The Cognitive Mechanisms Underlying Testimonial Injustice in Collaborative Decision-Making

Michelle M. Wang (michellewang@princeton.edu) & Natalia Vélez
Department of Psychology, Princeton University

Abstract

At its best, collaboration allows for humans to overcome individual limitations by pooling the knowledge of many minds. However, realizing the full potential of collaboration is difficult because some team members' perspectives may be systematically excluded based on their social identities. Here, we use a classic marble-urn paradigm to empirically test the underlying causes of and specific conditions that give rise to *testimonial injustice*. Across two experiments (one preregistered, one pilot; N = 147), we find that men are consistently (a) more confident than women in expressing their testimony, and (b) less likely than women to revise their beliefs after incorporating others' testimony, specifically when the evidence participants have access to is moderately strong (Experiment 1) and highly diagnostic (Experiment 2) but not definitive. Our work provides important insights into when and how collaborations fail, as well as the psychological root causes that underlie persistent gender discrepancies in group decision-making contexts.

Keywords: collaboration; decision making; testimonial injustice; gender; social groups; statistical inference

Introduction

Collaboration provides a way for us to make better decisions by pooling the knowledge and perspective of many minds (Dreber et al., 2015). However, fully integrating everyone's perspectives provides challenges, as working as a team also introduces systematic biases in who gets heard. A key source of bias in this area is gender: In workplace contexts, for example, women are penalized for speaking out more (Brescoll, 2011), their contributions are often downplayed (Heilman & Haynes, 2005), and they are asked to—and ultimately perform—more tasks that are low in visibility and promotability than their male counterparts (Babcock et al., 2017; Nelson et al., 2023). These inequalities are known as **testimonial injustice**, defined here as inequities in whose perspective gets voiced and heard.

Testimonial injustice has pernicious consequences both for teams *and* for the people whose perspectives are overlooked. It harms *teams* because teams end up systematically missing out on a particular set of perspectives. For example, because women's perspectives were historically underrepresented in U.S. healthcare research, both as researchers and as study participants, early cardiovascular disease research was largely male-centric. By adopting a one-size-fits-all approach, early treatment for cardiovascular disease failed to account for human variation in artery size, hormonal cycles, and body fat composition—variation that would have been considered earlier if women's perspectives were included. This oversight led to more misdiagnoses and less effective treatment plans for both male and female patients, causing harm to patients across the board, as well as to researchers (George et al., 2015; Mosca et al., 2011). Testimonial injustice, of course, also leads to harmful consequences for the *individuals* whose perspectives are overlooked, because the decisions made by the team may not align with their priorities. For example, early research on voice-recognition and speech-to-text technologies overwhelmingly used male engineers and male voices for testing and development. As a result, these technologies were more effective in recognizing male than female voices (Bajorek, 2019), leading to lower accuracy and poorer user experience for women.

The goal of the present research is to understand the cognitive mechanisms underlying testimonial injustice. Our overarching hypothesis is that testimonial injustice may arise from both discrepancies in people's (1) confidence to express their own testimony and (2) willingness to revise their testimony based on others' input. Critically, these two possibilities are not mutually exclusive, and they may affect different stages of group decision-making and planning. For example, suppose that an intern has noticed a mistake during an important team meeting. In this scenario, both causes we listed may prevent her from getting her message across. On one hand, the intern may be less likely to share her initial judgment to begin with, because she does not expect her input to be well received based on her age, gender, and status (Martin & Ruble, 2004; Mead, 2012; Medina, 2013; Ridgeway, 2009; Tice, 1992). On the other hand, even if the intern *does* share her testimony, she might be quick to revise her original judgment after hearing others' opinions—even when her own evidence was reasonably strong—by placing greater weight on others' input than her own (Carli, 1989; Eagly et al., 1981; Eagly & Chrvala, 1986).

Both possibilities are supported by decades of research on group decision-making and planning in businesses, courtrooms, and healthcare settings. Prior work finds converging evidence that social identities such as gender systematically bias how much information people disclose in meetings (Brescoll, 2011) and how likely they are to conform to other group members' input (Eagly, 1983). However, even after decades of research on biases in collaboration, we still do not know the evidentiary conditions that give rise to each of these possibilities, and consequently, how to best intervene on their underlying causes. Therefore, studying the conditions that contribute to testimonial injustice enables us to identify specific targets of intervention (e.g., encouraging

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

women to speak up more vs. implementing decision-making structures that limit undue belief revision).

