Beyond accuracy:

The reputational costs of independent judgment aggregation

Charles A. Dorison, Georgetown University

Bradley DeWees, United States Air Force

Julia A. Minson, Harvard University

Abstract

Most important life decisions are made collaboratively. But how should such collaborations be structured? Prior research dictates that making independent estimates before interaction maximizes judgment accuracy. In the present research, we examine the extent to which these prescriptions carry unintended reputational costs. Across six studies (N = 2,988) and three participant samples, we hypothesized and found that participants who followed an independent process (and thus first generated their own estimate) assessed their collaborators' judgments more negatively than those who evaluated an identical judgment without first generating their own estimate. This effect occurred because the independent process heightened disagreement, which was associated with reputational penalties. Study 1 demonstrated the basic effect. Study 2 revealed that the effect was mitigated when disagreement was extremely low. Study 3 showed that people interpreted disagreement in an egocentric manner: as disagreement increased, others' judgments – but not one's own – were seen as less accurate. Studies 4, 5A, and 5B demonstrated the robustness of the effect in complex decision-making scenarios with both lay people and national security experts. Finally, Study 6 revealed that following an independent judgment process led to negative evaluations of a partner's competence and decreased willingness to collaborate in the future. Our work thus uncovers a novel tension in collaborative judgment between what is often best for accuracy and what is often best for reputation management.

Keywords: Judgment and decision making, collaboration, reputation, accuracy

Statement of limitations

Our research examined whether independent judgment aggregation triggers reputational costs across multiple domains and multiple participant samples. Most importantly, our conclusions are limited because they are based on short interactions among strangers. Future research should examine the potential downstream effect of judgment aggregation on individuals with longstanding relationships. It could be the case that pre-existing relationships or other situational variables (e.g., power dynamics) might impact the results. It could also be the case that our results are dampened when participants have the chance to engage in longer collaboration. We also acknowledge that our current model of collaborative judgment is limited - surely there are other hybrid ways to combine judgments and other outcomes worthy of study. While we tried to focus on the main two judgment aggregation procedures studied in prior literature and relevant interpersonal evaluations, other procedures and outcome variables merit further testing. The analyses on interpersonal evaluations may also need to be replicated with behavioral or more long-term measures. Further, future work is needed to understand how to eliminate such effects. Finally, it is untested whether our results generalize outside of the United States or to other cultural contexts.

Important judgments and decisions in domains ranging from parenting to international diplomacy often rely on the contributions of multiple individuals. How should such collaborations be structured? Prior research on quantitative estimation offers a gold standard: to maximize judgment accuracy, collaborators should make independent assessments and only later combine them with those of other group members, lest social influence cause estimates to assimilate toward each other. Such assimilation would undermine the "wisdom of crowds," decreasing the accuracy of the final estimate (Galton, 1907; Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Minson, Mueller, & Larrick, 2017; Surowiecki, 2004).

A tacit assumption behind this recommendation is that accuracy is the focal (or perhaps only) goal of collaborative judgment and decision making (for a review, see Gigone & Hastie, 1997). And yet, such a narrow focus on accuracy might overlook other important outcomes. Decision makers care deeply about how they are perceived by others (for reviews, see Lerner & Tetlock, 1999; Schlenker & Weigold, 1992; Tetlock, 2000, 2002). A large body of research makes clear that decision makers want to be seen as trustworthy and competent (for reviews, see Baumeister & Leary, 1995; Goffman, 1959; Lerner & Tetlock, 1999; Mayer, Davis, & Schoorman, 1995; Schlenker & Weigold, 1992; Tetlock, 2000, 2002). In many situations (e.g., on a date or before a big promotion), people may care more about social evaluations than the accuracy of their judgments. We examine the extent to which prescriptions that maximize judgment accuracy may have unintended consequences for interpersonal evaluations.

Specifically, we test whether the process endorsed by prior research (aggregating independent estimates) leads collaborators to form more negative impressions of each other's inputs and interpersonal characteristics—and ultimately decreases their willingness to work

together in the future. We test this hypothesis across a variety of tasks and with both lay and expert samples.

We also assess the psychological mechanism through which such effects may occur. Specifically, we test whether the effect of producing independent judgments on peer evaluations is driven by systematic differences in the amount of disagreement that different task structures produce. Compared to a judgment process wherein individuals make their own estimates after seeing those of a peer (referred to as "dependent judgments" in prior work), independent judgments (where individuals make their own judgment before seeing anyone else's input) are likely to increase the level of disagreement between collaborators' estimates. Based on prior research, we theorize that individuals may then attribute such disagreement to flawed judgment on the part of their partner rather than the structure of the situation.

Theoretical Background

Offering independent assessments has thus been established as a "best practice" for maximizing the accuracy of collaborative judgments, *provided that* the final product represents an approximately equal weighting of the relevant inputs (Clemen & Winkler, 1986; Einhorn & Hogarth, 1975; Hogarth, 1978; Sniezek & Henry, 1989; Soll & Larrick, 2009). Equal weighting increases the likelihood that individual errors will cancel each other out (Larrick & Soll, 2006; Lorenz et. al., 2011; Soll & Larrick, 2009). Under most conditions, independent judgment aggregation thus outperforms attempts to identify and give priority to more accurate judgments. However, even though it is often the case that collaborators are responsible for both generating judgments and evaluating their quality, few studies have examined how judges evaluate each other's contributions.

One literature that indirectly sheds light on evaluation of peer judgment (rather than solely accuracy) is work on the Judge Advisor System. The key finding in this body of work is that judges underweight the estimates of others relative to their own – and thus also relative to the normatively appropriate benchmark of equal weighting (Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000; for a review, see Bonaccio & Dalal, 2006). Individuals typically adjust approximately 30% of the distance between their own and others' estimates, effectively treating their own judgments as more than twice as accurate as those of peers (Harvey & Fischer, 1997; Soll & Larrick, 2009).

However, the Judge Advisor System paradigm requires an individual to make a tradeoff between own and another person's judgment, thus making it difficult to establish whether people are evaluating their own judgment positively or evaluating another person's judgment negatively. Task order is also fixed: individuals almost always begin by making their own estimates, after which they are exposed to those of a peer and asked to revise their earlier estimate. Even though a few studies have varied task order (e.g., Koehler & Beauregard, 2006; Rader, Soll, & Larrick, 2015; Sniezek & Buckley, 1995; Yaniv & Choshen-Hillel, 2012), none examined the effects on interpersonal evaluations. Thus, questions regarding how the structure of the task might affect interpersonal evaluations remains largely unexplored.

Hypothesis Development

The present manuscript broadens the scope of analysis on collaborative judgment to include reputational consequences. Our predictions regarding the effects of judgment order on evaluations of peer input (and peers themselves) bridge two classic bodies of research in psychology.

First, we draw on work on the phenomenon of anchoring and insufficient adjustment (Tversky & Kahneman, 1974). Estimates under uncertainty are systematically influenced by seemingly irrelevant quantities ("anchors") that are cognitively available at the time of making a judgment (Epley & Gilovich, 2001, 2006; Frederick, Kahneman, & Mochon, 2010; Frederick & Mochon, 2012; Janiszewski & Uy, 2008; Loschelder, Friese, Schaerer, & Galinsky, 2016; Mochon & Frederick, 2013; Simmons, LeBoeuf, & Nelson, 2010). Traditionally, the primary outcome of interest is the extent to which the focal judgment made by the participant assimilates to the anchor presented by the researcher (for important exceptions in the context of negotiations, see Gunia, Swaab, Sivanathan, & Galinsky, 2013; Northcraft & Neale, 1986; Majer, Tröschel, Galinsky, & Loschelder, 2020).

But anchoring may have interpersonal consequences, as well. If peer estimates serve as anchors in a collaborative judgment context, independent estimates will be further apart from each other than estimates made sequentially or "dependently." Thus, decision-makers following the recommended practice of generating independent judgments are likely to experience a greater level of disagreement between their judgments and those of their counterparts, than decision-makers whose judgments assimilate toward each other. We thus theorize that participants who make their own estimates *after* considering the estimate of a peer will make estimates that are closer to that peer's estimate than participants who made independent estimates *before* considering a peer's estimate, with the latter approach resulting in higher levels of disagreement.

Second, we draw on theory and research on "naïve realism" (i.e., the objectivity illusion; Ross, Lepper, & Ward, 2010; Ross, 2018) to predict how individuals will interpret such disagreement. According to this research literature (Robinson, Keltner, & Ward, 1995; Ross et al., 2010; Ross & Ward, 1996; Ross, 2018), people are "naïve realists" who rarely stop to

question the extent to which their perceptions, beliefs, and judgments are shaped by their own cognitive machinery and the social situation in which they find themselves (Pronin, Gilovich, & Ross, 2004).

One important consequence of naïve realism is that people disparage others who disagree with them, judging them to be uninformed, unintelligent, or biased by malevolent motives (for review, see Ross, Lepper, & Ward, 2010). This phenomenon has been demonstrated with respect to political and social views (Kunda, 1990; Pronin, Gilovich, & Ross, 2004), the merit of scientific findings (Kahan, Peters, Dawson, & Slovic, 2017; Kahan et al., 2012; Lord, Ross, & Lepper, 1979), and even matters of taste (Blackman, 2014). Most relevant to the present research, several studies in the domain of quantitative judgment have demonstrated that people take less advice after exposure to estimates that are very different from their own, attributing dissimilarity in estimates to the flawed judgment of others (Liberman, Minson, Bryan, & Ross, 2011; Minson, Liberman, & Ross, 2011).

Drawing on this work, we thus predict that when engaged in collaborative judgment and decision-tasks following an independent rather than dependent process, people will interpret disagreement between estimates as a negative signal about the quality of their counterpart's estimate and their abilities and characteristics more broadly.

Theoretical Aims

Our work extends the research literature in several ways. We contribute to the literature on collaborative judgment and decision-making by highlighting the potentially negative interpersonal consequences of the classic advice to begin collaborative judgment tasks by first rendering independent estimates. Such recommendations may be incomplete in light of the additional consequences of temporal ordering for evaluation of the judgments and the peers who

offer them. We extend this literature to consider a broader suite of consequential interpersonal outcomes that might arise in this context and are of great importance to decision-makers.

