Are Frogs More Forgiving Than Acorns Anticipate?

How Response Pressure Confounds Can Generate (Apparent) Biases in Metaperceptions

Rebecca L. Schaumberg

ESMT Berlin

Joseph P. Simmons

University of Pennsylvania

Abstract

Much research suggests that people are remarkably poor at understanding how others judge them, particularly regarding the impact of their actions on others. People appear to succumb to a host of nominally distinct biases that take a similar form. They overestimate how harshly they will be judged for bad actions and underestimate how positively they will be judged for good actions. In this article, we demonstrate that these biases can arise from "response pressure" confounds that are baked into how metaperceptions are typically studied. Specifically, we show that socially appropriate responses differ between those asked to *predict* how another person will feel about their actions and those asked to *report* how they feel about another person's actions. Predicting that someone will feel extremely grateful for your good deed is less appropriate than reporting that you are extremely grateful for someone else's good deed, while predicting that someone will feel extremely upset for your minor transgression is more appropriate than reporting that you are extremely upset at someone else's minor transgression. So if some participants provide inauthentic but socially appropriate responses, apparent biases will emerge, even if they don't exist. Across 13 preregistered studies, we show that differences in response pressures between predicting "how others will judge me" vs. reporting "how I would judge others" can generate apparent biases and produce fanciful ones, such as people overestimating how negatively a frog would judge them if they were an acorn that fell on its head.

According to the social psychology literature, people are remarkably bad at understanding how their actions will be judged by other people, with a host of biases distorting these judgments. For example, people appear to overestimate how negatively others will judge them for small offenses and potential indiscretions. They exhibit a late-email-response bias (Giurge & Bohns, 2021), a deadline-request bias (Whillans et al., 2022), a declining-invitations bias (Givi & Kirk, 2024), an asking-sensitive-questions bias (Hart et al., 2021), a giving-critical-feedback bias (Abi-Esber et al., 2022), a speaking-honestly bias (Levine & Cohen, 2018), a constructive-confrontation bias (Dungan & Epley, 2024), a negotiating-job-offer bias (Hart et al., 2024), a poor-performance bias (Savitsky et al., 2001), a revealing-negative-information bias (Kardas et al., 2024), a regifting bias (Adams et al., 2012), a giving-late-gifts bias (Haltman et al., 2025), and a disengaging-from-a-passion-pursuit bias (Berry et al., 2025). Conversely, people appear to underestimate how positively others will judge them for small positive actions. They exhibit a compliment bias (Boothby & Bohns, 2021; Zhao & Epley, 2021a, 2021b), a small-acts-of-kindness bias (Kumar & Epley, 2023), a gratitude bias (Kumar & Epley, 2018), a reaching-out-to-an-old-friend bias (Liu et al., 2023), a giving-social-support bias (Dungan et al., 2022), and an allyship bias (Birnbaum et al., 2024).

This literature, which continues to document new and nominally distinct biases, not only describes people's apparent misunderstandings of their social world but also prescribes what they should do to overcome them. For example, Abi-Esber et al. (2022) advise, "The next time you hear someone mispronounce a word, see a stain on their shirt, or notice a typo on their slide, we urge you to point it out to them—they probably want more feedback than you think" (pp. 1383-1384). In a managerial setting, Birnbaum et al. (2024) suggest that "Organizations may also consider programs or trainings that highlight how allyship generates appreciation" (p. 13). And, Savitsky et al. (2001) write, "Our message, then, in its simplest form, is one of liberation: People can put aside some of their concern

about others' reactions when deciding what choices to make in life because the audience they are so concerned about is seldom as judgmental and attentive as they believe" (p. 55).

Indeed, this literature offers no shortage of advice. People are told to compliment strangers more often (Boothby & Bohns, 2021), to worry less about giving late gifts or sending late replies (Giurge & Bohns, 2021; Haltman et al., 2025), to request deadline extensions (Whillans et al., 2022), to reach out to old friends (Liu et al., 2023), to refuse more requests (Lu et al., 2023), and so on. In general, people are presumed to overestimate the negative impact of their bad actions and to underestimate the positive impact of their good ones (see Elsaadawy & Carlson, 2022, for a review), and they are advised to adjust accordingly.

