Author's Response to the Review of 'Discrimination that Matters: Replication with Extensions of "Perceived Discrimination and Political Behaviour" (2020)'

The author would like to thank the editor and reviewer for their constructive comments and suggestions that have helped improve the quality of this manuscript. The manuscript has undergone a thorough revision according to reviewer's comments. Please see below provided responses.

Response to the reviewer

Reviewer Comment 1.1 — On p. 7 the author states "Generalised linear models contain unobserved heterogeneity that is independent of the outcome variable and dependent on variances of omitted control variables (Wooldridge 2010; Mood 2010)", which is not correct for all glms (e.g. not for a gaussian glm with identity link).

Reply: The comment was taken into account and the sentence was changed to refer only to relevant logit models: 'Due fixed variance of the error term in logit models, the effect of the key treatment variable also contain degree of unobserved heterogeneity (Mood 2010; Wooldridge 2010).' Sentence is located on page 10.

Reviewer Comment 1.2 — However, the concern about the comparison of coefficients on the link scale and odds ratios can be warranted, but does not necessarily extend to comparisons of differences in probability! (see the same Mood 2010, but also Holm et al. 2015; Breen et al. 2018). If the analyzed units are correctly averaged over in the average treatment effect on the treated, these are comparable as well (cf. Kuha & Mills 2018 about the different meanings of unobserved heterogeneity). In particular, the scaling and comparison issue has nothing to do with maximum likelihood estimation (p. 26).

Reply: Write later when you're done with the other comments. - requests about the estimand etc.

Reviewer Comment 1.3 — Due to a misunderstanding of the extent of incomparability, the paper also refrains from interpreting the substantial impact of discrimination, which is falling behind the state of the art for logistic regression models using predicted differences, and (average) marginal or partial effects.

Reply: This point was taken seriously into account and for Studies 1, 2 and 4 average marginal effects are calculated in `R`. Complete results are available in the Appendix: Tables 3 and 4 (p. 5), Table 6 (p. 9), Table 8 (p. 11), Table 10 (p. 13), Table 12 (p. 15), Table 14 (p. 19), Table 16 (p. 21), Table 67 (p. 65), Table 69 (p. 66), Table 71 (67).

Reviewer Comment 1.4 — Also correctly averaged over odds ratios and risk ratios can be compared.

Reply: I consider this comment to be substantively the same as comment Comment 1.2, therefore the relevant response can be found in the reply for comment Comment 1.2.

Reviewer Comment 1.5 — This issue is compounded by interpreting the sign ("directionality") of the results, even when the confidence intervals extend to seemingly similar negative plausible values of the coefficients in the population (e.g., p. 16). A more careful formulation might be better: "The effect is still positive in the observed data".

Reply: This comment was taken into account with utmost attention as a follow-up comment to Comment 1.1, Comment 1.2 and Comment 1.3. Following the previous comments about the interpretation of substantive effects, directionality ceased to be the key element of interpretation, which consequently diminishes the problem pointed out in the comment. In cases where directionality was mentioned, careful formulations were used to describe the implications of directionality.

Reviewer Comment 1.6 — The original paper and the replication use various alternative versions of what the author calls the "control matrix", discussing them purely as potential confounders while they also change the comparison group in logistic regression (and some matching strategies) and could also be mediating and moderating factors (or outcomes of the political behavior), leading to an increased danger of overcontrolling and collider bias (Spector & Brannick 2011). Why should political discrimination be controlled by social discrimination and vice versa? Is integration in social networks an independent predictor and not an outcome of identity formation? It's also conceivable that social integration strongly moderates the effect of discrimination, making an additive model doubtful. These types of questions should be clarified in the paper.

Reply: This is a very insightful comment which helped in clarifying the role of specific controls. It is true that some of the control variables from original models, such as trust in parliament or political efficacy (vote as duty and satisfaction with democracy) could also be outcomes which exposure to societal and political discrimination could impact. The original research brought both fully developed models and simplified models that did not contain potentially confounding variables and results remained unchanged when these variables were excluded. I took this as a proof that four variables in question severe confounders and colliders. Studies 1 and 2 (by design) also utilised models that do not contain 4 potentially confounding and colliding variables. Additionally, theoretical mechanisms in the original paper might suggest mediation argument , but with variables such as discontent/frustration, depression, insecurity, inferiority and low self-esteem.
Societal and political discrimination can be used as controls in a single model because despite both being types of discrimination with similar behavioural patterns and outcomes, their origin is different, which makes them distinct and independent from each other. Independence of societal and political discrimination from each other, provides solid ground for using both variables in a single model. Combining different forms of discrimination in a single model is also a common practice in political science and social psychology (i.e. group and individual discrimination in well known studies such as: Bourguignon et al., 2006; Branscombe, Schmitt, and Harvey, 1999; Schildkraut, 2005). The explanation has been included in the paper.
The point about the integration in social networks was taken into account by removing this variable from the final estimation of models, due to the noted risk of reversed causality. However, there are sources that indicate that identity formation might happen after participating in specific social networks (see i.e. Dolberg and Amit, 2023).
Final point about social integration (community involvement) as a moderator of societal and political discrimination has also been taken into account. Models that include this variable were retested in the following manner: community involvement as moderator of both societal and political discrimination; community involvement as moderator of just societal discrimination, but not political discrimination;