Furthermore, prior research on anchoring, a powerful force in individual judgment, has not been extensively considered in the context of interpersonal processes. In the present work, we contribute to a growing body of research tying individual-level cognitive biases with their interpersonal consequences (for related work, see Tenney et al., 2019; Dorison & Heller, 2022; Dorison, Umphres, & Lerner, 2021; Jordan, Hoffman, Nowak, & Rand, 2016; Grossman, Eibach, Koyama, & Sahi, 2020). As such we bridge multiple levels of analysis by demonstrating the interpersonal effects of a robust individual phenomenon.

Research overview

We present the results of six studies examining the effect of task order on evaluation of peer inputs in collaborative judgments. Study 1 tests our basic hypothesis in the context of an everyday judgment. Study 2 examines disagreement as a situational moderator of our effect by experimentally manipulating low vs. high levels of disagreement. Study 3 investigates potentially biased interpretations of disagreement. Study 4 expands beyond the domain of quantitative estimation tasks to a complex medical decision-making scenario. Study 5 conceptually replicates Study 4 in a national security domain with both laypeople and an elite national security sample. Finally, Study 6 tests whether the effects of task order impacts perceptions of competence and willingness to engage in future collaboration.

Study 1

Study 1 provides an initial test of the basic effect. We examined whether independently generating an estimate before evaluating another's estimate changed participants' evaluation of the accuracy of a counterpart's estimate regarding a common consumer topic. We also varied the

level of effort associated with producing an estimate to assess whether any potential effect requires participants to engage in involved, deliberate processing (as would happen in most important contexts) or whether it would also emerge when people offer quick intuitive judgments. Finally, as individual difference moderators, we measured participants' domain expertise and the extent to which the domain was personally important to them.

Method

Open science practices statement. In keeping with best practices for fully reproducible science (Simmons, Nelson, & Simonsohn, 2012), we report all methodological decisions (e.g., determining sample size), manipulations, and measures. Study materials, data, pre-registrations, and analysis scripts are available here: https://osf.io/54ek8/?view_only=6858e03623754c9fa7f27d90fa6ba3d7. Studies 3, 5, and 6 were pre-registered.

Design and Participants. Study 1 employed a 2 (Judgment order: Independent, Dependent) X 2 (Process: Structured process vs. Intuitive estimate) fully between-subjects design. We manipulated judgment order such that participants did or did not generate their own estimates prior to evaluating the focal estimate (Task order: Independent, Dependent). We crossed this factor with the method of generating the estimate, such that participants either followed a structured 7-step process to estimate the quantity in question or provided a quick intuitive estimate. We determined our sample size by doubling the cell size from prior pilot studies to detect any potential interaction between our variables. We recruited 808 volunteers (M_{age} = 44, 59% female) from the Harvard Digital Lab for the Social Sciences (DLABSS), a forum for unpaid volunteers who wish to contribute to social science research. More information about the DLABSS pool is available here: http://dlabss.harvard.edu/about/.

Procedure. In this and all subsequent studies we obtained informed consent at the start of the study procedure. Participants then answered a demographic questionnaire (a precondition of volunteering in the DLABSS subject pool). We then randomly assigned participants either to give their own estimate of the lifetime cost of owning an average-sized dog and then to evaluate another estimate (Independent condition) or to evaluate the target estimate without generating their own (Dependent condition). We also randomly assigned participants either to use (and/or evaluate the result of) a systematic seven-step process for making the estimate (Process condition) or to simply make and/or evaluate an intuitive estimate (Intuition Condition). The instructions from the Process conditions are presented in the Appendix.

Our dependent variable was the participant's evaluation of another estimate, purportedly made by a peer. This target estimate was in fact an expert answer to the estimation task from the University of Pennsylvania School of Veterinary Medicine. In all conditions, participants reported the likelihood that the target answer was within 10% of the truth by checking a point on a Likert scale anchored at "0%; No chance" and "100%: Absolute certainty" with scale points arranged in 5% increments. This evaluation served as the primary dependent variable.

After making their estimates and/or evaluations, we asked a series of exploratory questions measuring whether the participant had domain expertise (i.e., "have you owned a dog in the past 5 years?") and how important this domain of knowledge was to them (i.e., "how important is knowledge about dog ownership to you?"). We recorded responses to the domain-importance question using a 5-point Likert scale anchored at "Not at all" and "Very much."

Results

Independent vs. Dependent. Our key hypothesis was that participants would make more negative evaluations of the target estimate if they had already independently generated their own

estimate. This was indeed the case: participants who generated their own estimate prior to evaluating the target estimate judged that estimate as less likely to be accurate ($M_{independent}$ = 39.04, *SD* = 26.98) than participants who did not generate their own estimate ($M_{dependent}$ = 45.34, *SD* = 26.67), *95% CI* [-10.01, -2.60], *t*(806) = -3.34, *p* < .001). As depicted in Figure 1, the effect of independent vs. dependent judgment aggregation was negative and significant whether participants evaluated a peer who used a structured 7-step process (*95% CI* [-12.25, -1.62], *t*(391) = -2.57, *p* = .011) or made an intuitive judgment (*95% CI* [-10.80, -0.47], *t*(413) = -2.15, *p* = .033). These effects were not moderated by domain expertise or domain importance (*p*s > .39).

Process vs. Intuition. There was a positive main effect of using a process – participants rated the accuracy of estimates derived via the 7-step process as being higher than the accuracy of estimates derived via intuition ($M_{process}$ = 44.10, *SD* = 26.98; $M_{intuition}$ = 40.41, *SD* = 26.92, *95% CI* [0.02, 7.46], *t*(806) = 1.97, *p* = .049). When we regressed target evaluation on judgment order, process, and their interaction, we found that the magnitude of the effect size for judgment order was nearly 70% larger than the effect size for process (d_{order} = -0.24, *95% CI* [-0.37, -0.10], $d_{process}$ = 0.14, *95% CI*[-0.005, 0.29]). However, although the size of the regression coefficient was directionally larger for judgment order (vs. for process), a linear hypothesis test revealed that the regression coefficients were not significantly different from each other (*p* = .33). There was no interaction between the two variables (*95% CI* [-8.70, 6.10], *t*(804) = 0.35, *p* = .730).

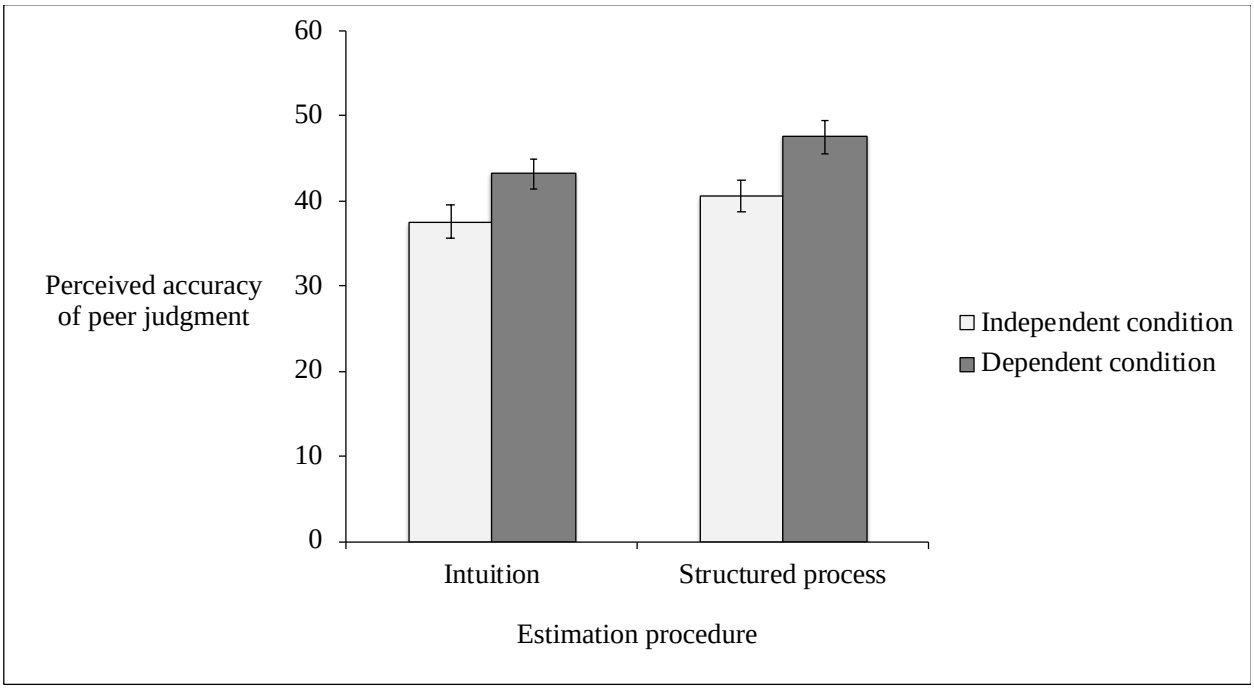

Figure 1. The vertical axis represents the judged likelihood that the target response was within 10% of the correct answer in Study 1. Participants who generated their own estimate prior to evaluating the target estimate judged that estimate as less likely to be accurate. Bars represent standard errors.

The role of disagreement. In line with research on naïve realism, we theorized that heightened disagreement underpinned differences in the evaluations of the target estimate. To provide an initial test of this hypothesis, we examined the relationship between disagreement and participants' evaluation of the target estimate in the Independent conditions (we did not collect estimates from participants in the Dependent conditions in this study). Consistent with our theorizing, participants evaluated the target estimate more negatively as a function of disagreement between their own estimate and the target estimate (*95% CI* [-19.70, -13.93], $t(396) = 11.45$, $p < .001$).[1] We test this hypothesis more stringently in the studies to follow.

[1] Because the distribution of disagreement was not normal, but possessed a long right tail, we also tested a robust regression (Rousseeuw et al., 2015), which again showed a similarly significant result ($p < .001$).

Discussion

Participants who generated their own estimate rated a target estimate as less accurate than participants who rated the exact same target estimate, produced in the same manner, but without having generated an estimate themselves. This effect held for both intuitive "snap" judgments as well as judgments derived through a step-by-step process and was practically meaningful in size.

These results are particularly intriguing, since from the standpoint of any given participant, it might be logically defensible to evaluate a target estimate as a function of how far it deviates from their own. On the other hand, evaluations of the accuracy of *identical information* should not change as a function of task order. In this manner, our results dovetail with classic studies in social psychology documenting the various ways in which people fail to account for "the power of the situation" in their assessments of others' behavior (Ross & Nisbett, 1991).

