Beyond accuracy:

The reputational costs of independent judgment aggregation

Charles A. Dorison

Kellogg School of Management

Bradley R. DeWees

United States Air Force

Julia A. Minson

Harvard Kennedy School

Study materials, data, pre-registrations, and analysis code are publicly available here.

Revise and Resubmit, *Organizational Behavior and Human Decision Processes*

Abstract

Most major decisions in organizations 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 seven 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 between what is often best for an organization (maximizing judgment accuracy) and what is often best for an individual decision maker (managing one's reputation).

Most major judgments and decisions in organizations are made collaboratively, often under conditions of extreme uncertainty. How well will a new product do on the market? How should we allocate scarce resources such as time or money? Should we take an aggressive, high-risk strategy or settle for the sure thing? Thus, managing such collaborations to achieve desired outcomes is a necessary leadership skill.

How should such collaborations be structured? Prior research on quantitative estimation offers a gold standard for how individuals should approach collaborative judgment and decision making. To maximize judgment accuracy, collaborators should begin by making 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. However, a narrow focus on judgment accuracy neglects other goals that may be important to decision makers. In the present research, we draw on the foundational assertion that people care deeply about their reputations and the impressions they leave on others. For example, a large body of research in psychology and organizational behavior makes clear that decision makers often care deeply about upholding their reputations 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., in the weeks before a big promotion decision), decision makers may in fact care more about the immediate impressions they leave on others than the future accuracy of their forecasts. We thus examine the extent to which prescriptions that maximize judgment accuracy may have unintended consequences for such reputational outcomes.

Across seven studies, we manipulated the order in which research participants made a judgment or decision of their own versus evaluated a judgment or decision produced by a peer. We consistently found that the order endorsed by prior research – offering one's own independent estimate first – led individuals to assess the judgments and decisions of others as being of lesser quality than individuals who evaluated peer input prior to offering their own judgment or decision. We observed this effect in a variety of estimation domains and with both lay and expert samples. Importantly, task order affected not only participants' evaluations of others' judgments, but also their evaluations of the individuals offering those judgments and their willingness to collaborate with those individuals in the future.

Why did such effects occur? We documented that the effect of independent vs. dependent order on evaluations was driven by systematic differences in the amount of disagreement that emerged as a function of task order. Compared to a dependent judgment order, independent judgment order increased the level of disagreement between partners' estimates. Importantly, individuals made biased inferences about the causes of disagreement, attributing it to flawed judgment on the part of the their partner rather than the structure of the situation.

Taken together, our research demonstrates that turning to the wisdom of crowds may come with a reputational cost: fellow crowd members appear less wise to those who first offered an opinion of their own. Our work thus uncovers a novel tension between what is often best for an organization (structuring choices to maximize judgment accuracy) and what is often best for an individual decision maker (structuring choices to manage one’s reputation).

Prior research on the effects of judgment order

In prior work, offering independent assessments constitutes the gold standard for maximizing collaborative judgment accuracy, *provided that* the final joint 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. Instead, prior research traditionally treats the accuracy resulting from different aggregation approaches implemented by a third party (i.e. the experimenter) as the focal outcome (for a review, see Gigone & Hastie, 1997).

A literature that indirectly sheds light on evaluation of peer judgment is work on the Judge Advisor System. The most common 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). Specifically, 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).

Explanations for this under-weighting of others' inputs include the idea that judges have access to their own reasons for an estimate but not others' reasons (Yaniv & Kleinberger, 2000). Additionally, judges may have an "egocentric bias." That is, individuals may believe that they are simply better judges than those from whom they receive advice (Krueger, 2003).[1] However, research testing the causal role of these mechanisms has found limited support. For example, people still underweight peer input even if they have no recall of the justifications supporting their own judgments and have few reasons to feel confident in them (Soll & Mannes, 2011). Yet, although the ultimate cause of underweighting of advice remains elusive, research paints a clear picture of individuals favoring their own judgments over of those of their peers.

Despite these robust findings, the advice taking literature cannot address the question of how task order affects evaluations of peer judgments. First, the Judge Advisor System paradigm always 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. Second, task order is usually 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.

The few studies that have varied judgment order produced mixed results. Rader, Soll, and Larrick (2015), for example, found that individuals who formed an independent judgment prior

[1] The latter explanation is related to ones that invoke "epistemic vigilance" (Sperber et al., 2010; Trouche et al., 2018) in that both portray individuals as believing they have better judgment than others. Explanations invoking vigilance, however, distinguish themselves by arguing that discounting is not a bias, but a functional strategy because it helps individuals guard against being accidentally or intentionally misinformed (Sperber et al., 2010).

to receiving advice were more likely to accept advice than judges who saw advice before making judgments of their own. The judges who first saw advice “pushed away” from it in forming their own judgments, suggesting that rendering one’s own judgment before evaluating that of a peer’s may lead to more favorable peer evaluations. In contrast, Yaniv and Choshen-Hillel (2012) compared advice utilization by participants who either (1) saw a picture of food and made their own estimate of the number of calories vs. (2) did not see the picture and did not make their own estimate. They found that participants in the second condition incorporated advice to a greater extent; however, because only participants in the first condition saw the picture, direct comparisons to the work by Rader and colleagues cannot be made. Other findings from studies that manipulated order in an advice-taking paradigm are inconclusive—Sniezek & Buckley (1995), for example—found no difference in the rate of advice taking between those who made their own judgments before seeing advice and those who saw advice first (see also Koehler & Beauregard, 2006).[2]

Importantly, however, none of these papers examined the effect of task order on interpersonal evaluations rendered by collaborators, and instead focused on more traditional measures of advice utilization and accuracy. Thus, questions regarding how task order might affect the way that collaborators evaluate each other and the downstream reputational consequences of these evaluations remain largely unaddressed by the prior literature.

Anchoring, disagreement, and naïve realism

[2] Sniezek & Buckley (1995) do find an effect of ordering on the use of what they call the “confidence utilization” strategy, or taking the advice of the advisor expressing the highest degree of confidence in her/his judgment. They find that individuals who first make judgments of their own are slightly less likely to use this strategy than individuals who first see advice. It is unclear how, if at all, this measure maps onto advisor evaluations.

In the present work, our predictions regarding the effects of judgment order on evaluations of peer input (and peers themselves) bridge two classic bodies of research in judgment and decision-making and social psychology. First, we draw on work on the phenomenon of anchoring and insufficient adjustment (Tversky & Kahneman, 1974) and its likely effects on divergence between judgments (i.e., disagreement). Second, we draw on theory and research on "naïve realism" (i.e., the objectivity illusion; Ross, Lepper, & Ward, 2010; Ross, 2018) in order to make predictions about how individuals are likely to interpret such divergence.

Anchoring and disagreement. Prior research has demonstrated that estimates under uncertainty are systematically influenced by seemingly irrelevant quantities ("anchors") that are cognitively available at the time of making a judgment. A large literature has explored the underlying causes and boundary conditions of such anchoring effects (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; Tversky & Kahneman, 1974). 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. Although the vast majority of research on anchoring has focused on individual judgment, there are clear implications for judgments that are made collaboratively, given that anchors may often be provided by peers. Yet, outside of the negotiations literature (e.g. Gunia, Swaab, Sivanathan, & Galinsky, 2013; Northcraft & Neale, 1986; Majer, Tröschel, Galinsky, & Loschelder, 2020) the effects of anchoring have not been systematically studied in interpersonal contexts.

In considering the effect of anchoring on collaborative judgment, we predicted that to the extent that peer estimates serve as anchors, 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 than decision-makers whose judgments assimilate toward each other as a result of anchoring.