community involvement as moderator of just political discrimination, but not societal discrimination. None of the interaction terms is statistically significant. A footnote was added on page 9 of the paper and results are available in Section 4 of the Appendix.

Reviewer Comment 1.7 — In general, it would be helpful to clarify the theoretical and empirical estimands (Lundberg et al. 2021) and if they differ between the original study and the replication. This also could be used to contrast how covariates are used in the regression model and matching and how this could support the argument for using the average treatment effect on the treated instead of the average treatment effect (p. 11).

Reply: To apply this comment fully, entire sections of Replication Design, Analysis, and Discussion and Conclusion were re-written. The ordering of studies was changed compared to the previous version of the paper. Study 3 from the previous version is now presented as Study 2 and vice versa. The idea is to present studies that use the same dataset and same methods before presenting studies that contain either different methods (Study 3) or different datasets (Study 4). Differences between estimands was used to set limitations in comparisons between studies. The reviewer suggested that estiimands could help in understanding differences in the use of covariates in matching and regression models. This point was clarified through limited comparability of results as suggested in the Replication Design section and subsequently in the discussion. Covariates in this case refer to treatment variables (discrimination variables). In case of control matrix, the same set of covariates was used in the regression models and in matching. As indicated matching was done based on the entire control matrix, which was later used in the regression models, prior to G computation of average treatment effect on the treated (ATT). The key argument for use of ATT instead of average treatment effect (ATE) is given in footnote 11, page 14 of the paper. It is said that ATT is more pertinent effect to estimate using matching procedure. The main reason for this is that the matching procedure re-weights the entire sample by taking into account that discriminated individuals represent a minority in the sample. ATT in this situation is much more convenient for detection of the effect of exposure to discrimination in the new sample. Estimating ATT utilises the available treated observation and with balance between treated and untreated data it also helps in estimating more reliable effects of discrimination. Additionally, since nearest neighbour and matching was used in all models, the literature also suggested that ATT is the effect that can be reliably estimated using this methods, while optimal pair and optimal full matching can be used for ATE and ATT respectively.

Reviewer Comment 1.8 — The different matching procedures (optimal full, nearest neighbor, and optimal pair) are not clearly described in the paper, leaving the reader without any potential explanation for the different results of these matching schemes. If there are good arguments to prefer any one of these matching algorithms, the other ones could also be relegated to the appendix.

Reply: This point has been taken into account. A footnote has been added on page 11 giving the source to the vignette with detailed explanation of differences in these methods. The vignette suggests that no specific quality of different matching methods should make one preferable when it comes to estimating the ATT. For that reason all reported results were left in the paper as they were.

Reviewer Comment 1.9 — Alongside complex robustness checks, I am missing tests of basic assumptions of the gen- eralized linear models used (Pregibon 1981, see Dunn & Smyth 2018: 297ff.), like (most importantly) linearity on the logits and the impact of influential outliers, and less pressingly homoscedasticity (but see Leamer 2010; King & Roberts 2015) and independence of errors. One approach requiring relatively little effort would be plotting the quantized residuals on the main predictors of interest.

Reply: This is a very valuable point and it thoroughly addressed. Entire Section 1 in the Appendix (pp. 3-5) has been updated with additional graphs and tables in the following order: Figure 2: Partial Residual Plots of Original Models that test for linearity of the logits; Figure 3: Difference in Fits of Original Models to check for influential outliers; Table 2: Breusch-Pagan Test Results for Relevant Original Models to check for heteroskedasticity. Additionally, Tables 3 and 4 bring average marginal effects of logit and multinomial models from the original paper. Results indicate that original results report linear effect with no significant outliers, although heteroskedasticity is present in the original models.

Reviewer Comment 1.10 — While basic assumption tests are generally important to better evaluate the validity of the orig- inal study, doing other robustness tests makes missing them more striking. This is particularly important when we take the highly skewed distribution of observed perceived discrimination into account (p. 10), which does not increase my confidence in the linearity of the effect. Especially the difference between the high percentage of people who do not perceive to be discriminated against and people who have a score of 1 is unlikely to be the same as between people on the highest level of discrimination. The very low number of people with very high values in perceived discrimination also makes a comparison of predicted probabilities between maximum and minimum highly uncertain and somewhat dubious. The matching procedure uses only a binary treatment indicator, which implies a different theoretical estimand that should be made more explicit in the paper.

