Appendix B: Supplementary analyses

Design balance and descriptive statistics

Participants were evenly randomized across posts (Chi-squared test, $\chi^2(9) = .347$, $p \approx 1$) and across conditions (Chi-squared test, $\chi^2(3) = .089$, p = .993). The median time to evaluate the Facebook post was 32 seconds in the control condition (*inform others* condition: 35 seconds; *peer performance* condition: 35 seconds; *monetary bonus* condition: 42 seconds; minimum overall time: 3 seconds, maximum overall time: 25 minutes). On a scale from 1 to 6 (3.5 response at chance level), the average accuracy score in the control condition was 3.87 (*SD* = 1.37; *inform others* condition 3.95, *SD* = 1.38; *peer performance* condition 4.01, *SD* = 1.45; *monetary bonus* condition 4.22, *SD* = 1.46). In the control condition, 61.2% of participants correctly guessed the scientific validity of the post (inform others: 63.2%, +2% compared to control; peer performance: 65.1%, +3.9%; monetary bonus: 70.5%, +9.3%).

Pre-registered analyses

As our accuracy scale was on a scale from 1 to 6, we also tested a simpler binary measure that tested whether participants were correct (e.g., giving a "valid" rating when the post contained valid claims) or incorrect (e.g., giving a "valid" rating when the post contained invalid claims). Results were consistent with the results in the main text, with the only difference that the difference between the peer performance and control condition was not significant at the 5% level ($M_{performance} - M_{control}$ = +4%, β = 0.17 (-0.07,0.41), z = 1.829, p = .101, $p_{\text{uncorrected}}$ = .067). This result may be attributable to both a lower statistical power of the test and a lower effectiveness of this incentive compared to monetary bonuses.

Robustness analyses

We replicate the main hypotheses, excluding participants who failed the attention check in the study ("If you are reading carefully, select Completely agree."). Thirty-three participants failed the attention check. Analyses excluding this subgroup provide almost identical results for the effect of the experimental conditions (peer performance: β = .20 [.02, .39], z = 2.566, p = .016; inform others: β = .11 [-.07, .30], z = 1.413, p = .158; monetary bonus: β = .50 [.31, .69], z = 6.278, p < .001). All other tests pertaining to search style frequency and their predictiveness of accuracy scores remain equally significant.

We also repeat the analyses using mixed-effects models, including post id as a random intercept. Again, results are almost identical for the effect of the experimental conditions and the frequency and effectiveness of search strategies like lateral reading and click restraint.

Psychometric correlates of non-monetary incentives

As an exploratory analysis, we tested whether general altruism as measured in the adapted version of the self-report altruism scale (Rushton et al., 1981) predicted the effectiveness of the message in the *inform others* condition. Similarly, we tested the predictive power of a reduced version of the social comparison scale (Gibbons & Buunk, 1999) for the effectiveness of the message in the *peer performance* condition. The short form of these two questionnaires, however, presented poor internal consistency for both scales

(self-report altruism scale, 6 items: α = .683; social comparison scale, 4 items: α = .697), therefore, the interpretation of results is problematic.

Correlational analyses suggest that neither scale predicts performance in the corresponding condition (altruism scale for *inform others* condition: β = .011 [−.020, .041], z = 0.680, p = .499; social comparison scale for *peer performance* condition: β = −.011 [−.050, .027], z = −0.570, p = .573). As an additional measure, we also test whether these scales moderate the effect of condition. We run two ordinal logistic regressions predicting accuracy scores, one comparing control with *peer performance* condition and the other comparing control with the *inform others* condition. For the *peer performance* comparison, we include the social comparison scale as main effect and in interaction with condition. For the *inform others* comparison, we include the altruism scale as main effect and in interaction with condition. The prediction in both cases is that if the underlying psychometric traits of social comparison and altruism influence the effectiveness of the incentive in the respective conditions they are associated with. In other words, the interaction term between scale and condition should be significant. What we find instead is that the interaction is non-significant in both tests (social comparison score × peer performance concern condition: β = −.02 [−.08, .03], z = −0.870, p = .381; altruism score × inform others: β = −.02 [−.06, .03], z = −0.820, p = .412).

The lack of significance of these exploratory analyses may be due to a number of factors, including the non-significant effect of the inform others condition combined with scales having questionable internal consistency (both having Cronbach's α < .70). The reduction of items for each scale might have contributed to reducing the scales' consistency and predictive power of the questionnaires, which in turn might have led to failures in observing a connection between social comparison concerns and accuracy in the peer performance condition. It is also possible that the amount of concern for social comparison required for the incentive to work is relatively low and, therefore, does not necessarily require high scores on the relevant psychometric scale.