We did not find evidence that domain importance or expertise moderated the effects of task order on evaluations. However, the amount of disagreement with the target estimate in the independent condition strongly predicted target evaluations. We begin to systematically test the role of disagreement in Study 2 and throughout our remaining studies.

Study 2

Study 2 examines whether the effect of task order on evaluation of peer judgment will be diminished in contexts with extremely low levels of disagreement. We also generalize results by examining a different estimation domain.

Method

Design and Participants. Study 2 employed a 2 (Judgment order: Independent, Dependent) X 2 (Disagreement: High, Low) fully between-subjects design.

We recruited 424 adult participants (M_{age} = 35, 53% female) via Amazon Mechanical Turk (mTurk). We offered participants $0.70 for survey completion. The stimuli in this survey dealt with the costs of childrearing. We directed the study advertisement at parents to recruit participants with some domain expertise and interest in the topic. At the end of the survey, we asked participants whether they were in fact parents, making it clear they would not be penalized for answering truthfully if they were not. 43 participants stated that they were not parents, leaving us with a final sample of 381 (M_{age} = 35, 55% female). Our results remain unchanged when we include data from participants who indicated that they were not parents (see below).

Procedure. Participants first stated how much they "trusted their own judgment" in estimating the costs associated with childrearing. We also asked how knowledgeable participants believed they were on this topic.

Participants then read that fellow survey takers made the same estimate that participants would see on the subsequent screen. Participants in the Independent condition read: "*After* making your own estimates, we would like you to evaluate the likelihood that another participant's estimate is correct" (emphasis not present in survey instructions). Participants in the Dependent condition were told that they would evaluate the likelihood that another participant's estimate was correct "*Before* making your own estimate." Participants in both conditions then went on to make their own estimates and evaluate a target estimate. The order of the tasks was determined by condition assignment.

In order to create more (vs. less) disagreement between the participants and the target estimate they evaluated we asked participants to make judgments of relatively large or small quantities. In the "High Disagreement" condition, participants estimated the average *total* cost of raising a child from birth to age 18. In the "Low Disagreement" condition, participants estimated

the average *monthly* cost of raising a child. For the target response, we used an estimate calculated by CNN Money ($233,610 total or $1,145 monthly; Vasel, 2017). These two versions of the question ensured that, on average, participants making estimates about total costs observed larger discrepancies between their own estimates and the target estimate than participants making the monthly cost estimate. Indeed, the High Disagreement condition produced over 30 times the average disagreement as the Low Disagreement condition ($137,442 vs. $4,388).

To account for the different magnitudes of the quantities being evaluated, we elicited evaluations of the target estimate in the High (Low) Disagreement conditions by asking participants how likely they believed it was that the target estimate was within $20,000 or $100 of the correct answer. We recorded these evaluations on a five-point Likert scale anchored at "Not at all likely" and "Very likely." These evaluations served as our primary dependent variable. As a manipulation check, participants also stated how much they agreed with the target estimate, which we also recorded on a five-point Likert scale anchored at "Did not agree at all" and "Agreed completely." We ended the study by collecting basic demographic data.

Results

In line with our prior results, participants who estimated lifetime childrearing costs (High Disagreement condition) rated their partner's estimate as less likely to be accurate after having made their own estimate ($M_{dependent}$ = 2.88, $M_{independent}$ = 2.53; *95% CI* [-0.68, -0.03], $t(210) = -2.15$, $p = .032$). By contrast, when participants estimated the monthly costs of childrearing (Low Disagreement condition), the effect of previously making an estimate themselves was no longer significant, and in fact in the opposite direction ($M_{dependent}$ = 2.49, $M_{independent}$ = 2.66; *95% CI* [-0.18, 0.53], $t(167) = 0.935$, $p = .351$). As a result, we observed a statistically significant interaction between judgment order and level of disagreement (*95% CI* [-1.01, -0.04], $t(377) = -$

2.14, p = .033), suggesting that the level of disagreement moderated the effect of task order. Results are presented in Figure 2. When we include all participants (not just those who indicated they were parents), the identical pattern of results emerged (interaction: *95% CI* [-0.96, -0.05], $t(420) = -2.18$, p = .030). We also conceptually replicate the results using the self-reported measure of disagreement, although here the interaction fails to reach traditional levels of significance (interaction: *95% CI* [-0.89, 0.04], $t(377) = -1.78$, p = .075). Taken together, the results provided strong evidence that while following an independent process led to negative evaluations of peer judgment when disagreement was high, this effect was mitigated in cases where disagreement was low.

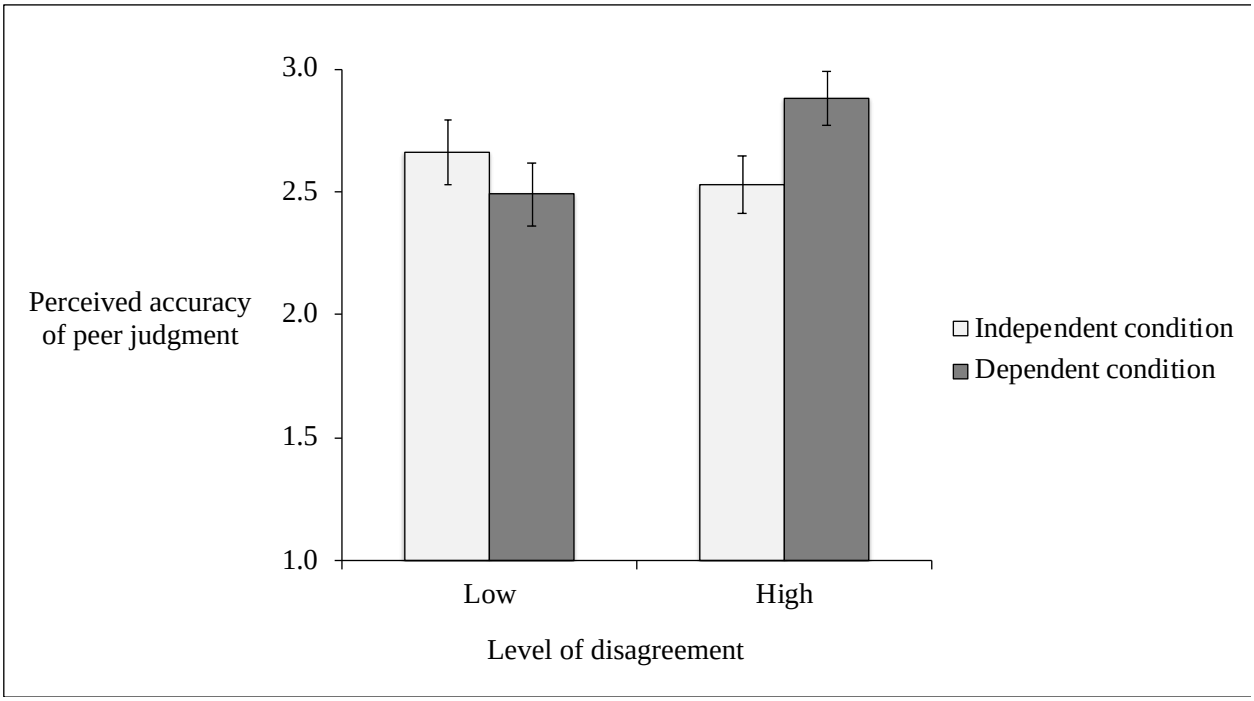

Figure 2. Judged likelihood that the target response was within 10% of the correct answer, measured using a five-point Likert scale (vertical axis). In the high disagreement condition, task order influenced evaluations. In the low disagreement condition, it did not. Bars represent standard errors.

Discussion

Study 2 demonstrates that the effect of judgment order emerged only for estimates where participants were likely to disagree with the estimate they were evaluating. The results provide

process evidence for the role of disagreement through a "manipulate the mediator" approach (Pirlott & MacKinnon, 2016). As in Study 1, participants seemed largely unaware of the situational features such as task structure or the magnitude of the quantity they were estimating on the discrepancy between their own estimates and those they were evaluating. Instead, they used this level of discrepancy as a valid signal on which to base their evaluations.

Study 3

Study 3 directly examines whether participants interpret disagreement in a biased manner. It may be the case, that when encountering disagreement, participants treat it as a signal that this estimation task must be particularly challenging, and thus any estimate is likely to be far from the truth. This logic however, would dictate that individuals evaluate their own judgments more negatively as disagreement increases, as well. If, instead, only peer evaluations suffer in the presence of disagreement, we can conclude the presence of naïve realism.

Method

Design and Participants. All participants generated their own judgment before viewing a judgment generated by a peer. Our design employed three independent variables. First, within-subjects, participants evaluated their own judgment and the other's judgment (Focus of Judgment: Self, Other). Second, between-subjects, we counterbalanced the order in which participants evaluated their own and another's judgment. To these factors we added a third, approximately continuous treatment for disagreement. We created this variable by randomly selecting an estimate from a previous pool and calculating the absolute value of the difference between this estimate and the participant's own estimate (more details follow in the Procedure section).

We pre-registered a sample size of 400 prior to exclusions. We recruited 401 participants through mTurk (M_{age} = 38, 50% female). Participants received $0.30 for completing the study and had the possibility of receiving a $0.50 bonus if their estimate fell within 10% of the correct answer. After implementing our pre-registered exclusion criteria, our final sample consisted of 302 participants (M_{age} = 38, 53% female).

Procedure. After informed consent, we told participants that they would estimate the number of M&Ms in a jar and that they would receive a $0.50 bonus if their estimate was within 10% of the truth. We administered three basic comprehension questions based on these instructions (e.g., "How many estimates will you make?"). Following these questions, participants read that they would see the estimate of another participant. Participants then read: "After you see the estimate of the other participant, you will evaluate the likelihood that your [own estimate/the other's estimate] AND [the other's estimate/your own estimate] is correct." We counterbalanced the order in which participants evaluated their own versus the other participant's estimate. Participants then estimated the number of M&Ms in a clear container.