In this paper, we add to the growing list of biases that appear to distort people's inferences about how others will judge their actions. For instance, we document a mentioning-someone's-work-in-the-Glorpatch-forum bias, in which people underestimate how grateful others would be if their work is mentioned in an alien forum; a chewing-up-a-pillow bias, in which people, imagining themselves as dogs, overestimate how upset their owners would be if they chewed up a pillow; and a landing-on-a-frog's-head bias, in which people, imagining themselves as acorns, overestimate how upset a frog would be if they landed on its head.

But, as we will show, the fact that these biases emerge even in such foreign or fantastical contexts is neither a testament to their pervasiveness nor grounds for further prescriptions. It is instead the sign of a problem: The methods used to establish these effects contain built-in confounds that can generate evidence for biases that may not actually exist. Ask about the positive impact of any small deed, and the methods will make it look like people underestimate its influence. Ask about the negative impact of any small infraction, and the methods will make it look like people overestimate its influence. For

almost any conceivable action, the method alone will generate the appearance of a bias. To produce 30 more biases, one simply needs to apply the method to 30 more actions.

Action-Based Metaperceptions: What I Think Other People Think About My Actions

Metaperceptions are our beliefs about what other people think of us. They can concern our traits (e.g., how competent we think others believe we are), our relationships (e.g., how close we think someone feels to us), our interactions (e.g., how much we think someone liked us during a conversation), or our actions (e.g., what we think others think about something we did or did not do).

In this article, we focus on *action-based* metaperceptions, forecasts of how others will respond to our behavior, such as how much others will appreciate our compliments or how harshly they will judge us for our mistakes.[1] We focus on this class of metaperceptions for a number of reasons. First, this area of work heavily relies on the methodological approach that we critique here: interpreting differences between predictions and reports as evidence of predictors' miscalibration. Second, this is a large and rapidly growing literature, with more than 20 papers published in the past four years (since 2021), most in top-tier outlets (see Table S1 in the Supplementary Online Materials [SOM]). And collectively this literature has been taken as evidence that people are chronically poor judges of how others see their actions, provided the foundation for the claim that people are undersocialized (Epley et al., 2022; Pei et al., 2025), and used to recommend changes in how people should think and behave. Because this literature is both theoretically influential and practically prescriptive, it is essential to know whether its empirical foundations are sound.

We contend that they are not.

Metaperception Bias or Jerk Avoidance?

[1] For simplicity, we use *metaperceptions* to refer to the action-based kind, adding "action-based" only when distinguishing this class from the broader literature.

To understand the problems plaguing research on biases in metaperceptions, consider the standard design used to study them. One group of participants—the predictors—predicts how others would respond to their actions, while another group—the reporters—reports how they would or do actually respond. For example, predictors might be asked, "How upset would your colleague be at you for making a mistake?" while reporters are asked, "How upset would you be at your colleague for making a mistake?" Or predictors might be asked, "How grateful do you think this other person would be for the nice thing you did for them?" while reporters are asked, "How grateful would you be for this nice thing someone did for you?" A discrepancy between predictions and reports is taken as evidence of a metaperception bias, a systematic miscalibration between how we think others will react to our actions and how they actually do.

At first glance, this method may seem straightforward and unobjectionable. But it violates a core principle of experimental design: that only one thing should vary across conditions. If you are testing whether pen color affects penmanship, the pens must be identical except for color. If the red pen is a ballpoint and the blue pen is a highlighter, any difference in handwriting could be due to either pen color or pen type. The study would be confounded, and any observed difference uninterpretable.

The dominant approach to studying biases in metaperceptions suffers from the same problem. In these studies, what differs between conditions is not only the participant's role – predictor or reporter – but also the question they are asked. Predictors might be asked, "How much will others appreciate what you did?" while reporters are asked, "How much do you appreciate what was done for you?" These are not the same questions. And so the role that participants adopt – predictor or reporter – is confounded with the question they are asked.

This confounding of question with condition is a problem because people's answers to questions are not always perfectly accurate or truthful, and the ways in which they are inaccurate or untruthful

may vary across questions, and thus across conditions. Different questions can invoke different social desirability concerns, conversational norms, levels of confusion, scale (mis)interpretations, and/or response biases (e.g., Grice, 1975; Krosnick, 1999; Krosnick et al., 1996; Krumpal, 2013; Orne, 1962; Paulhus, 1991; Schwarz, 1999; Tedeschi et al., 1971; Thomas & Kyung, 2019; Zorowitz et al., 2023). As a result, observed differences between conditions may reflect discrepancies in how people interpret these questions or distort their answers to these questions rather than genuine metaperception biases.

