Supplementary Materials for:

Social truth queries: Development of a new user-driven intervention for countering online misinformation

Madeline Jalbert*, Morgan Wack, Pragya Arya, & Luke Williams

*Corresponding author. Email: mjalbert@uw.edu

This file includes:

Supplementary Materials Section 1: Fixed Effects (t-test and ANOVA) Analysis

Analytic Approach

In each experiment, we investigated how the presence of social truth queries compared to no replies (Experiments 1-3) and neutral, non-truth-related replies (Experiment 2-3) on Twitter posts containing misinformation impacted judgments of truth and sharing.

In Experiment 1, we ran a paired-samples *t*-test comparing the mean truth and sharing ratings of posts containing misinformation when they appeared with a truth-related reply compared to when they appeared without a reply. In Experiments 2 and 3, we ran a repeated measures ANOVA comparing ratings across reply type (truth-related reply, neutral reply, or no reply) for both truth and sharing judgments. We followed up significant main effects of reply type with a simple effects analysis using a Bonferroni adjustment for multiple comparisons to investigate which reply conditions were significantly different from each other. For these analyses, we report a *p*-value adjusted for these multiple comparisons that can be compared to the standard alpha level of .05.

Data for all studies was analyzed using SPSS (v 28.0.1.1). Repeated measures *d* effect sizes were calculated using Comprehensive Meta-Analysis Software (Version 4) from mean differences and standard deviations of the differences, taking into account the correlation between repeated and new claims and corrected for small sample size.

Experiment 1

Truth Ratings

There was a significant effect of the presence of a reply on truth ratings, with Tweets with the replies judged to be less true, $M = 2.91$, $SD = 0.75$, than Tweets without replies, $M = 3.11$, SD

= 0.74, mean difference = -0.20, 95% CI [-0.36, -0.27], t (99) = -2.30, p = .024, d = -0.26, 95% CI [-0.49, -0.04].

Sharing Ratings

There was a significant effect of the presence of a reply on sharing likelihood ratings, Tweets with the replies had a lower reported sharing likelihood (M = 2.49, SD = 1.15) than Tweets without replies, M = 2.70, SD = 1.15, mean difference = -0.21, 95% CI [-0.41, 0.00], t (99) = -2.02, p = .046, d = -0.18, 95% CI [-0.35, 0.00].

Experiment 2

Truth Ratings

There was a significant overall effect of reply type on truth rating, F (2, 596) = 7.23, p < .001, partial eta^2 = .024. In this analysis Tweets that appeared with a truth-related reply were not judged to be significantly less true, M = 3.12, SD = 1.06, than Tweets that appeared with no reply, M = 3.27, SD = 1.04, mean difference = -0.15, 95% CI [-0.35, 0.05], p = .194, d = -0.14, 95% CI [-0.29, 0.01], although the effect was directionally the same as in the mixed effects analysis. However, Tweets that appeared with a truth-related reply were judged to be significantly less true than Tweets that appeared with neutral, non-truth-related replies (M = 3.43, SD = 1.03), mean difference = -0.30, 95% CI [-0.10, -0.51], p < .001, d = -0.46 [-0.58, -0.33]. There was no significant difference between the no reply and neutral conditions, mean difference = 0.15, 95% CI [-0.03, 0.34], p = .127.

Sharing Ratings

Turning to sharing, we also found a significant overall effect of reply type on sharing ratings, F (2, 600) = 5.36, p = .005, partial eta^2 = .018. This time, there was not a significant

effect of there being a lower intent to share Tweets appearing with truth-related reply, $M = 2.46$, $SD = 1.31$, compared to Tweets that did not appear with a reply, $M = 2.63$, $SD = 1.32$, mean difference = -0.17, 95% CI [-0.36, 0.02], $p = .094$, $d = -0.19$, 95% CI [-0.30, -0.07], but Tweets that appeared with a truth-related received significantly reduced intent to share ratings compared to Tweets appearing with neutral, non-truth related replies, $M = 2.72$, $SD = 1.40$, mean difference = -0.26, 95% CI [-0.44, -0.07], $p = .004$, $d = -0.16$, 95% CI [-0.23, -0.08]. There was again no significant difference between Tweets that appeared with no replies and those that appeared with neutral non-truth-related replies, mean difference = 0.09, 95% CI [-0.11, 0.29], $p = .858$.