In this paper, participants observe marbles drawn from an urn and infer which urn produced the marble based on their firsthand observations and testimony from a partner who observes a different sample. This marble-urn task is a foundational paradigm of cognitive science, because it can be extended to formally describe how people combine firsthand evidence with a variety of social cues, including their partner's testimony, confidence, emotional expressions, and intent (Campbell-Meiklejohn et al., 2017; Oey et al., 2023; Wu et al., 2024). Here, this paradigm serves as the ideal setup to isolate the impacts of different potential social sources of bias on group decision-making using one set-up. Experiment 1 tests how participants incorporate their own evidence with their partner's testimony in different stages of decision-making (expressing and revising testimony). Experiment 2 pilot tests a more fine-grained version of this study by varying evidence strength parametrically rather than categorically. All experimental procedures and analytic approaches were preregistered: https://aspredicted.org/bw4zg7.pdf.

Experiment 1

Methods

Participants 107 adults (53 women, 54 men) participated in this study online on Prolific. Following preregistered exclusion guidelines, an additional 3 participants were excluded due to low self-reported attention.

Procedure Participants worked with a simulated agent that was presented as another participant (i.e., their partner) to guess which of two urns—a Majority-Red or Majority-Blue urn—was being drawn from (Fig. 1A). They were told that each participant would receive a different sample of 10 marbles, which provided *Weak*, *Moderate*, or *Strong* evidence about which urn it was drawn from (Fig. 1B). Participants then provided an initial prediction (Red/Blue) and confidence rating (High/Low; Fig. 1C) and were given the opportunity to revise their prediction after viewing the simulated agent's response, which varied by (a) whether it matched the participant's response, and (b) the simulated agent's level of confidence.

Participants completed 24 trials, counterbalancing evidence strength (Strong/Moderate/Weak), majority marble color (Red/Blue), partner response (Congruent/Incongruent), and partner confidence (High/Low).

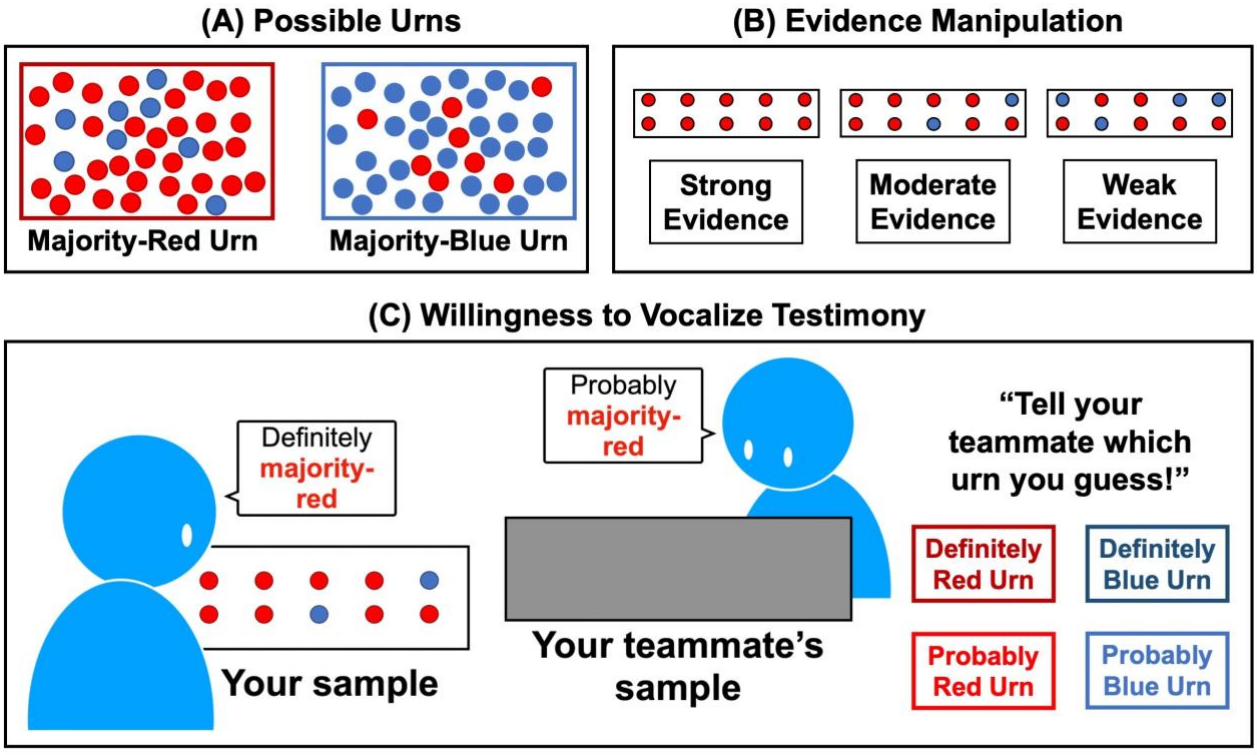

Figure 1: Experiment 1 procedure. (A) Samples were drawn from a *Majority-Red* or *Majority-Blue* urn. (B) On each trial, participants received strong, moderate, or weak evidence about which urn the marbles were drawn from. (C) Participants were asked to infer which urn the marbles were drawn from and indicate their confidence, both before and after seeing their partner's response.