Indeed, we contend that predictors and reporters usually face different social pressures regarding what counts as an acceptable answer. When people are asked to predict how others will react, they often feel pressure to appear modest, humble, or accountable. When people are asked to report their own reactions, they feel pressure to appear kind, polite, or forgiving. These opposing pressures shape people's scale responses, irrespective of their true perceptions or beliefs.

In this paper, we refer to these forces as *response pressures*, defined as pressures to respond like a "good" person should or to avoid responding like a "bad" person would (e.g., Krumpal, 2013).[2] We show that the different questions asked in different conditions elicit different response pressures, and that those pressures alone can produce the pattern of results that is taken as evidence for biases in metaperceptions. Just as you cannot tell whether penmanship differences arise from pen color or pen type, you cannot tell whether differences in appreciation ratings are caused by role or response pressures. What appear to be biases may just be (some) people trying to avoid responding like a jerk.

[2] A large literature suggests that people "distort their answers towards the social norm in order to maintain a socially favorable self-presentation" (Krumpal, 2013, p. 2028; see also Crowne & Marlowe, 1964; Paulhus, 1991). Indeed, this distortion isn't done merely to impress others, but also "to maintain a positive self-image, to maximize self-worth and to reduce cognitive dissonance resulting from divergence between social norms, self-perception and self-demands on the one hand, and reality on the other hand" (Krumpal, 2013, p. 2030). Thus, even when providing confidential responses in the privacy of their own homes, (at least some) people may distort their answers so as to comply with social norms and expectations.

Figure 1 illustrates how this response pressure confound can play out. Suppose you are asked to imagine paying someone a compliment or giving them a small gift, and you are assigned to the *prediction* condition (Figure 1, Panel A). When asked, "How grateful do you think the recipient will be for the nice thing you did?" (1 = *not at all*; 5 = *extremely*), you may hesitate to select "extremely," as claiming that your simple compliment could really make someone's day may appear self-centered, egotistical, or presumptuous. In this circumstance a norm of modesty may pressure some participants away from indicating that their actions would have a large impact on someone else, even if they truly believe that they would.

Now consider the *reporting* condition (Figure 1, Panel B). Suppose you are asked to imagine that someone gave you a compliment or a small gift, and you are then asked, "How grateful are you for what the person did?" (1 = *not at all*; 5 = *extremely*). You might avoid selecting a low level of gratitude, as doing so could appear rude, ungrateful, or impolite. In this circumstance, social norms may lead some participants to express their appreciation, even if they don't feel especially moved by the gesture. In other words, response pressures in the reporting condition may push some people to exaggerate their feelings of gratitude.

Similar response pressures may shape people's responses to negative actions. Imagine you make a small mistake at work that inconveniences your colleague (Figure 1, Panel C). When asked to make a *prediction* – "How upset would your colleague be for what you did?" (1 = *not at all*; 5 = *extremely*) – you may avoid selecting "not at all," because doing so could make you appear unaccountable or unwilling to take responsibility for your actions. Conversely, if you are the target of that mistake (Figure 1, Panel D) and you are asked to *report* your feelings – "How upset would you be for your colleague's mistake?" (1= *not at all*; 5 = *extremely*) – you may avoid indicating that you are extremely upset, for that may appear rude or unforgiving.