On the next screen, participants saw a reminder of their own estimate and the estimate of another participant (the order of presentation was again counterbalanced). We built the pool of others' estimates (664 in total) by culling the middle 80% of estimates from pilot studies that had used the same stimulus. Participants then evaluated their own judgment and the judgment of the other (or the reverse order) by indicating how likely it was that the estimate was within 10% of the correct answer. We recorded responses on a five-point Likert scale anchored at "Not at all likely" and "Very likely." The next screen concluded the study with questions regarding the participant's gender and age.

Results

Overall, participants evaluated their own estimates as more likely to be accurate than those offered by others (*95% CI* [0.39, 0.67], $t(301) = 7.37$, $p < .001$). However, our key question of interest was how judges evaluate both their own and others' estimates after encountering varying levels of disagreement. If judges penalized themselves at the same rate as they penalized others, we would expect to see approximately the same (negative) relationship between disagreement and evaluation for both own and others' judgments. If, however, judges interpreted disagreement in a self-serving way, we would expect to see a negative relationship between disagreement and evaluations of others' judgments but *not* evaluations of own judgments. We would thus observe a significant interaction between focus of judgment (self, other) and amount of disagreement on evaluations of accuracy.

We tested a hierarchical linear model to account for the fact that each participant evaluated two estimates (own and a peer's). In line with our predictions, we observed a significant interaction (*95% CI* [$2.3*10^{-4}$, $6.5*10^{-4}$], $t(300) = 4.16$, $p < .001$; see Figure 3). Specifically, we found a clear negative relationship between level of disagreement and evaluations of others' judgments (*95% CI* [$-7.0*10^{-4}$, $-3.4*10^{-4}$], $t(300) = -5.64$, $p < .001$). However, no relationship emerged between the amount of disagreement and accuracy evaluations of one's own judgments (*95% CI* [$-2.4*10^{-4}$, $0.89*10^{-4}$], $t(300) = -.89$, $p = .376$). Individuals interpreted higher levels of disagreement to mean that their counterpart was incorrect; however, they did not apply the same standard to their own judgments. The order in which participants evaluated their own versus another's judgment had no main effect on evaluations, nor did it interact with other variables. Thus, the results provided evidence that rather than attribute disagreement to the difficulty of the task or to their own inaccuracy, participants attributed it to flawed judgment on the part of a peer.

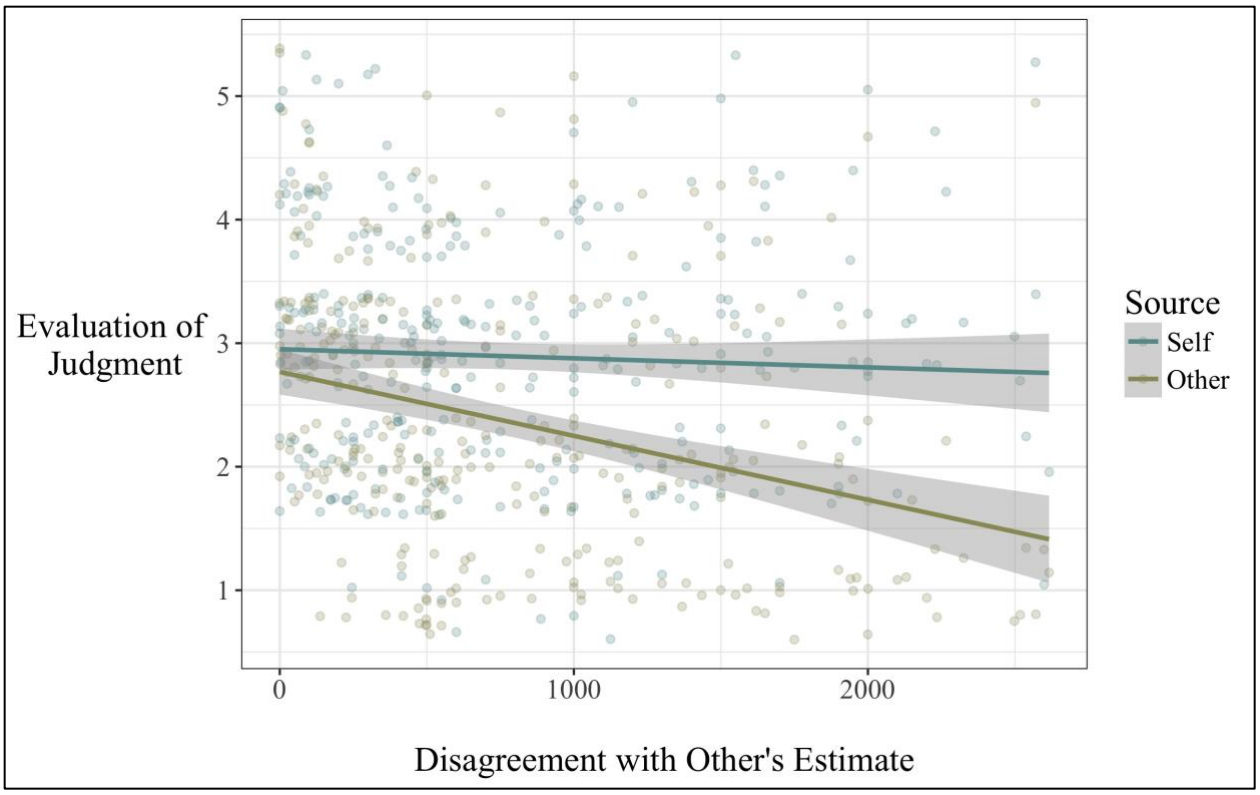

Figure 3: Participants judged the likelihood that either their own (blue) or another's (gold) estimate was correct on a scale that ranged from 1 - 5 (vertical axis; points above and below 1 and 5 are due to jittering the display to allow for visualization of all raw data). As disagreement increased (horizontal axis), the differing slopes suggest that participants interpreted disagreement as a sign that the other's estimate was incorrect, rather than as a sign that both own and other's estimates were less likely to be correct.

Discussion

Study 3 results revealed that the effects of order on evaluations of peer judgments are underpinned by self-serving interpretations of disagreement. In line with research on naïve realism, participants viewed higher levels of disagreement as an indicator that the other's judgment—but not their own—was incorrect.

Study 4

Studies 1-3 provided consistent evidence for an effect of task order on evaluation of peers' quantitative estimates and for a key role of disagreement in driving such effects. Study 4 builds upon the prior studies in two ways. First, Study 4 departs from evaluations of simple quantitative judgments to evaluations of complex decisions with no apparent correct answer.

Rather than measuring participants' evaluations of judgment accuracy, we instead measured participants' evaluations of (1) the overall quality of a particular course of action; and (2) the quality of the reasoning behind it. Second, Study 4 again extends our investigation into the central role of disagreement by examining its role as a statistical mediator.

Method

Design and Participants. We employed a single-factor, two-level (Independent, Dependent) between-subjects design. We aimed for a sample size of 400 and successfully recruited 399 participants through mTurk (M_{age} = 35, 46% female). Our recruitment message told participants that they would "make decisions about ethical dilemmas," and that the survey would involve reading and writing. Compensation for the study was $1.00. We offered bonus payments of $0.50 if the supervisor's favored option matched that of the participant (see below).

Procedure. We adapted materials from "The Kidney Case" (Austen-Smith, Feddersen, Galinsky, & Liljenquist, 2014), a simulation designed for teaching students about biases in ethical decision-making. Participants took on the role of a member of a Kidney Transplant Review Board. Their task was to determine the allocation of one kidney among four deserving candidates (we simplified the task from the eight candidates presented in the original exercise). Each description of the four transplant candidates offered a compelling reason for being selected as the kidney recipient (e.g., one candidate was a veteran, another a single parent, another a philanthropist, etc.). The complete descriptions of the candidates are presented in the Appendix.

We told all participants that their recommendations would be paired with another participant's recommendation in the survey and that an mTurk worker in a future survey would play the role of "supervisor" and evaluate the two recommendations. After evaluating the recommendations, the supervisor would make a recommendation of her/his own. If the

supervisor made the same recommendation as the participant, the participant would receive a $0.50 bonus. Importantly, the bonus did not depend on whether the supervisor agreed with the other mTurker (i.e., the target). We specified that the supervisor would *not* see the participant's evaluations of the target, but only the transplant recommendations provided by both.

Participants' random assignment to condition (Independent vs. Dependent) determined whether they made their own kidney allocation recommendation prior to evaluating the recommendation of another participant. In the Independent condition, participants read the four candidate profiles and then selected a single candidate to receive the kidney. After making their selection, we asked them to write a few sentences to explain their choice.

Independent condition participants then saw the choice ostensibly made by another mTurker along with a brief explanation for that decision. In reality, participants were randomly assigned to see one of the four transplant candidates. This ensured that, by chance, 25% of participants evaluated a target who chose the same candidate as they did and 75% evaluated a target who made a different choice. As a manipulation check, we asked participants to consider how similar the target's answer and reasoning were to their own using a 7-point Likert scale anchored at "Not at all" and "Extremely."

We then asked participants two sets of questions that constituted our main dependent variables. First, we asked a series of four questions evaluating the target's choice in terms of being intelligent, thoughtful, ethical, and moral. These four questions were elicited on 7-point Likert scales anchored at "Strongly disagree" and "Strongly agree." The four items achieved high reliability (Cronbach's $\alpha = .93$) and we thus combined them into a composite rating representing the participants' overall evaluation of the target's choice. Second, we asked

participants a single item regarding whether they would support the target's choice using a 7-point Likert scale anchored at "Strongly disagree" and "Strongly agree."

Participants in the Dependent condition engaged in the same tasks, though in a different order. They viewed the four candidate profiles and then, rather than choosing a candidate of their own, saw the other participant's choice and justification. After viewing the target response, they answered the same questions regarding the similarity of this choice and reasoning to their own; the morality, ethicality, thoughtfulness, and intelligence of the target response; and their willingness to support the target choice. Only after making these evaluations did they select a kidney recipient and provide an explanation for their own decision.

Results

We first examined whether participants in the dependent condition considered the target's answer and reasoning as being more similar to their own than did participants in the independent condition. This was the case for both subjective similarity of the answer ($M_{dependent}$ = 4.60 vs. $M_{independent}$ = 4.07, $t(397) = 2.52$, $p = .012$) and subjective similarity of the reasoning ($M_{dependent}$ = 4.60 vs. $M_{independent}$ = 4.16, $t(397) = 2.11$, $p = .035$).