Results

Does gender affect willingness to share testimony? Using the *lme4* package in RStudio, we fit a preregistered model that predicted participants' initial confidence ratings based on fixed effects of participant gender, evidence strength, and their interactions, and a random effect of participant ID.

As expected, participants expressed higher confidence overall when their evidence was stronger (main effect of evidence strength: $\chi^2(1) = 1507.0$, $p < .001$).

More central to our key research question, this effect was further qualified by a predicted two-way interaction between participant gender and evidence strength ($\chi^2(1) = 7.01$, $p = .030$; see Fig. 2). Additional post-hoc analyses indicated that men expressed marginally higher confidence than women when they received moderately strong evidence ($b = 1.00$, $SE = .58$, $z = 1.73$, $p = .084$), but no gender differences emerged when the evidence was strong or weak ($ps > .35$).

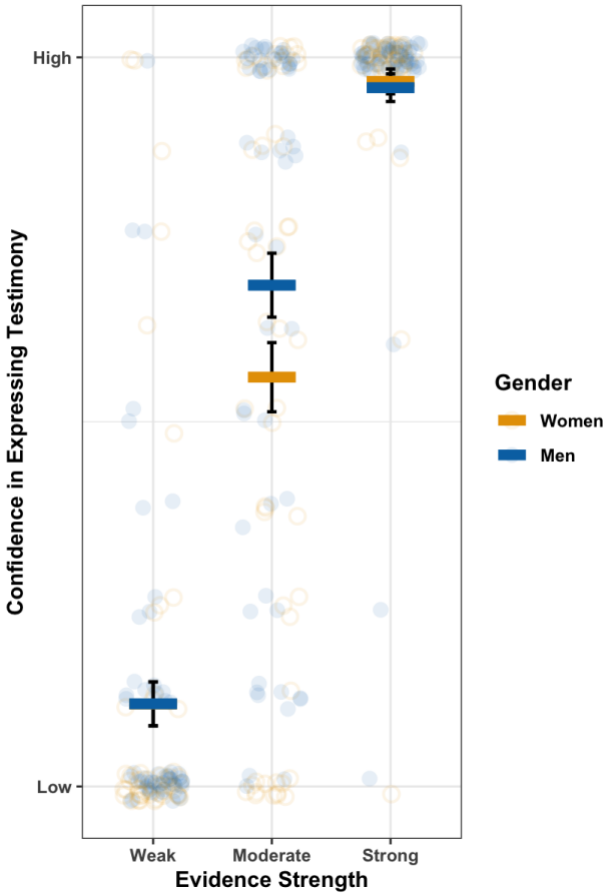

Figure 2: Participant confidence when expressing testimony by participant gender and evidence strength. Small shapes reflect individual averages, and large lines reflect group averages. Error bars reflect 95% CIs.

Does gender affect belief revision based on others' testimony? Using the *lme4* package in RStudio, we fit two preregistered models that predicted (a) whether participants revised their testimony, and (b) their degree of confidence change[1] based on fixed effects of participant gender, evidence strength, partner confidence levels, and their interactions, and a random effect of participant ID.

For the binary outcome of whether participants changed their testimony, the model again revealed a significant interaction effect between participant gender and evidence strength ($\chi^2(1) = 8.39$, $p = .015$), where women were more likely to revise their initial testimony when their own evidence was moderately strong ($b = 1.34$, $SE = .59$, $z = 2.28$, $p = .023$), but no gender differences emerged when the evidence was strong or weak ($ps > .10$; see Fig. 3).

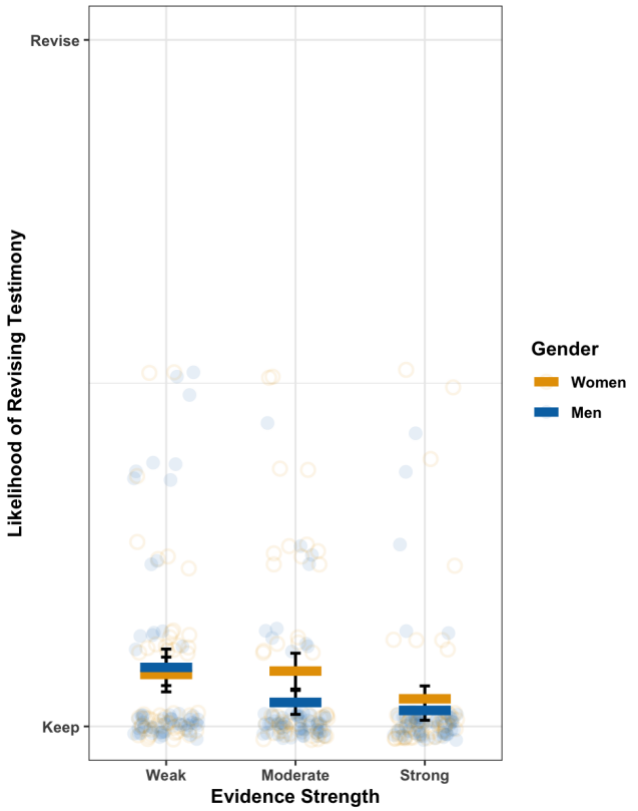

Figure 3: Whether participants revise their testimony after observing their partner's testimony by participant gender and evidence strength. Small shapes reflect individual averages, and large lines reflect group averages. Error bars reflect 95% CIs.