Reply: This is also a very valuable comment. Results of Partial Residual Plots (mentioned in the reply to Comment 1.9) indicate that effects of societal and political discrimination are linear, despite skewed distributions with respect to different levels of discrimination. These assumptions test indicate that comparison of predicted probabilities between maximum and minimum a meaningful strategy. Explanations regarding different estimands for matching procedures are provided in re-written Replication Design section in the paper (already explained in reply to Comment 1.7).

Reviewer Comment 1.11 — Like the original study, the paper confuses the terms likelihood (of the data given the model) with probability (of events in a stochastic process, e.g. voting). In addition, hypotheses 3 and 4 are phrased as if in-group attachment and engagement would have been measured on a continuous scale, while they are modeled as binary indicators too. If the main interest is in the unobserved latent variable, it should be measured directly (Kuha & Mills 2018).

Reply: This is a very insightful comment. Based on the comments, I updated formulation of my original hypotheses based on this comment, although I left hypotheses of original paper in their source formulation, with a footnote indicating there is a confusion in the formulation. In the second sentence, due to lack of specification, I can only assume that the reviewer refers to original hypotheses, not the original hypotheses from Studies 2-4. The intention of the robustness checks was not to change original hypotheses. The original formulation uses the term 'enhances' which doesn't seem to imply a continuous scale in the undercurrent of these hypotheses. Therefore, this comment has been taken into account, but no further changes in the paper were inserted, since it was not the intention of Study 1 to change or alter the original set of hypotheses. In general we agree it would have been much better if ethnic-based engagement and in-group attachment had been measured on a continuous scale.

Reviewer Comment 1.12 — The paper currently includes references to tables in the appendix, which are not part of the manuscript. They should either be added in the text (where they are important), to the appendix of the manuscript or referenced as online supplementary material with a fixed identifier instead. In the spirit of the webinar, it would also be helpful to have a short clarification if the original study could be reproduced, accordingly pointing to supplementary material.

Reply: This is a very useful point. Following this point, I double checked and synchronised the tables and graphs from the paper and the appendix. I also made sure that each table in the paper document has a fixed identifier. To account for reproducibility of the original results, I added a sentence at the beginning of the analysis with a link to the original replication files and code in the footnote.

Reviewer Comment 1.13 — Non-significant effects should not be interpreted as no effect (e.g., p. 19 & p. 29) without equivalence tests if the effect is with high confidence below a specific threshold.

Reply: This is a very insightful comment. I changed the entire interpretation of results to account for non-significant-effects as no effect, considering that no equivalence tests were conducted.

Reviewer Comment 1.14 — Two additional minor issues should be resolved as well: It is unclear which confidence intervals are reported, e.g. in Table 1 (probably 95%-CIs?) and the number of decimal points in tables should not go beyond the uncertainty of the estimates (e.g. at most 2, maybe just 1 decimal point in most cases).

Reply: This is a very useful comment. I updated tables in the paper with relevant information indicating that 95% confidence interval were reported in these tables. Considering that usual effects detected in the political and social psychology are small, the decision at the beginning was to keep 3 decimal places in reporting results. On the other hand, since the original paper reported results with 3 decimal points, therefore I deemed that it is meaningful to follow the same strategy of reporting as in the original paper.

Reviewer Comment 1.15 — The paper discusses arguments supporting effects that were not consistently observed in the replication (p. 29), which is somewhat confusing. They can either

be deleted, used in the original derivation of the hypotheses, or more clearly be connected with a discussion of why they might not be observable with the methods used in the replication.

Reply: This is a very important comment. All arguments supporting effects that were not consistently observed were deleted from the paper. Instead, the discussion has been rewritten to reflect only the observed effects. Following Comment 1.13 observed effects were regarded as consistently statistically significant effects in a single study. Considering that some studies share conceptually similar variables, a larger-scale comparison was also attempted at the end of the discussion.

Reviewer Comment 1.16 — While the author understands his matching analysis as a robustness test, the procedure allows causal inference, which was not the focus of the original paper. Hence, applying matching is more than a simple robustness check and could be much more prominently discussed and expanded upon. These more in-depth causal analyses would probably lead to a longer paper, which could be remedied by a second recommendation: The replication paper currently lacks a clear focus, it is not entirely clear which question it tries to answer, analyzing many interesting aspects at the same time. Focusing either on (1) robustness, (2) causal inference, or (3) generalizability outside the original outcomes and sample could alleviate this. An orientation towards which goal the replication has, could be helpful here (e.g. generalization with different methods and data, Freese & Peterson 2018). Additional material could be provided in the appendix.