Experiment 3

Truth Ratings

Once again, there was a significant overall effect of reply type on truth ratings, F (2, 596) = 39.72, $p < .001$, partial $eta^2 = .117$. Simple effects analysis revealed that once again, Tweets that appeared with truth-related replies were rated as significantly less true, $M = 3.15$, $SD = 0.67$, than Tweets that appeared with no replies, $M = 3.45$, $SD = 0.64$, mean difference = -0.30, 95% CI [-0.39, -0.20], $p < 0.001$, $d = -0.29$, 95% CI [-0.45, -0.13], and Tweets that appeared with neutral, non-truth related replies ($M = 3.45$, $SD = 0.67$), mean difference = -0.30, 95% CI [-0.39, -0.20], $p < .001$, $d = -0.45$, 95% CI [-0.59, -0.32]. There was again no significant difference in truth ratings between Tweets that appeared with no replies compared to Tweets that appeared with neutral replies, mean difference = 0.00, 95% CI [-0.09, 0.09], $p > .999$.

Sharing Ratings

There was a significant overall effect of reply type on sharing ratings, F (2, 596) = 14.87, $p < .001$, partial $eta^2 = .047$. Mirroring the findings for truth ratings, Tweets that appeared with

truth-related replies were rated as significantly less likely to be shared, $M = 2.18$, $SD = 0.98$, than Tweets that appeared with no replies, $M = 2.35$, $SD = 1.04$, mean difference = 0.16, 95% CI [0.07, 0.25], $p < 0.001$, $d = -0.13$, 95% CI [-0.24, -0.01], and Tweets that appeared with neutral, non-truth related replies, $M = 2.33$, $SD = 1.04$, mean difference = -0.18, 95% CI [-0.26, -0.09], $p < .001$, $d = -0.17$, 95% CI [-0.24, -0.10]. There was again no significant difference in sharing ratings between Tweets that appeared with no replies compared to Tweets that appeared with neutral replies, mean difference = -0.02, 95% CI [-0.10, 0.07], $p > .999$.

Supplementary Materials Section 2: Cognitive Reflection Test Analysis

Results

In each of our experiments, we additionally assessed whether the impact of user truth queries on truth and sharing judgments depended on an individual's tendency to utilize intuitive (vs. analytical) processing, as assessed by performance on the Cognitive Reflective Test (CRT). After making either truth or sharing ratings, participants completed a seven-item Cognitive Reflection Test (CRT): The first three items were a reworded version of Frederick (2005) via Shenhav et al. (2012), followed by the four-item CRT by Thomson & Oppenheimer (2016).

For each experiment, we then conducted an additional analysis including CRT and its interaction in each of our main models. CRT scores were calculated by adding up the number of correct analytical responses out of the seven CRT questions and then centering the responses by subtracting the mean score.

Across our studies, we failed to find that individual tendencies in utilizing intuitive vs. analytical processing moderated the impact of truth queries (vs. no replies or neutral replies) on truth and sharing ratings. There was no significant interaction of CRT and reply condition for truth ratings in Experiment 1, b = -0.08 (95% CI [-0.17, 0.01]), t (691.43) = -1.66, p = .098, Experiment 2, b = 0.01 (95% CI [-0.06, 0.09]), t (1486.2) = 0.33, p = .744 for the interaction with truth queries vs. no replies, b = -0.03 (95% CI [-0.10, 0.04]), t (1486.1) = -0.81, p = .42, for the interaction truth queries vs. neutral replies, or Experiment 3, b = -0.01 (95% CI [-0.05, 0.03]), t (6873.5) = -0.26, p = 0.794 for the interaction with truth queries vs. no replies, b = -0.02 (95% CI [-0.06, 0.02]), t (6874.8) = -1.17, p = 0.242 for the interaction with truth queries vs. neutral replies.