Beyond the main hypotheses, participants also showed evidence of reasonable patterns of belief updating. The model revealed main effects of evidence strength (where participants were less likely to revise their testimony with

[1] The difference between their revised testimony (if applicable) and original testimony. For example, a change from Definitely Red to Definitely Blue indicates a degree of confidence change of 3 scale points, while no revision indicates a degree of confidence change of 0.

stronger evidence; $\chi^2(1) = 19.23$, $p < .001$) and partner confidence level (where participants were more likely to revise their testimony when their teammate indicated higher confidence; $\chi^2(1) = 11.02$, $p < .001$). These main effects were further qualified by a two-way interaction between evidence strength and partner confidence level ($\chi^2(1) = 7.39$, $p = .025$), where participants were more likely to revise their testimony when their teammate indicated higher confidence for weak ($b = .58$, $SE = .29$, $z = 2.03$, $p = .042$) and moderate ($b = 1.33$, $SE = .39$, $z = 3.44$, $p < .001$) evidence strengths, but not when they received strong evidence themselves ($p > .80$).

For participants' graded degree of confidence change, the model again revealed a significant interaction between participant gender and evidence strength ($\chi^2(1) = 6.48$, $p = .039$), where women indicated a larger confidence change in their testimony when their own evidence was moderately strong ($b = .07$, $SE = .03$, $t = 2.28$, $p = .024$), but not when the evidence was strong or weak ($ps > .15$).

Again, separate from the main analyses, the model also revealed main effects of evidence strength (where participants indicated lower levels of confidence change with stronger evidence; $\chi^2(1) = 15.08$, $p < .001$) and partner confidence level (where participants indicated higher levels of confidence change when their teammate indicated higher confidence; $\chi^2(1) = 13.90$, $p < .001$). These main effects were further qualified by a two-way interaction between evidence strength and partner confidence level ($\chi^2(1) = 7.21$, $p = .027$), where participants indicated lower levels of confidence change when their teammate indicated higher confidence for weak ($b = .05$, $SE = .02$, $t = 3.02$, $p = .003$) and moderate ($b = .06$, $SE = .02$, $t = 3.47$, $p < .001$) evidence strengths, but not when they received strong evidence themselves ($p > .95$).

Discussion

These data suggest that testimonial injustice in collaborative decision-making arises from two mechanisms: discrepancies in people's (a) confidence to express their own testimony and (b) willingness to revise their testimony. Across both these factors, we found that compared to men, women were marginally less confident in expressing their testimony and significantly more willing to revise their testimony when their own evidence was moderately strong (not when it was definitively strong or weak). However, the current experiment used a coarse manipulation of evidence strength. In order to more precisely estimate this potential gender difference, we next ran a more fine-grained version of the present experiment where evidence strength varied parametrically.

Experiment 2 (Pilot Study)

Methods

Participants 40 adults (20 women, 20 men) participated in this pilot study online on Prolific.

Procedure The experimental procedure was identical to Experiment 1, except that evidence strength varied parametrically rather than categorically, so participants completed 40 trials instead of 24, which were counterbalanced by evidence strength (1-10 red marbles), majority marble color (Red/Blue), partner response (Congruent/Incongruent), and partner confidence (High/Low). Evidence strength was coded as majority color proportion, which varied between 50% (i.e., 5 red and blue marbles) and 100% (i.e., 10 red marbles).

The following analyses use preliminary pilot data and are presented for exploratory purposes only; they were not preregistered.

Results

Does gender affect willingness to share testimony? Using the *stats* package in RStudio, we fit a model that predicted participants' initial confidence ratings based on fixed effects of participant gender, evidence strength, and their interactions.

Again, as expected, participants expressed higher confidence overall when their evidence was stronger (main effect of evidence strength: $\chi^2(1) = 1036.9$, $p < .001$).

More central to our key research question, the model again revealed a two-way interaction between participant gender and evidence strength ($\chi^2(1) = 12.19$, $p = .032$; see Fig. 4). Additional post-hoc analyses indicated that men expressed higher confidence than women only for samples where the majority color proportion was 90% ($b = 1.35$, $SE = .45$, $z = 3.03$, $p = .002$), but not others ($ps > .20$).