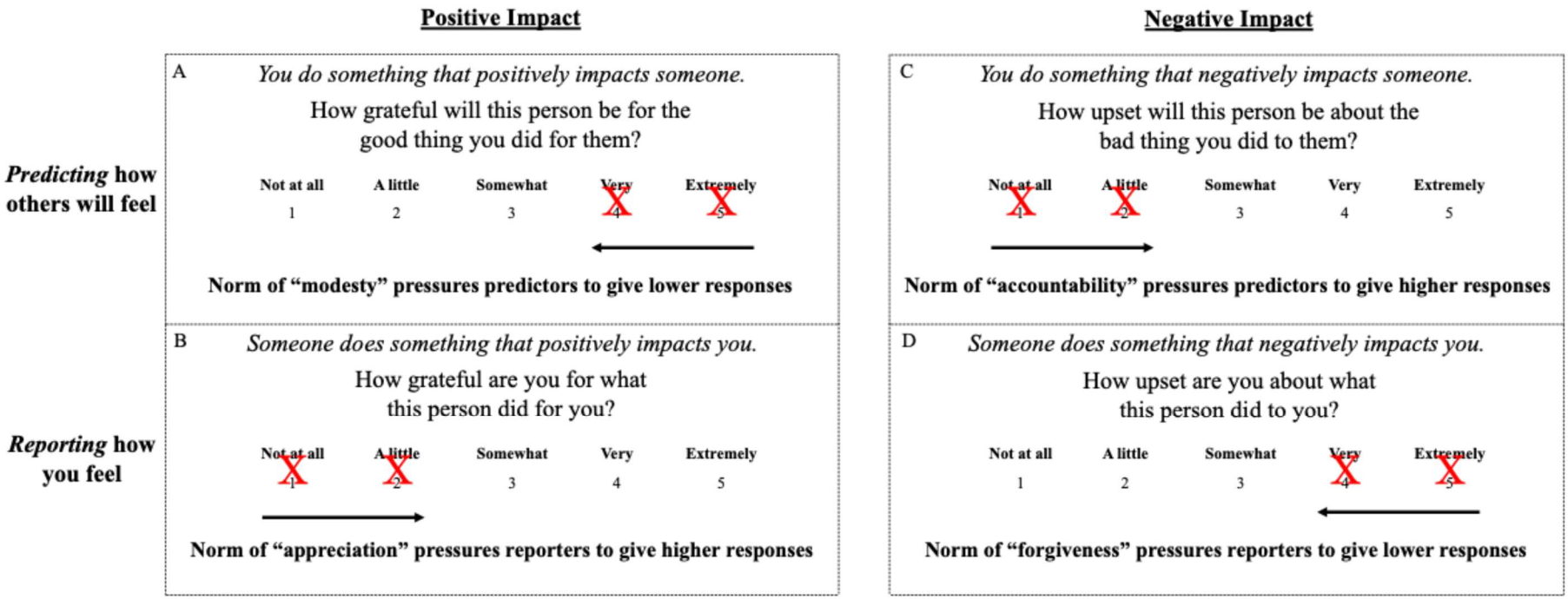
Taken together, these examples illustrate how response pressures may differ between conditions that ask people to predict the impact of their actions on others and conditions that ask people to report the impact of others' actions on themselves. For positive actions, response pressures push predictions downward ("It'll have a small impact") and self-reports upward ("It had a big impact"). For negative actions, the pattern reverses. Response pressures push predictions upward ("It'll have a big impact") and self-reports downward ("It had a small impact").[3] And so response pressures alone can generate results that look like biases in metaperceptions.

Importantly, apparent biases can emerge even if only some of the participants experience these pressures. Consider, for example, a 200-person experiment, in which there is truly no difference between predictors and reporters – they all feel an average of '3' on a 5-point scale – but in which 10% of participants act on these pressures, so that pressured predictors answer '5' and pressured reporters answer '1'. In this case, the average prediction will be 3.2 and the average report will be 2.8. So even with just 10% of participants responding in a pressured way, it will appear as though predictors overestimate reporters' unhappiness. Thus, whenever a context elicits *some* level of socially pressured responding, the standard method can generate evidence of biases in metaperceptions, even when no such biases exist.

[3] This narrative is almost surely a bit oversimplified. For example, for positive actions, pressured predictors may avoid both the high end of the scale – so as not to appear immodest – and the very low end – so as to not impolitely suggest that the recipient will be completely ungrateful. For negative actions, pressured predictors may also avoid both the low end of the scale – so as not to appear unaccountable – and the very high end – so as to not impolitely suggest that the target of their actions will be overly harsh. All told, predictors may be pressured to give more moderate responses, while reporters may be pushed by social norms toward one end of the scale.

Figure 1

How response pressures can differentially influence how predictors vs. reporters answer questions in metaperception studies.



We Don't Actually Know Reporters' True Inner States

In metaperception research, predictors' estimates are compared to reporters' responses, and any discrepancy is presumed to reflect error on the part of the predictors. Indeed, a core assumption of this literature is that predictors' estimates are fallible, whereas reporters' responses provide the unbiased ground truth. But as outlined in Figure 1 – and as our findings later demonstrate – response pressures can bias the answers of both predictors and reporters. As a result, reporters' responses cannot be taken at face value, and so they cannot serve as a benchmark against which to judge whether predictors' forecasts are right or wrong.

