Dear Dr. Karhulahti,

We would like to thank you for the opportunity to submit a second revision of our article, titled “Unveiling the Positivity Bias on Social Media: A Registered Experimental Study On Facebook, Instagram, And X.” to *PCI Registered Reports*.

We are grateful to the editor and reviewers for their thoughtful scrutiny and valuable feedback. We revised the manuscript, including the theoretical introduction and the methodology of the main study. We also uploaded the data analysis script for the main study to OSF.

Please find below a point-by-point response to all the comments raised. In addition, we highlighted the changes to our manuscript within the document by using yellow colored text.

Sincerely,

The authors

Table of Contents

Editor - Veli-Matti Karhulahti

General comment:

Dear Alexandra Masciantonio and co-authors,

Thank you for all careful revisions. Three reviewers returned to assess the work and they all agree that the plan is significantly improved and closer to ready. They have some further notes, however. I summarize the main points and add some of my own.

Response:

Once again, we appreciate the time the editor devoted to our manuscript, as well as the suggestions we received.

Comment 1:

Theory

The additions to the introduction are a major step forward, yet all three reviewers still believe the theoretical part could be improved. Below, I try to follow the study's rationale as clearly as I can. Perhaps this also helps with clarifications in the final edits.

Your construct is "positivity bias in social media self-presentation", which has the first appearance in Reinecke & Trepte (2014): "We thus suggest that positive forms of authenticity are shown more frequently on SNS and are more likely to receive reinforcement in the SNS context than negative forms of authenticity. SNS users are thus more likely to engage in positive authenticity than in negative forms of authentic self-presentation. We propose the term 'positivity bias in SNS communication' to refer to this phenomenon." As mentioned earlier, this seems to potentially happen in two ways:

PB1: selectively posting more positive daily experiences (and/or not posting negative daily experiences).

PB2: exaggerating the positivity of any posted experience (both positive and negative).

- You want to test PB2 (2nd construct pathway), and this your H1.
- You justify H1 with Goffman's theory: people present themselves differently (tactically) in social contents.
- The auxiliary hypothesis is that all social media are contexts where presenting oneself positively brings more rewards (on average), hence tactical positivity.

I think all this makes sense. Perhaps more explicitly organizing the text to follow a chain of reasoning in subsections (as suggested by one reviewer) could be useful. What could be expanded is the last point, i.e. the assumption that positive self-presentation brings additional rewards in social media (vs F2F). This could be done easily by referring more to the vast literature on likes, retweets, etc. which gamify and quantify social media interaction (not present in F2F).

Response:

We have revised the theoretical introduction.

Firstly, we have added titles as suggested by reviewer 1 (comment 1).

Secondly, we have added references regarding the control on self-presentation being easier online, for example:

> ***"A long line of research has shown that individuals can adapt their communication more easily in online environments, particularly as a result of asynchrony (Walther, 1996)." (p. 4).***

Finally, we edited the section on Face Theory (see comment 4 of reviewer 1).

Comment 2:
In turn, H2 is based on the idea that social media are different.

- Social media differ by architecture, affordances and social-cultural context.
- Such differences can further reward and punish positive self-presentation.
- Instagram excessively supports positive self-presentation (vs X/FB), hence tactical positivity.

This also makes sense. As the reviewers imply, what could be further explained is again the last section: what are those specific features in Instagram that cause its users self-present more positively? Is it the very fact that all posts are images? As someone who checks Instagram once or twice per month, the vast amount of posts I see go "here's my cat", "here's my coffee", etc. I find it difficult to imagine how one would even be able to post a "negative" image (e.g., "I had a terrible argument today, here's a picture of it" or "my family member is ill, here's a picture of it"). In other words, because the UX is designed for visual content, this seems to make it extremely clumsy for negative sharing. This likely pushes the social-cultural context to be even more positive. This makes your H2 logical for me, but it could be spelled out for other readers too, theoretically, in more detail.