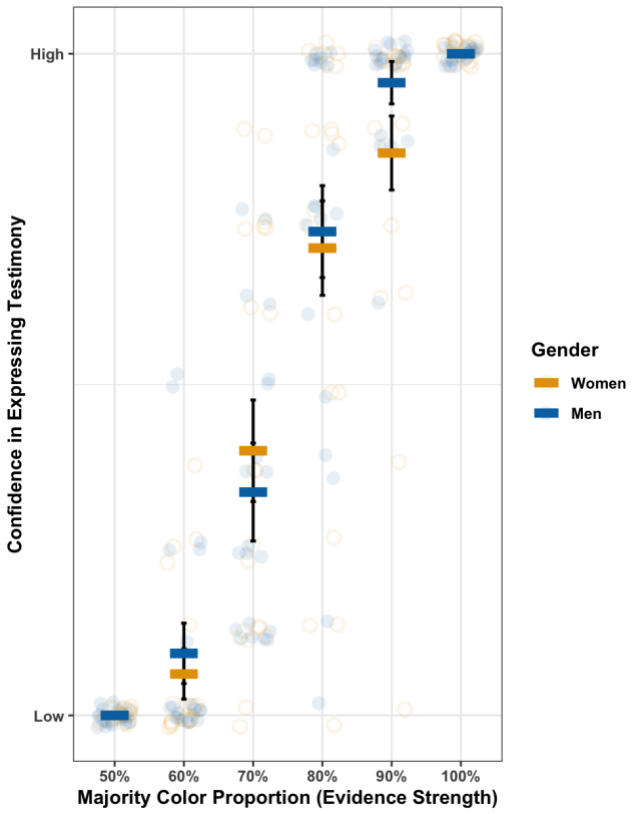

Figure 4: Participant confidence when expressing testimony by participant gender and the majority color proportion of the sample. Small shapes reflect individual averages, and large lines reflect group averages. Error bars reflect 95% CIs.

Does gender affect belief revision based on others' testimony? Using the *stats* package in RStudio, we fit two models that predicted (a) whether participants revised their testimony, and (b) their degree of confidence change based on fixed effects of participant gender, evidence strength, partner confidence levels, and their interactions.

For the binary outcome of whether participants changed their testimony, the model again revealed a significant interaction effect between participant gender and evidence strength ($\chi^2(1) = 14.57$, $p = .012$), showing that women were more likely to revise their initial testimony for samples where the majority color proportion was 90% ($b = 1.75$, $SE = .56$, $z = 3.12$, $p = .002$), but that no gender differences emerged for other samples ($ps > .30$; see Fig. 5).

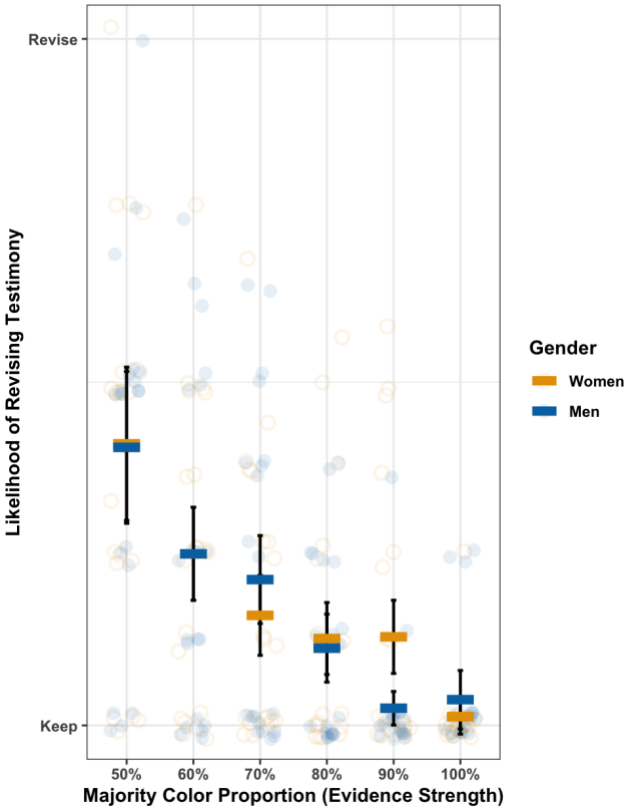

Figure 5: Whether participants revise their testimony after observing their partner's testimony by participant gender and the majority color proportion of the sample. Small shapes reflect individual averages, and large lines reflect group averages. Error bars reflect 95% CIs.