We also failed to find a significant interaction of CRT and reply condition on sharing ratings in any experiment, Experiment 1: $b = 0.05$ (95% CI [-0.04, 0.15]), t (691.3) = 1.06, p = .292, Experiment 2: $b = 0.02$ (95% CI [-0.06, 0.09]), t (1497.07) = 0.46, p = .643 for truth queries vs. no replies; $b = -0.01$ (95% CI [-0.08, 0.06]), t (1496.32) = -0.22, p = .826 for truth queries vs. neutral replies, Experiment 3: $b = -0.01$ (95% CI [-0.05, 0.02]), t (6876.6) = -0.85, p = 0.398 for truth queries vs. no replies; $b = -0.01$ (95% CI [-0.04, 0.03]), t (6875.1) = -0.33, p = 0.740 for truth queries vs. neutral replies.

There was no significant main effect of CRT on truth judgments in Experiment 1, b = 0.00 (95% CI [-0.08, 0.08]), t (205.63) = -0.09, p = .932, Experiment 2, $b = -0.04$ (95% CI [-0.10, 0.01]), t (1172.2) = -1.47, p = .142, or Experiment 3, $b = -0.01$ (95% CI [-0.05, 0.03]), t (698.7) = -0.32, p = 0.747. However, there was a significant main effect of CRT on sharing judgments, with a higher CRT was associated more generally with lower ratings of sharing across posts, Experiment 1: $b = -0.12$ (95% CI [-0.24, 0.00]), t (140.2) = -2.01, p = .046, Experiment 2, $b = -0.14$ (95% CI [-0.21, -0.07]), t (672.2) = -3.99, $p < .001$, and Experiment 3 b = -0.09 (95% CI [-0.15, -0.04]), t (385.5) = -3.19, p = 0.002.

Discussion

Overall, the effects of social truth queries were not consistently moderated by an individual's CRT scores, indicating that truth queries were similarly effective at reducing truth and intent to share misinformation across people who have more analytical and more intuitive processing styles. This is consistent with prior findings that the impact of accuracy nudge do not appear to be consistently moderated by CRT scores (Pennycook et al., 2020). However, it is also worth considering that our participants (who were recruited from MTurk and Prolific) may have

been familiar with the CRT items used, and thus including alternative format measures of cognitive reflection in future studies could provide better insight (Woike, 2019).

Supplementary Materials Section 3: Political Orientation

In each of our experiments, we additionally assessed whether the impact of user truth queries on truth and sharing judgments depended on an individual's political orientation, as measured on a seven-point unnumbered scale from "extremely liberal" (coded as 1) to "extremely conservative" (coded as 7). For each experiment, we then conducted an additional analysis including political orientation and its interaction in each of our main models. Political orientation ratings were centered by subtracting the mean score from each value (M = 3.27, 3.33, and 2.83 for Exp. 1-3 respectively).

Results

There was no significant interaction of political orientation and reply condition (truth queries vs. no replies) for truth ratings in Experiment 1, b =-0.06 (95% CI [-0.15, 0.03]), t (691.3) = -1.36, p = 0.174, Experiment 2, b = 0.01 (95% CI [-0.08, 0.09]), t (1486.1) = 0.18, p = 0.860, for the interaction with truth queries vs. no replies, b = 0.08 (95% CI [0.00, 0.16]), t (1486.2) = 1.86, p = 0.063, for the interaction truth queries vs. neutral replies, or Experiment 3, b = -0.02 (95% CI [-0.06, 0.03]), t (6875.5) = -0.79, p = .430, for the interaction with truth queries vs. no replies, b = -0.04 (95% CI [-0.09, 0.00]), t (6875.9) = -1.89, p = .058, for the interaction with truth queries vs. neutral replies. Although p-values were close to significance in a couple of these interactions (ps = .063 and 0.058), the corresponding b were in opposite directions (0.08 and -0.04). Thus, political orientation did not appear to consistently moderate the impact of truth queries across our experiments.