Btw, you might want to check a paper from a few weeks ago by Avalle et al. that quite comprehensively finds no differences in negative behaviors between social media platforms. But they only studied text so it's still consistent with H2 (if the auxiliary hypothesis is based on the visually driven design of Instagram): Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., ... & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. Nature, 628(8008), 582-589.

Response:

We thank the editor for his feedback. Firstly, we have changed the selection study on Prolific, so that participants regularly use social media to interact with other users (see comment 4 of reviewer 2). We have also made it possible to describe a video and not just an image.

For the article by Avalle et al (2024), we thank the editor for sharing. We think it would be interesting to discuss this paper in the discussion, not least because they do not have Instagram, but also because it highlights the need to go further and look at the characteristics of the platforms rather than just the platforms themselves. To this end, we offered more information on the characteristics of platforms in terms of architecture, affordances and socio-cultural context (see comment 1 of reviewer 1).

Comment 3:

Methods 1

Two reviewers are concerned about your selected effect r=.21. I see their worries. The term "SESOI" often causes confusion, so I try to open it (please allow this explanation, hopefully it doesn't sound too instructive -- it helps myself better see the context). SESOI tends to have two different meanings/implications:

a) effect for planning statistical power, and
b) smallest pragmatically or theoretically Meaningful Effect (ME).

Optimally, (a) and (b) are the same, but that's rare because ME is often difficult to calculate and justify. Most studies don't have ME but use a heuristic like meta-analysis to plan for statistical power. That can be ok, but doesn't allow making general inference about whether the effects are meaningful or not.

In the present case: it's ok to power this study for .21 if you believe the effects in the field are usually this size. But as reviewers note, this doesn't mean that .15 or .2 would necessarily be meaningless effects to find (or >.21 meaningful to find!). Here .21 is just a number to help estimate needed statistical power. One can choose .21 for power and say: "*In this study, we have power for .21. If we find >.21, this is informative but not necessarily a meaningfully large effect. If we find <.21, we are powered to rule out larger effects—but also smaller effects might still be meaningful for theory and practice.*"

That said, as noted earlier (and by reviewer), in this study you could also easily calculate ME based on what's the smallest agreeable difference in valence. Assuming you have high agreement, a step from "positive" to "very positive" is practically meaningful because your raters were able to observe a consistent difference (if the scale would be wider, e.g. -10 to +10, consensus would become unlikely). Although we don't know beforehand what the standardized effect of a raw step is, it's likely going to be more than .21 (which would be logical, considering we're talking about sensing differences in degrees of positivity).

In sum: you can use .21 as ad hoc SESOI, as long as you're clear about the above limitations. Or you can further improve the design and define ME, e.g., based on the agreement of raters. The latter would be similar to what Anvari and Lakens call the smallest subjectively experienced difference: Anvari, F., & Lakens, D. (2021). Using anchor-based methods to

determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, 96, 104159.

Response:

We would like to thank the editor for his suggestions. We have decided to provide further justification for our approach in response to reviewer 1's comment 6 and reviewer 2's comment 2:

> ***"Based on these parameters, our power analysis for a repeated measures ANOVA Indicated that we require at least 219 participants for Hypothesis 1 (within-subject effects), and 270 participants for Hypothesis 2 (interaction effects). In this study, we are statistically powered to detect an effect size of .21. If we observe an effect greater than .21, it will be informative, though not necessarily indicative of a large effect. Conversely, if the effect size is less than .21, while we can rule out larger effects, smaller yet significant effects could still be theoretically and practically meaningful." (p. 17).***

Comment 4:

Methods 2

One reviewer makes an important note about your coding plan. Indeed, you need to predefine a threshold for agreement. We have recently developed guidelines for transparently coding open-ended data, which seem very relevant here: https://osf.io/preprints/psyarxiv/86y9f (to be clear, I'm not asking to cite the paper but review for more detailed answers). To avoid kappa-hacking and other such common concerns, it's important to spell out things like:

- What's a sufficient threshold of agreement to proceed with the test at Stage 2?
- What will be done if agreement isn't sufficient after first coding round? As re-coding same data with same raters will cause huge bias, will guidelines be updated and new coders trained?
- Will the confirmatory design be changed to an exploratory one if it turns out that high agreement is impossible to obtain?
- Considering that ratings are categorical ("positive", "very positive" etc.), would Fleiss' kappa be more informative than ICC?

