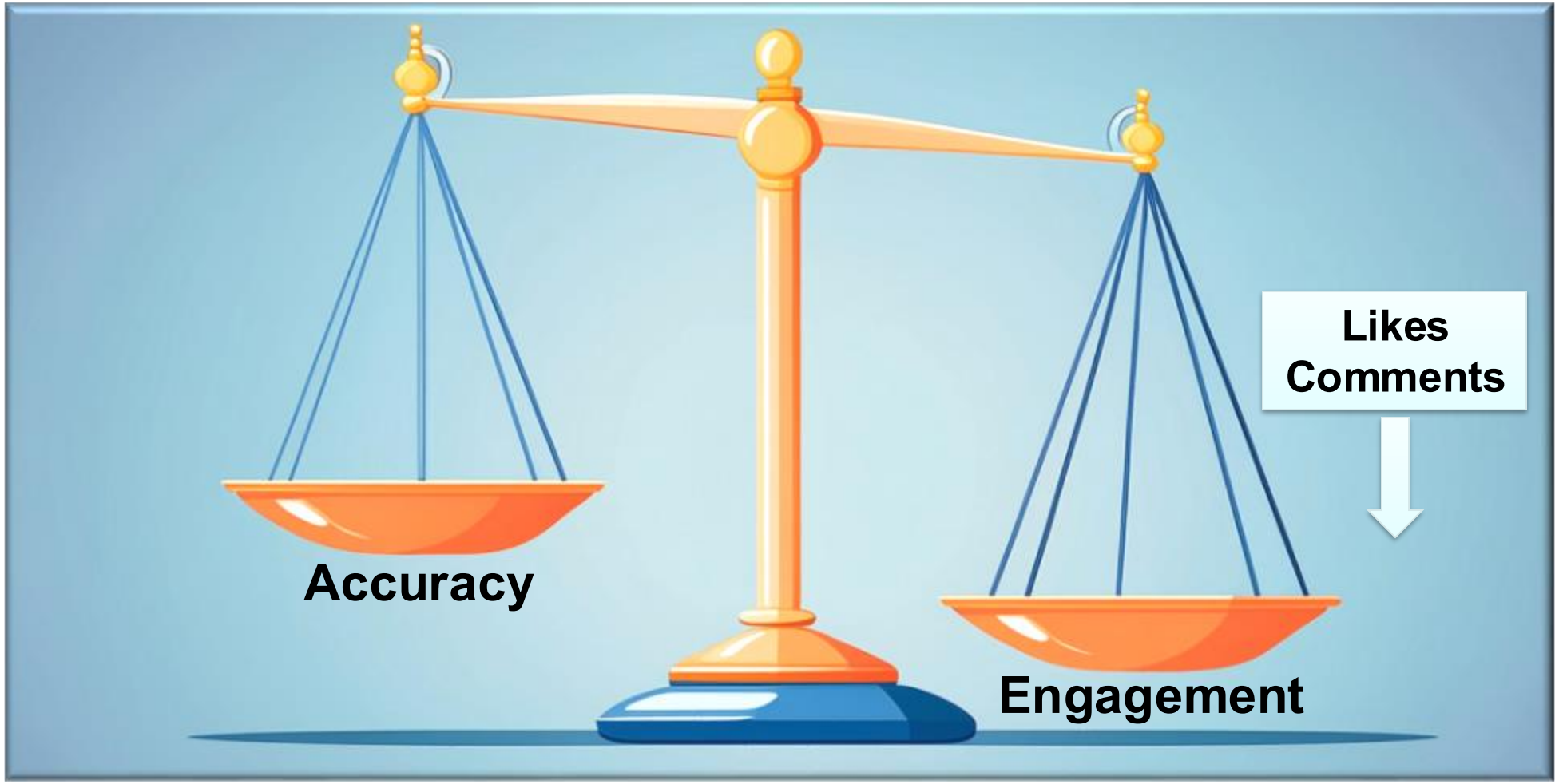
Certifiably True: The Impact of Self-Certification on Misinformation

Swapneel Mehta, Postdoctoral Associate

*Nichols, Aaron D., Nina Mazar, Tejovan Parker, Swapneel Mehta, Gordon Pennycook, David Rand, and Marshall Van Alstyne (2024), “Certifiably True: The Impact of Self-Certification on Misinformation.”

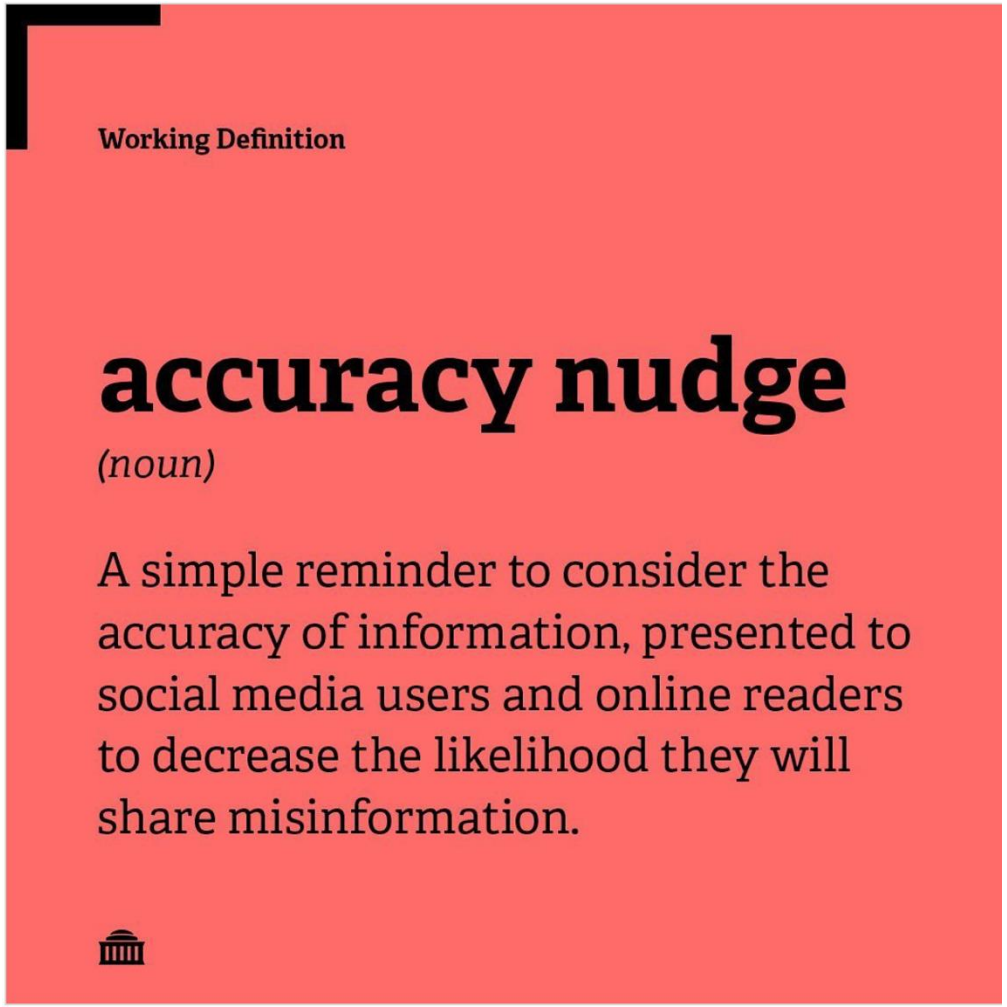
BU Questrom School of Business

The Social Media Dilemma



Ren, Dimant, and Schweitzer 2023

Accuracy Nudges



Pennycook G, Rand DG. Nudging Social Media toward Accuracy. *Ann Am Acad Pol Soc Sci.* 2022 Mar; 700(1):152-164.

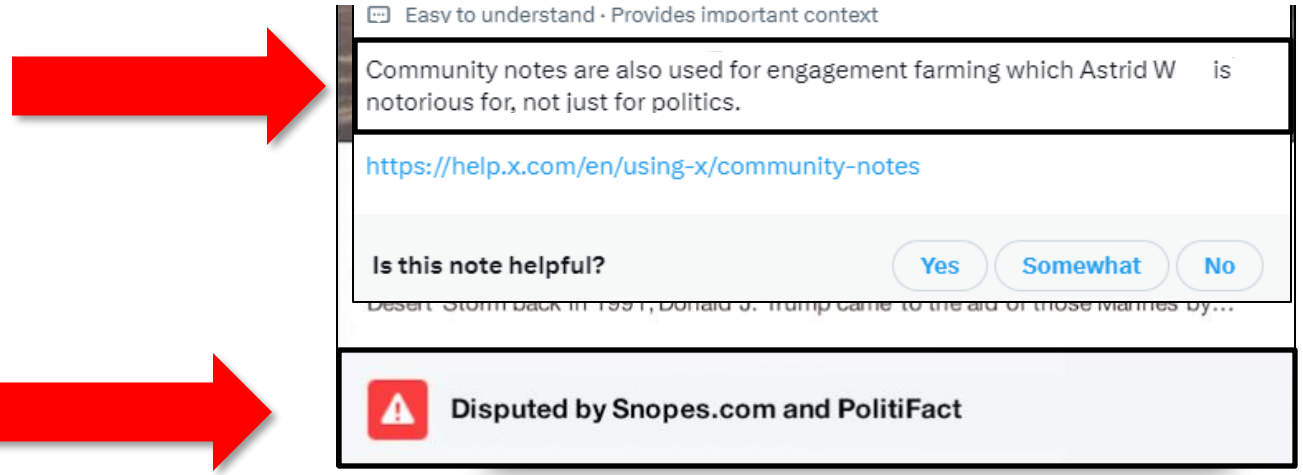
Fact-Checking, Labeling, and Debunking

BBC

Meta to replace 'biased' fact-checkers with moderation by users

Meta is abandoning the use of independent fact checkers on Facebook and Instagram, replacing them with X-style "community notes" where...

