Peer Review Comments

Article: Ambridge, B., et al. (2018). Effects of Both Preemption and Entrenchment in the Retreat from Verb Overgeneralization Errors: Four Reanalyses, an Extended Replication, and a Meta-Analytic Synthesis. Collabra: Psychology, 4(1): 23. DOI: https://doi.org/10.1525/collabra.133

Article type: Original Research Report

Editor: Fernanda Ferreira

Article submitted: 10 January 2018

Editor decision: Accept Submission

Revision submitted: 09 March 2018

Article accepted: 22 May 2018

Article published: 02 July 2018

Responses for Version 1

Reviewer A:

1) General comments and summary of recommendation

Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:

- Are the methodologies used appropriate?

- Are any methodological weaknesses addressed?

- Is all statistical analysis sound?

- Does the conclusion (if present) reflect the argument, is it supported by data/facts?

- Is the article logically structured, succinct, and does the argument flow coherently?

- Are the references adequate and appropriate?:

This paper reanalyzed four previous studies and conducted a new study to investigate the effects of preemption and entrenchment. Different from the previous papers, the authors chose to use chi-square statistics as oppose to frequency of the constructions to measure preemption and entrenchment. Because their reanalysis of the previous studies and analysis of the new study yielded somewhat inconsistent patterns and were not conclusive as to whether preemption and entrenchment has independent effects, they conducted a meta-analysis with all data sets.

In general, I find the paper well-written. The authors provided clear and adequate explanations of the issues with the previous studies and how they made their choices for statistical models. I also appreciate the amount of effort and consideration they put into the corpus work which can benefit any research on this topic.

I have a few concerns about the statistical choices and the implications. First, the authors made the point that difference score can have several problems. While the concerns are very reasonable, I wonder, in their raw score analyses, how they factor out random noise brought in by pragmatic and semantic factors (if the sentence is degraded because of semantic infelicitous).

Another concern is that their preemption and entrenchment predictors are calculated out of chi-square values. This means that this score will be higher for verbs with higher frequency compared

to a lower frequency verb (assuming that the proportions of the constructions are constant). That is, if I understand the implementation correctly, the predictor score is higher in the case of 50 vs. 30 compared to 5 vs. 3. I am wondering whether this is a consequence we want and whether this is consistently with the empirical findings if any. A related but probably a tangential concern is: how much weight does this model give to errors in the input? Is it consistently with the pattern in verb learning?

In addition, the authors mentioned that there is collinearity between the predictors. In their model comparison, they removed predictors. I wonder how different coefficients changed during this procedure and whether they tested the effects the collinearity brought into their final model (such as VIFs). This may help us understand to what degree we should be concerned about the collinearity between the predictors.

2) Figures/tables/data availability:

Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:

The use of figures and tables are effective. There is a minor concern: Does the last column in Table 7 "All uses of pour/all other verbs except ground or figure locatives" mean "All other uses except figure locatives"? Repeating the row names seems a little confusing.

3) Ethical approval:

If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.

If not, please detail where you think a further ethics approval/statement/follow-up is required.:

Ethical approval has been obtained and consent has been collected for the original study in the paper.

4) Language:

Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:

The paper is well-written and the terminologies are clearly defined.

Reviewer H:

1) General comments and summary of recommendation

Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:

- Are the methodologies used appropriate?

- Are any methodological weaknesses addressed?

- Is all statistical analysis sound?

- Does the conclusion (if present) reflect the argument, is it supported by data/facts?

- Is the article logically structured, succinct, and does the argument flow coherently?

- Are the references adequate and appropriate?:

I am very enthusiastic about the publication of this paper, and believe it is an especially strong fit for Collabra. The methodologies used are robust and well-justified, and the studies themselves create a strong case that moves a literature beyond a 'stuck point' toward a more integrated model of verb learning and generalization. It is particularly nice to see the reanalysis of previous work under a shared format combined with self-replication.

My primary lingering concern is with some of the framing; in particular, I felt that the biggest and most striking conclusion from the research was hidden partway through the discussion of the models:

In particular, on the assumption that learners are not literally calculating chi-square statistics, a successful account is likely to be one that yields preemption and entrenchment as effects that fall naturally out of the learner's attempts to communicate meaning, rather than one that treats these effects as mechanisms in their own right.