There was a main effect of political orientation in Experiment 1, b = 0.08 (95% CI [0.01, 0.16]), t (206.8) = 2.10, p = 0.037 and Experiment 3, b = 0.06 (95% CI [0.02, 0.11]), t (708.8) =

2.83, p = .005, with more conservative participants giving higher overall ratings of truth. This effect did not reach significance in Experiment 2, b = -0.01 (95% CI [-0.08, 0.05]), t (1156.1) = -0.35, p = 0.727

Turning to sharing judgments, there was a significant interaction of political orientation and reply condition in Experiment 1, b = -0.12 (95% CI [-0.23, -0.02]), t (696.3) = -2.31, p = 0.021. As participants became more conservative, they were less influenced by the presence of social truth queries on their reported likelihood of sharing. However, there was no significant interaction of sharing by reply condition in Experiment 2, b = -0.04 (95% CI [-0.12, 0.04]), t (1496.2) = -0.97, p = 0.332, for the interaction with truth queries vs. no replies, b = -0.02 (95% CI [-0.11, 0.06]), t (1496.3) = -0.55, p = .586, for the interaction truth queries vs. neutral replies, or Experiment 3, b = 0.01 (95% CI [-0.04, 0.05]), t (6873.7) = 0.32, p = .075, for the interaction with truth queries vs. no replies, b = 0.00 (95% CI [-0.04, 0.04]), t (6873.1) = 0.04, p = .969, for the interaction with truth queries vs. neutral replies.

There was a main effect of political orientation on sharing in Experiment 2, b = 0.12 (95% CI [0.04, 0.20]), t (651.1) = 2.85, p = .005, and this effect was close to significance in Experiment 1, b = 0.13 (95% CI [0.00, 0.26]), t (141.0) = 1.93, p = 0.056, with more conservative participants saying that they would be more likely to share the posts overall. This effect was not significant in Experiment 3, b = 0.04 (95% CI [-0.03, 0.12]), t (382.2) = 1.23, p = .220.

Discussion

By and large, the impact of truth queries was similarly effective in reducing belief in and intent to share false information across individual differences in political orientation. Political

orientation did not moderate the impact of truth queries on truth judgments in any experiment. The one instance in which political orientation moderated the impact of truth queries was for sharing ratings in Experiment 1, with truth query replies reducing the reported likelihood of sharing less compared to no replies as participants became more conservative. However, this effect did not reach significance in Experiments 2 and 3, and there was additionally no significant interaction of sharing and reply conditions when comparing truth query replies to neutral replies.

The potential for the effectiveness of truth queries to be robust across political orientations is promising, especially given recent concerns that accuracy nudges may not be effective for conservatives (Rathje et al., 2022). However, given that our sample was primarily liberal ($M < 4$ on our 1-7 scale for each experiment) and our materials were non-political, future investigations with a more representative sample of political ideologies utilizing polarized information would provide further insight regarding the effectiveness of truth queries for different individuals.