Reply: This is a very valuable and insightful point. To address the first remark about matching being an approach beyond simple robustness check, I adapted the organisation of the paper by rearranging the analysis and conceptually designating matching analysis as an extension. The key justification for such move was discussed in the Replication Design through highlighting differences between estimands in robustness checks and extensions.
The second remark offers also an important guidance on how the paper might have been conceptualised and reorganised. The Replication Design section is rewritten, with three aims in mind: 'The aim of Studies 1 and 2 is to check robustness of original results with respect to different controls and outcome variables. Study 3 intends to extend original conclusions by testing for causal effects of different types of discrimination on political behaviour. Study 4 aims to extends the original analysis on different population of newly arrived immigrants (in UK and the Netherlands).' Even though there are no specifically formulated questions, from the aforementioned quote it is evident that this replication intends to extend out knowledge of the effects of societal and political discrimination by checking the robustness of original results, but also by applying causal inference on the same sample as the original study and also to check if results of the original study can be replicated on a different population with the same methods used in the original study. Due to limitations of the available data and impossibility of their extension, the approach taken through multiple studies and without central focus reflects the exploratory character of this replication. '

Reviewer Comment 1.17 — Alternatively, the separate aspects could be combined by a central thread discussing if the original paper's claims have causal validity. The current elements could be understood as tests of various threats to causal inference, also allowing for a final discussion of

some limitations that have been ignored so far (potential endogeneity, moderating and mediation factors, and especially the danger of misattribution of cause and effect in the cross-sectional analysis). This reorganization would probably necessitate also using matching analysis to analyze the Dutch data-which is currently noticeably missing.

Reply: This is also a very valuable point and it has been thoroughly considered and taken into account. As Comment 1.7 indicated, and which was later implemented in Study 3, there is a difference in estimand in this study, and the original paper and extensions in Study 1. For that matter, discussion of results refrained from framing results in Study 3 as, making a strong causal claim about the mechanism in the orginal paper. However, Study 3 acknowledges that just the exposure to discrimination, regardless of the intensity of discrimination, does have some causal effects with respect to original mechanism. Current elements could indeed be understood as tests of various threats to causal inference, but because these were not pursued from the beginning in the aims of the study and were not conceptualised from the beginning as such, this opportunity was not pursued. Trying so subsume all four studies as checks of threats to causal inference would also risk diminishing the significance of standalone results of each study. It is also possible that such framing of different studies could indicate that the original study tests the causal mechanisms using causal inference, which is not the case, because it offers a causal mechanism, but it does not use the causal inference methods to test it. For that matter, as an appropriate method (not pursued in this paper) for testing such mechanism could also be a causal mediation analysis (Imai, Keele, and Tingley, 2010). To address the last remark, the matching analysis of the Dutch data was considered infeasible from the very beginning, considering a very small ratio of respondents who reported discrimination in any form (check figures 21 and 22 in the Appendix).

Reviewer Comment 1.18 — The paper currently suffers from missing parts of sentences, grammatical errors, and other orthographic inaccuracies. Additional editorializing would greatly improve the quality of the paper and might also help with some smaller issues of unclearness.

Reply: Thanks for the last point. To address this point I re-read the paper and edited the manuscript thoroughly. I also used the edited document that the reviewer sent after sending the letter to address all grammatical and orthographic errors in the paper. I am very grateful for the effort Reviewer put in doing the editorial work, which is beyond the scope of the Review itself.

References

Bourguignon, David, Eleonore Seron, Vincent Yzerbyt, and Ginette Herman. 2006. "Perceived group and personal discrimination: differential effects on personal self-esteem". *European Journal of Social Psychology* 36 (5): 773–789.

Branscombe, Nyla R., Michael T. Schmitt, and Richard D. Harvey. 1999. "Perceiving Pervasive Discrimination Among African Americans: Implications for Group Identification and Well-Being". *Journal of Personality and Social Psychology* 77 (1): 135–149.

Dolberg, Pnina and Karin Amit. 2023. "On a fast-track to adulthood: Social integration and identity formation experiences of young-adults of 1.5 generation immigrants". *Journal of Ethnic and Migration Studies* 49 (1): 252–271.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis". *Psychological Methods* 15 (4): 309–334.

Schildkraut, Deborah J. 2005. "The Rise and Fall of Political Engagement among Latinos: The Role of Identity and Perceptions of Discrimination". *Political Behavior* 27 (3): 285–312.