Disagreement between judgments is not in itself a negative (Janis, 1972). In fact, disagreement in a particular form underpins the superior accuracy of group versus individual judgments; in most cases, estimates that are further from each other are more likely to err on opposite sides of the truth. When aggregated, such errors cancel each other out (Larrick & Soll, 2006). This effect, however, speaks solely to the accuracy of aggregated judgments without addressing other outcomes that may be of importance to collaborators, such as interpersonal evaluations and associated reputational concerns.

In many, if not most, consumer and organizational contexts, decision-makers engage in both the generation of judgments and decisions, as well as in an unstructured process of evaluating and weighting them to arrive at a final product. In the course of this process, estimators are likely to express opinions about the accuracy of each other’s beliefs and assumptions, ultimately arriving at a set of conclusions about the judgment at hand, but also about each other. The prior literature on anchoring has not been extended to shed light on these downstream interpersonal dynamics that might arise from estimates being closer to or further from each other. The present work examines exactly this.

Disagreement and naïve realism. In the present work, we draw on research on the phenomenon of “naïve realism” (Robinson, Keltner, & Ward, 1995; Ross et al., 2010; Ross & Ward, 1996; Ross, 2018) to predict that judges’ evaluations of each other’s inputs are likely to be influenced by the level of disagreement between their initial estimates. Several research streams have demonstrated that individuals treat their own perceptions of the world around them as an accurate and unbiased representation of an underlying reality. According to this work, people are “naïve realists” who do not 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; Ross, 2018). While this approach may be largely functional in perceiving physical objects and events, believing oneself to be an objective judge of ambiguous social stimuli has been demonstrated to give rise to a suite of psychological biases (Griffin & Ross, 1991; Ross et. al., 2010).

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 reviews, see Ross, Lepper, & Ward, 2010; Ross, 2018). This phenomenon has been demonstrated with respect to political and social views (Kunda, 1990; Pronin, Gilovich, & Ross, 2004) and even matters of taste (Blackman, 2014). Relatedly, individuals evaluate the merit of scientific findings as a function of whether they conform to their prior beliefs (Kahan, Peters, Dawson, & Slovic, 2017; Kahan et al., 2012; Lord, Ross, & Lepper, 1979). Most relevant to the present research, several studies in the domain of quanitative 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).

In line with the naïve realism literature, we predicted that disagreement will, in turn, lead individuals to disparage peers' judgments—and the peers themselves. In other words, we predicted that rather than interpreting disagreement as a signal of uncertainty or task difficulty, individuals will interpret it egocentrically as a negative signal about their countepart's judgment, and by implication their counterpart's other, more fundamental, qualities. Of note, we test this process account not only through statistical mediation, but also through statistical moderation. That is, we predicted that the effect of task order on evaluations of peer judgment would be mitigated in cases where task order would not influence disagreement.

Theoretical contributions

Taken together, we predicted (1) that following an independent judgment process would heighten disagreement and (2) that heightened disagreement would be associated with negative evaluations of counterparts' judgments (and counterparts themselves).

Our work extends the research literature in several ways. First, we contribute to the literature on collaborative judgment and decision-making by highlighting the potentially negative reputational consequences of the classic advice to begin collaborative judgment tasks by first rendering independent estimates. The prescriptive advice that the prior literature has offered is predicated on the assumption that the focal goal is to maximize judgment accuracy. While accuracy is no doubt an important goal, it is not the only one. Such recommendations may be incomplete in light of the additional reputational 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 that may be of great importance to decision-makers.

Second, 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, 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.

Third, the key questions examined by research on naïve realism revolve around how individuals evaluate others' positions relative to their own. As such, naïve realism research does not make predictions regarding situations when an individual has not yet formed their own opinion. In the present research, we extend this work by highlighting the effects of biased attributional processes on the formation of new judments.

Finally, our work has important organizational implications because task order and process design represents a lever that decision makers and organizational leaders can easily manipulate in order to improve collaboration effectiveness. Yet, the prior literature has only examined this variable from the standpoint of judgment accuracy, without offering recommendations on how organizations might also manage the interpersonal dynamics that arise in these common contexts. Our research begins to fill this gap by offering additional insight into maximizing the benefits of collaborative judgment and decision-making.

Research overview

We present the results of seven studies examining the effect of task order on evaluation of peer inputs in judgment and decision-making tasks across a range of domains. In Study 1, we tested our basic hypothesis and found that, in the domain of consumer judgments, individuals evaluated a peer's estimate as less accurate when they had first made their own independent estimate. In Study 2, we tested disagreement as a moderator of our effect, finding that the effect of task order on evaluations of peer judgment occurs in contexts with moderate or high levels of disagreement, but not extremely low levels of disagreement. In Study 3, we explicitly tested our predictions based on naïve realism and showed that people do indeed interpret disagreement between their own and another's estimates as a signal of error on the part of the other individual (but not themselves). In Study, 4 we expanded beyond the domain of quantitative estimation tasks and demonstrated our phenomenon in a complex medical decision-making scenario. In Study 5, we conceptually replicated Study 4 in a national security domain with both a lay sample and an elite national security sample. Finally, in Study 6, we demonstrated that the effects of task order extend beyond the evaluation of peer inputs and impact the interpersonal evaluations of the peers themselves, including willingness to collaborate with specific individuals in the future.

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.

Study 1

Study 1 provides an initial test of the basic effect: does independent vs. dependent judgment order influence perceived accuracy of a partner's estimate? Participants produced a simple consumer estimate (the lifetime cost of owning a dog) and evaluated the accuracy of an estimate ostensibly produced by a peer. We tested whether independently generating an estimate before evaluating another's estimate changed participants' evaluation of the accuracy of the target estimate. In addition to examining task order, we varied the level of effort associated with producing an estimate. This comparison of effort allowed us to test whether any potential effect requires participants to engage in involved, deliberate processing (as would happen in many professional contexts) or whether it would also emerge when people offer quick intuitive judgments. Furthermore, we measured participants' domain expertise and the extent to which the domain was personally important to them in order to examine whether these variables moderated the effect of task order on evaluations.

Method

Design and Participants. Study 1 employed a 2 (Judgment order: Independent, Dependent) X 2 (Process: following a structured 7-step process vs. making a single 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 or provided a quick intuitive estimate (instructions available in the Appendix). We determined our sample size by doubling the cell size from prior pilot studies in order 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 sample 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 randomly assigned participants either to give their own estimate of the lifetime cost of 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 simply make and/or evaluate an intuitive estimate (Intuition Condition).

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. The target answer was in fact an expert answer to the estimation task from the University of Pennsylvania School of Veterinary Medicine.

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. The results supported this hypothesis: 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]). Put another way, the mere fact of generating one's own estimate led to a larger change in participants' evaluations of a peer's contribution than whether that peer used a thoughtful process or simply guessed the answer. There was no interaction between the two variables (*95% CI*[-8.70, 6.10], *t*(804) = 0.35, *p* = .730).