Response:

We would like to thank the editor for his help. We have responded to reviewer 2 by better specifying the coding plan:

> ***"As with the pilot study, the texts' number of words and the number of emojis will be counted. Three researchers will qualitatively analyze all the texts to estimate their valence on a 7-point scale (-3 = Very negative; 3 = Very positive). We will use***

Intra Class Correlation Coefficient to verify inter-rater reliability. 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." (p. 19)

Coding will be performed using a 7-point scale (-3 = Very negative; 3 = Very positive), making Intra Class Correlation Coefficient preferable to Fleiss' kappa.

Comment 5:

Brief notes

- Please report the effect sizes of your pilot (especially that related to H1, seems to be around d ≈ .2 ?)
- In the design table "Interpretation" H1/H2 cannot be corroborated solely based on significance but also needs a sufficiently large effect. E.g., if .21 is decided as interesting here, one could reject H0 when >.21. Note that null tests haven't been planned (e.g., equivalence test) so the design doesn't allow obtaining evidence for no effect (H0), just evidence for effect (H1) or no evidence for effect (no H1).
- I'm not aware of data on this but for the record, I believe that the chronology for most people's Instagram posts is reverse to the present study design: not selecting a memory but an available image that was created as an affect response ("oh, my food/pet looks nice"). Perhaps worth discussing at Stage 2.
- Related to the above: it feels that most Instagram posts have recently become short videos (ala Tiktok). You might consider allowing participants to describe either a picture or video. I agree with reviewers that this option should be for all participants, all platforms.
- I agree with the reviewers about oversampling because open-ended data often include many low-quality responses. Please ensure that all discarded responses will also be shared as part of the available data.
- Please include the R code for all confirmatory Stage 2 analyses for final check to obtain Stage 1 IPA.
- As reviewers note, the screening question "Which of the following social media sites do you use on a regular basis (at least once a month)?" allows participants who only read and don't post. It could be useful to ask their posting habits at some point to have an idea how many lack posting experience. I agree that if a user lacks posting experience on a platform, they seem unable to exress platform-specific bias.
- Like one reviewer, I don't follow this: "the questionnaire will only be able to be answered on a smartphone to get as close as possible to a real-life situation." Why is only smartphone real-life? I personally use social media only via laptop.
- For transparency, please add a footnote on page 7 to clarify that some of the RQs were revised during Stage 1 review. If you prefer, you don't need to refer to the main tests as preregistered (all RRs are).

- Regarding H2 (“The posts’ valence is dependent on the social media”), do you really mean absolute valence or increase in valence from T1 to T2?

It’s many things, but this is an experiment where small details matter a lot. I hope you find the reviewers' feedback useful and my additional notes of some value. If you disagree with something or believe that some suggestions are unclear or wrong, you may rebut them as usual or contact me directly for additional clarifications.

Response:

To respond to all the comments:

- **We have reported the effect sizes of the pilot study.**
- **We have specified that we will not rely solely on significance in Table 1 (see comment 9 of reviewer 2).**
- **We will discuss this point in Stage 2.**
- **Participants will be able to describe a video or photo.**
- **We plan to oversample (see reviewer 2's comment 2).**
- **The R script for the analyses has been uploaded to the OSF project.**
- **We have modified the selection study as previously mentioned, and participants will have to use the platforms regularly to interact with other users.**
- **Participants will be able to respond on the device of their choice.**
- **We have added a footnote to clarify that the research questions have changed.**
- **We meant absolute valence.**

Thanks again to the editor for his time.

Reviewer 1 - Julius Klingelhoefer

General comment:

I would like to thank the authors for the opportunity to review the resubmitted manuscript. I share the editors' positive outlook on the manuscript and believe that the revised manuscript and the proposed study have been improved in central ways. I have a couple of smaller suggestions that the authors may want to consider, specifically regarding the theoretical outlook and how the control group is implemented, but overall, I believe that the proposed study looks very promising and would recommend that after consideration of the feedback from this round of reviews, the study should go to data collection.