Separate from the gender-related analyses, the model revealed main effects of evidence strength (where participants were less likely to revise their testimony with higher majority color proportions; $\chi^2(1) = 133.22$, $p < .001$) and partner confidence level (where participants were more likely to revise their testimony when their teammate indicated higher confidence; $\chi^2(1) = 64.02$, $p < .001$). These main effects were again qualified by a two-way interaction between evidence strength and partner confidence level ($\chi^2(1) = 16.05$, $p = .007$), where participants were more likely to revise their testimony when their teammate indicated higher confidence for samples where the majority color proportion was 60% ($b = 1.89$, $SE = .32$, $z = 5.99$, $p < .001$), 70% ($b = 1.51$, $SE = .34$, $z = 4.44$, $p < .001$), and 80% ($b = 1.24$, $SE = .41$, $z = 3.02$, $p = .003$), but not others ($ps > .15$).

For participants' graded degree of confidence change, the model again revealed a significant interaction between participant gender and evidence strength ($\chi^2(1) = 11.15$, $p = .048$), where women indicated a larger confidence change in their testimony when their own evidence contained 90% of the majority color marbles ($b = .15$, $SE = .05$, $t = 3.06$, $p = .002$), but not for the other marble distributions ($ps > .15$).

Again, regarding the secondary analyses, the model revealed main effects of evidence strength (where participants indicated lower levels of confidence change with stronger evidence; $\chi^2(1) = 107.51$, $p < .001$) and partner confidence level (where participants indicated higher levels of confidence change when their teammate indicated higher confidence; $\chi^2(1) = 60.59$, $p < .001$). These main effects were further qualified by a two-way interaction between evidence strength and partner confidence level ($\chi^2(1) = 34.35$, $p < .001$). Participants revised their initial judgment in response to a high-confidence judgment from a partner when their own evidence was weak (i.e., 80% or fewer majority color marbles; $ps < .05$), but not when their own evidence was strong (i.e., 90% or more majority color marbles; $ps > .50$).

Discussion

These preliminary pilot data provide additional preliminary support for the possibility that testimonial injustice arises from both discrepancies in people's confidence to express their own testimony in the first place, as well as their willingness to revise their testimony when incorporating others' input.

Furthermore, consistent with the findings from Experiment 1, this more fine-grained version of the study provided preliminary evidence that these gender discrepancies are most robust when participants' own evidence is strong but not definitive. Along with the findings from Experiment 1, this result holds important theoretical and practical implications, as the evidence people encounter in science, law, medicine, and other decision-making contexts is also most often moderately strong but not definitively conclusive (Guyatt et al., 2011; Hozo et al., 2008). Together, these data support the possibility that testimonial injustice arises in these contexts not because women's evidence is weaker than men's, but because women are less confident in expressing their judgments and more willing to revise them even when their own evidence is relatively strong.

General Discussion

Successful collaboration requires integrating the perspectives of team members with different—and often unique—information. Yet, even in well-intentioned groups, some perspectives may fail to shape collective decisions, leading teams to overlook valuable information and miss important insights. Here, we found that, even when accessing reasonably strong evidence, women were consistently less confident in *expressing* their testimony and more likely to *change* their initial judgment when considering others' input.

Importantly, these discrepancies were not uniform across levels of evidence strength. Specifically, women and men were equally confident in their testimony when they had access to definitive or inconclusive evidence, but women were less confident than men when their own evidence was moderately strong (Experiment 1). To further clarify the specific conditions that give rise to these discrepancies, we piloted a more fine-grained manipulation of evidence strength, which provided preliminary evidence that such gender discrepancies are most robust when evidence is highly diagnostic (Experiment 2).

We observed similar discrepancies in women and men's willingness to revise their initial testimony based on input from a partner. Specifically, these discrepancies were most pronounced when evidence was moderately strong (Experiment 1) and highly diagnostic (Experiment 2). Together, these results suggest that testimonial injustice is most likely to emerge when team members are making judgments based on reasonably strong yet not definitive information, as is often the case in real-world scientific and legal contexts (Guyatt et al., 2011). Together, these findings shed light on both the cognitive mechanisms and evidentiary conditions that give rise to testimonial injustice.

Why did gender discrepancies emerge when participants had access to reasonably strong evidence? One possibility is that there might be gender discrepancies in epistemic standards (i.e., what evidence "counts" as strong *enough* to shape beliefs). For example, when assessing performance, women consistently report higher evidence thresholds than men, requiring stronger evidence to rate their performance positively, feel confident about their abilities, and pursue advanced opportunities (Abraham et al., 2024; Beyer & Bowden, 1997; Cimpian et al., 2020; Marshman et al., 2018). Indeed, in the pilot data from Experiment 2, men reported comparable levels of confidence and likelihood to revise after observing samples containing 90% and 100% of the majority color marbles, while women's scores at 90% were closer to those at 80%. Although these preliminary observations should be considered cautiously, this pattern is consistent with the possibility that women apply higher epistemic standards when considering whether the evidence they have in hand is strong enough to warrant absolute confidence.