Consider a simple example. Suppose you think I will be "quite upset" with you for missing a meeting, and the truth is that I am in fact "quite upset." But when I am asked how upset I am, I want to seem forgiving and reasonable, and so I say that I am "not upset at all." Because the metaperception literature takes reporters' responses as the unbiased truth, researchers may conclude that you are miscalibrated, that you overestimated how bad it is to miss meetings, and they may suggest that you should worry a bit less about missing future meetings. But that is wrong, because in this example your forecast was right. It was bad to miss the meeting, but I just didn't say so. The bias was in my report, not in your prediction.

Of course, the idea that self-reports do not perfectly reflect people's inner states is not new. Psychologists have long known that self-reports can be shaped by social and experimental pressures, and researchers have developed tools to try to mitigate these problems – for example, the bogus pipeline (Jones & Sigall, 1971), the bona fide pipeline (Fazio et al., 1995), and the implicit association test (Greenwald et al., 1998). Yet, in the study of biases in metaperceptions – in the comparison between predictions and reports – these concerns are forgotten or neglected. Reporters' ratings are taken as fact, and predictors are judged against them.

Claims about judgmental inaccuracy and bias require an accurate and unbiased criterion. But reporters' self-reports about their inner states cannot be trusted to provide one. As a result, researchers cannot use these reports as the criterion against which to evaluate predictors' estimates. A discrepancy between predictions and reports cannot be taken as evidence that predictions are biased. Predictions may be biased one way, biased the opposite way, or not biased at all.

Contributions of the Current Research

Psychologists have long known that small shifts in wording, framing, or response format can systematically shape how people answer questions. That fact is not new. What has been underappreciated is what this implies for research traditions that rely on comparing responses to different questions across experimental conditions, as is the case for the study of action-based metaperceptions. If the questions differ in ways that inadvertently alter participants' responses, then any apparent effect may be nothing more than an artifact of those differences. In such designs, the confounding of questions and condition can create the very pattern that is later interpreted as a meaningful psychological result.

Psychologists also know that self-reports do not provide a transparent window into internal states. That, too, is well established. What has not been recognized is how serious this limitation becomes when those same self-reports are treated as the ground truth for evaluating accuracy. Yet this is exactly what the metaperception literature does. It treats reporters' answers – answers potentially shaped by social pressures, conversational norms, and measurement demands – as the objective benchmark against which predictors' "bias" is judged. If reporters' responses are themselves biased, then any gap between predictions and reports may reflect biased reports rather than biased predictions. In other words, the dominant method relies on an accuracy criterion that is known to be inaccurate.

These concerns are somehow not part of the metaperception canon. None of the most recent integrative reviews (e.g., Donnelly et al., 2022; Elsaadawy & Carlson, 2022; Elsaadawy et al., 2023; Epley et al., 2022; Grutterink & Meister, 2022; Vorauer et al., 2025) mention the methodological vulnerabilities we document here. None raise the possibility that predictor–reporter differences might arise because different questions are asked in different conditions, or because reporters' answers may themselves be shaped by social pressures. In these reviews, discrepancies between predictions and reports are consistently interpreted as predictors' mistakes, while reporters' judgments are treated as valid reflections of how they actually feel. In short, the validity of comparing predictors' and reporters' responses has gone unexamined in the field's recent major summaries.

Overall, our results show that metaperception studies may not at all be measuring how well people understand others. They may instead be measuring the correspondence between two socially shaped responses, a difference that can be mistaken for a psychological bias.

As we explain in the General Discussion, these issues are not unique to action-based metaperceptions. Any paradigm that varies the questions across conditions risks consequentially confounding question with condition, and any paradigm that treats one self-report as "the truth" and another as "the bias" inherits the same vulnerability. This does not mean that the response pressures we document will always occur, but it does mean that observed differences may reflect properties of the questions or reporting demands rather than the psychological processes under study.

Research Overview

We conducted 13 studies to investigate whether apparent biases in metaperceptions may arise simply because different experimental conditions elicit different response pressures. We organized our studies in two parts.