Please find my more detailed comments below.

Response:

We appreciate again the reviewer's time and constructive feedback. We have taken into account his suggestions, the responses of which are detailed below.

Comment 1:

Theory

I think the additions and changes to the theory section improve the manuscript, yet, I believe the theory section still could be a bit clearer and slightly more well-structured in some areas:

- As it is structured and discussed now, the different theoretical approaches seem somewhat disconnected from each other. I think relating the approaches to each other and organizing/systematizing them would make it more clear what the theoretical contribution of the paper is. E.g., I would suggest explicating more how architecture, affordances, and social-cultural context relate to self-presentation. In some respects this link becomes clear, e.g., through the discussion of norms, but with shareability, for example, it is not really clear in which way the authors would expect higher shareability within a platform to affect post valence.
- Further, I believe that the revised paper would benefit from using advance organizer paragraphs and/or including one additional heading level.
- Particularly, "afford" is used before being properly introduced as an approach. I think explicating the theoretical approaches more clearly would also address this issue with clarity.

Response:

We have rewritten the theoretical framework in line with the reviewer's feedback. We have detailed the parts related to architecture, affordances and socio-cultural context. In particular, we have made the architectural implications more explicit:

"Social media architecture refers to the underlying design and structural elements that govern functionalities, user interactions, and data flow within social media platforms. It includes various features that are crucial when examining the positivity bias (Bossetta, 2018). The most important is the connection mode, which pertains to the type of relationships that can be formed between users. Facebook operates on a bidirectional connection mode, where interactions occur mostly among users who are mutually recognized as "friends". This model promotes an intimate and reciprocal interaction environment, allowing users to control who sees their content (Vitak & Kim, 2014). In contrast, Instagram and Twitter/X employ a unidirectional connection mode, where one can follow another without requiring reciprocation. This type of connection fosters a dynamic where users often engage with a broader, less personal audience (Marwick & boyd, 2011). As a result, interactions on these platforms can be less about mutual exchange and more about broadcasting to followers, which may encourage users to present more curated, idealized versions of their lives to attract likes, shares, and new followers." (pp. 5-6)

We have done the same for the affordances:

"Affordances address not the objective features of platforms, but how users perceive them (boyd, 2010). They can be defined as "the perceived actual or imagined properties of social media, emerging through the relation of technological, social, and contextual, that enable and constrain specific uses of the platforms" (Ronzhyn et al., 2022, p. 3178). We suggest that two affordances are particularly relevant to positivity bias.

Visibility concerns the perceived degree of visibility of the published content (Treem & Leonardi, 2013). On Facebook, the visibility of content is generally lower compared to Instagram and Twitter/X due to its more enclosed, bidirectional nature. This might reduce the pressure to maintain a universally appealing image. In contrast, the higher visibility on Instagram and Twitter/X, driven by their unidirectional following system, amplifies the reach of posts and potentially the need to maintain a positive, attractive persona (Boczkowski et al., 2018).

Shareability pertains to the ease with which content can be shared across the platform and how suitable the content is perceived to be for different formats (Masciantonio et al., 2024). Facebook users often share a mix of text and images, which allows for more nuanced self-expression and personal storytelling (van Dijck, 2013). Instagram, being predominantly image-focused, encourages users to post visually appealing content, which often involves high levels of stylization and impression management (Boczkowski et al., 2018). Twitter/X, known for its textual content, promotes brevity and wit, often leading to oversimplified or emphatic statements." (p. 6).

We also added headings to add clarity to the introduction.

Comment 2:

The authors propose in the revision that the baseline (control) condition should be relating an event to a group of friends. This seems reasonable (see also my later comment). However, the theoretical section mostly focuses on social media in general and later on the platforms, but the comparison to a friend group or other types of points of comparison are not discussed or only briefly touched upon. I think introducing these comparisons earlier and discussing existing empirical evidence on such comparisons and theoretical mechanisms that explain the differences would make it more obvious and provide better arguments as to why this was chosen as the point of comparison for the social media platforms.