1 week ago



e.g., Hsu, Thompson (2023), Clayton et al. 2023

The Current Toolkit Helps Improve Discernment, but it Lack Bite!

- Inoculation & nudges are not targeted enough and may be hard to scale
- Labeling and Debunking: “Implied Truth Effect” (Warnings on subset of misinformation increase perceived accuracy of headlines without warnings; Pennycook et al. 2020)
- Burden of misinformation is on consumers (“victims”) instead of culprits (sharers)
- No direct penalization of misinformation sharers
- Unlikely to deter those who report they intentionally share false information

Self-Certification: A Mechanism to Reduce Misinformation

Allows users to certify truthfulness of their shared content

- Backed by users' money (!); points; reputation/social capital

Certifications can be challenged

- Lose money if false, gain money if true

Uses economic and behavioral theories to combat misinformation

- Market externalities[1], signaling[2], and screening[3]
- Nudges accuracy[4], Wisdom of Crowds[5]

e.g., [1]Cooter 1989; [2]Akerlof 1970; [3]Spence 1973; [4]Pennycook et al. 2021; [5]Martel et al. 2024

Can Self-Certification clear Markets of Misleading Information?

- Reduce sharing of false claims
- Increase sharing of true claims!
- Reduce sharing of sensational false claims

Experiment 1 – Overview

Social media users (*N* = 1,490 participants; 29,800 responses)

- Cloud Research Connect
- M_{age} (SD) = 44.19 years (15.31); Female = 49.7%; Prefer Democrat = 52.8%

Shown 20 news headlines and made sharing decisions

- Half true (false), Half interesting (boring)
- Randomly selected from 202 pre-tested headlines

Choosing to share affected bonuses

- Sharing interesting headlines (+$0.05); boring headlines (-$0.05)

Across 3 between-ss conditions:

- Control-Sharing (*n* = 500), Costless Certification (*n* = 503), Costly Certification (*n* = 487)

If you saw this article on social media, what would you choose to do with it?

Social Media Baseline

If you saw this article on social media, what would you choose to do with it?

Not Share	Share
$0	**Boring** -$0.05 **Interesting** +$0.05

Costless Warrant / Cheap Talk

If you saw this article on social media, what would you choose to do with it?

Not Share	Share	Warrant as True and Share
$0		

Boring	-$0.05
Interesting	+$0.05

Accuracy Nudge But No Internalization of External Costs!

Warrant / Costly Certification

If you saw this article on social media, what would you choose to do with it?

Not Share	Share	Warrant as True and Share
$0	**Boring** -$0.05 **Interesting** +$0.05	⬆

Simulating outcome after warrants are challenged (with $0.10 collateral).
Certainty: 100% of warrants are challenged.

Accuracy Nudge AND Internalization of External Costs!

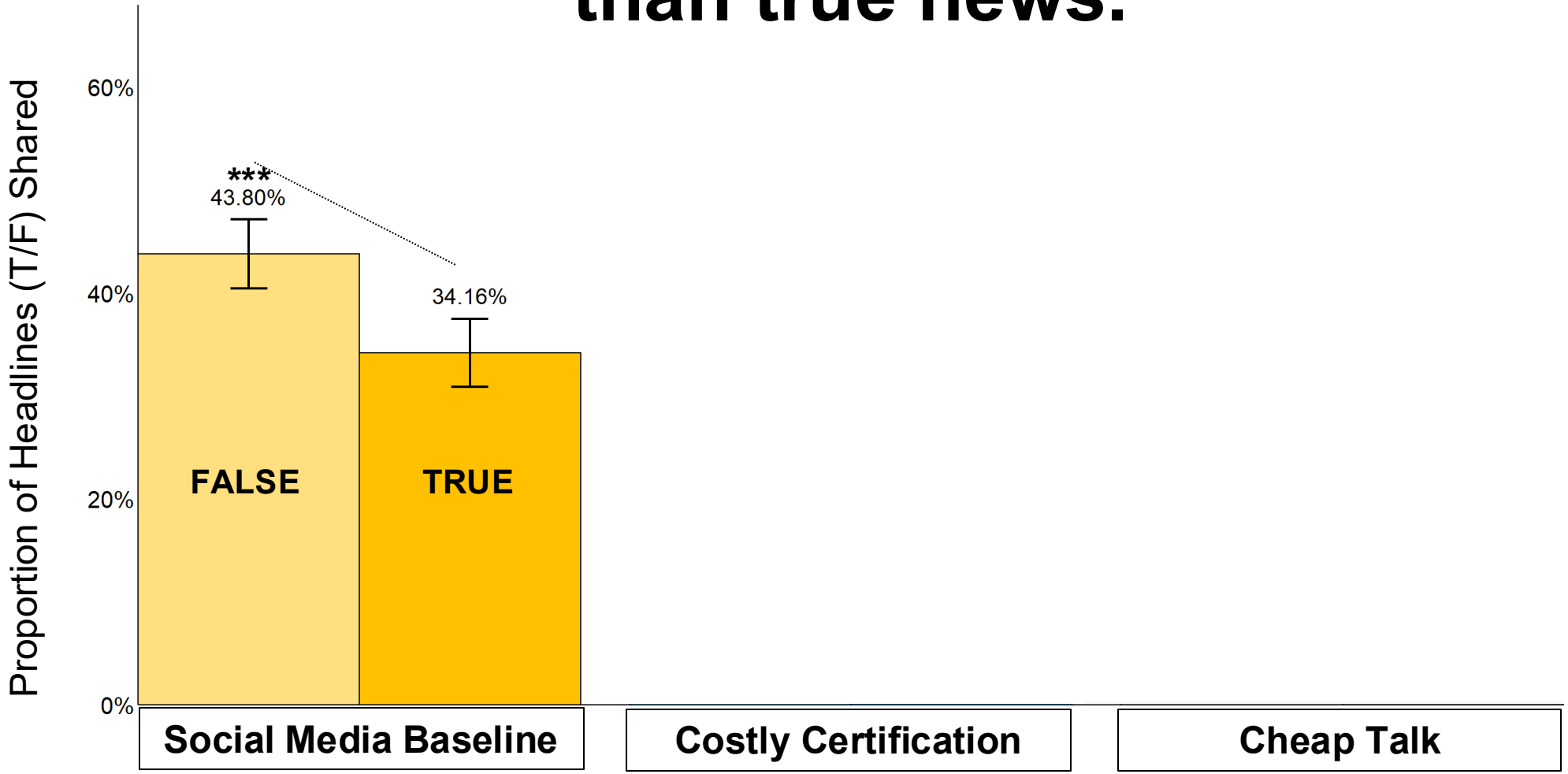
Warrant / Costly Certification

Example True Headline	Example False Headline
 USATODAY.COM **Trump ally Lindsey Graham must testify in Georgia grand jury investigation, federal judge rules**	 WFXRTV.COM **Florida schools to hire vets without teaching experience**
If you saw this article on social media, what would you choose to do with it?	**If you saw this article on social media, what would you choose to do with it?**
Not Share	Not Share
Share	Share
Warrant as true and Share	Warrant as true and Share

Can Self-Certification clear Markets of Misleading News?