We next tested our key confirmatory hypothesis: whether participants who generated their own choice prior to evaluating the target's choice evaluated the target's decision more negatively. Analyzing our composite measure of how moral, ethical, intelligent, and thoughtful participants thought the target's choice was revealed lower evaluations in the Independent (M = 4.11, SD = 1.87) than in the Dependent condition (M = 4.51, SD = 1.75; *95% CI* [0.04, 0.75], $t(397) = 2.17$, $p = .031$). The effect of judgment order was negative and statistically significant for three of the four items in the composite: participants in the Independent condition thought that target's choice was less intelligent (*95% CI* [0.06, 0.86], $t(397) = -2.26$, $p = .025$), less

reasonable (*95% CI* [0.04, 0.81], $t(397) = -2.15$, $p = .033$), and less moral (*95% CI* [0.04, 0.82], $t(397) = -2.14$, $p = .033$). The negative direction held when we asked participants to rate how ethical the target's choice was, though the results did not reach significance (*95% CI* [-0.13, 0.65], $t(397) = -1.31$, $p = .19$). Further, participants in the Independent condition were less likely to support the target's preferred transplant choice ($M = 4.21$, $SD = 2.22$) than were participants in the Dependent condition ($M = 4.71$, $SD = 2.15$; *95% CI* [0.08, 0.94], $t(397) = 2.31$, $p = .021$).

A key remaining question was to what extent, if at all, the effect of judgment order (Dependent vs. Independent) on evaluations were underpinned by disagreement, as predicted by theory and research on naïve realism. To begin assessing this question, we first examined whether participants were more likely to agree with their partner in the Dependent condition. In line with prior research on anchoring, this was in fact the case: While participants in the Independent condition agreed with their targets 26.3% of the time (a proportion indistinguishable from chance), 50.5% of Dependent condition participants agreed with their targets, a number highly unlikely to be due to chance given random selection from four possible options ($p < .001$).

Second, we examined whether these different levels of objective agreement mediated the effect of task order on participants' likelihood of endorsing the target's choice and their evaluations of the target. To do so, we tested two mediation models using the Lavaan package in R (Rosseel, 2012). In both models, the independent variable was condition (1 = Independent, 0 = Dependent) and the mediating variable was disagreement (1 = target's choice was different than the participant's, 0 = target's choice was the same as the participant's). Finally, the dependent variable was either the composite evaluation of the choice (Model 1) or support for the choice (Model 2). In both models, we observed a significant indirect effect through disagreement (*b*s = -

0.52 and -0.66, respectively, *p*s < .001), providing evidence consistent with the hypothesis that disagreement with the target's choice underpinned negative evaluations.

Of note, the large and robust indirect effects of task order on evaluations through disagreement were somewhat at odds with the relatively more modest total effects of judgment order on evaluations. An exploratory examination of the underlying data revealed a surprising pattern that helped explain this discrepancy. Among participants who agreed with the target choice (n = 152), participants in the Independent condition reported more positive composite evaluations ($M_{independent}$ = 6.19 vs. $M_{dependent}$ = 5.76, *95% CI*[0.06, 0.80], $t(150) = 2.33$, $p = .021$) and greater support for the target's recommendation ($M_{independent}$ = 6.74 vs. $M_{dependent}$ = 6.33, *95% CI*[0.08, 0.73], $t(150) = 2.47$, $p = .015$) than participants in the Dependent condition. This pattern suggested that the experience of agreement was interpreted and evaluated somewhat differently when people arrived to that agreement independently.

In sum, on the one hand, participants in the Dependent condition were more likely overall to agree with the target (and agreement in general led to more positive evaluations). However, because contingent on agreement, targets were evaluated more positively in the Independent condition, some of the positive effect of the Dependent task order on evaluations was counteracted. We test for replication of this pattern in Study 5 and consider its implications further in the General Discussion.

Discussion

Study 4 extends our investigation from simple quantitative judgments to a complex ethical dilemma. Nevertheless, participants who made their own decision first were less willing to endorse the course of action chosen by another participant and evaluated that same course of action less positively. The effect of task order on evaluations of peers' judgments was driven by

the likelihood that a participant would agree with the target. Participants who made independent decisions disagreed more often; in turn, disagreement underpinned evaluations of the choice itself.

Study 5

Study 5 tests the generalization of our effect in a new domain with both a lay (Study 5A) and elite expert (Study 5B) sample. We report the two studies in parallel, noting only where they significantly diverged in method or result.

Method

Design and Participants. As in Study 4, we employed a single-factor, two-level (Independent, Dependent) design. We aimed for a sample size of 400 for the lay sample and ended up recruiting 402 participants (M_{age} = 38, 44% female). For th expert sample, we pre-registered[2] that we would collect a sample of 500 experts or recruit for one month, whichever came first. In the end, we were able to collect data from 164 participants after one month (M_{age} = 36; 20% female). Recruitment of this expert sample began with the authors' personal and professional contacts in national security, and expanded to include current and former members of the U.S. Department of Defense (military and civilian), members of the Department of State, Congressional staff members, academics with research interests in national security, and staff members of the White House National Security Council. 78% of the sample reported having military experience, with ranks ranging from junior enlisted to brigadier general. Civilians in government included GS-13s, -14s, and -15s, which are individuals at the upper end of the civilian rank scale equivalent to mid- through senior-officer military ranks.

[2] **Note to reviewers**: we accidentally made this pre-registration public. We apologize for the inconvenience. In the pre-registration, we selected aspredicted.org's "it's complicated" option for whether data had already been collected. We selected this option because we had collected data on the lay sample, but not the expert sample. No experts had taken our survey when we pre-registered the study.

Our recruitment message to lay participants stated that they would make decisions in a national security context; our message to the expert population stated that we were conducting "a research project on decision-making in national security environments." In our expert recruitment message, we stressed that we were looking for individuals who had national security experience, and that completing the survey was strictly voluntary.

Compensation for the mTurk study was $1.00 with a possible bonus payment of $0.50. For the expert sample, we offered the possibility of a bonus (a $100 Amazon e-gift card), but specified that the bonus was not guaranteed for completing the survey.

Procedure. We developed the procedure in consultation with members of the National Security Fellows program at the Harvard Kennedy School of Government, a program reserved for individuals at high levels of military command or civilian leadership in national security. Within the constraints of an embedded survey experiment, the scenario was consistent with the limited information, uncertainty, and high stakes inherent in many national security decisions (Snyder, Bruck, & Sapin, 1962).

For both samples, participants gave informed consent and then assumed the role of an operations staff member for the commander of United States military forces in Africa. They read that the commander had recently received intelligence on the location of a threatening terrorist, whom we referred to as "Combatant X." Participants read that their task consisted of four steps: 1) reviewing background information on Combatant X; 2) considering possible courses of action; 3) recommending and explaining a course of action; and 4) evaluating the course of action proposed by another member of the staff. We reversed the third and fourth steps in this process based on condition: participants in the independent condition first recommended a course of action before evaluating a course of action proposed by a peer, whereas participants in the

dependent condition evaluated the course of action proposed by a peer before recommending their own course of action.

After reviewing these initial instructions, participants read the main body of the scenario, which was identical across conditions. The scenario stated that the commander had recently received intelligence on the possible location of Combatant X. The scenario stressed that Combatant X was considered one of America's deadliest enemies. It also stressed, however, that Combatant X's suspected compound was in a heavily populated area, which posed the risk of civilian casualties if U.S. forces were to attack. It was unclear whether the local government was aware of and supporting Combatant X's shelter. This uncertainty posed a difficult diplomatic problem for the United States, which had interests in maintaining good relations with the local government as well as in capturing the terrorist.

At the request of the commander, participants reviewed four decision options. The options, which we presented in randomized order, included "embedding a conspirator," "waiting for movement," "assisting the host nation," and "independently attacking" (see Appendix for complete descriptions of each option). Each option offered compelling reasons for being selected, as well as clear risks in terms of loss of life or diplomatic tensions.

Our treatment occurred after participants reviewed the options. Participants in the Independent condition selected their top option to propose to the commander and then provided an explanation of their choice. Participants then reviewed the recommendation of their purported partner, which consisted of one randomly selected option and a corresponding explanation (which we composed to be similarly persuasive across all the options).

As in Study 4, participants in the Independent condition began with a manipulation check by indicating how similar they perceived the partner's recommendation to be to their own. Then,

they completed the two primary dependent variables. First, they indicated how intelligent, thoughtful, ethical, and moral the partner's recommendation was. Also as in Study 4, these assessments of the partner's recommended option achieved a high level of reliability (lay sample Cronbach's $\alpha = .92$; expert sample: $\alpha = .89$). We thus combined them to form a global measure of a participant's evaluation of the partner's recommendation. Second, participants also stated whether they believed that the partner's chosen option was the "best overall" option on a 7-point Likert scale. Participants in the Dependent condition completed the exact same tasks, only with the order of selecting an option of their own and evaluating the recommendation of another reversed.

We concluded the study with demographic questions. We asked participants in the lay sample to report their age, gender, and political orientation. We asked participants in the expert sample whether and how much experience they had in the national security arena, what rank they had attained in their most recent national security job, their highest level of education, their age, gender, and political orientation. If participants made the same choice as the commander (in actuality, a retired national security professional), they were entered into a raffle for a $100 bonus.

Results

As in prior studies, our key confirmatory hypothesis was that participants in the Independent condition would make more negative evaluations of their peer's judgment than would participants in the Dependent condition. As depicted in Figure 4, and in line with our prior results, this hypothesis was supported: participants' composite evaluations in the lay sample were more negative in the Independent condition as compared to the Dependent condition ($M_{independent} = 4.33$, $SD = 1.70$; $M_{dependent} = 5.01$, $SD = 1.63$; *95% CI*[0.35, 1.00], $t(400) = 4.05$, $p <$

.001). The same pattern emerged among the experts, who were similarly prone to change their evaluation of a decision based on the ordering manipulation (composite rating $M_{independent}$ = 3.59, SD = 1.64 $M_{dependent}$ = 4.13, SD = 1.62; *95% CI*[0.03, 1.05], $t(162) = -2.11$, $p = .037$). Of note, the mean difference in evaluations in the expert sample (0.54) was approximately 80% of the size of the mean difference in evaluations in the lay sample (0.68).