Relatedly, women also report lower confidence in general than men, with men exhibiting overconfidence in a variety of domains, including for academic assessments, medical knowledge, and investment decisions (Bandiera et al., 2022; Bengtsson et al., 2005; Lundeberg et al., 1994; Vajapey et al., 2020). We propose that these gender discrepancies in confidence may be amplified in moderate-to-strong evidence conditions due to differences in the perceived costs of error (Byrnes et al., 1999; Croson & Gneezy, 2009; Eckel & Grossman, 2008). From a utility calculus perspective (Jara-Ettinger et al., 2016), decision-makers weigh the cost of expressing confidence and being wrong against the cost of expressing uncertainty and potentially revising their judgment. When evidence is weak, caution is warranted for all decision-makers regardless of gender; when evidence is definitive, the risk of error is minimal for everyone. However, when evidence is strong but not conclusive, the tradeoff between asserting and risking error is more salient. It is precisely in this ambiguous middle ground—where confidence could be justified but error remains possible—that gender differences in the anticipated social costs of overconfidence may have the strongest influence on collaborative decision-making.

Another possibility, not mutually exclusive from the others, is that women may place greater emphasis than men on the social context in which the task occurs. For example, women on average have more interdependent self-construals than men (Cross & Madson, 1997; Guimond et al., 2006), which might cause them to be more responsive to others' input during group decision-making processes, particularly under conditions of ambiguity. This heightened sensitivity to social considerations is likely amplified when participants are certain about the *direction* of their belief (e.g., choose the majority-red or majority-blue urn), but there remains uncertainty about how *strongly* to assert that belief (e.g., indicate high or low confidence) in a highly social context.

Given these potential causes underlying the interactive effects of gender and evidence strength on participants' expression and revision of testimony, future research that explores the relative weight of these causes more directly would be fruitful for identifying specific points of intervention. In addition, by comparing these data to those of an ideal observer model, future modeling efforts can also inform whether participants are appropriately calibrated to the evidence that they receive (e.g., if gender discrepancies in testimony expression exist because men are overconfident or because women are underconfident).

Furthermore, aside from gender discrepancies in the expression and revision of testimony, future implementations of the present setup can also shed light on additional processes that give rise to testimonial injustice in group settings. For example, a multiplayer version of the present setup can directly test the possibility that testimonial injustice also arises via discrepancies in not just internalized beliefs about one's own social identities, but also how these identities are perceived by others. These future iterations will complement the current research in characterizing a more comprehensive picture of the various root causes of testimonial injustice. Until then, this research serves as a first empirical step towards understanding the specific conditions that give rise to testimonial injustice and how it might be effectively mitigated.

Acknowledgments

We thank members of the Princeton CoLab for their helpful feedback and discussion.

References

Abraham, L., Hallermeier, J., & Stein, A. (2024). Words matter: Experimental evidence from job applications. *Journal of Economic Behavior & Organization*, *225*, 348–391. https://doi.org/10.1016/j.jebo.2024.06.013

Babcock, L., Recalde, M. P., Vesterlund, L., & Weingart, L. (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, *107*(3), 714–747.

Bajorek, J. P. (2019). Voice recognition still has significant race and gender biases. *Harvard Business Review*, *10*, 1–4.

Bandiera, O., Parekh, N., Petrongolo, B., & Rao, M. (2022). Men are from Mars, and women too: a Bayesian meta-analysis of overconfidence experiments. *Economica*, *89*, S38–S70. https://doi.org/10.1111/ecca.12407

Bengtsson, C., Persson, M., & Willenhag, P. (2005). Gender and overconfidence. *Economics Letters, 86*(2), 199–203. https://doi.org/10.1016/j.econlet.2004.07.012

Beyer, S., & Bowden, E. M. (1997). Gender differences in seff-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, *23*(2), 157–172. https://doi.org/10.1177/0146167297232005

Brescoll, V. L. (2011). Who takes the floor and why: Gender, power, and volubility in organizations. *Administrative Science Quarterly*, *56*(4), 622–641. https://doi.org/10.1177/0001839212439994

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin, 125*(3), 367–383. https://doi.org/10.1037/0033-2909.125.3.367

Campbell-Meiklejohn, D., Simonsen, A., Frith, C. D., & Daw, N. D. (2017). Independent neural computation of value from other people's confidence. *Journal of Neuroscience*, *37*(3), 673–684. https://doi.org/10.1523/JNEUROSCI.4490-15.2016

Carli, L. L. (1989). Gender differences in interaction style and influence. *Journal of Personality and Social Psychology, 56*(4), 565–576. https://doi.org/10.1037/0022-3514.56.4.565

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science, 368*(6497), 1317–1319. https://doi.org/10.1126/science.aba7377

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474. https://doi.org/10.1257/jel.47.2.448

Cross, S. E., & Madson, L. (1997). Models of the self: Self-construals and gender. *Psychological Bulletin, 122*(1), 5–37. https://doi.org/10.1037/0033-2909.122.1.5