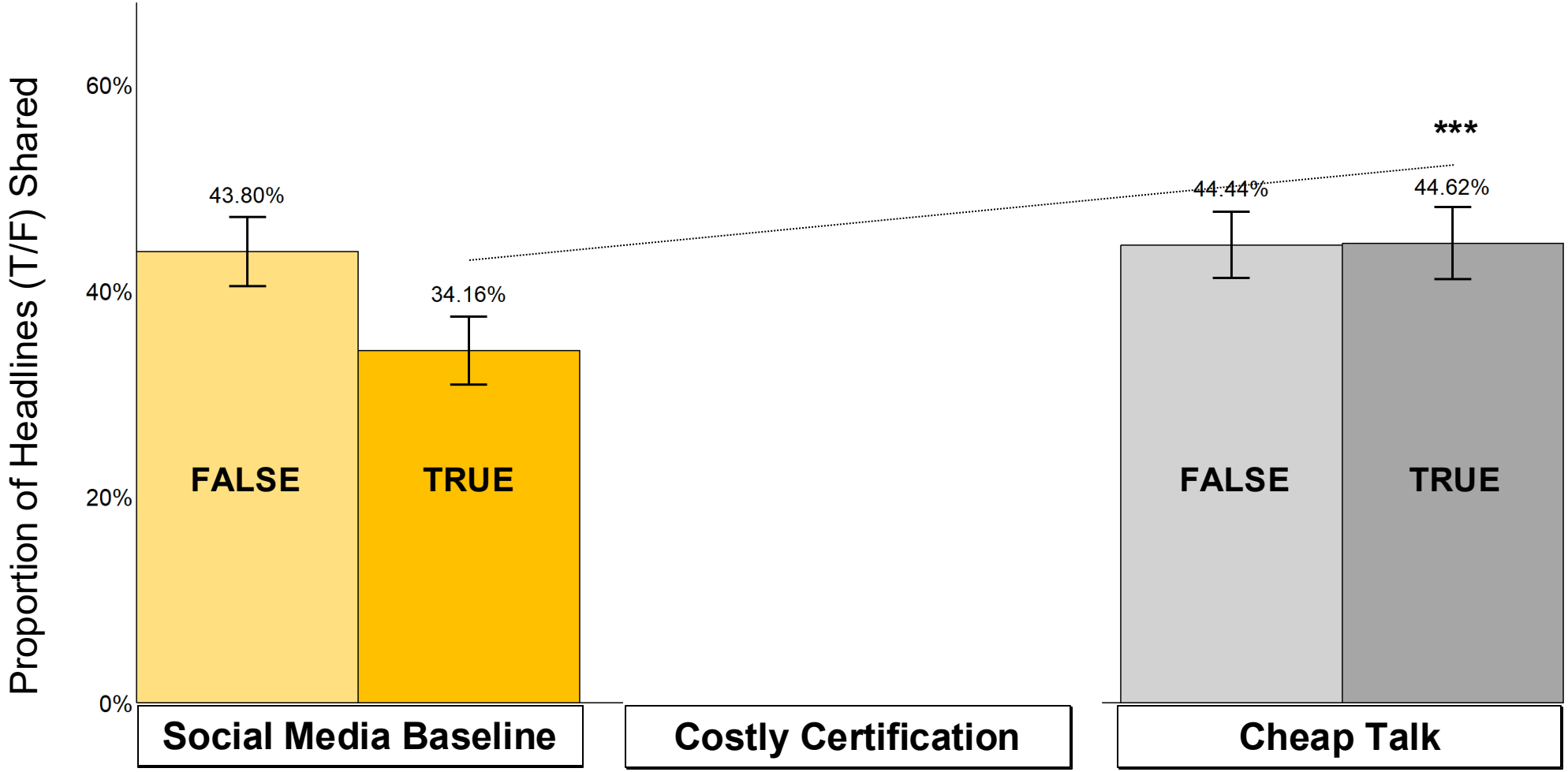
Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

In the Baseline condition, false news is shared more than true news.



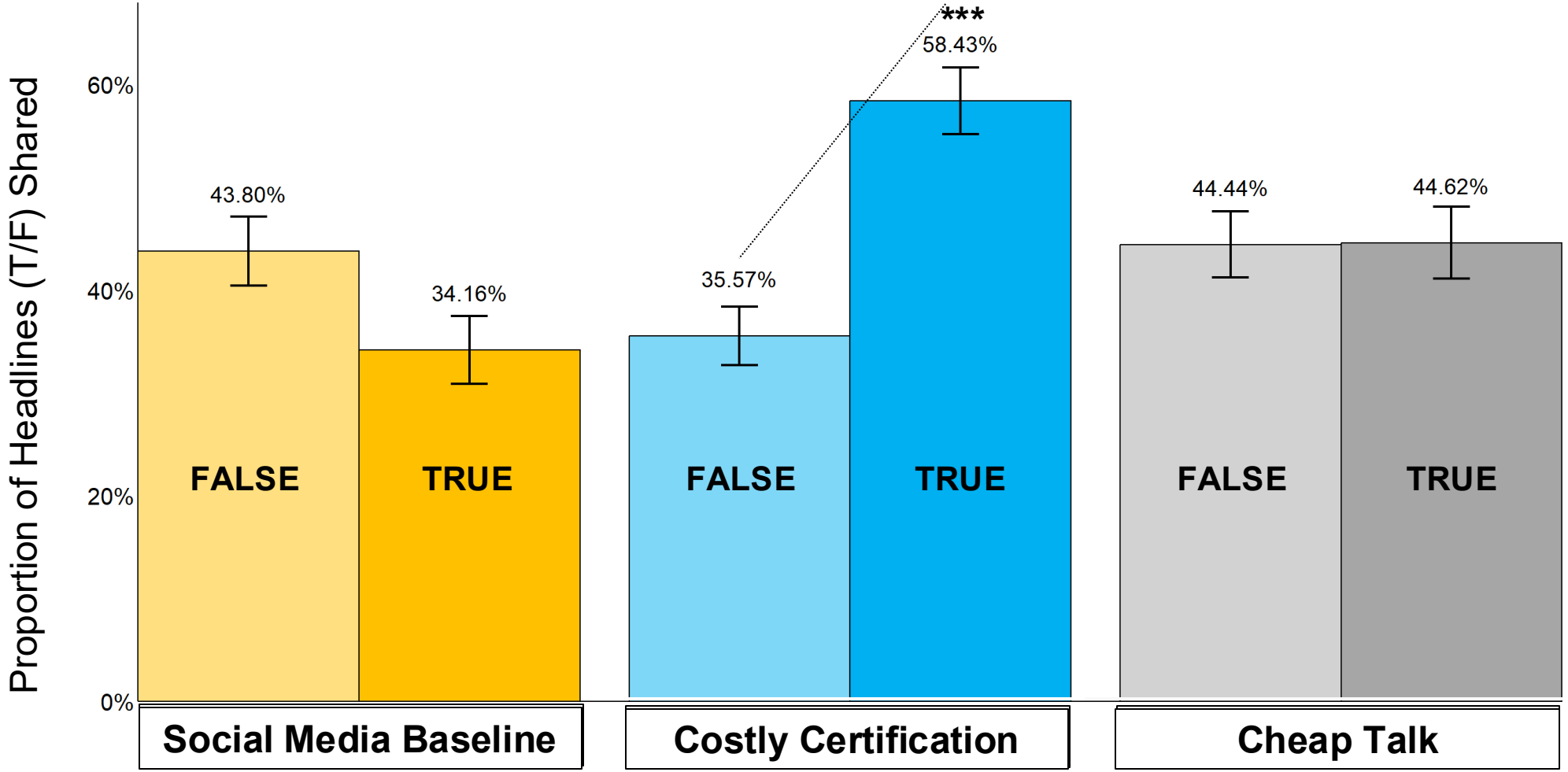
Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

In the Cheap Talk condition, true news is shared more than in the social media baseline condition.