Treating these inter-related phenomena as part of a more general learning strategy (that doesn't require each to be built in separately) is an important step, and I feel the abstract should reflect this conceptual change, rather than framing this as finding support for 'both entrenchment and preemption' - the authors convincingly (for me) overturn the notion that these two effects are particularly descriptive of human learning (vs. results of a broader learning algorithm).

My only other complaint concerns the models; they seem to be primarily citing the success of these approaches rather than directly comparing them on this larger dataset (which I would say is outside the scope of this paper), and so I wonder how much they add to the paper, or specifically how much the authors wish to stake out conclusions with respect to these modeling choices given their analysis of the behavioral data. In particular, there are many classes of learning algorithms in addition to the Rescorla-Wagner model that weight cue strength form positive & negative evidence - what specifically does the RW approach capture about this data? If the point is that a very general cue-weighting approach can & should be used to understand verb-argument structure learning, I am firmly on board, and would encourage the authors to focus on that instead of the particular implementation of one such model.

2) Figures/tables/data availability:

Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:

All data is available, and figures/charts are used well. I have no concerns here.

3) Ethical approval:

If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.

If not, please detail where you think a further ethics approval/statement/follow-up is required.:

Informed consent was given by all participants and the research was approved by the relevant universities.

4) Language:

Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:

The text is very well written, and relatively jargon-free. I would recommend additional explanatory text for non-language readers, possibly including a larger number of examples or diagrams/vignettes to clarify the nature of entrenchment & preemption. This paper is a beautiful model of careful, replicable science that should serve as a model to many scientists, so I'd encourage the authors to aim for maximum accessibility with this in mind.

Reviewer I:

1) General comments and summary of recommendation

Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not nov-

elty, topicality, or scope. A checklist of things to you may want to consider is below:

- Are the methodologies used appropriate?

- Are any methodological weaknesses addressed?

- Is all statistical analysis sound?

- Does the conclusion (if present) reflect the argument, is it supported by data/facts?

- Is the article logically structured, succinct, and does the argument flow coherently?

- Are the references adequate and appropriate?:

The present paper reports on five re-analyses of prior studies testing the relationship between corpus-derived measures of preemption and entrenchment on grammaticality judgments, a large-scale replication of one of those studies and a meta-analysis across the 6 studies. In general, I'm enthusiastic about this paper. I think the approach is well-founded and I appreciate that the authors rely on a meta-analytic synthesis of the five studies, rather than bending over backwards to interpret every non-significant effect. Another strength of this paper is its comparison of three models of learning that might account for the empirical results. I have some concerns about the paper in its present form, however. Most of these relate to the content and presentation of statistical analyses.

Major Issues

1) One of my major concerns is the justification of the chi-square statistic as a measure of contingency. In the categorical data analysis literature, the chi-square statistic is considered inappropriate for summarizing the degree of association between two categorical variables (see Agresti, 2013, Chapter 3). This is because the chi-square statistic is influenced by the raw frequencies in each cell. As a result, a verb that occurs 500 times in the double object dative and 500 times in the prepositional object dative construction would yield a different chi-square statistic than one that occurs 50 times in the double object dative and 50 times in the prepositional object dative. The odds ratio, a widely used measure of effect size when analysing contingency tables, is insensitive to raw frequencies and, therefore, would be identical in the two situations above. My initial inclination was to request the author's to use odds ratios, rather than chi-squared statistics as their measure of contingency. This would have the added bonus of likely reducing the collinearity between pre-emption and entrenchment scores, since any correlation due to verb frequency would be removed. If the authors are interested only in the degree to which a given verb's distribution differs from all verbs' distributions, then the odds ratio is a more appropriate measure.

However after some thought I realized that, in this study, the chi-square statistic's sensitivity to frequencies seems desirable, since, from a statistical learner's perspective, a moderate association in a large dataset will often be more informative than a strong association in a small dataset. If this is what the author's want, I think it's important to clarify that the chi-square test is quantifying how unlikely the observed deviation between a given verb's distribution and all verbs' distribution is under the hypothesis that the given verb shares the same distributional properties of all other verbs. The p-value, rather than the chi-square statistic per se, might be more interpretable in this context, since a .01 change in the p-value has a much clearer meaning than a 1 unit change in the chi-square value. But I don't think the results would change much at all.