Furthermore, in the lay sample, participants in the Independent condition were less likely to believe that the proposed option of a partner was the best possible option (M = 4.07, SD = 2.07) compared to participants in the Dependent condition (M = 4.91, SD = 1.85; *95% CI*[0.46, 1.22], $t(400) = 4.30$, $p < .001$). In the expert sample, the relationship held directionally, though the difference was not statistically significant (M = 3.97, SD = 2.07, vs. M = 4.26, SD = 1.96; *95% CI*[-0.34, 0.91], $t(162) = -0.91$, $p = .362$). In retrospect, we suspect that this dependent variable – whether the target's response was the best *possible* option – was especially conservative among experts, who would be better able to imagine other possibilities not included in the four that we had presented. For example, in an open-ended text box, one of our expert participants wrote "I only chose this option because it is the most logical of the options you have provided. In reality, I would have chosen none of the options."

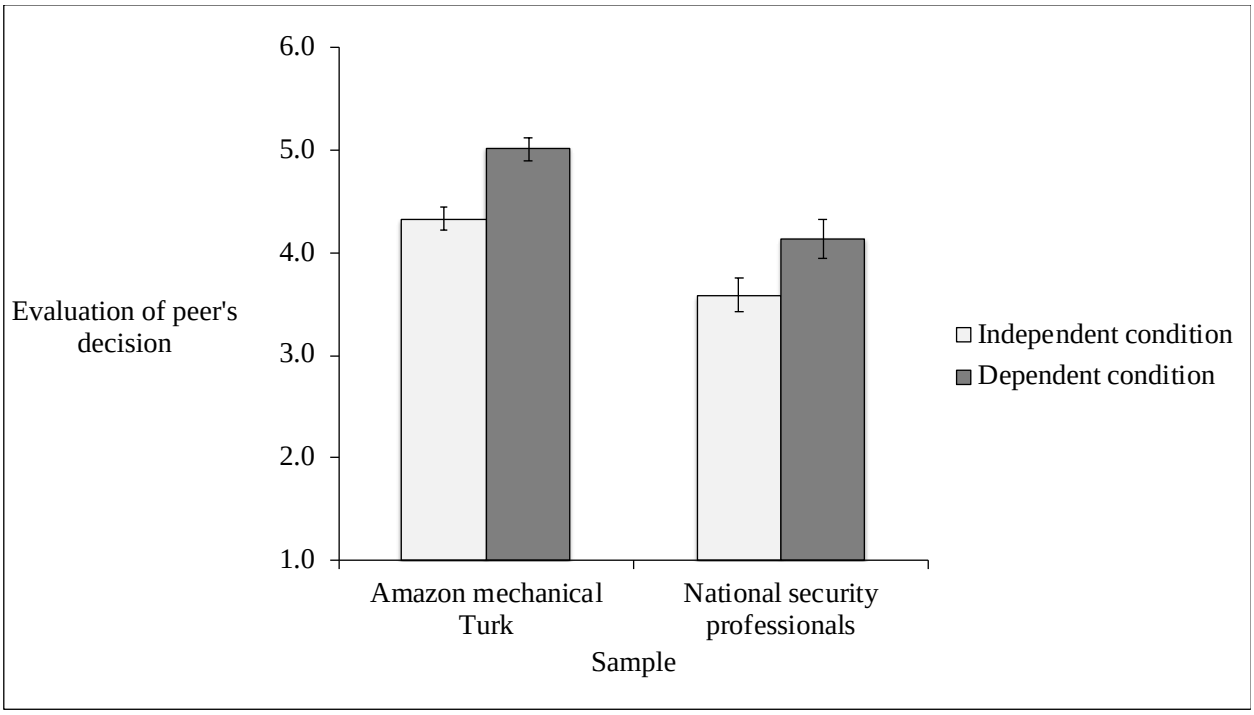

Figure 4. The vertical axis represents the evaluation of the target's response. Participants who generated their own choice prior to evaluating the target choice judged that choice more negatively. Bars represent standard errors.

The role of disagreement. As in Study 4, we next examined to what extent the effect of task order on evaluations were underpinned by disagreement. First, we examined whether participants were more likely to agree with their partner in the dependent condition. In line with prior research on anchoring, this was the case in the lay sample ($M_{independent}$ = 21% vs. $M_{dependent}$ = 47%, $p < .001$) and slightly smaller in size, and only marginally significant, in the expert sample ($M_{independent}$ = 18% vs. $M_{dependent}$ = 30%, $p = .070$).

Next, we examined whether these different levels of objective agreement mediated the effect of task order on participants' likelihood of endorsing the target's choice and general evaluations. To do so, we fit four total mediation models using the Lavaan package in R (Rosseel, 2012). Models 1-2 included the lay sample and Models 3-4 included the expert sample. In all four models, the independent variable was condition (1 = Independent, 0 = Dependent) and

the mediating variable was disagreement (1 = target's choice was different than the participant's, 0 = target's choice was the same as the participant's). Finally, the dependent variables were either the composite evaluation of the target's choice (Models 1 and 3) or whether the target's choice was evaluated to be the best possible option (Models 2 and 4). In the lay sample (Models 1 and 2), we found evidence of a significant indirect effect through disagreement (*b*s = -0.76 and 0.60, respectively, *p*s < .001). In the expert sample, we found a similar pattern of results, although the indirect effects were marginally significant (*b*s = -0.39 and -0.33, respectively, *p*s = .070 and .069).

Building on Study 4, we also found that agreement was evaluated differently in each condition. In the lay sample, we observed a significant interaction between our treatment and agreement on the composite evaluation of the target (*95% CI*[-1.50, -0.38], $t(398) = -3.31$, $p = .001$).[3] Both simple effects of this interaction were significant. Participants in the Independent condition gave higher composite ratings to targets in cases of agreement ($M = 6.59$, $SD = 0.60$) than did participants in the Dependent condition, ($M = 6.01$, $SD = 0.90$; *95% CI*[-0.87, -0.27], $t(138) = -3.76$, $p < .001$). The opposite pattern emerged in cases of disagreement: participants in the Independent condition gave lower evaluations to disagreeing targets ($M = 3.72$, $SD = 1.35$) than did participants in the Dependent condition ($M = 4.09$, $SD = 1.61$; *95% CI*[0.01, 0.73], $t(260) = 2.02$, $p = .044$). In the expert sample, the pattern was in the same direction, but not statistically significant given the smaller sample size for both agreement ($M_{independent} = 5.98$ vs. $M_{dependent} = 5.84$) and for disagreement ($M_{independent} = 3.06$ vs. $M_{dependent} = 3.38$). We discuss this pattern further in the General Discussion.

Discussion

[3] In the lay sample, 140 of 402 participants chose the same answer as the target. In the expert sample, 41 of 164 participants chose the same answer as the target.

For both laypeople and national security experts, independent (vs. dependent) judgment aggregation led to lowered evaluations of the intelligence and morality of the options offered by peers. This effect appeared to be driven by the greater likelihood of disagreement in the Independent condition.

Study 6

Study 6 advances our investigation by testing whether the effect of task order generalizes to evaluations of the peers themselves and has downstream consequences for collaboration. Critically, Study 6 also allows us to directly assess the effect of task order on accuracy.

Method

Design and participants. Study 6 featured a within-subjects design. All participants completed estimates on two topics (described in detail below), one topic for which they made a judgment and then viewed a partner's judgment (Independent task) and one for which they saw a partner's judgment and then offered their own (Dependent task). We counterbalanced both the order of the two estimation topics and the order in which participants completed the Dependent vs. Independent tasks. We hypothesized that participants would evaluate the partner from the Dependent task more positively than the partner from the Independent task, because heightened disagreement associated with the Independent task would be associated with negative interpersonal evaluations of the partner.

We recruited 400 participants via mTurk. Only one participant failed the attention check at the beginning of the survey, leaving us with a total of 399 participants for analysis after pre-registered exclusions (M_{age} = 40, 43% female).

Procedure. After completing the attention check, participants learned that in the study they would work with other mTurkers to estimate the frequency of particular policy opinions

reported by participants in a prior pilot study. In this earlier pilot study, we had asked participants their opinions regarding different policies related to the COVID-19 pandemic. Participants then read that they would estimate the proportion of prior pilot study participants that supported a particular policy. We were able to assess the accuracy of the estimates because we knew the true proportion of prior pilot study participants who reported each opinion.

We further told participants that while completing the estimation tasks, they would also see the estimates of other online participants (i.e., their partners). Critically, we informed them that while they would be assigned two different partners for the first two topics, they would be able to select which partner they wanted to keep for the third and final topic. This design enabled us to test the impact of task order on participants' desire to collaborate with another individual in a realistic, incentivized setting. Participants answered three questions to show that they understood the instructions. They were not allowed to advance in the study until they answered correctly.

On the next page, participants learned that their estimates were incentivized. Specifically, participants read that if any of their own or their partners' estimates were within 10% of the true answer, they would receive a spot in a lottery in which they could win $20 to split with their partner.

Participants then estimated the proportion of prior pilot study participants who had given specific answers to two questions (with order counterbalanced). One question asked whether given the shortage of COVID-19 vaccines at the time of the study, healthy people should (a) try to get the COVID-19 vaccine as soon as possible so that doses don't go to waste and herd immunity is achieved sooner, or (b) delay getting the vaccine until other, more vulnerable people get vaccinated. The other question pertained to whether the government should punish unsafe

behavior during the pandemic (e.g., implement fines for not wearing a face covering or for gathering in large groups).

For both questions, participants also saw the estimates of their partner. The estimates were randomly selected from a distribution of responses from the same pre-survey in which we asked the initial set of pilot study participants to state their views on the policies—and also make estimates regarding the views of their peers. As noted above, we were able to identify the accuracy of the estimates because we had the opinions of the prior pilot study participants.

The critical manipulation was the order in which participants made their own estimate versus saw the estimate of their partner for each estimation topic. For one of the topics, participants saw the estimate of their partner before making their own estimate (i.e., the Dependent task). For the other topic, participants saw the estimate of their partner after making their own estimate (i.e., the Independent task). After making their two estimates, participants evaluated their partners. We reminded participants of their own and their partners' estimates on the two prior questions and of the fact that they may be entered into a lottery for $20 depending on the accuracy of their own and their partners' answers.