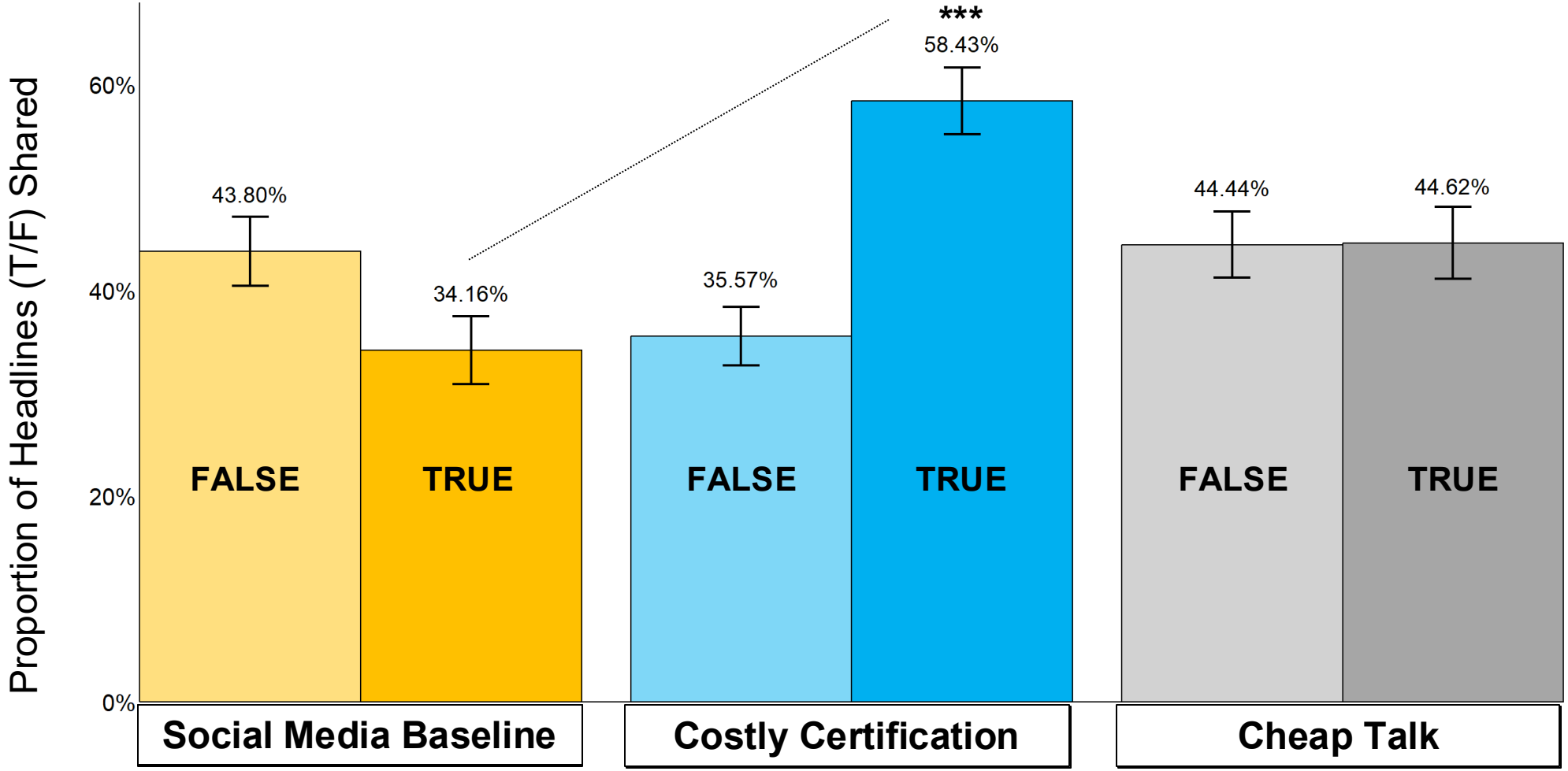
Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

In the Costly Certification condition, true news is shared more than false news.



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

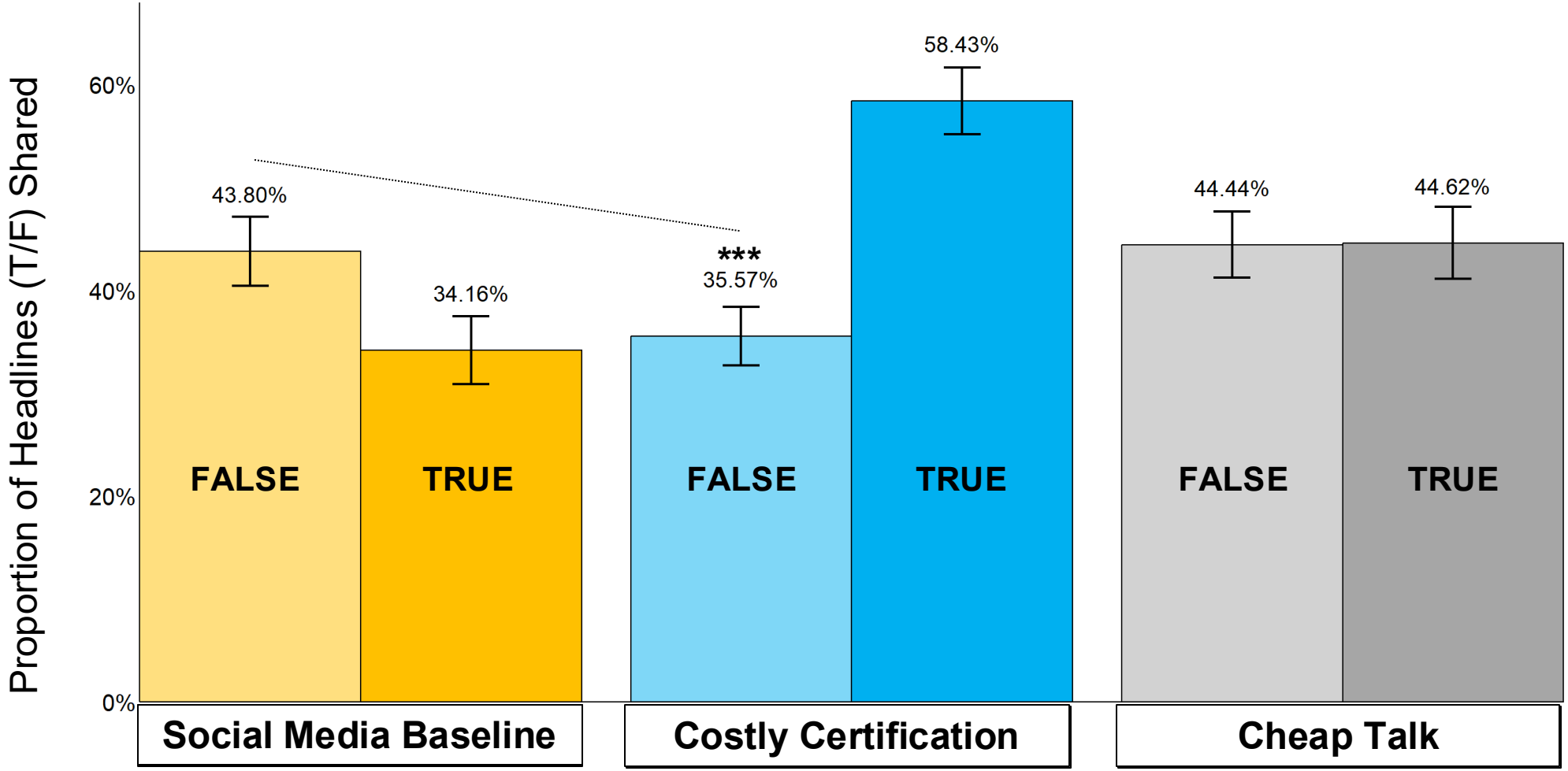
In the Costly Certification condition, true news is shared more than in the Social Media Baseline cond.



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

Lying is less likely when it appears more self-diagnostic (*JCR:* Gai and Puntoni 2021)

In the Costly Certification condition, false news is shared less than in the Social Media Baseline cond.



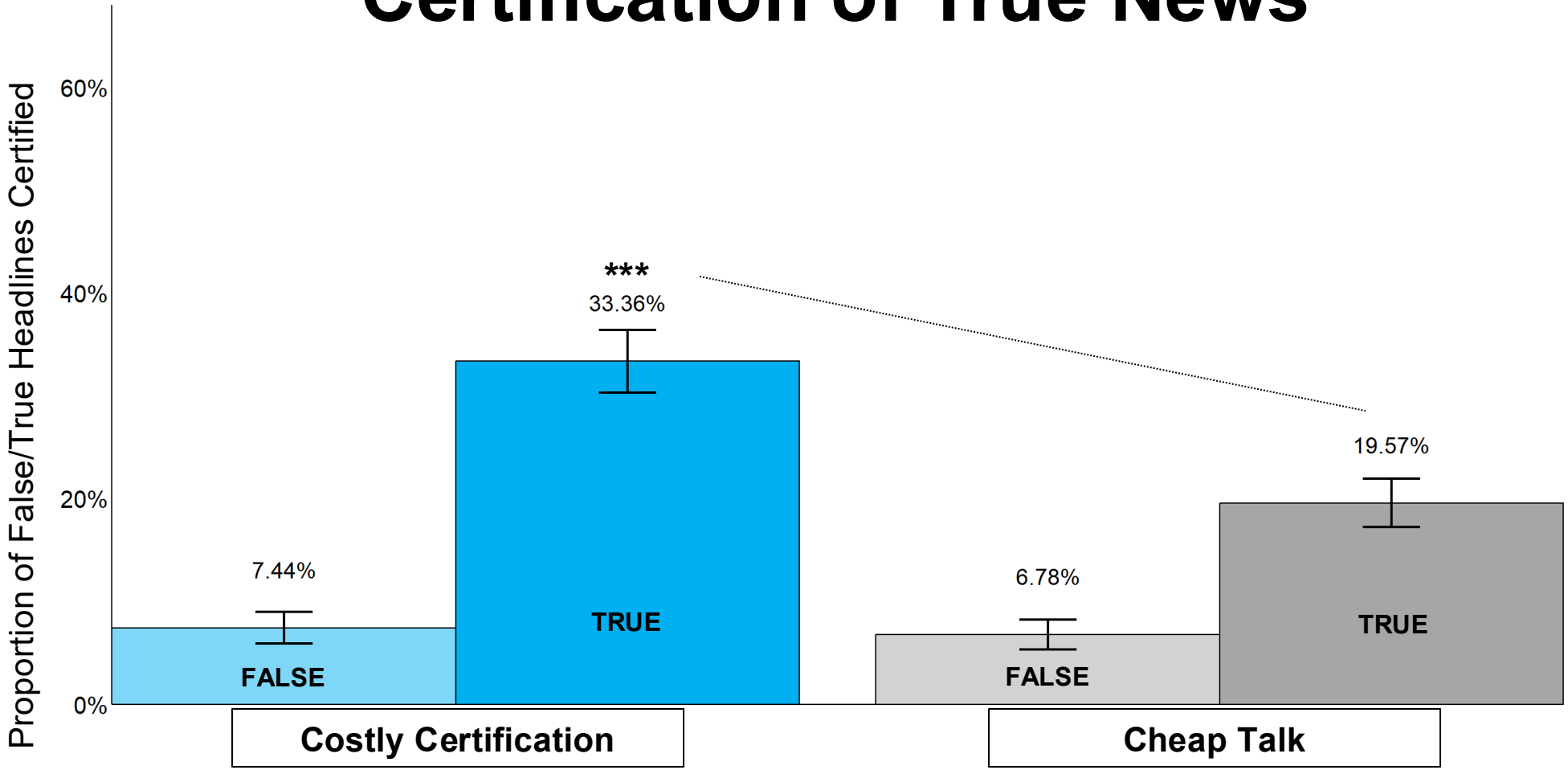
Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