2) The authors rely on two methods for making inferences about the effects of their entrenchment and preemption variables. First, they fit Bayesian linear mixed models with single predictors and examine the 95% credible interval for regression coefficients; second, they fit linear mixed effects models with and without each predictor and test whether its inclusion lead to additional reductions in the deviance, as indicated by the likelihood ratio test. I think their approach is justifiable, but there is an important assumption here, which should be made explicit---that any reduction in the deviance is due to an effect in the hypothesized direction. It is in principle possible that the effect of preemption (or entrenchment) toward the given construction could be significantly negatively related to grammaticality judgments once entrenchment (or preemption) has been included. I think this is unlikely to be true and I understand why, given the collinearity between predictors, the authors do not report on coefficients from these models. However, I don't want readers to get the wrong impression that a likelihood ratio test indicates something about the directionality of an effect.

3) The authors present a number of figures to summarize their results, which I think could be reduced dramatically. First, they present graphs of each of their maximum posterior estimate and

credible interval for each model. This takes up a lot of space, and I think no information is gained by using a visualization rather than a table. Additionally, if tables were used rather than figures, the authors could also include conventional diagnostic indices for MCMC, such as number of effective samples and the Gelman-Rubin convergence diagnostic (called Rhat in rethinking), for each statistic. Second, the authors plot the predicted values from each single-predictor model against the observed (mean) rating for each construction. I think this is a great idea; however I don't think they need to do it for the semantic predictors since those do not appear to be of substantive interest. To be honest, I'm not sure why the single variable semantic predictor models were included. It would also be helpful to plot the credible intervals around the regression line.

4) The authors present a meta-analytic synthesis of the studies. In general, I think this is an excellent idea, especially given the somewhat inconsistent pattern of results across the studies. However, I have a concern with their approach. Effect sizes are nested within studies in their analysis, and this is not accounted for using the random-effects meta-analysis. As the authors point out, explicitly modelling the effect of study wouldn't have been possible (although this wouldn't have accounted for correlated sampling errors due to the same participants being used to calculate multiple effect sizes; just correlated effect sizes from similar methods within studies), and I agree this would be difficult. An alternative approach, however, is robust variation estimation (Hedges, Tipton and Johnson, 2010). This is easily achieved in the metaphor package using the robust function (which also allows for a small sample size correction, which the authors might want to look into).

References

Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.

Hedges, L., Tipton, E., & Johnson, M. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. Research Synthesis Methods, 1(1), 39-65. 10.1002/jrsm.5

2) Figures/tables/data availability:

Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:

There appears to be adequate data and code to reproduce all figures and analyses.

3) Ethical approval:

If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.

If not, please detail where you think a further ethics approval/statement/follow-up is required.:

Ethics approval was reported for new study and I assume it was reported in prior papers that are re-analyzed here.

4) Language:

Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:

The text is sufficiently clear.

Editor Decision for Version 1

Editor: Fernanda Ferreira, Editor
Affiliation: University of California, Davis, US
Editor decision: Revisions Required
Decision date: 27 March 2018

Dear Prof. Ambridge,

Thank you for submitting this interesting, well written paper. I now have three reviews of your manuscript, appended in full below. As you will see, all three believe this work should be published in the journal, and they have also provided a number of thoughtful, constructive comments which I would like you to address in a revision. The concerns relate to two aspects of the paper. First, the reviewers suggest alternative statistical approaches to your data, including a rethink of whether the chi-square statistic is the appropriate for your analyses. Fortunately, the reviewers' statistical comments are generally consistent with each other, so you do not need to sort among conflicting recommendations. The other concern, however, is about your treatment of learning models, and here the reviewers do not quite agree: Reviewer H believes that aspect of the paper doesn't contribute much because your data can't choose among them, but Reviewer I views this discussion as a strength. I'm inclined to think your discussion of the models is useful, but please review Reviewer H's ideas for improving the presentation.

I am confident you can address these points in a revision, so I hope you will resubmit this work to Collabra. If you do, I will either evaluate the paper myself or send it to one of the reviewers of this version to see whether they have any remaining concerns. Either way, I expect to be able to get back to you quickly with a final decision.

The full review information should be included at the bottom of this email. There may also be a copy of the manuscript file with reviewer comments available once you have accessed the submission account.

To access your submission account, follow the below instructions:

1) login to the journal webpage with username and password

2) click on the submission title

3) click 'Review' menu option

4) download Reviewed file and make revisions based on review feedback

5) upload the edited file

6) Click the 'notify editor' icon and email the confirmation of re-submission and any relevant comments to the journal.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been

obtained. This is the last opportunity for major editing;, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please do contact us.

