Below are brief point-by-point responses from the Transparent Replications team to the points made regarding our report draft, “Report #12: Evaluation of a study from “Sharing of misinformation is habitual, not just lazy or biased” (PNAS | Ceylan, Anderson, & Wood 2023.” The points initial points raised and responses by the authorship team are in black text, with the responses from the Transparent Replications team in blue text.

1. The study used a measure of news sharing habit that had an error in the question wording

    As you pointed out, there is a typo in one of the measures. However, all 4 items in the scale are highly correlated. If you drop this item, you will see that the results still hold. We also used other habits measures (reading habits and frequency of sharing). While reading habits are a weaker predictor compared to sharing habits but the results hold using any of these scales.

TR Team Response: It’s great to know that dropping this item doesn’t substantively impact the results. Our critique here is that a person reading the paper would not be aware of this issue in the analysis. In our view, the paper should have disclosed this issue to readers. Supplemental materials could have presented supporting analyses showing that this does not present a substantive problem for the results.

2. The study data were collected on Amazon Mechanical Turk with no quality checks

    We have replicated these results many, many times, and it is implausible that the result is due to noise instead of habit strength. We even built habits in Study 4 to demonstrate causality of the effect. We have included quality checks such as attention checks and elimination of duplicate ip addresses in subsequent research, and we have obtained comparable results to those in the set of studies published in PNAS.

TR Team Response: It’s great to know that data quality checks are used in subsequent research. Our critique here is that the results of this specific study should be interpreted with caution due to the lack of data quality checks.

3. The model did not converge

    We computed many models including participants and headline fixed effects and decided to report the most comprehensive and conservative model. We also computed a model without random effects, and with an optimizer (control = glmerControl(optimizer = "bobyqa"). In all these cases, models converged and results remained virtually identical. The consistent results despite different models attest to the robustness of the effect. We did not include

these in the web appendix because our focus was on reporting the other models including the various covariates requested by reviewers.

TR Team Response: It's great to know that other model versions did converge. Our critique here is that a reader looking at the paper would not be aware that the model results presented did not converge, and would not be aware of the other models that had been run. In our view, this information would be important for readers to have in order to evaluate the presented results.

4. The study was underpowered

> The power analysis we reported is for the focal effect, which is the interaction between habits scale and headline veracity. We are able to detect this effect even with 200 participants. Since we added a between-subjects variable (question order), we increased the sample size 4 times, which is in line with standard practices in the field. New approaches to power analysis with mixed effects offer various recommendations on how to calculate power. Even a recent paper suggests that power analysis does not lead to reliable results especially for mixed effect models (Pek, Pitt, and Wegener 2024).
>
> Pek, J., Pitt, M. A., & Wegener, D. T. (2024). Uncertainty limits the use of power analysis. Journal of Experimental Psychology: General, 153(4), 1139.

TR Team Response: Our report argues that the primary analysis is likely underpowered to detect a three-way interaction between the news sharing habit measure, the experimental condition, and the veracity of the news headline. We focused on the three-way interaction, rather than the two-way interaction your comment references, because the primary claim asserted in the Study 2 title concerns the three-way interaction: "Considering Accuracy Does Not Deter Habitual Sharing: Study 2."

As we discussed in the report, improving the statistical power of this study would likely require including more stimuli, not more participants. As shown in Figure 2 of Westfall et al. (2014), if a study has 16 stimuli, increasing the number of participants beyond 200 does very little to increase the statistical power.

We agree that Pek and colleagues (2024) make a good argument that calculating power for mixed effects models is complex and relies on assumptions about parameters that may be unrealistic or over-simplified. But, if one takes their view that power analyses are unreliable for mixed effects models, this is further reason to be very cautious about interpreting a null result as evidence of no effect, as Study 2 does in a few places (including its title). If one doesn't know how much statistical power they have, then they don't know what the chance of a false negative is.

The primary reason we were concerned about the statistical power for this model was that the null result on the three-way interaction was interpreted as direct evidence of no three-way interaction.