Lying is less likely when it appears more self-diagnostic (*JCR:* Gai and Puntoni 2021)

What do sharers choose to certify?

Next Up: Proportion of Certified False and True News

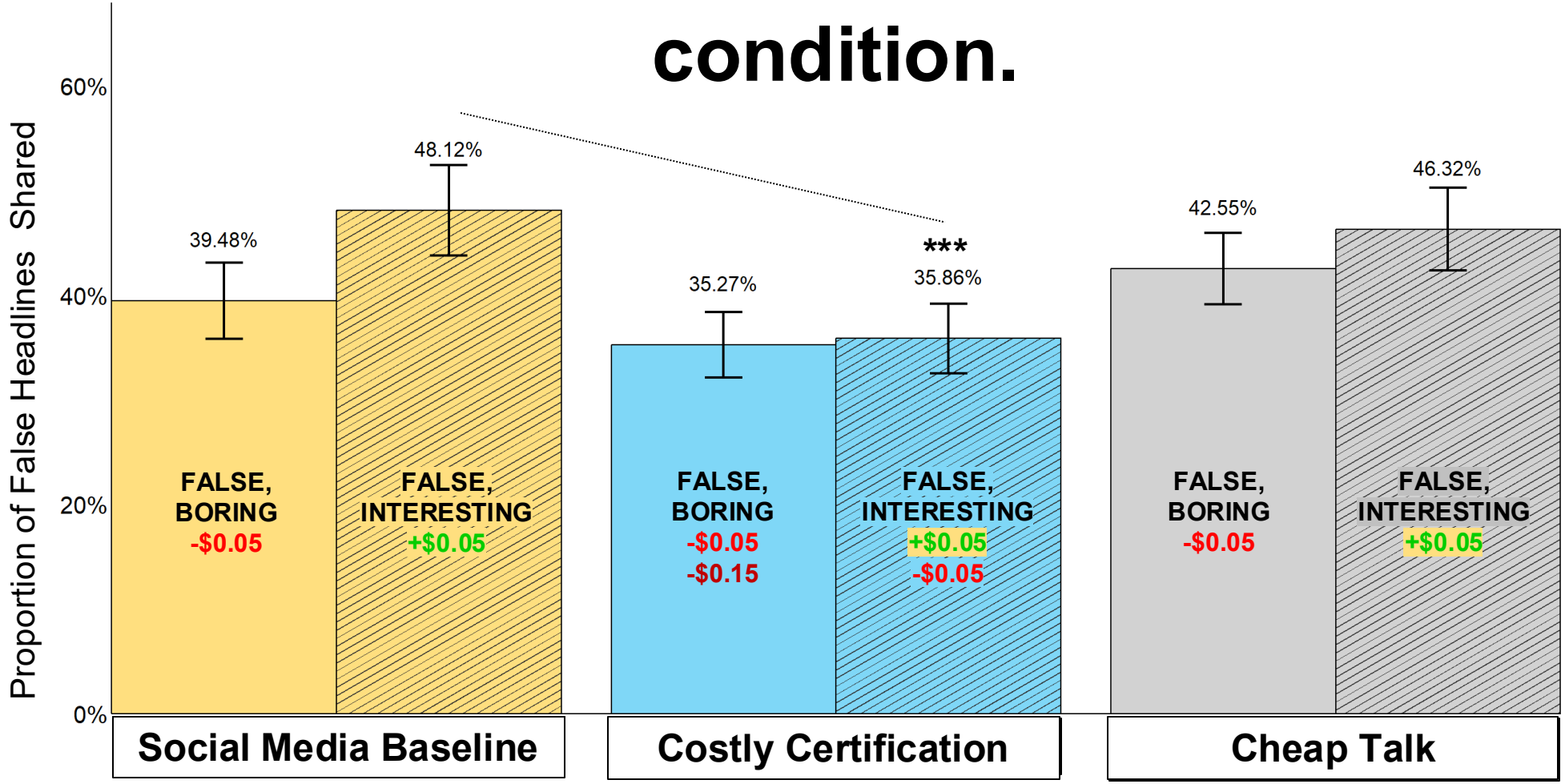
When Certification is Costly, we see an <u>Increased</u> Certification of True News



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

According to Past Research, False-Interesting News Is More likely to Go Viral than False-Boring News

In the Costly Certification condition, false-interesting news is shared less than in the Social Media Baseline condition.



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

Can Self-Certification clear Markets of Misleading Information?

- Reduce sharing of false claims
- Increase sharing of true claims!
- Reduce sharing of sensational false claims

Key Takeaways About Self-Certification to Fight Misinformation

- ***Self-Certification Mechanism***

  Option to signal content veracity, backed by collateral

- ***Headline News Sharing (Exp 1):***

  ↑ sharing overall

  ↑ true news; ↓ false news

- ***Headline News Consumption (Exp 2):***

  ↑ perceived accuracy

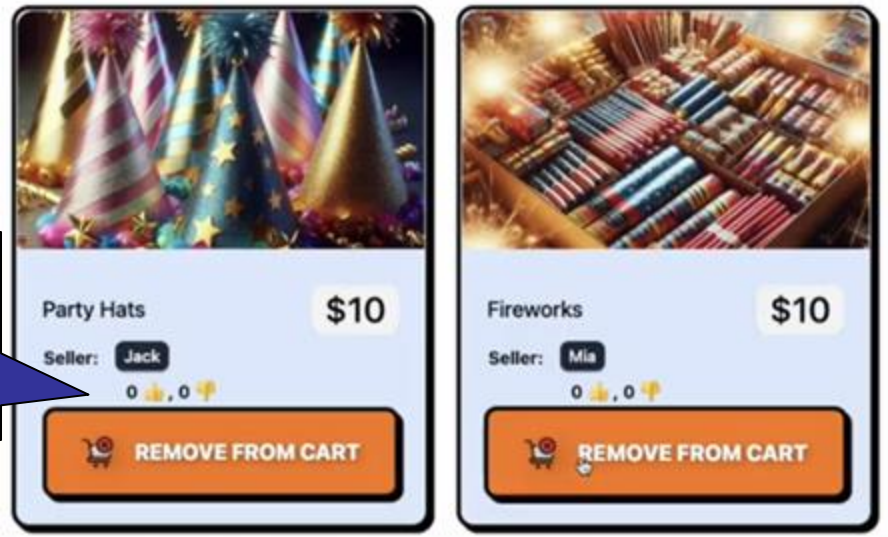
Truthmarket.com

Questions, Feedback, & Challenges
are encouraged ☺ swapneel@bu.edu

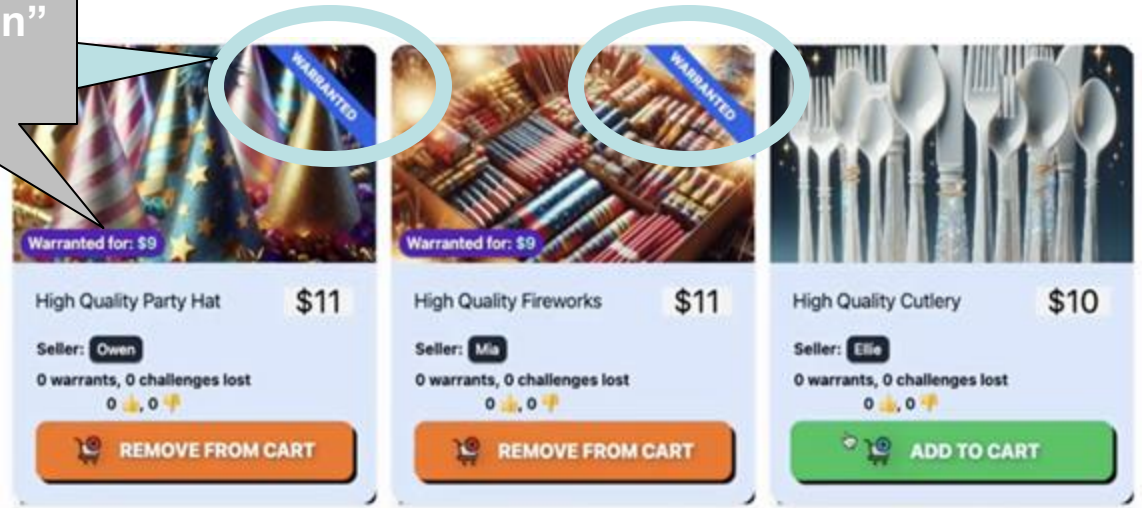
Nichols, Aaron D., Nina Mazar, Tejovan Parker, *Swapneel Mehta, Gordon Pennycook, David Rand, and Marshall Van Alstyne (2024), “Certifiably True: The Impact of Self-Certification on Misinformation.”