Response:

Regarding positivity bias, there is surprisingly little literature that compares it to positive self-presentation in "real life" (as distinct from qualitative literature). This is a point we shall return to in the discussion. We have, however, added a number of more general references in the theoretical introduction, for example:

> ***"A long line of research has shown that individuals can adapt their communication more easily in online environments, particularly as a result of asynchrony (Walther, 1996)." (p. 4).***

Comment 3:

More arguments within the theoretical background would be beneficial, for example, why are shareability and visibility specifically more relevant than other affordances?

Response:

As explained in the response to comment 1, we have discussed affordances in greater detail, including shareability and visibility.

Comment 4:

The statement "Indeed, this bias is rooted in the face theory […]" makes it seem like face theory is the only (or main) explanation for positivity bias. If this is what the authors want to say, this should be supported with arguments. If other mechanisms are presumed, this statement should be qualified.

Response:

We have revised the paragraph on Face Theory to make it more nuanced and to add more references:

> ***"This tendency towards positivity on social media can be rooted in a general desire for positive image and social approval (Pounders et al., 2016; Spottswood & Hancock, 2016). Indeed, the face theory postulates that individuals strategically manage their self-presentation to maintain their social identity and uphold their reputation in the***

eyes of others (Goffman, 1959). However, while positive self-presentation is not exclusive to social media, these platforms tend to enhance and amplify it." (p. 4)

Comment 5:

I think the authors should include a short definition of what they consider to be emojis, as the discussion and the footnote in the previous manuscript version make it clear that the definition is not necessarily straightforward.

Response:

We have defined emoji in the introduction:

"Emoji are small digital images used to express an idea, emotion, or concept in electronic communications. They can be used both as complementary cues in texts and as surrogates (Tandyonomanu & Tsuroyya, 2018)." (p. 4)

Comment 6:

Method

I think it is a good choice to identify a smallest effect size of interest. However, I do not believe that the choice of using r = .21 as the SESOI is appropriate here, or at least needs more justification. In my view this would mean that if you find an effect that is smaller than average media effects on self-disclosure, it would not be viewed as practically or theoretically relevant. This would mean that around 50% of effects would be considered not practically or theoretically relevant. An effect of r = .20 would mean that the hypothesis is not accepted (vs. r > =.21). Maybe I am interpreting this incorrectly and you can point out the implications that you were considering but based on Lakens' Paper, I would suggest that a power analysis based on an expected effect size of .21 may be more appropriate based on the response.

Response:

We thank the reviewer for his comment. We agree that our use of SESOI requires better explanation and clarification in the manuscript. We have discussed the implications in the method:

"Based on these parameters, our power analysis for a repeated measures ANOVA Indicated that we require at least 219 participants for Hypothesis 1 (within-subject effects), and 270 participants for Hypothesis 2 (interaction effects). In this study, we are statistically powered to detect an effect size of .21. If we observe an effect greater than .21, it will be informative, though not necessarily indicative of a large effect. Conversely, if the effect size is less than .21, while we can rule out larger effects, smaller yet significant effects could still be theoretically and practically meaningful." (p. 17).

Comment 7:

The use of a control group is a good change to the method. However, I am not quite sure whether this new operationalization may not be confounded with theoretically relevant characteristics. As you plausibly argue in the literature review, existing connections between the social media may influence how participants post on social media. However, when the control group is instructed to think about a group of friends, posting on a "follower"-style social medium (i.e., twitter, Instagram) may be different in at least two aspects, that is being a social media platform and type of shared pre-existing connection. This could reduce internal validity. However, it might be a sacrifice that the authors may deem appropriate for theoretical reasons or ecological validity.

Response:

We acknowledge the complexities that the reviewer has outlined concerning the control group and the influence of social media connections. We will address these points in detail in the discussion section of the paper, considering the trade-offs between internal validity and ecological validity.

Comment 8:

It needs to be explained what "understand the instructions" means.

