Dear Veli-Matti Karhulahti,

Once again, we thank both you and the reviewers for the time you devoted to our manuscript “Unveiling the Positivity Bias on Social Media: A Registered Experimental Study On Facebook, Instagram, And X”. We are pleased to be able to submit it for Stage 2 evaluation at *PCI Registered Reports*.

We have carefully considered all your feedback. However, we would like to point out that we encountered some methodological challenges in performing equivalence testing on our repeated measures model with covariates. While we included a simplified analysis in line with the reviewer’s request, the results we obtained seem counterintuitive (e.g., non-significant TOST results alongside significant NHST results). We remain open to any discussion or suggestions regarding this aspect to improve our approach.

Below are our point-by-point responses to each comment. In addition, on OSF, a revised version of the manuscript, with all modifications indicated in Track Changes mode, and a PDF cleaned version, are available.

Thank you for your consideration of our work. We look forward to your feedback.

Kind Regards,

The authors

Table of Contents

Editor - Veli-Matti Karhulahti

<u>General comment:</u>

Thank you for your carefully written Stage 2. We were lucky to have all 3 reviewers return to assess the full manuscript. In general, everyone is happy with the outcome, with a few identified issues to be revised before recommendation. Please consider the reviewers' feedback carefully. In addition, I list minor points that I noticed myself.

<u>Response:</u>

We sincerely thank the editor again for his thorough feedback and dedication to our manuscript. We have carefully revised the manuscript as suggested.

<u>Comment 1:</u>

In Stage 1 content p. 12 it says "does not allow us to answer RQ2" but my impression is that the data answers RQ2 negatively. You can correct this possible error in Stage 1 content.

<u>Response:</u>

We agree with the editor. We have revised the text as follows: "However, we found no significant differences between the three social media, which answers negatively to Pilot RQ2." (p. 12).

<u>Comment 2:</u>

On pages 8-9, I would also suggest changing "RQ1" (RQ2, RQ3, RQ4) to "Pilot RQ1" etc to help readers separate pilot and main RQs. If you choose to do this, please make the change coherent by editing the terms elsewhere in the ms too.

<u>Response:</u>

We have made this change throughout the manuscript for consistency.

<u>Comment 3:</u>

For coherence, you may wish to add a short sentence about Pilot RQ4 at the end of 3.3.4. (the only section in 3.3. that doesn't end with summary sentence)

<u>Response:</u>

We have added the following summary sentence: "Therefore, these analyses provide more information on the socio-cultural context of social media. The findings revealed that the norms for emotional expression do differ between Facebook, Instagram and Twitter/X (Pilot RQ4)." (p. 14).

Comment 4:

Please proofread the text (section 4 onwards) to ensure language is consistent (e.g. “The main research will therefore aims to address…” and especially the use of tense later on).

Response:

We have carefully proofread and corrected the language and tense throughout section 4 and the rest of the manuscript.

Comment 5:

On page 15, I suggest adding “RQ” at the end of the question: “how does the positivity bias manifest on social media, and does it vary depending on the type of social media platform? (RQ1)”

Response:

Instead of adding “RQ,” we decided to remove the question altogether to streamline the text: “The main research focused on understanding how the positivity bias manifests on social media and whether it varies depending on the type of platform.” (p. 16).

Comment 6:

For consistency, end of p. 16, perhaps change RQ to RQ2 and perhaps “We therefore additionally explore our pilot research question”

Response:

Since we did not retain the RQ phrasing in response to Comment 5, we did not change "RQ" to "RQ2."

Comment 7:

There are some parts in the discussion that I cannot fully follow. I list some of them here:

Response (after each comment for readability):

- “It may only reflect the increased control over self-presentation that social media conveys, compared to real-life interactions.” --> how would the obtained result of no difference reflect increased difference in social media vs real-life? I cannot follow the reasoning, please clarify or remove.

We clarified that this statement refers to findings from prior literature, not our results: “Therefore, what is often referred to as a positivity bias on social media (Reinecke & Trepte, 2014; Spottswood & Hancock, 2016; Utz, 2011) may not necessarily represent a

bias per se, but rather a consistent pattern of positive self-presentation and self-enhancement that occurs in both online and offline contexts. The literature findings may only reflect the increased control over self-presentation that social media conveys, compared to real-life interactions." (p. 23).

- "Therefore, instead of reflecting a cognitive bias towards positivity, the results may be indicative of the inherent design of social media platforms, which naturally lead users to present a more positive version of themselves." --> I cannot follow this. If people have, as the paper suggested earlier, positive bias in all social contexts, it's the people who have inherent bias (not the design). People would use any kind of social interaction design in ways that show them in a positive light to others, no? Please clarify or remove.