We built ”mini-Amazon” to run competitive online experiments

Control “Reputation” Market



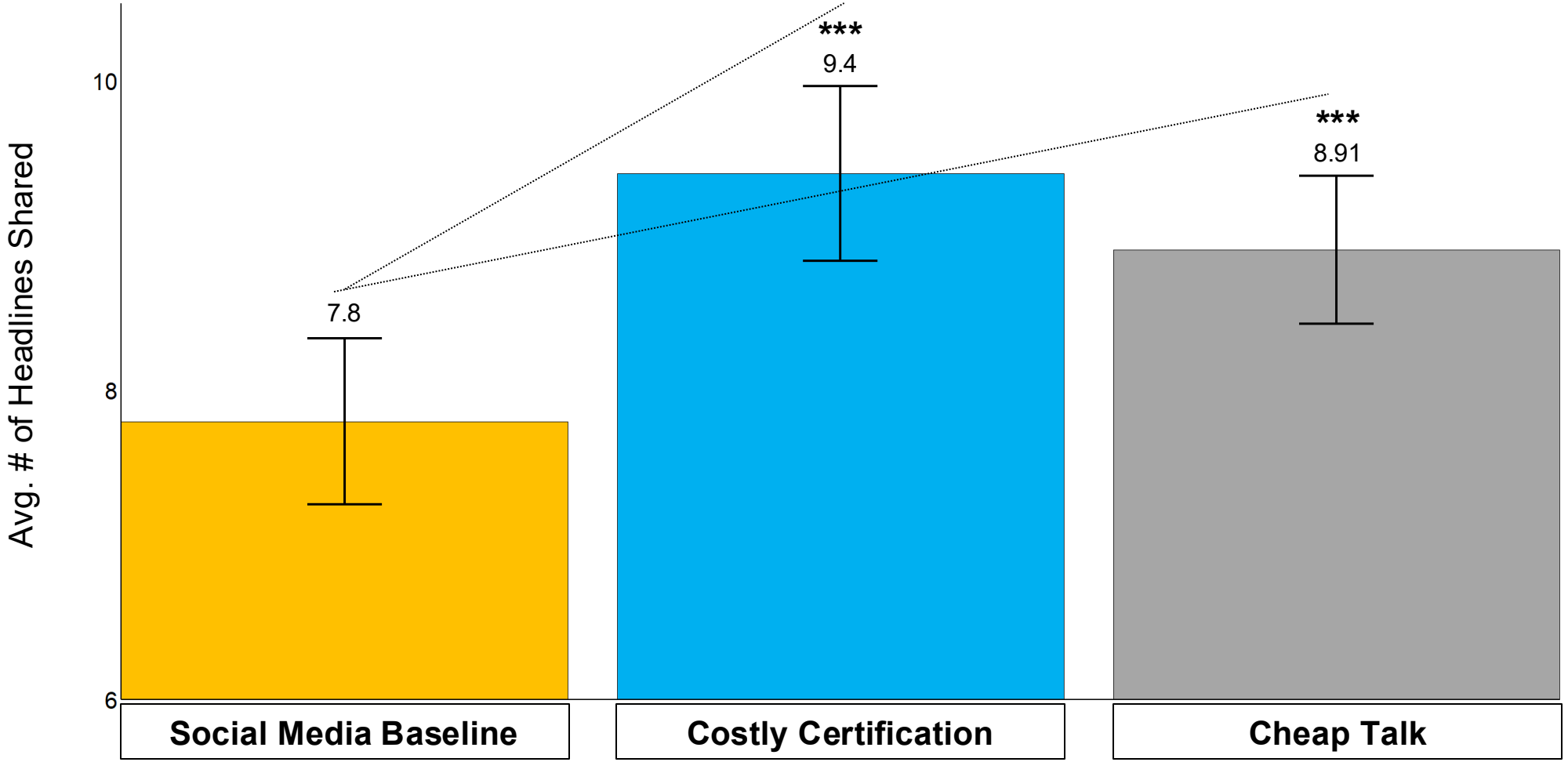
Treatment “Certification” Market



What News do Sharers Share?

Proportion of False vs. True News Shared

Having the option to certify the truthfulness (no matter if cheap talk or costly) increased total news shared



Note: Estimated means from linear model with robust SEs clustered on participant and headline. Error bars indicate 95% CIs. * p < 0.05; ** p< 0.01; *** p < 0.001.

Experiment 2. Stimuli – *Costless & Costly Certification*

Shared

Shared & Warranted as True

Costless & Costly Certification

To the best of your knowledge, how accurate is the claim in the above headline?
1 – Not at all accurate; 7 – Very Accurate

Fighting Misinformation with Self-Certification

Contributions:

- We test a novel misinformation intervention
  - + Producers can be held accountable
  - + User-Driven / Decentralized
  - + No Censorship (free speech is maintained)
  - + Scalable
- We use incentive compatible experimental designs

Key Limitations: Warrants are 100% challenged; no reputation building; no interactions with others; no variations in type and amount of collateral, only clearly true/false information, etc.

Experiment 2 – Overview

Participants (*N* = 2,003; 48,072 responses)

- Cloud Research Connect
- M_{age} (SD) = 39.97 years (12.77); Female = 49.9%; Prefer Democrat = 61.7%

Shown 24 news headlines and rated perceived accuracy of each

- Half true (false)
- Randomly selected from 172 headlines that were certified in Exp. 1

By condition, some headlines were displayed with labels

- Labels (or their absence) indicated if previous participants shared (and certified them)
- Conditions: Control (*n* = 505), Control-Sharing (*n* = 499),
  Costless Certification (*n* = 500), Costly Certification (*n* = 499)

Experiment 2. Stimuli – *Control*

Control

To the best of your knowledge, how accurate is the claim in the above headline?
1 – Not at all accurate; 7 – Very Accurate