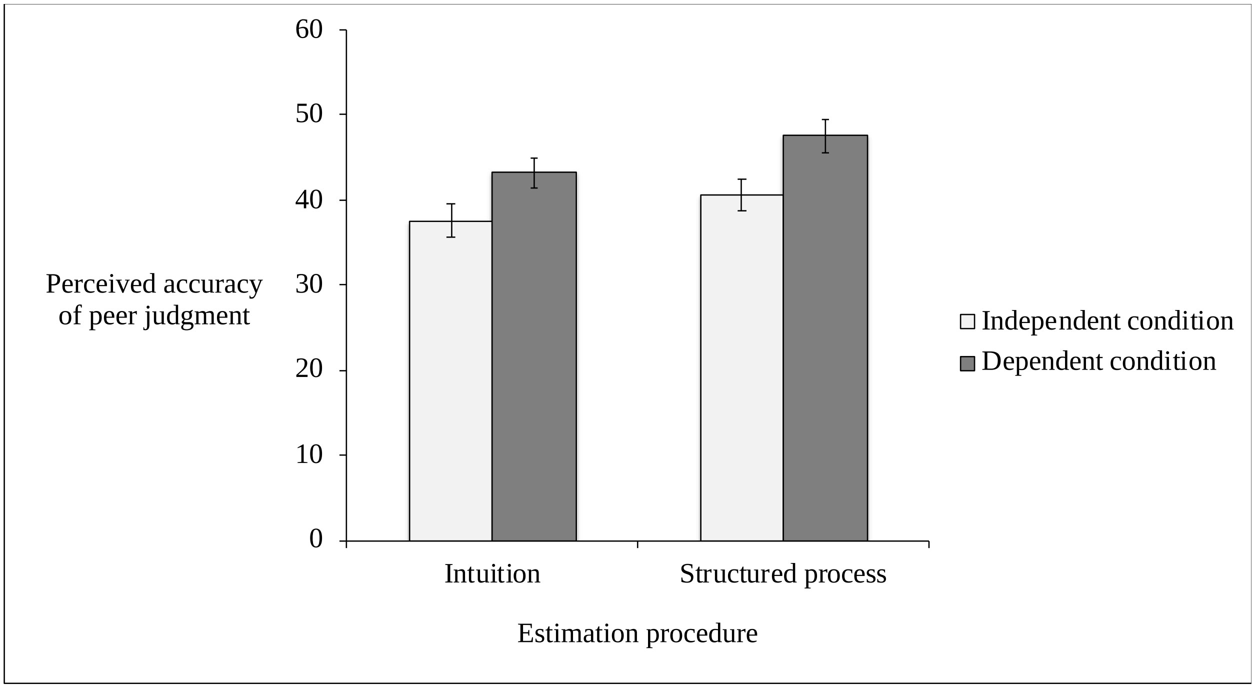


Figure 1. The vertical axis represents the judged likelihood that the target response was within 10% of the correct answer in Study 1. Participants who generated their own estimate prior to evaluating the target estimate judged that estimate as less likely to be accurate. Bars represent standard errors.

The role of disagreement. In line with research on naïve realism, we theorized that heightened disagreement underpinned differences in evaluations of the target's 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 condition (we did not collect estimates from participants in the Dependent condition 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).[3] We test this hypothesis more stringently in the studies to follow.

[3] 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

In Study 1, 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. A comparison of effect sizes showed how important task order can be: the mere act of generating a judgment had a 70% greater effect on evaluations than the difference between following a seven-step process relative to following one’s intuition.

We did not find evidence that domain importance or expertise moderated the effects of task order on evaluations, suggesting that explanations related to a motivation to defend one’s skill in personally important domains do not account for the effects of ordering. Suggestively, 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.

Study 2

We next examine the role of disagreement on the evaluation of peer judgments by randomly assigning participants to judgment tasks where high levels of disagreement are more or less likely. We hypothesized that the effect of judgment order depends on the level of disagreement. Specifically, we predicted that while participants who first generated an independent judgment (vs. did not) would judge the target estimate as less accurate when disagreement was moderate or high, this effect would dissipate when disagreement was extremely low. We also broaden our investigation to include a different familiar consumer estimate.

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 $0.70 for survey completion. The stimuli in this survey dealt with the costs of childrearing. We directed the study advertisement at parents in order to recruit participants with some domain expertise. At the end of the survey, we asked participants whether they were in fact parents, making it clear they would not be penalized for answering truthfully if they were not. 43 participants stated that they were not parents, leaving us with a final sample of 381 (*M*$_{age}$ = 35, 55% female). The statistical significance of results remains 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.

We varied the level of disagreement by asking participants to make judgments of relatively large or small quantities. Judgments of large (vs. small) quantities were intended to create more (vs. less) disagreement between the participants and the target estimate they evaluated. 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.

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 ($100) of the correct answer. We recorded these evaluations on a five-point Likert scale anchored at “Not at all likely” and “Very likely.” This evaluation 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

Manipulation check. We first examined the effect of judgment order on perceived agreement with the target estimate in the high vs. low disagreement conditions. As predicted, following a dependent (vs. independent) process increased perceived agreement in the High disagreement conditions ($M_{dependent}$ = 2.80, $M_{independent}$ = 2.44; *95% CI*[0.68, 0.06], *t*(210) = 2.35, *p* = .020), but did not increase perceived agreement in the Low disagreement conditions ($M_{dependent}$ = 2.41, $M_{independent}$ = 2.47; *95% CI*[-0.41, 0.30], *t*(167) = 0.31, *p* = .76)). That being said, the interaction between condition (high vs. low disagreement) and judgment order (independent vs. dependent) on perceived agreement was only marginally significant (interaction: *95% CI*[-0.89, 0.04], *t*(377) = -1.78, *p* = .075).

Evaluation of target's estimate. We next turned to our primary hypothesis: that the effect of judgment order on the evaluation of a target's estimate would depend on the level of disagreement. Specifically, we predicted that while participants who first generated an independent judgment (vs. did not) would judge the target estimate as less accurate when disagreement was moderate or high, this effect would dissipate when disagreement was extremely low.

This hypothesis was supported. 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 estimating 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), providing evidence 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).

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 (and directionally reversed) in cases where disagreement was low.

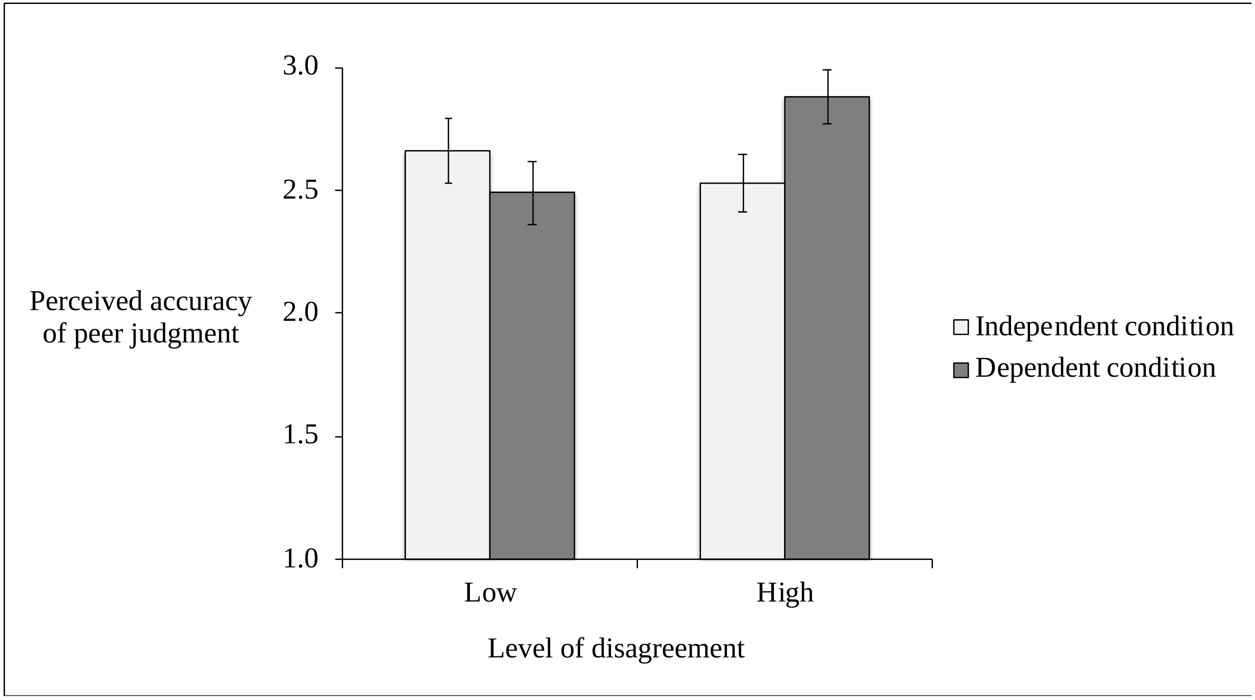


Figure 2. Judged likelihood that the target response was within 10% of the correct answer, measured using a five-point Likert scale (vertical axis). In the high disagreement condition, task order influenced evaluations. In the low disagreement condition, it did not. Bars represent standard errors.