We agree and have removed this statement.

- Unless I've misunderstood something earlier, it might be good to remove this from conclusions p. 26: "this may still reflect the inherent control that social media gives users over their self-presentation. Rather than demonstrating a distinct positivity bias, our findings suggest that the architecture and affordances of social media platforms encourage behaviors aligned with self-enhancement and positive self-presentation, offering users tools to manage their image effectively." --> I don't see any evidence was obtained that would support claims of encouragement by design. Only the unidirectional design was logically connected to the distinct nature of about X, right? (see below)

We agree with the editor. We revised this section to better align with our data: "While we did not observe significant differences in valence between events shared with friends and those posted on social media, this might be explained by the fact that our design gave participants the same degree of control in both conditions, which is not a reflection of real-life situations." (p. 27).

- What is said on p. 24 sounds logical: "Twitter/X's unidirectional mode of connection fosters less intimacy and reciprocity, which could increase the need for positive self-presentation." The results clearly support this new hypothesis, and it's an important finding. When it is noted later that "By focusing on these platform-specific characteristics…" I would encourage you to more explicitly pinpoint those specific features/mechanisms (if there were others beyond unidirectionality) that you believe were related to the finding. It would also be valuable to explain to readers what are the likely reasons for Avalle et al 2024 not finding such differences (there must be several aspects that explain it in their design vs yours).

We revised this section to focus on the connection mode: "By focusing on these platform-specific characteristics, like the connection mode, researchers can move beyond merely noting differences between platforms to understanding the underlying mechanisms driving these differences." (p. 25). We think that our findings do not allow us to go further into these characteristics, they mainly offer perspective. We also added a specification on the design of our study in comparison to that of Avalle et al. (2024): "While we found differences in our main study, some studies with other designs, such as opinion mining, did not reveal differences between platforms (Avalle et al., 2024)." (p. 25).

- Still on p. 24, it is said "This finding contrasts with prior research suggesting that users are more likely to post messages containing emoji on social media (Daniel & Camp, 2020), and messages featuring emoji are often perceived as more positive than those without (Novak et al., 2015)." Please clarify what the cited papers claim (e.g. users are more likely to post emoji on social media -- more likely compared to what?)

The paper we cited was an experiment, we specified it in a better way: "This finding contrasts with prior research suggesting that users consider posts with emoji to be easier to understand and more credible than messages without emoji or with an emoji inappropriate (Daniel & Camp, 2020), and messages featuring emoji are often perceived as more positive than those without (Novak et al., 2015)." (p. 25).

Comment 8:

One reviewer asks for the code so that we could reproduce the analysis. I can see you have data and code in OSF but there's no mention of it in the paper. Please add the links to the data and code in appropriate places.

Response:

The link to OSF was initially mentioned in "The Present Study" section. To improve accessibility, we have added a new "Data Availability Statement" section with the links to the OSF repository.

Comment 9:

That said, I was not able to use the provided code to reproduce the analysis and my colleague wasn't either. Could you please double check the code and, perhaps with a help of a friend or colleague, ensure that the code can be used by someone external to the study. E.g., as the paper preregistered ICC 0.75 and obtained exactly 0.75, post-publication readers might be interested in re-running the analyses [related: if some form of consensus negotiations were applied to improve rater consistency, please report those procedures, as discussed at Stage 1 -- recall transparency steps in data rating].

Response:

We reviewed and updated the code with the help of a colleague, who successfully ran the analyses. We uploaded a corrected R script instead of the prior HTML format and also included the R session version (4.2.1). The ICC of 0.75 was obtained without consensus, and these details are now included in the OSF repository.

Comment 10:

Two reviewers ask for a clarification about effect sizes. At Stage 1, we discussed this topic in detail. Adding more of this information to the discussion could clarify things to readers. The fact remains that we don’t know what ES are meaningful in practice. For me it was clear that your study was powered for .21 and you considered this ES interesting for that reason. That said, especially the main positive effect about X valence should be reported and its ES discussed.

Response:

Thank you for highlighting the importance of clarifying effect sizes. We revised the discussion to include some clarification: “Second, we selected an expected effect size of .21 based on prior meta-analyses (Ruppel et al., 2017), as a benchmark for powering our analyses. This value was chosen because it represents a theoretically plausible effect size in this domain, given the limited availability of direct comparisons across platforms. However, we acknowledge that this does not constitute a strictly defined Smallest Effect Size of Interest (SESOI). Instead, it reflects an expected effect size that was used for practical purposes in determining our sample size. The main positive effect on Twitter/X valence aligns closely with this SESOI, suggesting that our observed effects are within the range we deemed theoretically interesting at the outset. However, the meaningfulness of effect sizes, particularly in the context of social media research, depends on several factors beyond the magnitude of the observed effect.” (p. 26).