Experiment 2. Stimuli – *Control-Sharing*

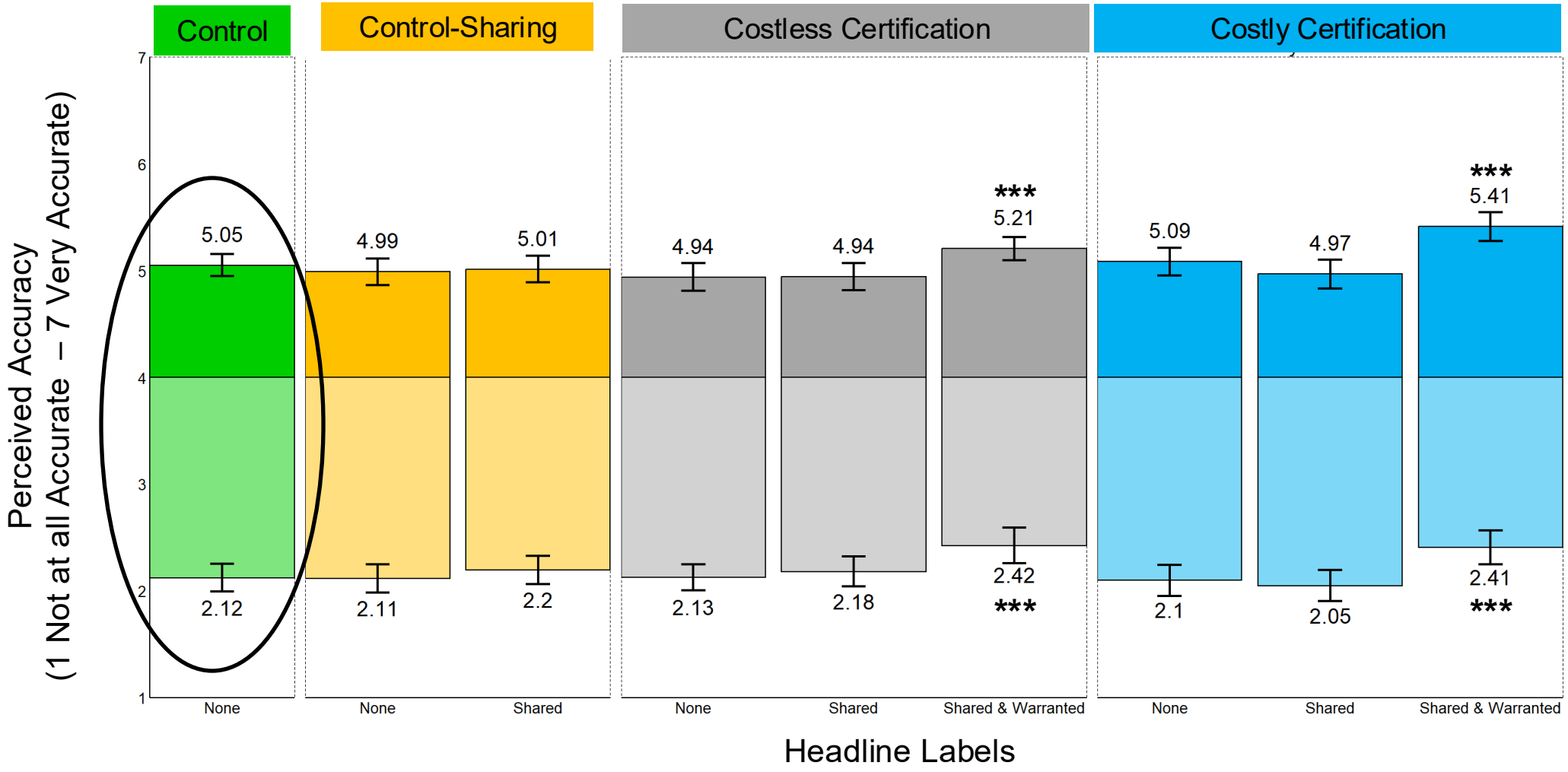
Shared

Control-Sharing

To the best of your knowledge, how accurate is the claim in the above headline?
1 – Not at all accurate; 7 – Very Accurate

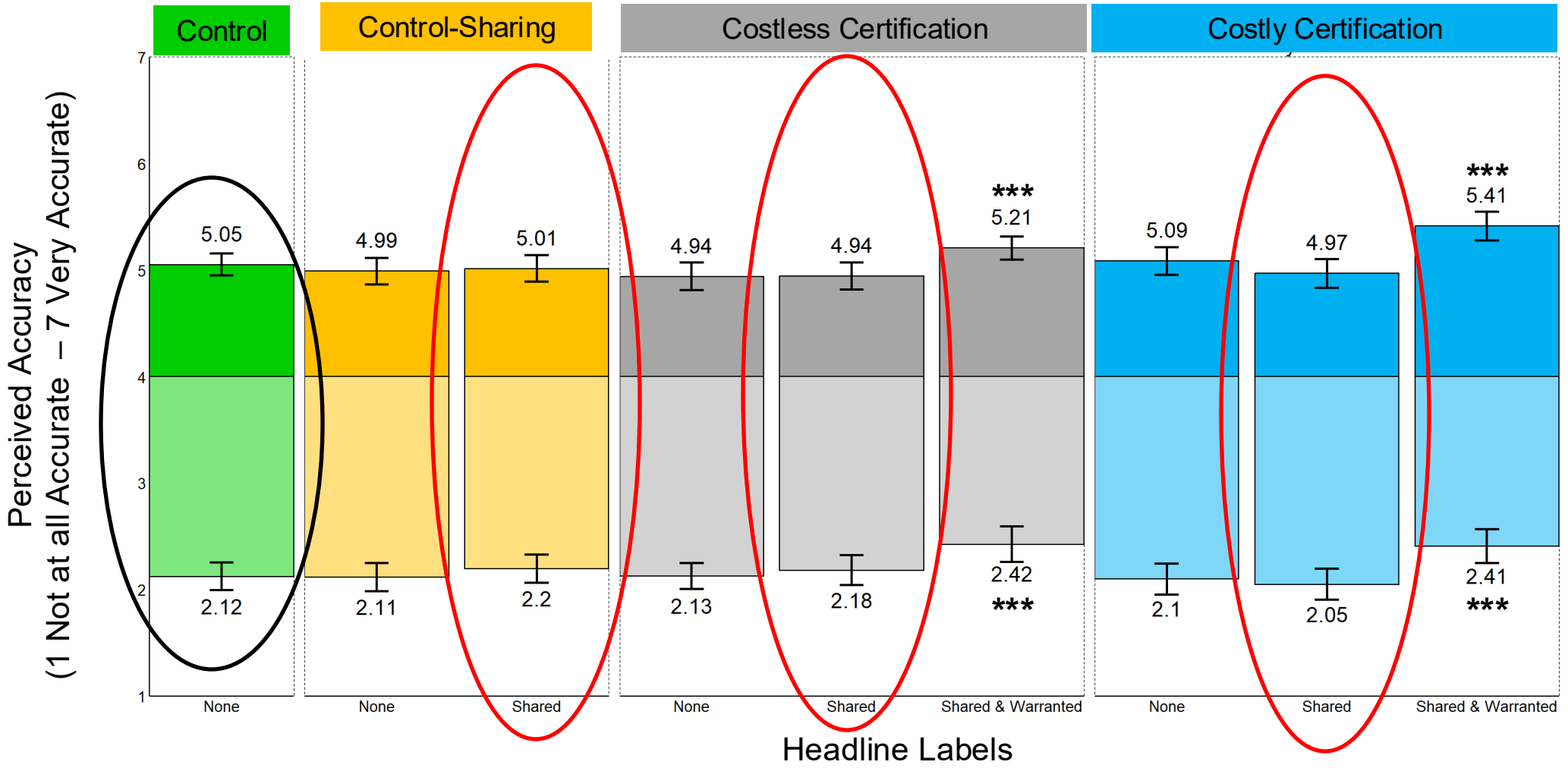
Perceived Accuracy of News Headlines

Comparisons Relative to the Control



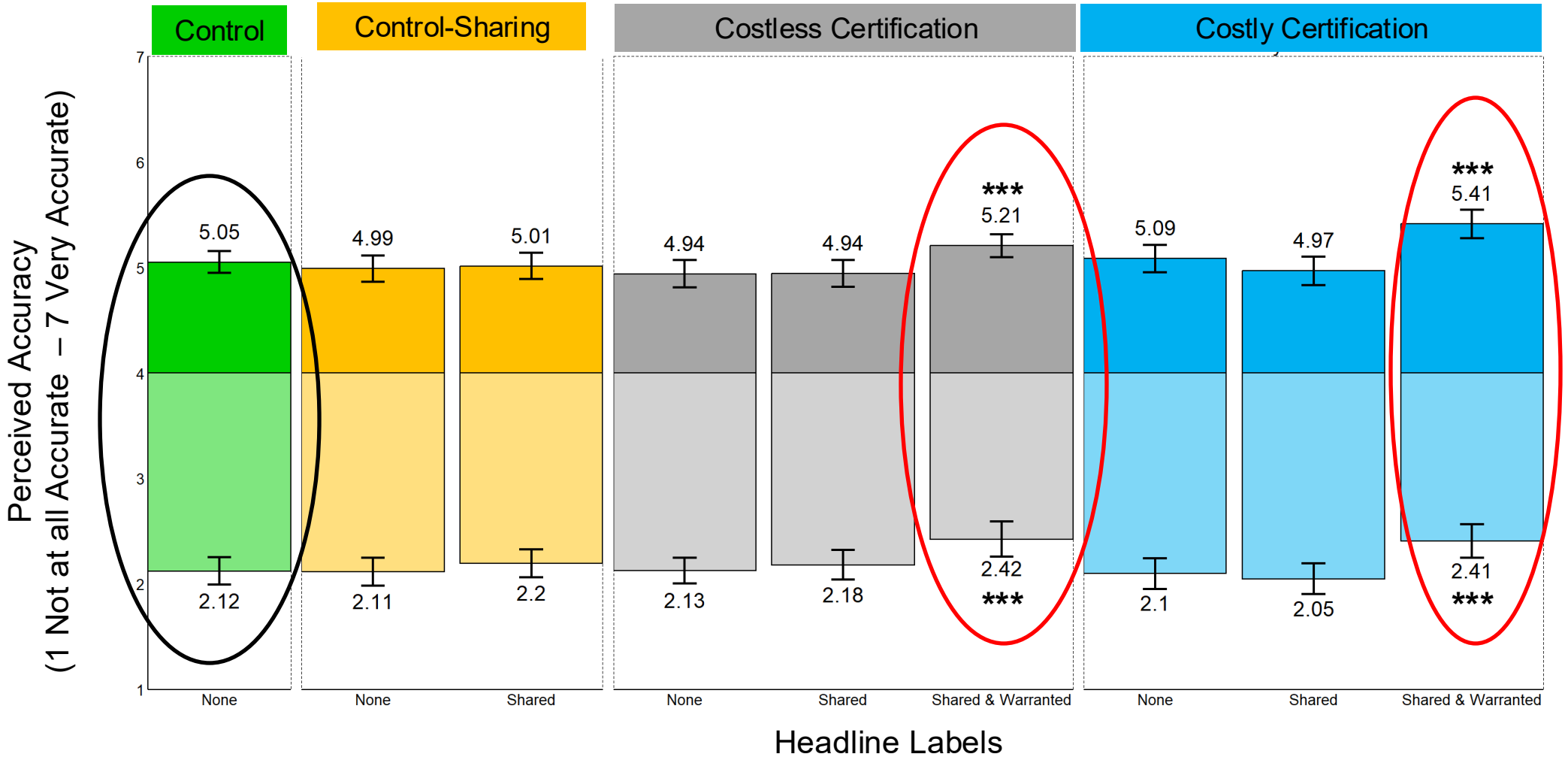
Note: Estimated means from linear model with robust SEs clustered on participant and headline ID. Error bars indicate 95% CIs. * p < 0.05; ** p < 0.01; *** p < 0.001.
All comparisons, unless otherwise indicated, are relative to their counterpart in the control