Discussion

Study 2 extended our investigation by manipulating the likelihood of disagreement orthogonally to judgment order. In line with our theorizing, the effect of judgment order emerged only for estimates where participants were likely to disagree with the estimate they were evaluating. In cases where disagreement was unlikely, judgment order had no effect on evaluations.

Study 3

Study 3 examines whether participants' treatment of disagreement as a negative signal regarding the accuracy of the target estimate is normatively appropriate. Given that the only information participants possess at the time of making the evaluation are the two estimates, it may seem reasonable to evaluate the target as a function of how far it deviates from one's own judgment. However, if that is the case, then the same standard of accuracy should apply to evaluations of one's own judgment, as well. To the extent that disagreement points to the difficulty of the judgment and the potential for error, individuals should evaluate both their own judgments and the judgments of their peers more negatively as disagreement increases. If, instead, only peer evaluations suffer while evaluations of own judgments remain intact, we can conclude that individuals are evaluating judgments in a biased manner, in line with predictions 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 of estimates 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&M's 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&M's in the 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 the independent condition in pilot studies that had used the same stimuli. 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 the other individual whose judgment they were evaluating 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 their partner's inaccuracy.

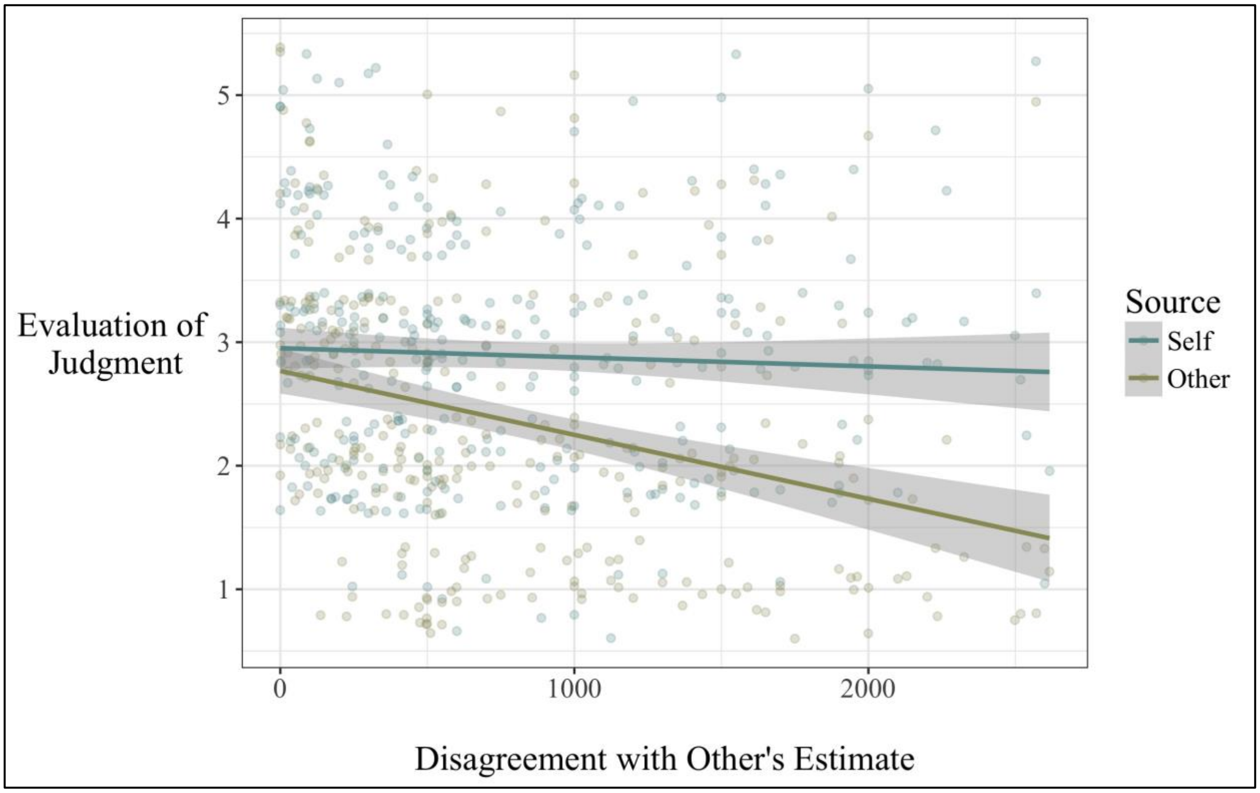


Figure 3: Participants judged the likelihood that either their own (blue) or another's (gold) estimate was correct on a scale that ranged from 1 - 5 (vertical axis; points above and below 1 and 5 are due to jittering the display). As disagreement increased (horizontal axis), the differing slopes suggest 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 provides support for the idea 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 was more likely to be wrong, rather than as an indication that the other's judgment *and* their own judgment were equally at risk of being incorrect.

Study 4

In the context of quantitative estimation, Studies 1-3 provided consistent evidence 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. Studies 1-3 also provided evidence for a key role of disagreement in driving such effects (e.g., as a statistical moderator in Study 2).

Study 4 builds upon the prior studies in two key ways. First, Study 4 examines our phenomenon in a different context: a medical domain laden with complex ethical considerations. Specifically, Study 4 asked participants to make and evaluate hypothetical choices and supporting reasons for which of several deserving individuals should receive a life-saving kidney. As with many decisions, there was no clearly correct choice. Because there is no way to evaluate the "accuracy" of such a choice, we instead solicited 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.

Although in many situations individual decision-makers have no way of knowing a priori which choice or which judgment is best, people must still work collaboratively to choose and execute a plan. In such cases, the individual who proposes the plan that is ultimately chosen often accrues reputational credit. To simulate this dynamic, we informed participants that a future set of participants, whom we referred to as “supervisors,” would make a choice of their own in the scenario. If the supervisor made the same recommendation as the participant, the participant would receive a financial bonus.

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.

Procedure. We adapted materials from “The Kidney Case” (Austen-Smith, Feddersen, Galinsky, & Liljenquist, 2014), a simulation designed for teaching 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 or not 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 participants 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

Manipulation check. We first examined whether participants in the dependent condition considered the target's answer and reasoning as more similar to their own than did participants in the dependent 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. $M_{independent} = 4.16$, $t(397) = 2.11$, $p = .035$).

Evaluation of target's choice. 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).

The role of disagreement. 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 general evaluations. To do so, we fit 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 found evidence of 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.

Thus, 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 discuss its implications further in the General Discussion.