5. Our central analysis was not included in the preregistration (but was claimed to be)

    This is an interesting claim. Our central prediction was for a two-way interaction. We did not expect that this effect would be modified by question order, and thus we did not specify the three-way interaction in the preregistration. Instead, we outlined the core, central effect we expected to be significant. We are unaware of any guidelines specifying that nonsignificant effects need to be preregistered.

TR Team Response: Our main concern was that the primary statistical model reported in Study 2—the mixed effects model predicting sharing behaviors by the news sharing habit measure, headline veracity, experimental condition, and all two-way and three-way interactions between those variables—was not preregistered. Yet, the paper says, “We preregistered all hypotheses, primary analyses, and sample sizes (except Study 1).”

Additionally, the main substantive claims in the paper about the results of study 2 focused on the results of the three-way interaction. From reading the paper, it appears that the point of study 2 was to determine whether or not accuracy primes worked for people who are high habitual sharers. This framing starts with the subheading that introduces study 2: “Considering Accuracy Does Not Deter Habitual Sharing: Study 2,” and is mentioned several more times in the discussion of study 2. Since that claim is the main focus of study 2, and the paper says “We preregistered all hypotheses, primary analyses, and sample sizes (except Study 1)”, it seems likely to us that readers would infer that there was a preregistered hypothesis about whether or not accuracy primes worked for high habit participants.

Because the paper says, “We preregistered all hypotheses, primary analyses, and sample sizes (except Study 1),” we think readers are likely to assume any given hypothesis or analysis was preregistered, unless it’s specifically labeled as not preregistered (e.g., “exploratory”; “post-hoc”; “non-registered”).

6. Predicted probabilities 1SD below the mean

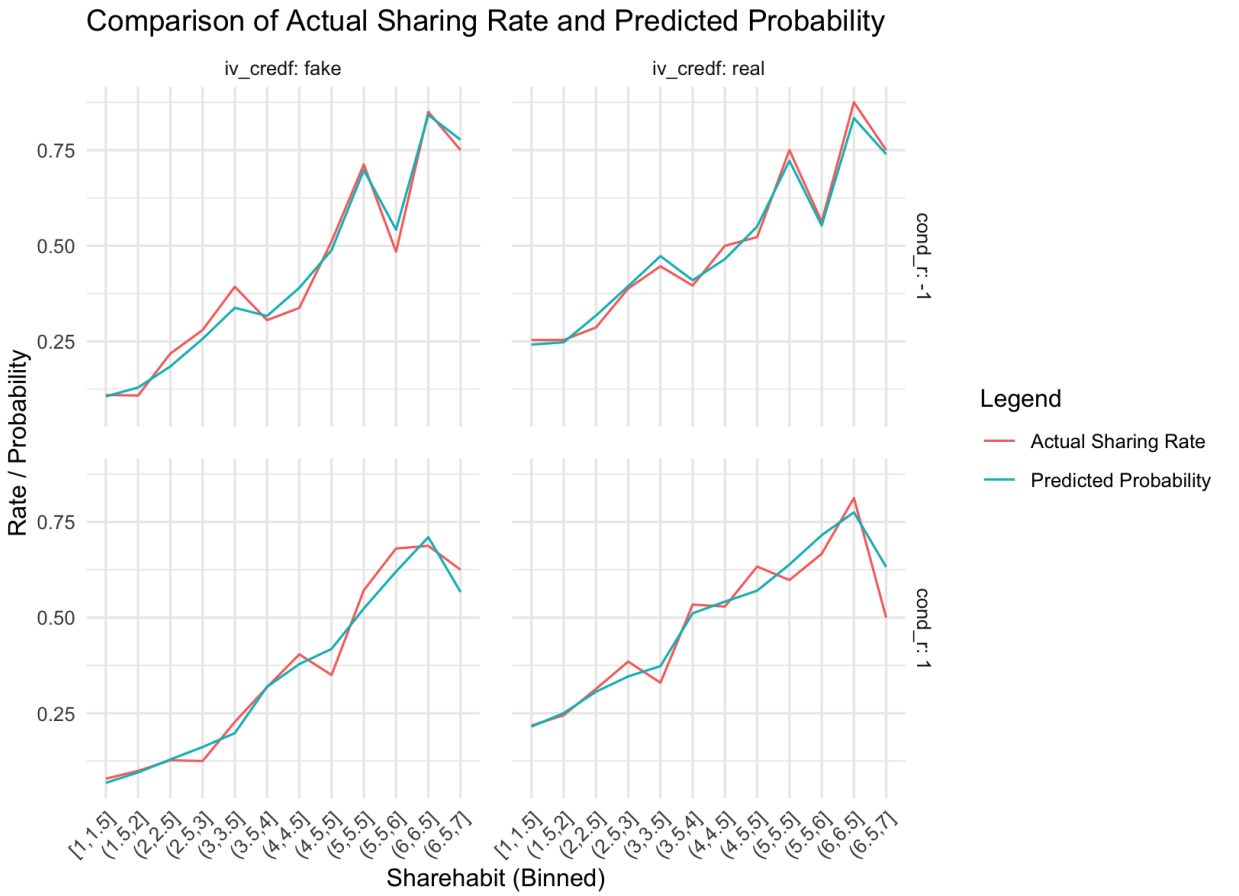
    Since we used the prediction model, we can still report 1 SD deviation below the mean. Technically, this is an appropriate way to analyze the data. More importantly, the results are nearly identical if we compare the predicted probabilities at sharing habit 1 and sharing habit 0.87 (-1SD below the mean). An alternative way to analyze the data would be determining Johnson Neyman points. But this approach would not have changed any of our conclusions, as shown by the results below