No Impact of “Shared” Labels



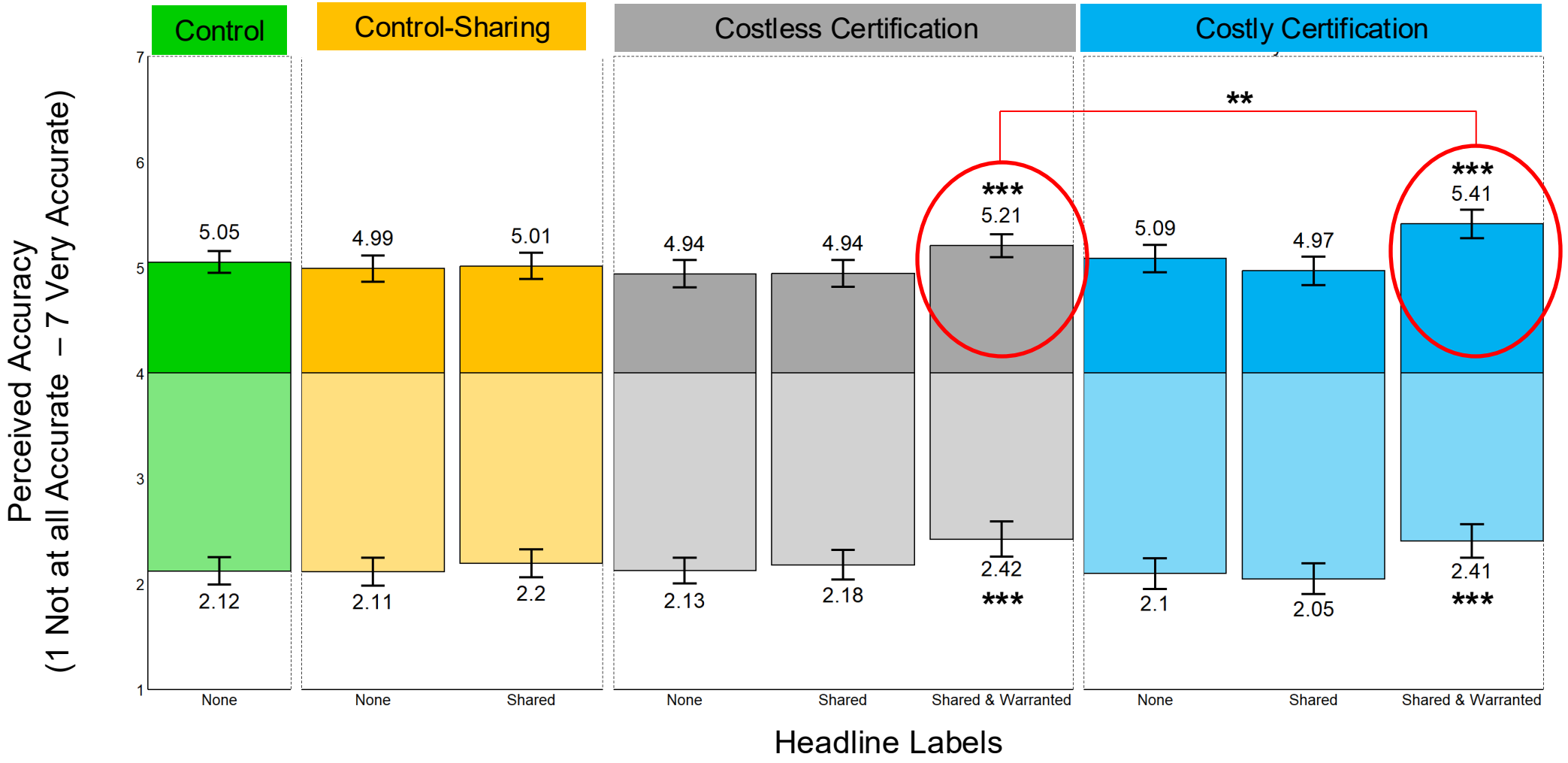
Note: Estimated means from linear model with robust SEs clustered on participant and headline ID. Error bars indicate 95% CIs. * p < 0.05; ** p < 0.01; *** p < 0.001.
All comparisons, unless otherwise indicated, are relative to their counterpart in the control

Certification ↑ Perceived Accuracy of True and False



Note: Estimated means from linear model with robust SEs clustered on participant and headline ID. Error bars indicate 95% CIs. * p < 0.05; ** p < 0.01; *** p < 0.001.
All comparisons, unless otherwise indicated, are relative to their counterpart in the control

Costly ↑ Perceived Accuracy of True Claims More



Note: Estimated means from linear model with robust SEs clustered on participant and headline ID. Error bars indicate 95% CIs. * p < 0.05; ** p < 0.01; *** p < 0.001.
All comparisons, unless otherwise indicated, are relative to their counterpart in the control

References

Akerlof, George A. (1970), “The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism,” *The Quarterly Journal of Economics*, 84(3), 488–500, http://www.jstor.org/stable/1879431.

Cooter, Robert D. (1989), “The Coase Theorem,” in *Allocation, Information and Markets*, ed. John Eatwell, Murray Milgate, and Peter Newman, London: Palgrave Macmillan UK, 64–70, https://doi.org/10.1007/978-1-349-20215-7_6.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig, Ayesha Ali, Joe Bak-Coleman, Sarit Barzilai, Melisa Basol, Adam J. Berinsky, Cornelia Betsch, John Cook, Lisa K. Fazio, Michael Geers, Andrew M. Guess, Haifeng Huang, Horacio Larreguy, Rakoen Maertens, Folco Panizza, Gordon Pennycook, David G. Rand, Steve Rathje, Jason Reifler, Philipp Schmid, Mark Smith, Briony Swire-Thompson, Paula Szewach, Sander van der Linden, and Sam Wineburg (2024), “Toolbox of Individual-Level Interventions against Online Misinformation,” *Nature Human Behaviour*, 8(6), 1044–52, https://www.nature.com/articles/s41562-024-01881-0.

Littrell, Shane, Casey Klofstad, Amanda Diekman, John Funchion, Manohar Murthi, Kamal Premaratne, Michelle Seelig, Daniel Verdear, Stefan Wuchty, and Joseph E. Uscinski (2023), “Who Knowingly Shares False Political Information Online?,” *Harvard Kennedy School Misinformation Review*, https://misinforeview.hks.harvard.edu/article/who-knowingly-shares-false-political-information-online/.

Martel, Cameron, Jennifer Allen, Gordon Pennycook, and David G. Rand (2024), “Crowds Can Effectively Identify Misinformation at Scale,” *Perspectives on Psychological Science*, 19(2), 477–88, https://doi.org/10.1177/17456916231190388.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand (2021), “Shifting Attention to Accuracy Can Reduce Misinformation Online,” *Nature*, 592(7855), 590–95, https://doi.org/10.1038/s41586-021-03344-2.

Rathje, Steve, Jon Roozenbeek, Jay J. Van Bavel, and Sander Van Der Linden (2023), “Accuracy and Social Motivations Shape Judgements of (Mis)Information,” *Nature Human Behaviour*, 7(6), 892–903, https://www.nature.com/articles/s41562-023-01540-w.

Ren, Zhiying (Bella), Eugen Dimant, and Maurice Schweitzer (2023), “Beyond Belief: How Social Engagement Motives Influence the Spread of Conspiracy Theories,” *Journal of Experimental Social Psychology*, 104, 104421, https://linkinghub.elsevier.com/retrieve/pii/S0022103122001408.

Spence, Michael (1973), “Job Market Signaling,” *The Quarterly Journal of Economics*, 87(3), 355–74, http://www.jstor.org/stable/1882010.

Zahidi, Saadia (2024), *The Global Risks Report 2024*, 1, World Economic Forum, https://www.weforum.org/publications/global-risks-report-2024.