Discussion

Study 4 documented the effect of making one's own choice in the domain of ethical decision-making, where participants considered life and death scenarios with no obvious correct answer. In line with our prior results, 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 judgment order on evaluations of peers' judgments was driven by the likelihood that a participant would agree with the target – specifically, participants who made independent decisions were half as likely to agree with the target they were evaluating as participants who made dependent decisions. In turn, disagreement underpinned evaluations of the choice itself (although agreement was perceived more positively in the Independent condition).

In our next study, we continued examining important judgment domains with no clear correct answers. Specifically, we studied the effects of task order in a national security decision with both lay and expert samples.

Studies 5A & 5B

In Studies 5A and 5B, we tested the generalization of our effect in a new domain with both a lay and expert sample. Specifically, we recruited both mTurk participants (Study 5A) and elite national security experts (Study 5B) and tested the effect of task order on evaluations of complex national security choices. To do so, we developed our own scenario, loosely based on the Obama administration's decision to raid the suspected compound of Osama Bin Laden in

Abbottabad, Pakistan in 2011.[4] In terms of structure, the scenario represented a conceptual replication of Study 4: decision-makers faced a set of options that were each appealing in their own way, making for a difficult decision. In terms of content, it represented a realistic situation in which elite national security decision-makers might find themselves. We developed the scenario 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).

We first conducted this experiment on a lay sample recruited from mTurk. We then ran the identical experiment with a sample of professionals with extensive national security expertise. This sample broadened the external validity of our findings and provided a more stringent test of our hypothesis, given that experts should be less susceptible to the effects of a simple manipulation like the ordering of tasks in a judgment and decision sequence. Below, 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 pre-registered[5] that we would collect a sample of 500 experts or recruit for one month, whichever

[4] We based our scenario on public reporting of the raid, not official details (e.g., Mahler, 2015).

[5] **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.

came first. We recruited 402 participants for the lay sample (M_{age} = 38, 44% female). For the expert sample, 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' contacts in the National Security Fellows program, described above, 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.

Our recruitment message to lay participants stated that they would make decisions in a national security context; our message to the expert population – delivered via e-mail by leveraging the personal and professional contacts of the authors – 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.

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.

Procedure. 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).

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 were highly correlated in both samples and thus achieved a high level of reliability (lay sample Cronbach’s α = .92; expert sample: α = .89). We thus we 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 national security areas, 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 options because it is the most logical of the options you have provided. In reality, I would have chosen none of the options." To the extent that our expert participants were able to imagine alternatives beyond those listed in the study regardless of condition, it would dampen any effect of task order on evaluations of whether the target's response was the best possibility.

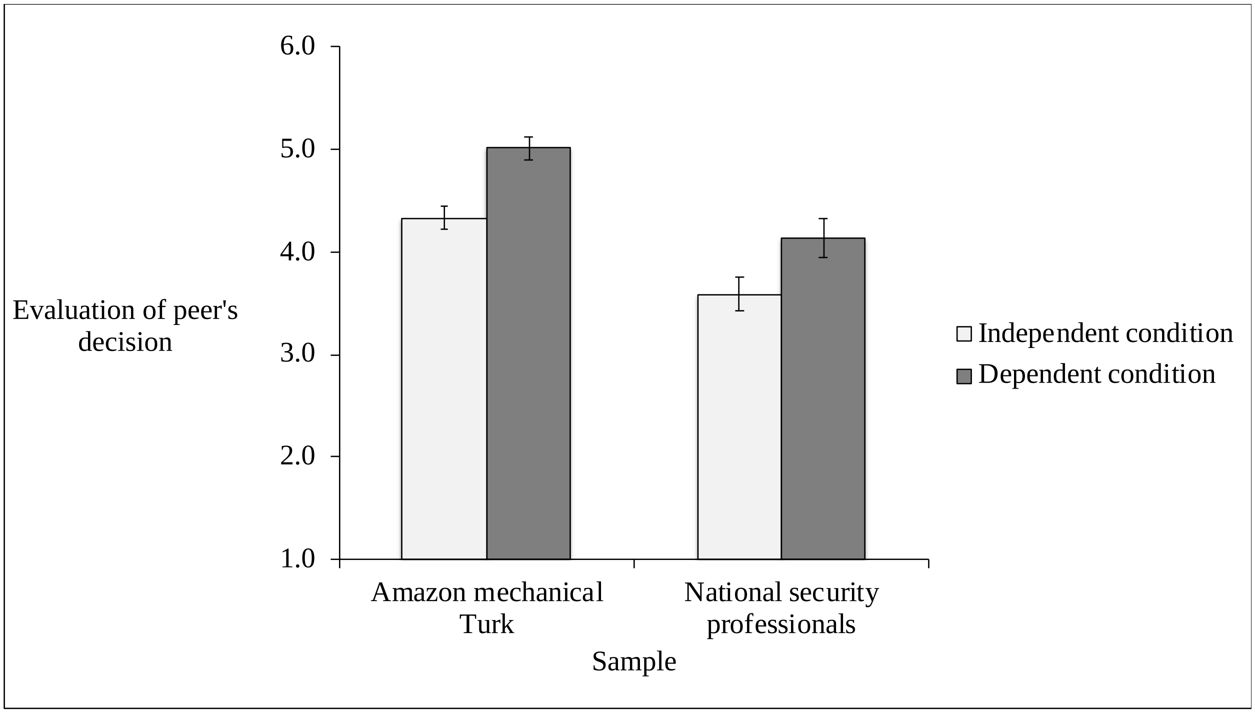


Figure 4. The vertical axis represents the evaluation of the target's response. Participants who generated their own choice prior to evaluating the target choice judged that choice more negatively. Bars represent standard errors.

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 as 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 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).[6] Both simple effects of this interaction were significant. Participants in the Independent

[6] 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.

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

Study 5 examined the effects of task order using a realistic scenario, in an important setting, with both lay and expert participants. For both lay and expert participants, independent (vs. dependent) judgment aggregation led to lowered evaluations of the intelligence and morality of the options offered by peers. In both samples, this effect appeared to be driven by the greater likelihood of disagreement in the Independent condition.

This study provided a replication of our effect from Study 4 both in a different context as well as with individuals experienced in the decision domain. Although our expert sample was considerably smaller due to the challenges of recruiting experienced professionals with a security background, we observed a pattern that was quite similar—although perhaps slightly weaker—to that produced by lay participants.

Study 6

Studies 1-5 provided robust evidence that task order influences evaluation of peer judgments and decisions. In Study 6, we test whether such effects generalize to evaluations of the peers themselves and have downstream consequences for collaboration. To do so, we examined whether participants evaluated a partner more positively when they had completed a task together using a dependent (vs. independent) judgment order. Specifically, participants reported how competent they believed their partners to be and chose whether they wanted to work with them on a future, incentivized task. As in prior studies, we also explored to what extent disagreement would underpin such effects. Critically, Study 6 also allowed 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 in which they made a judgment and then viewed a partner's judgment (Independent task) and one in 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, partly because heightened disagreement associated with the Independent task would be associated with negative attributions about 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 preferences of participants from a prior study. In this earlier 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 participants that supported a particular policy. Thus, we were able to assess the accuracy of the stimates because we knew the true proportion of prior participants who held each opinion.

We further told participants that while completing the estimation tasks, they would 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. 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 incentived. Specifically, participants read that if any of their 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 participants who had given specific answers to two questions (with order counterbalanced). As noted above, we were able to identify the accuracy of the estimates because we had the opinions of the prior participants. One question asked whether 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 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 participants to state their views on the policies—and also make estimates regarding the views of their peers.