Participants evaluated their partners in two ways, which served as the primary dependent variables in this study. First, participants indicated which partner they would prefer to work with for the third incentivized estimate. Participants indicated their preference on an 11-point Likert scale from -5 (definitely work with Partner A) to 5 (definitely work with Partner B), where a score of zero indicated indifference. We counterbalanced whether the labels "Partner A" and "Partner B" were associated with the partner with whom they worked on the Dependent vs. Independent task.

Second, participants evaluated the relative competence of the two partners. Participants indicated which partner they thought was more reasonable, intelligent, and knowledgeable on 11-point Likert scales ranging from -5 (definitely Partner A) to 5 (definitely Partner B), where a score of zero indicated indifference. The perceived competence index achieved a high level of reliability (alpha = .96). For both partner choice and perceptions of competence, we re-coded responses such that positive responses always indicated a preference for the partner from the Dependent task and negative responses indicated a preference for the partner from the Independent task.

Participants then made final estimates for each of the two questions, allowing us to evaluate both advice utilization and accuracy resulting from the Dependent and Independent task orders. Finally, participants learned that they did not need to complete a third estimate, but that they would be entered into the lottery for the bonus nonetheless. They then reported their gender and age.

Results

Partner choice. We first examined our key hypothesis: whether completing a task following Dependent (vs. Independent) sequence has implications for participants' future collaboration intentions. To test this hypothesis, we regressed partner choice in an empty regression (i.e., a one sample t-test comparing the mean to zero). We predicted, and found, an intercept significantly greater than zero, indicating support for our hypothesis that participants would prefer to collaborate on a future task with the partner from the dependent (rather than independent) task (b = .53, se = .18, t = 3.06, p = .002). To put these results in perspective, we can assess what percent of the time participants showed an absolute preference for each partner, defined as an overall score greater than zero (indicating a preference for the partner from the

dependent task), an overall score less than zero (indicating a preference for the partner from the independent task), or an overall score exactly equal to zero (indicating indifference). Participants preferred the partner from the dependent task 47% of the time, the partner from independent task 34% of the time, and were indifferent the remaining 19% of the time.

Perceived competence. We next examined whether the preference for partners from the Dependent task would extend to evaluations of more global partner characteristics. It is possible for example, that people prefer to work with those they agree with but recognize that this is a transient preference and does not actually arise from perceptions of greater competence on the part of agreeing partners. To test this hypothesis, we conducted an identical regression to the one above, replacing partner choice with perceived competence as the dependent variable. Results revealed a consistent pattern: participants perceived the partner with whom they worked on the dependent (vs. independent) task more positively (b = .51, se = .13, t = 3.96, $p < .001$), even though the advice they received was randomly drawn from the same distribution, and thus equally accurate on average. Results were also similar in magnitude to partner choice: participants perceived the partner from the dependent task as more competent 47% of the time and the partner from the independent task as more competent 32% of the time (the remaining 21% of participants were indifferent).

The role of disagreement. A key remaining question was to what extent, if at all, the effects of judgment task (dependent vs. independent) on partner choice and perceived competence would diminish when controlling for disagreement.

To begin answering this question, we first assessed whether we replicated the anchoring effect from our prior studies. Specifically, we examined whether disagreement was higher on the independent task compared to the dependent task. To do so, we subtracted the absolute value of

disagreement from the dependent task from the absolute value of disagreement from the independent task (i.e., Independent – dependent). If our theorizing was supported, the mean value of this variable should be greater than zero. Replicating a voluminous prior literature (and our previous studies), we found this to be the case: mean = 8.12, *se* = 1.10, $t = 7.39$, $p < .001$. Put another way, participants' initial estimates on the dependent task clearly and robustly assimilated toward the partner's estimate (i.e., the anchor).

We next examined whether effects on partner choice and perceived competence would diminish when controlling for this difference in disagreement that resulted from the anchoring effect. To do so, we re-ran the two regression models above, but with disagreement added as a control variable. In both models, we found a significant coefficient for disagreement (*bs* = .08 and .06 for partner choice and competence, respectively, both *p*s < .001) and saw that the intercept for the task type (independent vs. dependent) was no longer significantly different from zero (*p*s > .45 in both cases). Together, these two regressions indicated that differences in the level of disagreement in the independent vs. dependent task fully explained the relationship between judgment order and partner preference/perceived competence. Thus, instead of attributing differences in disagreement to task structure (or to their own inaccuracy or incompetence), participants attributed the disagreement to the incompetence of their partner, and were less willing to work with them in the future.

Advice-taking and accuracy. Finally, we assessed two exploratory dependent variables: advice-taking and judgment accuracy. While our prior analyses regarding disagreement focused on the absolute difference between participants' *initial* estimates and the advice they received, participants also had the opportunity to make a *final* estimate, which served as the foundation for these final two sets of analyses.

First, to measure advice-taking, we took the absolute value of the difference between participants' final estimates and the advice they received (for a similar procedure, see Rader, Soll, & Larrick, 2015; See, Morrison, Rothman, & Soll, 2011). We found that on the dependent task participants utilized advice to a greater extent as revealed by the fact that their final estimates were closer to the advice that they received ($M_{independent}$ = 17.15 vs. $M_{dependent}$ = 12.89, $t(398) = 4.14$, $p < .001$). This result replicates the disagreement result above, even though participants in the independent condition also had the opportunity to update their estimates based on advice.

Second, to measure accuracy, we first took the difference between participants' final estimates and the correct answer, as determined by responses from a pilot survey (the same survey that was used to generate the advice). We found that error was slightly higher for the Independent task compared to the Dependent task ($M_{independent}$ = 16.69 vs. $M_{dependent}$ = 15.07, $t(398) = 2.12$, $p = .034$). Of note, this finding contradicts the finding from the prior literature that recommends independent judgment aggregation to increase accuracy.

Why is this the case? An additional analysis shed light on this question. Specifically, our data allow us to calculate the level of error that would have been possible on the Independent task if participants had adhered to the recommendation of averaging their initial estimate with the advice they received to generate their final estimate. In line with the prior literature, had participants followed this procedure, they would have substantially reduced their average error ($M_{independent}$ = 16.69 vs. $M_{independentaveraged}$ = 13.72, $t(398) = 6.71$, $p < .001$). Indeed, averaging on the Independent task would have led to marginally lower error compared to the Dependent task ($M_{dependent}$ = 15.07 vs. $M_{independentaveraged}$ = 13.72, $t(398) = 181$, $p = .071$). This is partly due to the fact that on the Independent task participants' own estimate and the advice was more likely to

"bracket" (i.e., fall on directionally opposite sides of) the correct answer ($M_{independent}$ = 39.6% bracketed vs. $M_{dependent}$ = 23.4% bracketed, $t(398) = 4.97$, $p < .001$). Thus, we find that the psychological factors highlighted in this research rob participants of some of the accuracy benefits available through independent judgment aggregation.

General Discussion

Prior research offers a clear prescription for maximizing collaborative judgment accuracy: collaborators should render independent judgments prior to any interaction, which can then be aggregated—giving each approximately equal weight—to cancel out individual errors. Six studies demonstrate that this approach has an interpersonal downside: participants who follow an independent process assess their collaborator's judgment more negatively than those who evaluated an identical judgment without first generating their own. These effects applied to both quantitative estimates (Studies 1, 2, 3 and 6) and complex decisions with no correct answers (Studies 4 and 5). Furthermore, the effect was limited to others' judgments, but not one's own (Study 3). This pattern suggests that participants did not interpret disagreement as a signal to the difficulty of the task, but instead a signal of their partner's poor judgment. The phenomenon emerged whether participants believed that the judgment they were evaluating was the result of a simple guess or the result of a structured judgment process (Study 1) and for both laypeople and experienced professionals (Study 5). Finally, such effects generalized to evaluations of collaborator competence and willingness to work with them in the future (Study 6).

Participants' evaluations were largely driven by disagreement. Because participants anchored on peers' judgments, those who made their own judgments or decisions before evaluating those of others observed a greater amount of disagreement between themselves and their peers. In line with prior research on naïve realism, this disagreement ultimately accounted

for the differential evaluations produced by our task order manipulation. This was supported when tested by statistical moderation and statistical mediation.

Studies 4 and 5 also produced one unexpected set of results. Specifically, contingent on agreement, participants evaluated peer decisions made in the independent sequence more positively than they did in the dependent sequence. It is possible that people recognize the accuracy benefits of independent judgments and thus understand that agreement in the independent sequence is a stronger signal of accuracy than agreement in the dependent sequence. Alternatively, participants in the independent sequence may be particularly appreciative of the feeling of reduced uncertainty experienced when receiving corroborating advice (c.f., Gino, Brooks, & Schweitzer, 2012; Raghunathan & Pham, 1999). Finally, our results are consistent with work by Rader et al. (2015) who examined the impact of estimation order on the utilization of modal advice. Intriguingly, Study 2 of Rader et al. found that confidence in others' judgments was higher in the independent versus dependent judgment order. Because Rader et al. offered participants advice from the center of the estimate distribution, participants were often in close agreement with the advice they received, thus leading to a set of results that parallels the one we find here in cases of agreement. Future research should explore the factors that amplify versus dampen these effects.

Theoretical Implications

Our work has theoretical implications for both psychology and related fields. Most important, our work extends traditional thought regarding the benefits of independent judgment aggregation. Traditionally, research focused almost exclusively on judgment accuracy as the focal outcome of interest. As much as decision-makers are concerned with judgment accuracy, they are also often intensely concerned with how others evaluate them (Schlenker & Weigold,

1992; Tetlock, 2000, 2002). The present work highlights the need to broaden the scope of analysis to include interpersonal evaluations.

Second, our research extends classic work on anchoring by demonstrating that this phenomenon can have additional consequences for complex judgment and decision-making processes. Specifically, we show that anchoring can affect both actual and perceived disagreement between the judgments of group members, a process that ultimately affects the group members' assessments of each other's contributions and characteristics. Furthermore, we highlight the role of fundamental cognitive biases such as anchoring, previously studied primarily at an individual level, in shaping interpersonal processes.