Comment 11:

Please add conflicts of interest information, and a link to your original registered Stage 1 paper.

Response:

We added a section for conflict of interest (p. 28) and the following sentence: “The main research was pre-registered, meaning that the study's hypotheses, methodology, and analysis plan were reviewed and publicly documented prior to data collection through the Peer Community In Registered Reports (PCI RR) initiative, ensuring greater transparency and rigor in the research process (https://rr.peercommunityin.org/articles/rec?id=666).” (p. 8)

Comment 12:

Finally, if there are notable changes that were made to Stage 1 content, please list them in the next response. We don’t have a tracked changes file to help us review the changes. E.g., comparing versions, I can notice that this section has been removed: “We set a predefined threshold for agreement at 0.75, indicating good reliability among raters. If the initial coding round does not meet the ICC threshold, will also conduct additional training sessions for new coders to ensure a clearer understanding of the coding criteria and reduce biases in subsequent coding rounds. The same procedure will be used to assess the valence of the description of the image/video associated with the post.” This is important information and should be brought back.

<u>Response:</u>

A tracked-changes version was present, the file name is "Manuscript_After Stage 1_After data collection_tracked changes". For this revision, we again uploaded a tracked-changes version of the manuscript to the OSF repository under the name "Manuscript_After Stage 1_After data collection-revised_tracked changes".

Reviewer 1 - Anonymous

<u>General comment:</u>

Thank you for providing the opportunity to review the stage 2 version of the manuscript "Unveiling the Positivity Bias on Social Media: A Registered Experimental Study On Facebook, Instagram, And X". This manuscript presents an interesting examination of the positivity bias on social media. The authors have effectively maintained consistency with their approved Stage 1 hypotheses and methodologies, and the results are clearly presented and well-supported by the data.

<u>Response:</u>

We sincerely thank the reviewer for their thoughtful feedback and for recognizing the strengths of our manuscript.

<u>Comment 1:</u>

A few suggestions for improvement include providing more descriptive statistics, such as the frequency of social media use.

<u>Response:</u>

We have addressed this suggestion by including descriptive statistics on social media use: "On average, participants' frequency of use of Instagram was higher (M = 6.22, SD = 1.27) than their frequency of use of Facebook (M = 5.58, SD = 1.76) and Twitter/X (M = 5.42, SD = 1.77)." (p. 18).

<u>Comment 2:</u>

Additionally, H2a should be also mentioned in the discussion section.

<u>Response:</u>

We have incorporated this point into the discussion section: "Specifically, we were not able to confirm H2a, but in line with H2b, posts on Twitter/X had more negative valence than those on Instagram and Facebook." (p. 24).

<u>Comment 3:</u>

I also noted a minor issue in Table 1 where the statement "H1 is disconfirmed" may actually refer to H2.

<u>Response:</u>

We have corrected the statement in Table 1 to ensure accuracy.

Comment 4:

Regarding Figure 4, the color choices could be adjusted for better visibility, as the grey and light blue shades may be hard to distinguish.

Response:

We have updated Figure 4 and Figure 5, replacing the light shades with darker, more distinguishable colors.

Reviewer 2 - Marcel Martončik

General comment:

I would like to thank you for the opportunity to review this manuscript after Stage 2. I commend the authors for their research implementation and results presentation – specifically for strictly adhering to the preregistered plan in both presenting and discussing the results. I appreciate that the interpretations are firmly grounded in the data and avoid speculation. I am confident that this manuscript will make a valuable contribution to knowledge in this field.

Response:

We thank the reviewer for their encouraging feedback and for recognizing our efforts in adhering to the preregistered plan and presenting data-driven interpretations.

Comment 1:

1) Could you please explain what "Did not participate" means in the enrollment section of Figure 3? Does this refer to participants who were excluded during prescreening?

Response:

We appreciate the reviewer's question. "Did not participate" refers to individuals who met the inclusion criteria and were invited to the study but chose not to participate.

Comment 2:

2) While I'm not entirely certain about this point and offer it merely for consideration, I would suggest removing the phrase "As anticipated" from the exploratory section (p. 21), as it might create a misleading impression that this analysis was preregistered.

Response:

We agree with the reviewer and have removed the phrase "As anticipated" to avoid any potential misunderstanding.