Sharing habit	Predicted probabilities of sharing – Fake	Predicted probabilities of sharing – Real
Control condition (share-first)		
0.87 (reported)	0.05415136	0.15707102
1	0.07847644	0.19628033
Treatment condition (accuracy-first)		
0.87 (reported)	0.04381390	0.16599442
1	0.04824283	0.17620217

TR Team Response: Our critique here is about whether readers will be able to accurately interpret the presented results. In our view, readers are likely to believe that the references to "low habit sharers" refer to a group of actual participants because of how the information is presented in the paper. Similarly, readers of the paper are likely to interpret the graphs as being representations of the data itself, rather than predictive model results because of how they are labeled and described in the paper. In our view the paper would be improved by more clearly indicating to readers that predictive model results are what is being presented.

7. Predicted probabilities vs. actual sharing

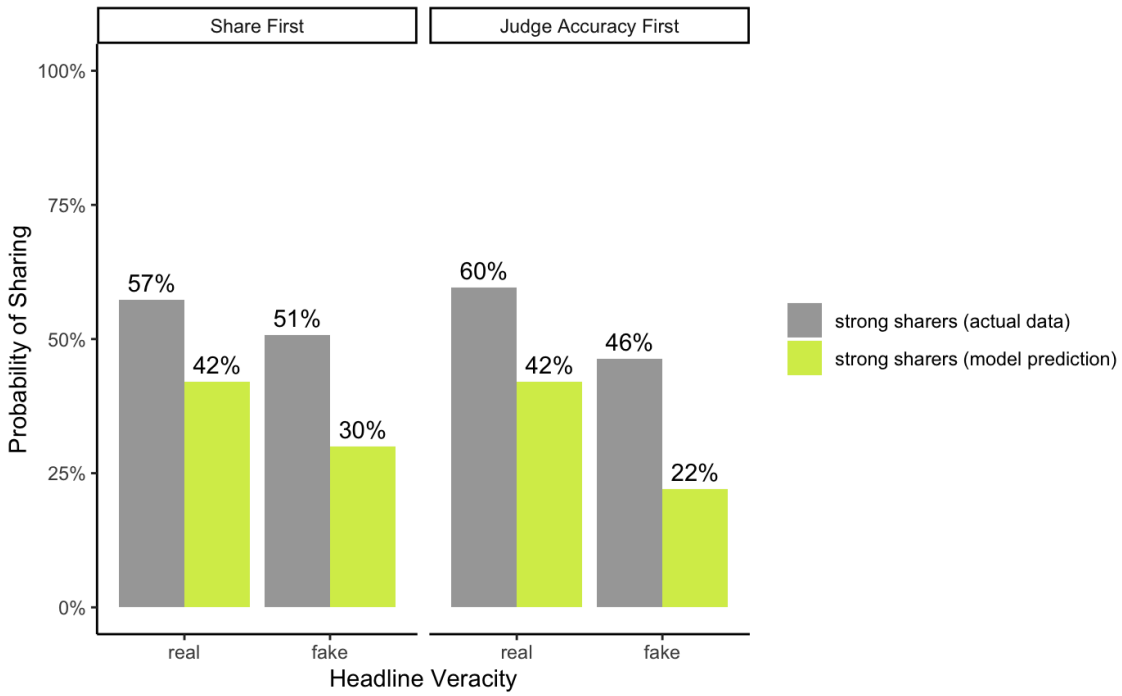
   We reported predicted probabilities, and these are clearly marked in our graphs. However, I plotted predicted and actual sharing at every habit bin. As you can see, on average, they are aligned, which simply means that our model successfully recovered the data. They are aligned across the different question order conditions and for real and fake headlines. If anything, in the accuracy first condition (cond_r = 1), actual sharing seems slightly ahead of predicted sharing especially for fake headlines at high levels of habits. In general, the area under the curve is pretty similar for predicted and actual values.



(Ceylan, email published with permission, 1/26/2025)

TR Team Response: In our evaluation of this paper we believed that the results presented in the paper were actual data, until we specifically examined the R code and saw that “emmeans” was being used to generate the data for the graphs. The captions and discussion in the paper itself did not clearly indicate to us what was being done, and we believe it is likely that other readers would have the same experience we had. In our view, more clearly labeling the graphs as predicted probabilities would have improved the clarity of the paper.