Third, our work extends research on the phenomenon of "naïve realism" and the way individuals assess the merit of judgments, decisions, and viewpoints espoused by others. Prior work has demonstrated that people disparage perspectives that they differ from their own. However, in many contexts people are confronted with the ideas of others when they have not yet had the chance to formulate their own stance. Our data suggest that naïve realism continues to operate in this context, via the assimilation process referenced above. When individuals must assess the judgments and decisions of others without first independently generating their own view, those target judgments *appear* to be more similar to one's own, as of yet unformed, judgments. Participants then proceed to make biased attributions for the observed disagreement or lack thereof, seemingly oblivious to the role that the structure of the situation played in shaping their assessments.

Future research should examine alternative approaches to structuring group processes that preserve the independence of members' inputs while avoiding potential interpersonal pitfalls documented here. In the case of quantitative estimation, such a process might involve simple

mathematical aggregation of independent estimates. Such a process would ensure that the inputs of collaborators receive equal weight, even in cases of severe disagreement, when the individuals involved may be the least willing to employ an averaging strategy. Furthermore, committing to a weighting strategy a priori may reduce the attention that collaborators devote to evaluating each other's judgments, which may in and of itself reduce the potential for conflict. In the case of more complex decisions, when statistical aggregation is not possible, one could imagine appointing a group leader to evaluate and aggregate the views of group members. This would again ensure that the aggregation strategy is not biased by any individual's personal stance on the problem.

Statement of Limitations

Our work has limitations that serve as foundations for future research. Most importantly, our conclusions are based on short interactions among strangers. Future research should examine the potential downstream effect of judgment aggregation on individuals with longstanding relationships. It could be the case that pre-existing relationships or other situational variables (e.g., power dynamics) might impact the results. We also acknowledge that our current model of collaborative judgment is limited - surely there are other hybrid ways to combine judgments and other outcomes worthy of study. While we tried to focus on the main two judgment aggregation procedures studied in prior literature and relevant interpersonal evaluations, other procedures and outcome variables merit further testing. The Table of Limitations (Table 1) addresses these and other considerations.

Conclusion

Many of the most important decisions in life are made collaboratively. But how should such collaborations be structured? Prior research on collaborative judgment and decision making

has focused on accuracy as the focal outcome of interest, with scant attention paid to other goals decision-makers might pursue. Traditionally, such phenomena are examined through an intrapersonal lens, with a narrow focus on maximizing future expected value of the choice itself. The present research suggests that an expanded focus on the interpersonal consequences for the decision maker would be a fruitful avenue for future research.

Table 1*:* Assessment of limitations.

Dimension	Assessment
Internal validity	
Is the phenomenon diagnosed with experimental methods?	Yes
Is the phenomenon diagnosed with longitudinal methods?	No
Were the manipulations validated with manipulation checks, pretest data, or outcome data?	Yes. The reported studies came on the heels of several pilot studies. Furthermore, all studies drew directly on a large prior research literature on anchoring for their experimental manipulations of independent vs. independent judgment aggregation. Study 4 included a manipulation check.
What possible artifacts were ruled out?	Across studies, we test and rule out the effect of knowledge of a particular domain, the effect of intuitive versus effortful estimation processes, different attributions for disagreement, and generalize the effect across topics and decision tasks.
Statistical validity	
Was the statistical power at least 80%?	In all studies, we achieved at least 80% power to detect the effect size found in pilot studies for the main hypothesis test of interest. While we would have preferred to achieve 80% power for the

	smallest effect size of interest (often as small as Cohen's D = 0.10), we were resource constrained and thus powered our studies based on our estimated effect size (Lakens, 2022).
Was the reliability of the dependent measure established in this publication or elsewhere in the literature?	All studies used face-valid dependent measures operationalizing interpersonal evaluations extensively used in the prior literature.
If covariates are used, have the researchers ensured they are no affected by the experimental manipulation before including them in comparisons across experimental groups?	Not applicable
Were the distributional properties of the variables examined and did the variables have sufficient variability to verify effects?	Yes
Generalizability to different methods	
Were different experimental manipulations used?	Our investigation focuses specifically on task order as the independent variable which only has two possible versions: making estimates independently or dependently.
Generalizability to field settings	
Was the phenomenon assessed in a field setting?	No
Are the methods artificial?	Yes
Generalizability to times and populations	

Are the results generalizable to different years and historic periods?	This was not explicitly tested, however the dynamics of disagreement appear fairly consistent across the existing psychological literature spanning the last 50 years. Thus, we expect ongoing generalizability.
Are the results generalizable across populations (e.g., different ages, cultures, or nationalities)?	Results relied on US samples, although the samples varied in their age and level of professional expertise.
Theoretical limitations	
What are the main theoretical limitations?	There are a few key theoretical limitations. our conclusions are limited because they are based on short interactions among strangers. Future research should examine the potential downstream effect of judgment aggregation on individuals with longstanding relationships. It could be the case that pre-existing relationships or other situational variables (e.g., power dynamics) might impact the results. It could also be the case that our results are dampened when participants have the chance to engage in longer collaboration. We also acknowledge that our current model of collaborative judgment is limited - surely there are other hybrid ways to combine judgments and other outcomes worthy of study. While we tried to focus on the main two judgment aggregation procedures studied in prior literature and relevant interpersonal evaluations, other procedures and outcome variables merit further testing. The analyses on interpersonal evaluations may also need to be replicated with behavioral or more long-term measures. Further, future work is needed to understand how to eliminate such effects. Finally, it is untested whether our results generalize outside of the United States or to other cultural contexts.

References

Austen-Smith, D., Feddersen, T., Galinsky, A., Liljenquist, K. (2010). The kidney case. Evanston: Kellogg School of Management, Dispute Resolution Research Center.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529.

Blackman, S. F. (2014). *Seeing the subjective as objective: Naïve realism in aesthetic judgments* (Working Paper, Princeton University).

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Clemen, R. T., & Winkler, R. L. (1986). Combining Economic Forecasts. *Journal of Business & Economic Statistics*, *4*(1), 39–46. https://doi.org/10.2307/1391385

Dorison, C. A., & Heller, B. H. (2022). Observers penalize decision makers whose risk preferences are unaffected by loss–gain framing. *Journal of Experimental Psychology: General*.

Dorison, C. A., Umphres, C. K., & Lerner, J. S. (2022). Staying the course: Decision makers who escalate commitment are trusted and trustworthy. *Journal of Experimental Psychology: General*, *151*(4), 960.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*(2), 171–192. https://doi.org/10.1016/0030-5073(75)90044-6

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and

adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12*(5), 391–396.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311-318.

Frederick, S., Kahneman, D., & Mochon, D. (2010). Elaborating a simpler theory of anchoring. *Journal of Consumer Psychology, 20*(1), 17–19.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141*(1), 124–133.

Galton, F. (1907). Vox populi. *Nature, 75,* 450-451.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*(1), 149–167.

Gino, F., Brooks, A. W., & Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: Feeling anxious motivates individuals to seek and use advice. *Journal of Personality and Social Psychology, 102*(3), 497.

Grossmann, I., Eibach, R. P., Koyama, J., & Sahi, Q. B. (2020). Folk standards of sound judgment: Rationality Versus Reasonableness. *Science Advances*, *6*(2), eaaz0289.

Gunia, B. C., Swaab, R. I., Sivanathan, N., & Galinsky, A. D. (2013). The remarkable robustness of the first-offer effect: Across culture, power, and issues. *Personality and Social Psychology Bulletin, 39*(12), 1547-1558.

Harvey, N., & Fischer, I. (1997). Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes, 70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and*

Human Performance, *21*(1), 40–46.

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science*, *19*(2), 121–127.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658-8663.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54-86.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*(10), 732.

Koehler, D. J., & Beauregard, T. A. (2006). Illusion of confirmation from exposure to another's hypothesis. *Journal of Behavioral Decision Making*, *19*(1), 61-78.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 33267.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111-127.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin. 125*(2), 255.

Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads." *Journal of Experimental Social Psychology*, *48*(2), 507-512.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude

polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020-9025.

Loschelder, D. D., Friese, M., Schaerer, M., & Galinsky, A. D. (2016). The too-much-precision effect: when and why precise anchors backfire with experts. *Psychological Science*, *27*(12), 1573–1587.

Majer, J. M., Trötschel, R., Galinsky, A. D., & Loschelder, D. (2020). Open to offers, but resisting requests: How the framing of anchors affects motivation and negotiated outcomes. *Journal of Personality and Social Psychology*, *119*(3), 582.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin*, *37*(10), 1325-1338.

Minson, J. A., Mueller, J. S., & Larrick, R. P. (2017). The Contingent Wisdom of Dyads: When Discussion Enhances vs. Undermines the Accuracy of Collaborative Judgments. *Management Science*.

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*(1), 69–79.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing

decisions. *Organizational Behavior and Human Decision Processes*, *39*(1), 84-97.

Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology*, 66, 29-38.

Pronin, E., Gilovich, T., & Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*, *111*(3), 781–799.

Rader, C. A., Soll, J. B., & Larrick, R. P. (2015). Pushing away from representative advice: Advice taking, anchoring, and adjustment. *Organizational Behavior and Human Decision Processes*, *130*, 26-43.

Raghunathan, R., & Pham, M. T. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes*, *79*(1), 56-77.

Robinson, R. J., Keltner, D., Ward, A., & Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, *68*(3), 404–417.

Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application. *Handbook of Social Psychology*.

Ross, L. (2018). From the fundamental attribution error to the truly fundamental attribution error and beyond: My research journey. *Perspectives on Psychological Science*, *13*(6), 750-769.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social

conflict and misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Values and knowledge* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1-36.

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*, *43*(1), 133-168.

See, K. E., Morrison, E. W., Rothman, N. B., & Soll, J. B. (2011). The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes*, 116(2), 272-285.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology*, *99*(6), 917–932.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN 2160588*.

Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28. https://doi.org/10.1016/0749-5978(89)90055-1

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, *296*(10.5555), 1095645.

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396.

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the politics of the beholder?. *Administrative Science Quarterly*, *45*(2), 293-326.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

Vasel, K., (2018). It costs $233,610 to raise a child. *CNN*. http://money.cnn.com/2017/01/09/pf/cost-of-raising-a-child-2015/index.html.

Yaniv, I., & Choshen-Hillel, S. (2012). Exploiting the Wisdom of Others to Make Better Decisions: Suspending Judgment Reduces Egocentrism and Increases Accuracy. *Journal of Behavioral Decision Making*, *25*(5), 427–434. https://doi.org/10.1002/bdm.740

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909