Response:

By this we meant that participants who did not follow the instructions (i.e. did not write the texts) would be withdrawn from the study. We have changed the sentence in the manuscript:

> ***"As with the pilot study, participants who will not give consent to take part in the study, who will not respond to the entire study, or who will not follow the instructions, will be removed from the study." (p. 17).***

Comment 9:

"To reflect the fact that Instagram is an image-oriented social media, they will also be asked an optional question: 'If you plan to use an image or photo to accompany this post, please describe it briefly here'." I assume this will be there for all platforms? Additionally, will participants only be asked about a post on the timeline ("traditional" posts) vs. a post to the story? To me this was not clear from these instructions and stories may differ substantially from "traditional" posts in the use of visual elements, e.g., GIFs, Stickers, etc.

Response:

This question will be asked for all platforms. We have also added a specification in the instructions to make it clear to participants that this is a post and not a story:

"Write a post below as you would in real life. Please note that this must be a post, and not a story."." (p. 18).

Comment 10:

Would it make sense to randomly vary whether the control vs. experimental condition will be introduced first to account for potential order effects in this within-participant factor?

Response:

It does indeed make sense and we will do it as specified now in the manuscript.

Comment 11:

Style

Latin letters should be italicized in results (e.g. r = x)

Response:

We made the changes.

Comment 11:

The authors say that sample size would be rounded to 300. I think this is somewhat misleading and I would suggest talking about oversampling, e.g., to account for participants who do not meet attention checks or other inclusion criteria.

Response:

We have modified the sentence in the manuscript (and increased it by 50 participants in response to reviewer 2' comment):

"To account for potential non-adherence to instructions by some participants, we also plan to oversample, aiming for a total of 350 participants." (p.17).

Reviewer 2 - Marcel Martončik

General comment:

Dear Authors,

I would like to greatly appreciate and thank for the thorough and thoughtful responses to the comments from myself and the other reviewers and the related edits of the manuscript. Your effort in addressing each point is commendable and has significantly improved the manuscript. I appreciate the time and effort you've put into making these revisions.

Having read the revised version of the manuscript, I'd like to highlight the following points:

Response:

We thank the reviewer for his time and comments. We have addressed all of his comments below.

Comment 1:

1) How will the dependent variable be calculated? In the measurement section, I found only this: 'Three researchers will qualitatively analyze all the texts to estimate their valence on a 7-point scale (-3 = Very negative; 3 = Very positive).' What happens if their assessments do not match? Will an aggregate score be calculated (what kind of?), or will they need to reach a unified conclusion through a reconciliation process, or is there another method?

Response:

We thank the reviewer for his comment; we have added more details on the procedure:

> ***"As with the pilot study, the texts' number of words and the number of emojis will be counted. Three researchers will qualitatively analyze all the texts to estimate their valence on a 7-point scale (-3 = Very negative; 3 = Very positive). We will use Intra Class Correlation Coefficient to verify inter-rater reliability. 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." (p. 19)***

Comment 2:

2) I am considering the appropriateness of applying the effect from a meta-analysis by Ruppel et al. (2017) to the context of this study. Hypotheses 1 and 2 compare the positivity of posts between different forms of media. However, the justification of SESOI is based on differences in self-disclosure (Ruppel et al., 2017). How closely related are these constructs? Is it possible to expect a similar effect in positivity as in self-disclosure? In addition, I agree with the

suggestion from the Recommender, "because you're using human raters to observe differences in posts: we thus already know that one step up in the scale is noticeable and meaningful," that having a 1-point difference (as an unstandardized unit) would be much more meaningful.

My subsequent question pertains to the difference between the hypotheses. Should the SESOI be the same for H1 and H2, even though they relate to different phenomena?

Minor comment/suggestion: During the pilot, a large number of participants were excluded for various reasons (e.g., misunderstanding the instructions). Perhaps it would have been advisable to plan for more oversampling…?