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 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 keep 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 Partner A and 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 (inidicating 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 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), despite the fact that the advice they received was randomly drawn from the same distribution, and thus equally accurate on average. Results were also similar in magnitude: 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 preference and perceived competence would diminish when controlling for disagreement, as predicted by theory and research on naïve realism.

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 thus 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). 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, despite the fact that participants in the independent condition had the opportunity to update to the advice they received.

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 normative benchmarking 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 lower advice utilization that we observe in the Independent task robs participants of the level of accuracy that would have been attainable to them, if not for the psychological factors highlighted in this research in which participants in the independent task perceived their partner as less competent (and thus less worthy of updating to their advice).

Discussion

Study 6 broadened our investigation by demonstrating that the effect of task order extended to global evaluations of peers, rather than just peer judgment. Participants not only judged the peer from the Independent task less positively, they also had less interest in working with this person on a new task. Our data again reinforced the key role of disagreement in driving such evaluations.

Importantly, Study 6 also allowed us to test the implications of task order for judgment accuracy. We find that despite the powerful anchoring effects that we document in every study, Dependent estimates were equally accurate as independent estimates because participants utilized advice to a greater extent on the Dependent task.

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 so as to cancel out individual errors. In seven studies, we demonstrate that following such a strategy creates reputational costs: participants who follow an independent process (and thus first generate their own judgment) assess their collaborator's judgment more negatively than those who evaluated an identical judgment without first generating their own judgment. These effects applied to both quantitative estimates (Studies 1, 2 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). 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 novices and experts (Study 5). Finally, such effects generalized to evaluations of collaborator competence and willingness to work with them in the future (Study 6).

In line with our theorizing, participants' evaluations were largely driven by disagreement (i.e., the extent to which target judgments diverged from the participants' own views as a function of task order). On quantitative estimates used in Studies 1, 2, and 6, when participants evaluated a peer's input prior to generating their own assessment, their own judgments assimilated toward the target, in line with the prior literature on anchoring and insufficient adjustment. On complex medical and national security decision tasks used in Studies 4 and 5, participants who had not made their own decision prior to evaluating that of a peer were more likely to make the same decision as the one they had evaluated. As a consequence, participants who did make 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 different evaluations produced by our task order manipulation. This was true in the case of statistical moderation and statistical mediation.

Although one could argue that questioning the merit of a judgment that deviates from one's own makes logical sense, Study 3 demonstrates that people go beyond the dictates of logic in punishing judgments that disagree with their own. When we asked participants to evaluate their own judgments and those of randomly selected partners, we observed that disagreement had a negative effect only on evaluations of judgments produced by others, but not on evaluations of judgments produced by the self. Thus participants did not interpret disagreement as a signal to the difficulty or uncertainty inherent in the task, but instead derogated their counterpart's judgment quality.

Importantly, we also observed our effect of task order on reputation with both lay and expert samples. In Study 5, national security professionals engaged in a decision-task that was likely highly familiar to many of them. Given that the task featured a high degree of uncertainty (similar to analogous real-world situations), one might expect that the evaluations of experts would remain immune to our order manipulation. Instead, similar to our lay samples, the experts assimilated their own decisions to those of an unknown "peer" after seeing the peer's choice, resulting in different degrees of disagreement between conditions. The perception of greater disagreement again led experts who had first offered their own decision to evaluate those of a fellow national security professional more negatively.

Studies 4 and 5 also produced one unexpected set of results. Specifically, contingent on agreement, participants evaluated peer decisions made in the independent order *more* positively than they did in the dependent sequence. Although we did not predict this finding, we can speculate about several interesting interpretations. First, 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 that they may experience when receiving corroborating advice. Uncertainty is aversive (e.g., Gino, Brooks, & Schweitzer, 2012; Raghunathan & Pham, 1999) and is likely to have been experienced more acutely in the independent rather than dependent sequence. 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 el 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 distriubution, 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.

In sum, we observe two forces with opposing consequences for evaluations of peers and their judgments: On the one hand, participants in the independent task sequence were more likely overall to disagree with the target (and disagreement in general led to more negative evaluations). However, because in cases of agreement targets were evaluated more positively in the independent task sequence, some of the negative effect of the independent task order on evaluations was counteracted. Future research should explore the factors that amplify versus dampen these effects, including, for example, the severity and salience of disagreement.

Theoretical Implications

Our work has theoretical implications for multiple literatures. First, our work extends traditional thought regarding the benefits of independent judgment aggregation. Historically, research focused almost exclusively on judgment accuracy as the focal outcome of interest. The present work highlights the need to broaden the scope of analysis to include reputational outcomes. As much as decision-makers are concerned with judgment accuracy, they are also often intensely concerned with their reputations and relationships (Schlenker & Weigold, 1992; Tetlock, 2000, 2002). Importantly, we demonstrate that evaluations of peer judgments and decisions affected by task order also extend to the evaluations of the peers themselves. Our work broadens the prior research on collaborative judgment to include the interpersonal consequnces of task structure.

Second, our work also holds implications for accuracy benefits documented in prior literature. When we evaluated the impact of task order on the more traditional accuracy measures, we find that the independent estimation sequence did not produce the expected accuracy benefits. In fact, the dependent sequence produced slightly more accurate estimates. This was due to the fact that participants who were exposed to peer judgments prior to rendering their own estimates actually incorporated peer input to a greater extent. That being said, if participants in the independent condition had followed the normative benchmark and averaged their initial estimates with advice, they would have been able to achieve greater accuracy. Thus, we find that rather than being unambiguously superior, the independent judgment sequence comes with a set of trade-offs because aggregating independent estimates can lead to negative interpersonal evaluations and lower advice utilization than is normatively appropriate.

Third, our research extends classic work on anchoring by demonstrating that this phenomenon can have additional consequences for complex judgment and decision-making processes beyond the well-documented assimilation effects. Specifically, we show that, because of assimilation, 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.

Fourth, our work extends research on the phenomenon of "naïve realism" and the manner in which individuals assess the merit of judgments, decisions, and viewpoints espoused by others. Prior work has demonstrated that people disparage ideas and viewpoints to the extent 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 have to 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 task order played in shaping their beliefs.

Finally, our work is related to prior research on the “Judge Advisor System” which has examined how weight on advice (WOA) varies with different levels of disagreement between judge and advisor. Different studies have reached different conclusions, with some studies concluding that weight on advice shows a negative linear relationship with disagreement (Liberman, Minson, Bryan, & Ross, 2011; Minson, Liberman, & Ross, 2011), and other studies concluding that WOA shows a curvilinear relationship (low WOA at high and low disagreement, with WOA peaking at mid-range levels) (Schultze, Rakotoarisoa, & Schulz-Hardt, 2015). These conclusions may result in part from the adjustment measure: when two estimates in a JAS experiment are close together, any small adjustment by the participant in the direction of the advisor eliminates a large proportion of the distance between the two estimates. Our work advances this discussion by showing that when participants are free to use an easily interpretable self-report measure to evaluate the quality of others’ estimates, we see a strong linear relationship between evaluation and disagreement.

Practical implications

Our work has important implications for the structure of collaborative judgment and decision-making processes in organizations. The prior literature has advised group-members to arrive at judgments and decisions independently in order to minimize the assimilation of estimates toward one another and preserve “the wisdom of crowds” (Gigone & Hastie, 1997; Minson, Mueller & Larrick, 2017; Sunstein & Hastie, 2015). Our current work demonstrates that while this approach is undoubtedly correct from the perspective of increasing the diversity of estimates, it might introduce a hidden reputational cost.