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Eagly, A. H. (1983). Gender and social influence: A social psychological analysis. *American Psychologist, 38*(9), 971–981. https://doi.org/10.1037/0003-066X.38.9.971

Eagly, A. H., & Chrvala, C. (1986). Sex differences in conformity: Status and gender role interpretations. *Psychology of Women Quarterly*, *10*(3), 203–220. https://doi.org/10.1111/j.1471-6402.1986.tb00747.x

Eagly, A. H., Wood, W., & Fishbaugh, L. (1981). Sex differences in conformity: Surveillance by the group as a determinant of male nonconformity. *Journal of Personality and Social Psychology, 40*(2), 384–394. https://doi.org/10.1037/0022-3514.40.2.384

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of Experimental Economics Results*, *1*, 1061–1073. https://doi.org/10.1016/S1574-0722(07)00113-8

George, J., Rapsomaniki, E., Pujades-Rodriguez, M., Shah, A. D., Denaxas, S., Herrett, E., ... & Hemingway, H. (2015). How does cardiovascular disease first present in women and men? Incidence of 12 cardiovascular diseases in a contemporary cohort of 1 937 360 people. *Circulation*, *132*(14), 1320–1328. https://doi.org/10.1161/CIRCULATIONAHA.114.013797

Guimond, S., Chatard, A., Martinot, D., Crisp, R. J., & Redersdorff, S. (2006). Social comparison, self-stereotyping, and gender differences in self-construals. *Journal of Personality and Social Psychology, 90*(2), 221–242. https://doi.org/10.1037/0022-3514.90.2.221

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., ... & Schünemann, H. J. (2011). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, *64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Heilman, M. E., & Haynes, M. C. (2005). No Credit Where Credit Is Due: Attributional Rationalization of Women's Success in Male-Female Teams. *Journal of Applied Psychology, 90*(5), 905–916. https://doi.org/10.1037/0021-9010.90.5.905

Hozo, I., Schell, M. J., & Djulbegovic, B. (2008). Decision-making when data and inferences are not conclusive: risk-benefit and acceptable regret approach. In *Seminars in hematology*, 45(3), 150–159. https://doi.org/10.1053/j.seminhematol.2008.04.006

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Lundeberg, M. A., Fox, P. W., & Punćcohaŕ, J. (1994). Highly confident but wrong: Gender differences and similarities in confidence judgments. *Journal of*

Educational Psychology, 86(1), 114–121. https://doi.org/10.1037/0022-0663.86.1.114

Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., & Singh, C. (2018). Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm? *Physical Review Physics Education Research, 14*(2), 020123. https://doi.org/10.1103/PhysRevPhysEducRes.14.020123

Martin, C. L., & Ruble, D. (2004). Children's search for gender cues: Cognitive perspectives on gender development. *Current Directions in Psychological Science*, *13*(2), 67–70. https://doi.org/10.1111/j.0963-7214.2004.00276.x

Mead, G. H. (2012). Symbolic interactionism. *In A First Look at Communication Theory*, pp. 54–66. New York: McGraw-Hill.

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination*. Oxford University Press.

Mosca, L., Benjamin, E. J., Berra, K., Bezanson, J. L., Dolor, R. J., Lloyd-Jones, D. M., ... & Wenger, N. K. (2011). Effectiveness-based guidelines for the prevention of cardiovascular disease in women—2011 update: a guideline from the American Heart Association. *Journal of the American College of Cardiology*, *57*(12), 1404–1423. https://doi.org/10.1161/CIR.0b013e31820faaf8

Nelson, L. K., Brewer, A., Mueller, A. S., O'Connor, D. M., Dayal, A., & Arora, V. M. (2023). Taking the time: The implications of workplace assessment for organizational gender inequality. *American Sociological Review*, *88*(4), 627–655. https://doi.org/10.1177/00031224231184264

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General, 152*(2), 346–362. https://doi.org/10.1037/xge0001277

Ridgeway, C. L. (2009). Framed before we know it: How gender shapes social relations. *Gender & Society*, *23*(2), 145–160. https://doi.org/10.1177/0891243208330313

Tice, D. M. (1992). Self-concept change and self-presentation: The looking glass self is also a magnifying glass. *Journal of Personality and Social Psychology, 63*(3), 435–451. https://doi.org/10.1037/0022-3514.63.3.435

Vajapey, S. P., Weber, K. L., & Samora, J. B. (2020). Confidence gap between men and women in medicine: a systematic review. *Current Orthopaedic Practice, 31*(5), 494–502. https://doi.org/10.1097/BCO.0000000000000906

Wu, Y., Merrick, M., & Gweon, H. (2024). Expecting the unexpected: Infants use others' surprise to revise their own expectations. *Open Mind*, *8*, 67–83. https://doi.org/10.1162/opmi_a_00117