Comment 3:

3) Regarding the exploratory analyses, I would recommend adding effect size measures (presumably Cramer's V) when presenting the chi-square test results. Similarly, an effect size measure is missing for the following analysis: "a significant main effect was found for the second contrast (Twitter/X vs. other platforms)" - I assume this would be Cohen's d or Hedges' g. Additionally, descriptive statistics should be included here, as well as for the subsequent analysis regarding gender differences: "gender showed a significant negative effect ($F(1, 521) = 14.949$, $p = .001$, $\eta2 = .026$), with women reporting lower valence of the image than men".

Response:

We thank the reviewer for this helpful suggestion. Effect size measures have been added throughout the manuscript where applicable. We also corrected an earlier mistake, as women showed higher valence rather than lower. Here is an example of the revisions: "We also did not find a significant effect of the first contrast (Instagram vs. other platforms), $t(563) = 0.494$, $p = .622$, $d = .494$. However, a significant main effect was found for the second contrast (Twitter/X vs. other platforms), $t(563) = -2.97$, $p = .003$, $d = -2.97$, suggesting that Twitter/X's valence is lower than the average valence on Instagram and Facebook. […] Additionally, age was a significant positive predictor ($F(1, 563) = 5.99$, $p = .015$, $\eta^2 = .010$), and gender showed a significant negative effect ($F(1, 563) = 9.86$, $p = .002$, $\eta^2 = .016$), with women reporting higher valence ($M = .99$, $SD = 1.16$) than men ($M = .73$, $SD = 1.21$)." (pp. 20-21).

Comment 4:

4) I realized only now, and regret not noticing during Stage 1, that the Methods section lacks a Statistical analysis subsection. Unfortunately, as it cannot be modified after Stage 1, I would suggest (depending on the Recommender's advice) including information about the software used at least in the Results section. Following this, I would recommend adding the analysis script to enhance reproducibility. Additionally, sharing raw data and a codebook would help provide better insight into what was analyzed and how.

Response:

We have added the software and its version used for the pilot study (p. 11) and the main study (p. 19). A "Data Availability Statement" section has been added, stating that all data, codebooks, and analysis scripts are available on OSF (p. 8).

Comment 5:

5) For analyses that yielded non-significant results, I would recommend conducting equivalence testing to confirm the absence of effects larger than SESOI. This could be included in the exploratory analyses section.

Response:

We appreciate the reviewer's suggestion regarding equivalence testing. We attempted to perform equivalence testing for the repeated measures ANCOVA model but encountered methodological challenges due to the complexity of the design, including repeated measures structure, covariates, and a lack of established methods for this analysis in R.

Instead, we computed effect sizes and confidence intervals for the main effects and interactions. Additionally, we conducted equivalence testing for the main effect of time (H1) and the two contrasts using a simplified model (excluding covariates). These analyses are included in the OSF repository (section "4.6 Equivalence testing"; https://osf.io/akgdj/).

That said, we must admit that we are not entirely at ease with these analyses, as the results seem counterintuitive (e.g., non-significant TOST results but significant NHST results for the main effect and contrasts). If the reviewer would like to discuss these results further, we would welcome the opportunity, as we are still becoming familiar with equivalence testing methods. Alternatively, we would be open to withdrawing these analyses in order to avoid sharing false analyses.

Comment 6:

6) For descriptive purposes, when presenting results for H1 (the non-significant interaction), I suggest including results for main effects (not just contrasts). I appreciate how the non-significant result for H1 is discussed, avoiding speculation beyond existing data. However, precisely because the interpretation is based on the absence of an effect, I would strengthen this claim with equivalence testing.

Response:

We included results for the main effects throughout the manuscript as follows: "As we can see in Figure 4, the results did not reveal a significant main effect of time, $F(1, 563) = 0.001$, $p = .982$, $\eta^2 = .01$, nor of the interaction between time and social media, $F(2, 563) = 0.436$, $p = .647$, $\eta^2 = .001$. However, we found a significant main effect of the type of social media ($F(2, 563) = 3.939$, $p = .020$, $\eta^2 < .000$). Regarding the contrasts, we did not find a significant effect of the first contrast (Instagram vs. other platforms), $t(563) = 0.494$, $p = .622$, $d = .494$; but a significant effect was found for the second contrast (Twitter/X vs. other platforms), $t(563) = - 2.97$, $p = .003$, $d = - 2.97$." (p. 20).

Comment 7:

7) The effect size for the number of followers on Instagram and X is smaller than r = 0.1. For completeness, it would be appropriate to include the results for Facebook follower numbers in the Results section (despite being non-significant) to allow comparison of effect sizes. Is this effect size substantial enough to warrant the stated interpretation that it "demonstrates that Instagram and Twitter/X's unidirectional mode of connection fosters less intimacy and reciprocity"?