Lower evaluations of group member contributions and the reasoning behind them are likely to increase conflict. And although classic literature has theorized that task conflict and relationship conflict have opposing effects on group outcomes, more recent work has found both to be deleterious (De Dreu & Weingart, 2003; De Wit, Greet, & Jehn, 2012). Participants in our studies were willing to make inferences about a partner's intelligence, competence, and reasonableness based on disagreement around a simple estimate. This finding suggests, that especially in contexts with limited information, task conflict may easily transform into relationship conflict via the attributions that collaborators make for disagreement.

Future research should examine alternative approaches to structuring group processes that preserve the independence of members' inputs while avoiding potential relational pitfalls associated with group members assessing each other's contributions after independently generating their own views. 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 interpersonal costs.

In the case of more complex decisions, when statistical aggregation of inputs is not possible, one could imagine appointing a group leader who is not committed to a course of action 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.

Finally, it may be the case that the negative attributions for disagreement that we document here would be ameliorated if collaborators had an appreciation for the role of the judgment sequence in generating more divergent or more similar views. Educating decision-makers regarding the effects of independent estimation may enable them to make more appropriate attributions for disagreement, and thus enable them to benefit from the accuracy benefits of independent judgment without paying the interpersonal costs.

Conclusion

Many of the most important decisions in organizations 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. While accuracy is an important goal, it is not the only one. The present research reveals that while following the traditional prescription to use independent judgment aggregation can have benefits (i.e., for accuracy), it can limit individuals' abilities to maximize other outcomes (i.e., reputations).

Our results have implications not only for generation of quantitative judgments, but also for a wide range of judgment and decision making phenomena. 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 reputational consequences for the decision maker would be a fruitful avenue for future research.

References

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

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

De Dreu, C. K., & Weingart, L. R. (2003). Task versus relationship conflict, team performance, and team member satisfaction: a meta-analysis. *Journal of applied Psychology*, *88*(4), 741.

De Wit, F. R., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: a meta-analysis. *Journal of applied psychology*, *97*(2), 360.

Dorison, C.A., Umphres, C., & Lerner, J.S. (2021). Staying the course: Decision makers who escalate commitment are trusted and trustworthy. *Harvard University Working Paper*.

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.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In *Advances in experimental social psychology* (Vol. 24, pp. 319-359). Academic Press.

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.

Janis, I. L. (1972). *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin Company.

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.

Krueger, J. I. (2003). Return of the ego--Self-referent information as a filter for social prediction: Comment on Karniol (2003). *Psychological Review*, *110*(3), 585–590. https://doi.org/10.1037/0033-295X.110.3.585

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

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.

Mahler, J., (2015). What Do We Really Know About Osama bin Laden's Death? *New York Times*. https://www.nytimes.com/2015/10/18/magazine/what-do-we-really-know-about-osama-bin-ladens-death.html.

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.

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.

Schultze, T., Rakotoarisoa, A. F., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment & Decision Making*, *10*(2).

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

Snyder, R. C., Bruck, H. W., Sapin, B., Hudson, V. M., Chollet, D. H., & Goldgeier, J. M. (2002). *Foreign policy decision making*. New York: Palgrave Macmillan.

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.

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, *27*(1), 81-102.

Sperber, D., CléMent, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sunstein, C. R., & Hastie, R. (2015). *Wiser: Getting Beyond Groupthink to Make Groups Smarter*. Harvard Business Press.

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.

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2018). Vigilant conservatism in evaluating communicated information. *PLoS ONE*, *13*(1), 1–16. https://doi.org/10.1371/journal.pone.0188825

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

Appendix A – Answer Process in Study 1

Cost of dog ownership seven-step estimation process
Step 1: The person estimated how much dog food costs every year. Step 2: The person estimated how much veterinarian bills cost every year. Step 3: The person then added estimates from Steps 1 & 2 to get an estimate of annual costs (they were allowed to use a calculator). Step 4: The person estimated how many years the dog would live. Step 5: The person multiplied the estimate of annual costs (Step 3) by the estimate of the dog's longevity (Step 4). They were again allowed to use a calculator. Step 6: The person estimated the one-time costs of owning the dog, such as collar, bedding, leash, etc. Step 7: The person then added together answers from Step 5 (continuing costs) and Step 6 (one-time costs) to calculate a final estimate of the lifetime cost of owning an average dog.

Table 1: Seven-step process for estimating the lifetime cost of owning a dog. Participants in the uncommitted condition saw the language above; participants in the committed condition saw the same seven-step process, only directed at the participant him or herself (i.e. “Estimate how much dog food costs every year”).

Appendix B – Kidney Allocation

Candidate Descriptions, Kidney Case

Candidate A Description: “Candidate A is a 34-year-old army veteran and the recipient of the Decoration of Valor. As part of the lead company in a successful and pivotal ground attack, he was exposed to a chemical warfare agent, which is known to be nephrotoxic (i.e., directly damages the kidneys). Since returning from the war, Candidate A has experienced a slow but steady decline in renal function and has now reached end-stage renal disease.

He suffers from post-traumatic stress disorder/depression and therefore may be less likely than others to comply with the rigorous post-transplant regimen. However, he has shown improvement with counseling. His psychiatrist argues that much of his depression stems from his poor health and may be partially or completely resolved when he is restored to health after transplant.”

Candidate A Target Justification (i.e., justification given by target): "Candidate A needs a kidney because he sacrificed himself on behalf of his fellow soldiers. He is relatively young, and even though his depression may complicate his compliance, the transplant may help with the depression. We could give him back the life that his military service has taken away from him. If we do not give it to him, what sort of message would that send to others about the value of service?"

Candidate B Description: “Candidate B is a 44-year old woman who has only been on the wait list for 5 months. She has suffered from high blood pressure for several years, but as a busy, single mother with four children, she has been inconsistent in taking her medication. As a result, her high blood pressure has definitely hastened her renal failure.

However, the available kidney is an exact match and therefore Candidate B has the best chance for a successful transplant outcome. Because the kidney is a good match, she won't have to take as many immuno-suppressant medications as other potential recipients. This means she may be able to avoid many of the health complications that the average kidney recipient experiences.”

Candidate B Target Justification: "Candidate B is the best match for the kidney, so we know it will be put to good use. Avoiding any health problems from a kidney transplant is rare, and we should take advantage of it. A successful transplant means that she will go off the list and stay off the list. More importantly, she is a single mother, and there is no replacing a single mother of four children. Her children depend on her and we are in a place to make sure that they can continue to depend on her."

Candidate C Description: "When he was in his twenties, Candidate C donated a kidney to his brother. At the time, doctors informed him that there was a small risk that he might someday develop the same disease as his brother, but as his kidney was an excellent match, he was willing to do whatever he could to aid his brother's health.

Now, fifteen years later, Candidate C has unfortunately developed the same disease as his brother, and although his case is less severe, his only remaining kidney is overwhelmed and beginning to fail."

Candidate C Target Justification: "Candidate C is in this situation because he volunteered a kidney earlier in life. His family has seen enough pain due to kidney-related health issues; we can help stop that pain by giving him a kidney now. It's because of people like him that we have kidneys to donate in the first place -- that sort of sacrifice should be rewarded. If we do not give it to him, what sort of message would it send to prospective donors?"

Candidate D Description: "Candidate D developed renal failure due to adult poly-cystic kidney disease, which is usually an inherited disease. She watched her father die of the same disease and is concerned about her younger siblings who may develop the same condition.