Response:

We agree that the constructs of positivity and self-disclosure, while related, necessitate a clear distinction and justification in their application to SESOI. We will elaborate on this in the discussion and we have already provided some details on SESOI in the manuscript. In addition, Hypothesis 1 and Hypothesis 2 address different statistical effects (main and interaction effects, respectively), but they are fundamentally related to the same underlying phenomenon in our study:

> ***"Based on these parameters, our power analysis for a repeated measures ANOVA Indicated that we require at least 219 participants for Hypothesis 1 (within-subject effects), and 270 participants for Hypothesis 2 (interaction effects). In this study, we are statistically powered to detect an effect size of .21. If we observe an effect greater than .21, it will be informative, though not necessarily indicative of a large effect. Conversely, if the effect size is less than .21, while we can rule out larger effects, smaller yet significant effects could still be theoretically and practically meaningful." (p. 17).***

Concerning the comment about oversampling, we have taken the reviewer's suggestion into account:

> ***"To account for potential non-adherence to instructions by some participants, we also plan to oversample, aiming for a total of 350 participants." (p.17).***

Comment 3:

3) Thank you for sharing the power analysis calculation script. I think it is even more important to share in advance an analysis script for future analyses. There are many different ways of how ANCOVA can be calculated even in the same software and it would be useful to know the author's method of analysis.

Response:

We have now added the analysis script to the same OSF project.

Comment 4:

4) This is rather my reflection:

I wonder why users are not allowed to complete the survey on the device they prefer for using social media, but instead are forced to use a smartphone. Additionally, they are instructed to answer the question: 'We will ask participants on which devices they most often use social media (computer, tablet, or smartphone).'

In terms of theory, how relevant is it for a person to be familiar with a given social media platform? Should the positivity bias effect manifest itself on any social media platform regardless of whether the person is familiar with and uses it, or does it primarily manifest itself on the platform that the person prefers?

I wonder, what is the source of the differences in positivity bias between platforms? The theoretical background suggests (if I understand it correctly) that the different effects should be due to differences in features between platforms. Thus, in order for different positivity biases among platforms to manifest, the user must be aware of the specific features of a given platform (I will demonstrate a strong positivity bias on Facebook only when I often use Facebook and therefore know its features and similarly weaker positivity bias on Instagram because I am familiar with its features that are distinct from Facebook and therefore cause different effect). From this perspective, then, wouldn't it be more appropriate for each participant to choose their preferred platform (as opposed to random assignment) and have to write a post for that platform? Perhaps even more appropriate would be to control this variable completely and incorporate it into the research design (some participants would write a post for their preferred platform, some for the one they use minimally or not at all) - but I understand that this would increase the sampling requirements.

Response:

We thank the reviewer for his reflective insight. We agree with the importance of allowing participants to use the device of their preference to enhance the comfort and naturalness of their interactions. Thus, we will permit participants to complete the survey on their preferred devices to maintain an ecological validity.

To preserve experimental rigor, we have decided to continue assigning the social media platform rather than allowing participants to choose. However, we have changed the criteria for the selection study: participants will have to use all three platforms regularly to interact with other users. They should therefore have a better knowledge of the platforms.

Comment 5:

5) I am a bit confused about the hypotheses and research questions (RQ). The first RQ is not explicitly formulated as an RQ; it is listed at the top of page 14 as 'The main research will therefore aim to address the following fundamental question:'. The explicitly formulated RQ1 at the beginning of page 15 then has no associated hypothesis. If it is not even supposed to have

one, and is only part of the exploratory analyses, then I would suggest to explicitly label it as such.

Response:

We have added this question based on previous comments from Round 1. This question was problematic, so we have revised its presentation to make it clearer in the manuscript:

> ***"The main research will therefore aim to address the following problematic: how does the positivity bias manifest on social media, and does it vary depending on the type of social media platform?" (p. 15)***

Comment 6:

6) I apologize but I do not understand why participants are advised not to report very positive event if the goal is to prevent participants from reporting traumatic experiences.

