the changes within the journal page limit, we have moved the tables of descriptive statistics into the online supplement.
- We have thoroughly edited to tighten our prose to make room for the additions described above.