Our first set of studies pits our response-pressure explanation against the prevailing assumption that people exhibit numerous, act-specific metacognitive biases. If apparent biases in metaperceptions stem from the chronic differences in the pressures faced by predictors and reporters, then they should arise regardless of the act. In Studies 1a – 3, we find exactly that: systematic predictor-reporter differences emerge even for acts that are fantastical, ambiguous, or neutral – acts for which real biases seem unlikely to exist. These results cast doubt on the idea that people exhibit a long list of distinct, act-specific metaperception biases and instead suggest that methodological confounds alone may be sufficient to produce them.

Studies 4a – 6b test the response-pressure account directly. In Studies 4a and 4b, we manipulated response pressures by asking participants to answer as either very polite or very impolite individuals. We found that the standard predictor-reporter difference was larger when participants were told to answer politely, and fully reversed when told to answer impolitely. This result demonstrates that what counts as a "polite" response differs between predictors and reporters, and that merely attempting to be polite can generate apparent biases in metaperceptions.

In Studies 5a and 5b, we asked observers to evaluate a predictor or reporter based on how they supposedly responded to questions about a positive or negative act. We found that observers evaluated identical answers differently when they were framed as coming from a predictor or a reporter. Finally, in Studies 6a and 6b, we compared cases in which predictors and reporters faced divergent response pressures – pressures that pushed them toward different responses – to cases in which their pressures were more aligned – pressures that pushed them to give similar responses. When response pressures were more aligned, the standard "biases" in metaperceptions no longer appeared.

Before turning to the studies, it is important to emphasize what our work can and cannot show. We do not – and cannot – prove that there are no true biases in metaperceptions. What we show is that the

methods used to document such biases contain inextricable confounds, and that these confounds render prior research inconclusive. We do not – and cannot – know whether the biases that emerge are fully due to confounds or whether they are manifestations of something real. The biases may exist. Or they may not. But with current approaches, and without an unbiased standard against which to evaluate people's predictions, we simply cannot tell.

Transparency and Openness

We preregistered all studies and report all manipulations and measures. The preregistrations, materials, data files, and analysis code can be found on ResearchBox: https://researchbox.org/4201. Table S2 in the SOM details the preregistered exclusions for each study. We used ChatGPT to assist with editing the manuscript and with generating and refining some of the code used to analyze our data.

Studies 1a-1d: The Alien Communication Failure Bias, the Dog Chews Pillow Bias, and the Acorn Lands on Frog Bias

The purpose of Studies 1a–1d was to provide an initial test of the idea that the questions themselves may be producing the predictor–reporter difference, by examining whether that difference emerges regardless of the act under consideration. If the same apparent bias appears across wildly different behaviors, this would suggest that the phenomenon may not be act-specific, as the literature presumes. Our account, in contrast, implies that acts may be largely interchangeable. If the apparent bias is generated by a response-pressure confound that is built into the standard method, then it should emerge no matter which act is being studied, whether it is failing to reply to a friend's text or failing to communicate with aliens.

In our first three studies (Studies 1a-1c), participants either predicted others' reactions or reported their own in a variety of uncommon scenarios, such as how upset a frog would be if an acorn landed on its head. We then compared these patterns to those observed in more conventional contexts, such as

predicting how a colleague would react to critical feedback (Study 1d). Observing the same predictor-reporter difference across such disparate behaviors would support the conclusion that these apparent biases arise from chronically present methodological confounds rather than from psychologies that are specific to any particular act.

Method

Studies 1a-1d shared a similar design. In each study, adult participants recruited from Mechanical Turk (Studies 1a, 1c, and 1d) or Prolific (Study 1b) either predicted how negatively someone would judge them for a bad action or reported how negatively they would judge someone for that same action. Table 1 reports the scenarios and questions used in each condition. Studies 1a, 1c, and 1d used three items to measure predicted and reported negative judgments (Cronbach's α = .96, .95, and .95, respectively). Study 1b used a single question. Except for Study 4a, we did not collect demographic information for any studies reported in this manuscript.

Table S2 in the SOM details the preregistered exclusions for each study. The sample sizes reported below are those obtained after applying these exclusions.

In Study 1a (N = 315), participants imagined that a linguist was contracted by the government to communicate with aliens, but that the attempt was unsuccessful. Those in the *prediction* condition took the linguist's perspective and predicted how negatively a government official would judge them. Those in the *report* condition took the official's perspective and reported how negatively they would judge the linguist.