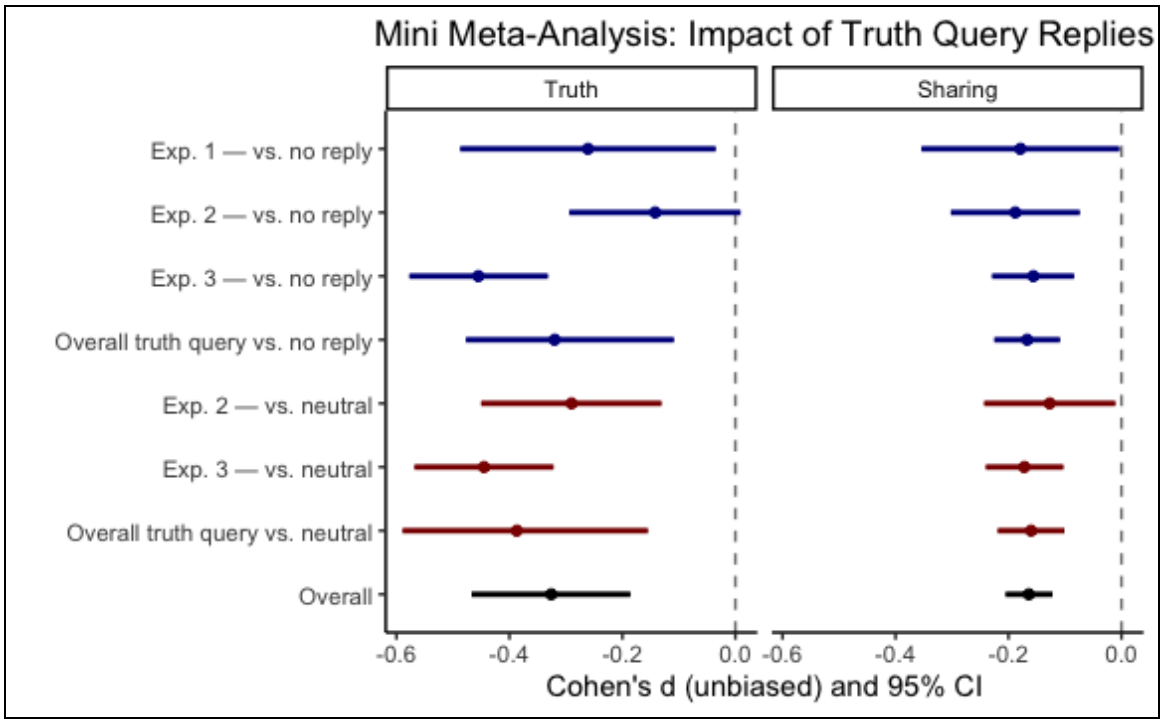

Figure S1. Forest plot of effect sizes (Cohen's *d* corrected for small sample bias) of truth query replies vs. no replies and neutral replies across experiments. Error bars represent 95% confidence intervals.

Table S1. Truth Ratings by Items that Appeared with Each Truth Query in Experiment 3.

Truth query	Truth query reply	No reply	Neutral reply	Comparing truth query reply and no reply conditions				Comparing truth query reply and neutral reply conditions			
	Estimated marginal means and 95% CIs			*df*	*t*	*b statistic and 95% CIs*	*p*	*df*	*t*	*b statistic and 95% CIs*	*p*
1. Does that make sense given everything else you know?	3.11 [2.71, 3.51]	3.37 [2.98, 3.76]	3.43 [3.04, 3.82]	1193.7	2.59	0.26 [0.06, 0.46]	.010	1193.8	3.19	0.32 [0.12, 0.52]	.001
2. Where did you learn this?	3.40 [2.97, 3.83]	3.72 [3.29, 4.14]	3.69 [3.26, 4.11]	1193.3	3.44	0.32 [0.14, 0.50]	<.001	1193.4	3.13	0.29 [0.11, 0.47]	.002
3. How do you know that?	3.18 [2.49, 3.88]	3.54 [2.85, 4.24]	3.52 [2.83, 4.22]	1193.0	3.72	0.36 [0.17, 0.55]	<.001	1193.1	3.53	0.34, [0.15, 0.53]	<.001
4. Do other people believe that?	3.24 [2.62, 3.86]	3.61 [2.99, 4.22]	3.66 [3.04, 4.28]	1193.5	3.82	0.36 [0.18, 0.55]	<.001	1193.6	4.37	0.42 [0.23, 0.60]	<.001
5. What evidence is there for that?	2.87 [2.39, 3.36]	3.36 [2.87, 3.84]	3.30 [2.81, 3.78]	1193.1	5.43	0.48 [0.31, 0.66]	<.001	1193.2	4.73	0.42 [0.25, 0.60]	<.001
6. Is there proof of that?	3.04 [2.50, 3.58]	3.35 [2.81, 3.89]	3.29 [2.76, 3.83]	1192.9	3.22	0.31 [0.12, 0.50]	.001	1192.9	2.62	0.25 [0.06, 0.45]	.009
7. Why would that be the case?	3.34 [2.99, 3.70]	3.55 [3.20, 3.90]	3.52 [3.17, 3.88]	1193.7	2.30	0.21 [0.03, 0.38]	.021	1193.8	1.99	0.18 [0.00, 0.35]	.047
8. How do you know this is true?	3.01 [2.66, 3.36]	3.09 [2.75, 3.43]	3.17 [2.83, 3.51]	1192.9	0.80	0.08 [-0.12, 0.29]	.424	1193.0	1.55	0.16 [-0.04, 0.36]	.121