Please could you have the revisions submitted by end of April? If you cannot make this deadline, please let us know as early as possible.

Thank you for sending us this very interesting manuscript. I look forward to seeing the next version.

Best wishes,

Fernanda Ferreira, Editor

University of California, Davis

Author's Response to Review Comments for Version 1

Author: Ben Ambridge
Affiliation: University of Liverpool, Liverpool, UK; ESRC International Centre for Language and Communicative Development (LuCiD), SE
Revision submitted: 09 April 2018

Hi - thanks for the reviews - please find our cover letter attached

Thanks

Ben

Attached document:

collabra-4-133-pr_Auth_Resp_1.docx

Responses for Version 2

Reviewer A:

1) General comments and summary of recommendation

Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:

- Are the methodologies used appropriate?

- Are any methodological weaknesses addressed?

- Is all statistical analysis sound?

- Does the conclusion (if present) reflect the argument, is it supported by data/facts?

- Is the article logically structured, succinct, and does the argument flow coherently?

- Are the references adequate and appropriate? *

I appreciate the authors' effort in responding to my comments. They have addressed most of my concerns. I appreciate their careful discussion on their statistical choice. I agree with the authors that the lack of "consistent approach to either model building or significance testing" in the previous studies is concerning and I am excited to see a paper like this one.

As I noted in my previous review, I also highly appreciate their work on the corpus and think it can be very useful for future research.

2) Figures/tables/data availability:

Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here). *

After the explanation and correction of the typo, I find the figures easy to understand. I agree with another reviewer that the tables are big and that makes them a little hard to read. I am sympathetic with the authors' debate on the balance between the amount of information and readability. I personally do not have a big issue with the figures/tables.

3) Ethical approval:

If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.

If not, please detail where you think a further ethics approval/statement/follow-up is required. *

The authors have sufficient ethics approval.

4) Language:

Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process. *

In their revision, the authors explained the critical concepts with sufficient examples. I found this very helpful.

Reviewer C:

1) General comments and summary of recommendation

Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:

- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate? *

Overall, I think this is a strong submission. I think the two paragraphs added to justify their use of the chi-square statistic over other measures of contingency is very helpful. At this point, my only one concern regards the meta-analysis. The syntax provided isn't a multi-level meta-analysis; it's a random effects meta-analysis with an unconventional random effects specification. The current model allows true effects to vary randomly across studies, but not within studies and, therefore, contains two random components: a between study variance, and a sampling error variance. A multi-level meta-analysis would allow effects to vary randomly between and within studies and would, therefore, contain three random components: a between study component, within study component and sampling error component. This could be achieved by modifying the syntax to random = ~1 | StudyID/XXXX, where XXXX is the name of a variable that labels rows in the dataset. This is a pretty common issue with metafor. The author actually mentions it on the package website: http://www.metafor-project.org/doku.php/analyses:konstantopoulos2011 Section: A common mistake in the three-level model.

You could try running a multilevel meta-analysis, but I doubt there is sufficient data. I also think the cur-

rent approach is justifiable, since it handles the dependence between observations from the same study. I would just rather it not be described as a multilevel meta-analysis so as not to give readers the wrong impression about such models are run.

2) Figures/tables/data availability:

Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here). *

The figures look good. The data are available.

3) Ethical approval:

If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.

If not, please detail where you think a further ethics approval/statement/follow-up is required. *

Ethics approval for new study reported.

4) Language:

Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process. *

Clearly written and jargon free.

Editor Decision for Version 2

Editor: Fernanda Ferreira, Editor
Affiliation: University of California, Davis, US
Editor decision: Accept Submission
Decision date: 22 May 2018

Dear Prof Ben Ambridge,

After review, we have reached a decision regarding your submission to Collabra: Psychology, "Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis", and are happy to accept your submission for publication, pending the completion of copyediting and formatting processes.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point. The accepted submission will now undergo final copyediting. You will be contacted once this is complete to answer any queries that may have arisen during copyediting and to allow a final chance to edit the files prior to typesetting. If you wish to view your submission during this time, you can log in via the journal website.

The review information should be included in this email.

Thank you for submitting this excellent work to Collabra. We hope you will continue to consider Collabra as an outlet for publishing the best research from your lab.

Kind regards,

Prof Fernanda Ferreira

University of California, Davis

fferreira@ucdavis.edu