The plot of predicted sharing vs. actual sharing that you provide above is very helpful, and we think it is valuable for readers of our report to see that the area under the curve is pretty similar for predicted and actual values. However, as we show in the report, the predicted and actual values for the specific levels of habitual sharing included in the primary plot for Study 2 (Fig. 4) differ quite substantially. This is one of the reasons why we think it’s important to make it clear that the plotted data are *predicted* probabilities.



In addition to this response, the authors also provided a follow up email response to our section in the report titled "the primary claims don't match the provided evidence." That response is below:

> We interpreted the lack of three-way interaction based on the data pattern I shared with you in the reactions document.
>
> You are making a thought experiment but frankly, you can just examine the pattern of the data.
>
> The data is showing us that everybody (both high and low habitual users) their sharing slightly, supporting the lack of three-way interaction.
>
> (Ceylan, email published with permission, 1/30/2025)

TR Team Response: Our critique here is that while you point out above that both high and low habitual users responded to the accuracy prime the same way, thus resulting in a lack of three-way interaction, the paper itself states that the lack of three-way interaction means that accuracy manipulation does not work on people with high sharing habits. This is stated in the subheading that introduces the study 2 results, and several additional places in the discussion of those results. Our report says the following:

> The main claim put forward in Study 2 is stated in the study's title: "Considering Accuracy Does Not Deter Habitual Sharing: Study 2." The

paper makes several similar claims when discussing the results from Study 2:

- “Thus, highlighting accuracy proved useful in reducing the spread of misinformation but not among the most habitual users.”
- “Priming accuracy concerns prior to sharing had only a modest impact on the discernment of everyone and did not ameliorate high habitual sharing of misinformation (Study 2).”
- “Once sharing habits have formed, they are relatively insensitive to changing goals through accuracy primes”

These are four examples of the claim being made in the paper that accuracy primes do not work for high habit participants. This is not in line with the evidence presented in the paper, which shows that accuracy primes have an impact on participants, and there was no statistically significant evidence that this differed by sharing habit level, as you explain above.