Comments on Álvarez-Benjumea, A., & Winter, F. (2018). Normative change and culture of hate: An experiment in online environments. European Sociological Review, 34(3), 223-237.

The research aimed to investigate the effect of descriptive and injunctive norms to determine how people respond to a number of sensitive topics on social media. It was assumed that individuals would be less likely to drop hateful comments and therefore, more likely to conform to the group norms when the abhorrent comments are moderately censored. The authors found that descriptive norms (censoring) are substantially more influential in decreasing the occurrence of hate speech than injunctive norms (actively confronting the violators). The authors sought to test their hypotheses by comparing hate speech scores of three different experimental conditions; extremely censored (allowing positive comments only), censored (deleting all derogatory comments, yet allowing both positive and neutral comments), and counter-speaking (reprimanding hateful comments by giving explicit punitive measures). After randomly assigning participants into three different groups, participants were asked to give comments on 9 controversial pictures that were previously selected in a preliminary study.

The research highlights important insights and carries a number of methodological merits. The findings supported multiple previous research that heavily censoring unwanted contents on social media might entice psychological reactance and therefore it’s a less effective strategy to reduce hostility on social media. However, past research showed that psychological reactance varies across cultures and its effect to entice forbidden behaviors is mediated by individuals’ prior experience of censorship, especially when it comes from a very powerful source (Ng et al., 2019). Cultural variation was unfortunately left out of the research design so that the authors were unable to take it into account. There was not enough evidence to conclude that reactance is a cause of a more prevalent hate speech (in extremely censored condition) as it requires two requirements; participants must believe that their freedom is threatened and perceive that this freedom is salient to them. Both were missing in the research design.

Although a discussion forum is the basis of most social media platforms, the actual social media space is more complex than the authors took into consideration in their experimental design. On Facebook or Twitter, for example, apart from writing comments, users are allowed to share content (by reposting or retweeting) and therefore, further spread its message to their circle of network. This seamless stream of information is profoundly influenced by the design of the platform according to the MAD model of moral contagion (Brady et al., 2019). It would be also interesting to look at whether participants would share the content and state their own emotional-moral expression after observing others’ responses to the same content.

As for measuring hate speech score, the authors thoughtfully opted for Krippendorf’s α to measure its reliability, which is a terrific decision since a more popular alternative (Cohen’sκ) is unable to cancel out the effect of the number of coders and the number of categories. The author chose to analyze their data using a linear mixed model which has many benefits compared to the ordinary ANOVA, yet using a within-subjects design by exposing the participants with all conditions (and applying complete counterbalancing) could be more powerful in detecting the effect than using unmatched between-group design. The findings could be improved by measuring right-wing authoritarianism (RWA) and treat it as a control variable. RWA is too important to be left out as people with high RWA tend to express prejudice towards outgroups, but are more vigilant to the violations of the norms so that they would be more likely to support hate-speech prohibition (Bilewicz et al., 2017). The authors could apply matched-group design by making sure that the RWA mean is equal across conditions or including it as a predictor in the model.

The report could have been more informative and reproducible if the authors had reported χ^2value, degree of freedom, confidence interval of each predictor-level estimate, exact p-values of each predictor and interaction term, likelihood ratio test (as a complementary of ICC), and both marginal and conditional R^2 of the model. The manuscript contains a minor inaccuracy as the authors wrote that Satterthwaite method was applied to determine p-value, while it is actually a method to *indirectly* estimate p-value by first, estimating the denominator degree of freedom, then calculating *F*-statistics from which p-value is estimated.

References

Bilewicz, M., Soral, W., Marchlewska, M., & Winiewski, M. (2017). When Authoritarians Confront Prejudice. Differential Effects of SDO and RWA on Support for Hate-Speech Prohibition. *Political Psychology*, *38*(1), 87–99. https://doi.org/10.1111/pops.12313

Brady, W. J., Crockett, M., & Van Bavel, J. J. (2019). *The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/pz9g6

Ng, A. H., Kermani, M. S., & Lalonde, R. N. (2019). Cultural differences in psychological reactance: Responding to social media censorship. *Current Psychology*. https://doi.org/10.1007/s12144-019-00213-0