After graduating from college, she joined a commercial real estate firm. She subsequently founded Capital Realty, currently one of the country's highest grossing privately owned real estate firms. 51 years old, she has already donated millions of dollars to support poly-cystic kidney disease research. Should she survive her disease through a kidney transplant, she is willing to convert her company to a non-profit entity and donate all future profits to kidney research."

Candidate D Target Justification: "Candidate D, who inherited her disease, is in no way responsible for the position she is in. She and the rest of her family have suffered more than enough as a result of this disease, and we are in a position to put a stop to it. Further, we should consider the effect that giving her kidney will have on others. If her life is saved, she will continue to give to kidney research, which could save the lives of many more people."

Appendix C – National Security Scenario

In today's scenario you will be a member of the operations staff for the commander of U.S. troops in Africa.

You, along with several other members of the staff, are meeting to discuss an urgent and important national security topic. Specifically, you have recently received intelligence indicating the possible location of a highly-valued military target, a person we will call Combatant X.

Your steps for today's task [**committed condition**]:

1. Review background information on Combatant X
2. Consider possible courses of action
3. Recommend a course of action and explain your recommendation
4. Evaluate the course of action proposed by another member of the staff (your "Partner").

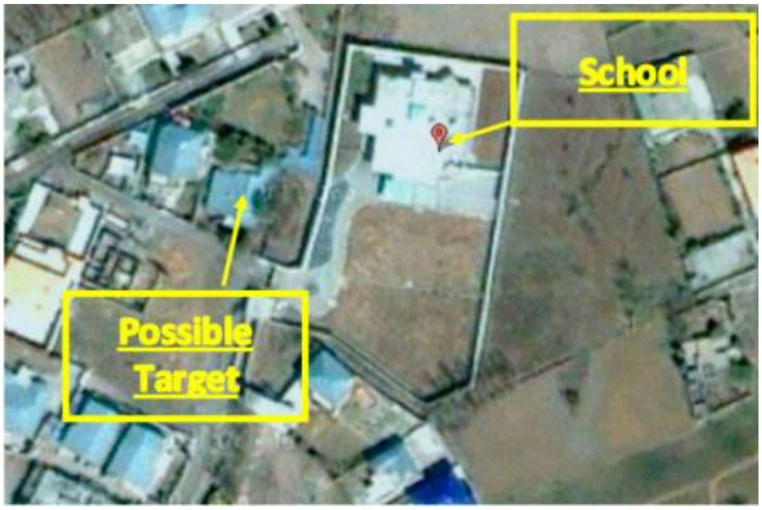
Earlier this week the Commander received an intelligence briefing that a high-value military target, Combatant X, may be hiding outside of Kampala, Uganda.

The government of Uganda is not aware (as far as U.S. intelligence knows) that Combatant X may be sheltering in their country. Though, if they were aware, it is unclear whether they would assist in capturing him, and they would be upset if the U.S. captured him without their permission.

This presents a difficult diplomatic and tactical problem, as the U.S. wants to maintain positive relations with the Ugandan government – their support in the fight against terrorism is very important over the long-term – but the U.S. also wants to capture Combatant X.

SOUTH SUDAN
Arua
Nile Albert
Gulu
Pakwach
Achwa
Lira
Lake Bisina
DEM. REP. OF THE CONGO
Masindi
Lake Albert
Lake Kyoga
Mbale
Ruwenzori
UGANDA
Jinja
KENYA
Kasese
KAMPALA
Katonga
Entebbe
Lake Edward
Lake George
Masaka
Mbarana
Lake Victoria
Kabale
RWANDA
KIGALI
TANZANIA

Intelligence analysts believe Combatant X may be hiding in a compound, pictured below. The compound is in the middle of a heavily populated area and is adjacent to a building that serves as a home and school for early-grade children (approximately an elementary school, though grades are not clearly defined). The possible target is too close to the school to allow for an airstrike of any kind, as that would entail unacceptably high civilian casualties.



Combatant X is considered one of the United States' deadliest enemies. The intelligence community has documented a long history of his leadership role in terrorist plots against the U.S. For over 10 years he has raised funds, recruited fighters, and planned attacks against U.S. personnel. Intelligence sources believe he travels with well-armed guards and that he has a well-established network of informants living in the area, making on-the-ground intelligence collection difficult.

At the request of the Commander, staff members have prepared a range of viable options. Please proceed to the next screen to review the options.

National Security Scenario Decision Options

“Embed a Conspirator” Option: The intelligence community has a highly-trained human intelligence agent in the area who could attempt to embed within Combatant X’s private network. If successful, the agent could orchestrate an event to capture Combatant X.

The advantages to this option are that it would limit human casualties and runs only a slight risk of upsetting the host nation (the U.S. can deny the existence of the intelligence agent). The disadvantages to this option are that it runs a strong risk of compromising the agent and alerting Combatant X, who, again, is believed to have several informants of his own in the area.

Embed a Conspirator Partner Justification (i.e., justification given by partner): "This should limit casualties, and probably won't upset Uganda's government. Both are really important to us. The other options have too many downsides like U.S. casualties in the independent option, upsetting Uganda in the wait option, and losing Combatant X in the assist option."

“Wait for Movement” Option: The U.S. has aerial intelligence assets stationed above the compound at all times. While constantly monitoring the feeds from those sources, the intelligence community could wait and see if Combatant X leaves the compound and moves to an area where an airstrike could target him without the risk of civilian casualties.

The advantages to this option are that it limits the risk of civilian casualties, and it also limits the risk to U.S. casualties, since personnel would not have to be sent to the compound for a raid. The disadvantages to this option are that it runs the greatest risk of losing track of Combatant X – if he does move, his best time to elude intelligence would be in the highly-populated area of the city (where an airstrike would not target him). This option may also upset the host nation if the U.S. conducts an airstrike without their permission.

“Wait for Movement Partner Justification: "This should limit civilian and U.S. casualties. Both are really important to us. The other options have too many downsides like U.S. casualties in the independent option, sacrificing the agent in the embed option, and losing Combatant X in the assist option."

“Assist Host Nation” Option: The U.S. could take what information they have to the host nation, and prompt them to conduct a raid of their own to capture Combatant X.

The advantages to this option are that it runs the least risk of upsetting the host nation – it might actually strengthen ties – and it poses the least risk to U.S. personnel, since they would not have to raid the compound themselves. The disadvantages to this option are that it runs the greatest risk of losing Combatant X. U.S. intelligence believes that Combatant X has informants within the host nation's government. There is also no guarantee of protection against civilian casualties if the host nation does launch a raid of its own.

"Assist Host Nation" Partner Justification: "This won't upset the Ugandan government and will limit U.S. casualties. Both are really important to us. The other options have too many downsides like U.S. casualties in the independent option, upsetting Uganda in the wait option, and sacrificing the agent in the embed option."

"Independent Action" Option: The U.S. could send military personnel to raid the compound and attempt to capture Combatant X. A military team is prepared to conduct the mission if desired.

The advantages to this option are that it gives the greatest chance of capturing Combatant X if he is, in fact, in the compound. The disadvantages to this option are that it puts the lives of U.S. personnel at risk, and will also very likely upset the host nation, who would not be aware of the attack beforehand. It also carries a moderate risk of civilian casualties – while military raids pose less risk than an airstrike, it is possible that civilians may be caught in the crossfire of a raid.

"Independent Action" Partner Justification: "This gives us the best chance at Combatant X and likely minimizes civilian casualties. Both are really important to us. The other options have too many downsides like sacrificing the agent in the embed option, upsetting Uganda in the wait option, and losing Combatant X in the assist option."