Response:

We added these results as follows: "To compare with the significant associations of the number of followers on Instagram and Twitter/X, the results for the number of followers on Facebook were $F(1, 563) = 0.538$, $p = .463$, $\eta^2 < .001$." (p. 20). We believe the difference in effect sizes is worth mentioning, as it contributes to understanding platform-specific patterns.

Comment 8:

8) Regarding the interpretation on p. 25, you state that a "possible explanation is that emoji usage may depend on the context: emoji might be more commonly used in private, interpersonal exchanges rather than in public social media posts." This interpretation assumes that participants interpreted the instruction "imagine sharing this event on [Facebook][Instagram][Twitter/X]. Write a post below as you would in real life. Please note that this must be a post, and not a story" as referring to a public post rather than a post for friends. However, this distinction isn't clearly implied by the instruction.

Response:

In the discussion, we refer to messaging apps such as Facebook Messenger, while the instruction clearly specifies a Facebook wall post, which is public, even if it is directed at friends. We clarified this distinction in the discussion: "One possible explanation is that emoji usage may depend on the context: emoji might be more commonly used in private, interpersonal exchanges (e.g. Facebook Messenger) rather than in social media posts (e.g., Facebook) (Cherbonnier et al., 2024)." (p. 25).

Comment 9:

9) The interpretation that follows in the discussion of the same RQ relies on platform norms, specifically emotional display norms. Could you please elaborate on this? Which norms are you referring to? What do existing empirical findings tell us about these norms?

Response:

We expanded this discussion as follows: "For example, Huang et al. (2022) have shown that the use of emoji is used as a sign of belonging to specific social circles on social media, helping to create boundaries between these circles. In this sense, emoji usage on social media may reflect broader emotional display norms, similar to those observed in face-to-face interactions." (p. 25).

Reviewer 3 - Julius Klingelhoefer

<u>General comment:</u>

Thank you for the opportunity to review this paper in its final form. I have read the preprint with great interest and I believe it is well written and well conducted. I only have some very minor comments, otherwise, I recommend the article for publication.

Overall, I believe this is really well done and I especially enjoyed the discussion of self-presentation, as it was clear and theoretically interesting.

<u>Response:</u>

We sincerely thank the reviewer for their positive feedback and thoughtful comments. We appreciate your recognition of the clarity and theoretical relevance of our discussion.

<u>Comment 1:</u>

p.12: η2 = 0.0001 has four digits, usually I would expect only three digits.

<u>Response:</u>

We reviewed and updated this value to three decimal places for consistency, as suggested.

<u>Comment 2:</u>

Some latin letters are not italicized even though they should be and some greek letters are italicized even though they should not be, e.g. N = on p. 13

<u>Response:</u>

We have carefully checked and corrected all instances of Latin and Greek letters to ensure proper formatting throughout the manuscript.

<u>Comment 3:</u>

I don't think the description of SESOI is clear. For the .21 effect size, the authors write "smaller effects could be practically meaningful", so .21 does not appear to be the smalles practically relevant effect size. Please also refer to my previous comments on SESOI, as I believe the power analysis is not based on a SESOI but on an expected effecte size (which is also good but it's not a SESOI). I would suggest clarifying this in 1-2 sentences. This seemst to persist both in the Method and Discussion sections.

<u>Response:</u>

We thank the reviewer for his comment. We agree that the effect size of .21 reflects an expected effect size derived from prior meta-analyses, rather than a strictly defined Smallest Effect Size of Interest (SESOI). We have clarified this distinction in the Discussion section: "Second, we selected an expected effect size of .21 based on prior meta-analyses (Ruppel et al., 2017), as a benchmark for powering our analyses. This value was chosen because it represents a theoretically plausible effect size in this domain, given the limited availability of direct comparisons across platforms. However, we acknowledge that this does not constitute a strictly defined Smallest Effect Size of Interest (SESOI). Instead, it reflects an expected effect size that was used for practical purposes in determining our sample size. The main positive effect on Twitter/X valence aligns closely with this SESOI, suggesting that our observed effects are within the range we deemed theoretically interesting at the outset. However, the meaningfulness of effect sizes, particularly in the context of social media research, depends on several factors beyond the magnitude of the observed effect." (p. 26).

<u>Comment 4:</u>

I believe there should be a space before and after minus signs (e.g, t = - 2)

<u>Response:</u>

We have updated all instances of minus signs to include spaces before and after them.