Table S2. Sharing Ratings by Items that Appeared with Each Truth Query in Experiment 3.

Truth query	Truth query reply	No reply	Neutral reply	Comparing truth query reply and no reply conditions				Comparing truth query reply and neutral reply conditions			
	Estimated marginal means and 95% CIs			*df*	*t*	*b statistic and 95% CIs*	*p*	*df*	*t*	*b statistic and 95% CIs*	*p*
1. Does that make sense given everything else you know?	2.13 [1.91, 2.36]	2.30 [2.09, 2.51]	2.35 [2.14, 2.55]	1193.0	1.97	0.17 [0.00, 0.33]	.049	1193.0	2.48	0.21 [0.04, 0.38]	.013
2. Where did you learn this?	2.41 [2.09, 2.72]	2.46 [2.15, 2.76]	2.40 [2.10, 2.70]	1192.8	0.55	0.05 [-0.13, 0.23]	.584	1192.9	-0.09	-0.01 [-0.19, 0.17]	.926
3. How do you know that?	2.18 [1.84, 2.52]	2.30 [1.97, 2.64]	2.38 [2.04, 2.71]	1193.3	1.5	0.13 [-0.04, 0.29]	.135	1193.3	2.37	0.20 [0.03, 0.37]	.018
4. Do other people believe that?	2.10 [1.75, 2.45]	2.21 [1.87, 2.55]	2.27 [1.92, 2.61]	1193.1	1.42	0.11 [-0.04, 0.27]	.156	1193.1	2.10	0.17 [0.01, 0.32]	.036
5. What evidence is there for that?	2.15 [1.90, 2.40]	2.45 [2.22, 2.69]	2.41 [2.17, 2.65]	1193.0	3.56	0.30 [0.14, 0.47]	<.001	1193.1	3.02	0.26 [0.09, 0.43]	.003
6. Is there proof of that?	2.01 [1.82, 2.19]	2.17 [1.95, 2.28]	2.17 [2.00, 2.33]	1193.3	1.37	0.11 [-0.05, 0.27]	.170	1193.4	2.01	0.16 [0.00, 0.32]	.045
7. Why would that be the case?	2.37 [2.06, 2.67]	2.53 [2.24, 2.83]	2.40 [2.11, 2.70]	1193.2	1.85	0.17 [-0.01, 0.35]	.065	1193.2	0.39	0.04 [-0.14, 0.21]	.699
8. How do you know this is true?	2.06 [1.70, 2.42]	2.30 [1.94, 2.65]	2.44 [2.09, 2.79]	1193.2	2.8	0.24 [0.07, 0.40]	.005	1193.2	4.48	0.38 [0.21, 0.54]	<.001

References

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732https://doi.org/10.3389/fpsyg.2019.02646

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Rathje, S., Roozenbeek, J., Traberg, C. S., Bavel, J. J. V., & Linden, D. S. van der. (2022). Letter to the editors of Psychological Science: Meta-Analysis reveals that accuracy nudges have little to no effect for U.S. Conservatives: Regarding Pennycook et al. (2020). PsyArXiv. https://doi.org/10.31234/osf.io/945na

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*, 423–428. https://doi.org/10.1037/a0025391

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Woike, J. K. (2019). Upon repeated reflection: Consequences of frequent exposure to the Cognitive Reflection Test for Mechanical Turk participants. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02646