Response:

We have indeed changed the instruction so that it only concerns painful events:

> ***"Please choose an event that is not too much painful". (p. 18).***

Comment 7:

7) I also wonder that each social media platform has different word limitations (e.g., there is a significant difference between Instagram and Facebook). I presume that frequent users of these platforms are aware of these limitations and may naturally adjust the messages they write for a particular platform – writing condensed messages (on a platform with a low word limit) where it is necessary to highlight the most salient elements of the event, etc. Should this be incorporated into the study? For instance, should the same word limit be used for all social media platforms, or should participants be asked after writing a post if they have taken into account the word limit of a given social media platform when writing a post, and consider that as a covariate?

Response:

We thank the reviewer for raising the issue of word limitations on different social media platforms, which indeed has a significant impact on how users craft their messages. We recognize that platforms like Twitter, despite having increased their character limit from 140 to 280, still impose constraints that are not present on platforms like Facebook or Instagram. In response to the reviewer's suggestion, our study design ensures that all participants are active users of the platforms involved, thereby familiar with the nuances of each platform's word limits. In addition, we have observed that the average length of posts typically does not reach the maximum limits set by these platforms, which suggests that while word limits are a factor, they may not severely restrict most users' expressions. Therefore, we have decided to not standardize word limits across platforms in our study.

This approach allows us to capture more naturalistic data on how users typically interact with each platform.

Comment 8:

8) Do the authors plan to use any threshold for inter-rater reliability? What is the planned procedure if the reliability is very low?

Response:

As explained in response to comment 1, we have added a threshold for inter-rater reliability.

Comment 9:

9) In Table 1, Interpretation given different outcomes should be based not only on the p-value but also on the size of the effects.

Response:

We have made the modifications in Table 1:

> ***"If there is a significant effect of the variable "Time" on valence ($p < 0.5$), H1 will be accepted. We will also take effect sizes into account, in line with the chosen SESOI, but it should be noted that equivalence tests have not been planned."***

Reviewer 3 - Anonymous

General comment:

Thank you for providing the opportunity to review the revised version of the manuscript "Unveiling the Positivity Bias on Social Media: A Registered Experimental Study On Facebook, Instagram, And X". I appreciate the authors' efforts to address my previous comments and make significant improvements to the content of the manuscript. I have a few minor suggestions to further improve the proposed research.

Response:

We thank the reviewer again for the time his/her devoted for our study and for his/her comments. We have provided answers to them below.

Comment 1:

General comments:

While the authors have added more information on emoji usage, I believe the argument and theoretical background on why it is important to consider the frequency of emoji usage could be strengthened. It would be beneficial to clarify the implications and insights gained from understanding the differential usage of emojis across different social media platforms.

Response:

In the theory section, we have added some additional information on the use of emoji:

> ***"Emoji are small digital images used to express an idea, emotion, or concept in electronic communications. They can be used both as complementary cues in texts and as surrogates (Tandyonomanu & Tsuroyya, 2018)." (p. 4).***

In the research questions, we also highlighted the fact that there is little research on the differences in emoji usage between platforms:

> ***"Finally, little is known regarding the relationship between emoji use and the positivity bias. The pilot study showed that the more positive the event, the more emoji participants used. However, when adding covariates, the association did not persist. Similarly, the pilot study did not reveal any platform-specific differences in emoji use; nevertheless, the ability to accompany a post with a photo or video might influence emoji usage. We therefore restate our research question:***
>
> ***RQ1: Does positivity bias have an influence on emoji use?"***

Comment 2:

As previously noted by the authors, asking participants to write about the same event twice (in this case once for a friend and once for social media) may introduce bias. To mitigate this, I recommend randomly assigning participants to write first about either sharing with a friend or posting on social media and then about the other event.

Response:

We agree with the reviewer's comment. We have changed the procedure to randomize sharing the event with a group of friends and on social media.

Comment 3:

Minor comments:

There is some inconsistency in language usage: The authors alternate between using "X" and "Twitter." It would be preferable to maintain consistency throughout the manuscript.

Response:

For consistency, we have made the changes in the manuscript by choosing Twitter/X everywhere.

Comment 4:

In section 2 “the present research” this sentence reads repetitive: “In support of open science, the research will be pre-registered on OSF.”

I wish the authors all the best with their planned research!

Response:

For greater clarity, we have made the paragraph lighter. We thank the reviewer again for his/her time.