In Study 1b (N = 411), participants in the *prediction* condition took the perspective of a dog who chewed up a pillow and predicted how upset their owner would be. Participants in the *report* condition imagined being the owner and reported how upset they would be.[4]

[4] In Study 1b, we preregistered to exclude participants who spent less than 40 seconds on the survey. In doing so, we failed to appreciate just how short the survey was, as 70% of our participants completed the survey in less time than that. To avoid

In Study 1c ($N = 389$), participants in the *prediction* condition imagined being an acorn that fell on a frog's head and predicted how negatively the frog would judge them. Those in the *report* condition imagined being the frog judging the acorn.

Finally, in Study 1d ($N = 577$), participants considered three ordinary transgressions: failing to reply to a friend's text, criticizing a colleague's presentation, or forgetting to pick up a neighbor's package (which was subsequently stolen). As in prior studies, some predicted how negatively the victim would judge them, while others reported their own judgment as the victim.

Results

Table 2 reports the descriptive statistics by condition for each study or, in the case of Study 1d, for each scenario. Table 2 also presents the results of independent samples t-tests comparing how negatively people predicted they would be judged (*prediction* condition) with how negatively participants reported they would judge others (*report* condition).

Across all scenarios – whether fanciful or mundane – the findings were the same (see Figure 2). Participants consistently predicted they would be judged more harshly than others reported judging them. This held for failing to communicate with aliens, chewing up an owner's pillow, and falling on a frog's head. And it also held for not texting back a friend, criticizing a colleague's presentation, and forgetting to pick up a neighbor's package that was stolen. The effect sizes ranged from moderate to large for both the fanciful scenarios and the mundane scenarios.

---

discarding so many observations, we decided to deviate from our preregistration and to ignore this exclusion rule in our reporting of the results in Table 2 and Figure 2. Importantly, excluding these participants does not meaningfully alter the results: Dogs predicted that their owners would be more upset ($M = 4.55$, $SD = 1.52$) than their owners said they would be ($M = 3.64$, $SD = 1.47$), $t(123) = 3.40$, $p = .0008$, $d = 0.61$.

Table 1

Scenarios and measures used in Studies 1a-1d

Study	Predict Condition	Report Condition
1a	Scenario There is an unusual development. Aliens have traveled from their home planet to earth. It is unclear their intentions or goals. Finding a way to communicate with them is vital. You are an expert in linguistics and languages. There is a group of government and military officials who have tasked you with finding a way to communicate with the aliens. They have asked you to use your skills with complex written languages to find a way to communicate with the aliens. You ultimately fail to find a way to communicate with the aliens. Measures 1. To what extent would the members of the group of government and military officials form a negative impression of you? 2. To what extent would the members of the group of government and military officials judge you negatively? 3. How much would the members of the group of government and military officials dislike you?	Scenario There is an unusual development. Aliens have traveled from their home planet to earth. It is unclear their intentions or goals. Finding a way to communicate with them is vital. You are part of a group of government and military officials who have tasked someone with finding a way to communicate with the aliens. This person is an expert in linguistics and languages. You have asked this person to use their skills with complex written languages to find a way to communicate with the aliens. This person ultimately fails to find a way to communicate with the aliens. Measures 1. To what extent would you form a negative impression of this person? 2. To what extent would you judge this person negatively? 3. How much would you dislike this person?
1b	Scenario Imagine that you are a dog. Your owner goes out to dinner, and while the owner is out you chew one of the owner's pillows so that it now has a small hole in it. Measure 1. How upset do you think your owner will be at you when the owner comes home from dinner and sees what you've done?	Scenario Imagine that you are the owner of a dog. You go out to dinner, and while you are out your dog chews one of your pillows so that it now has a small hole in it. Measure 1. How upset would you be at your dog when you come home from dinner and see what the dog has done?
1c	Scenario You are an acorn. There is a frog laying quietly under the tree. You fall from a high branch on a tree and you hit the frog on the head. Measures 1. To what extent would the frog form a negative impression of you? 2. To what extent would the frog judge you negatively? 3. How much would the frog dislike you?	Scenario You are a frog. You are laying quietly under a tree. There is an acorn on a high branch of the tree. The acorn falls. It hits you on the head. Measures 1. To what extent would you form a negative impression of the acorn? 2. To what extent would you judge this acorn negatively? 3. How much would you dislike this acorn?
1d	Scenarios *Scenario 1:* A friend texts you about a movie you recommended. You forget to text them back. *Scenario 2:* You gave your colleague critical, negative feedback on their presentation. *Scenario 3:* Your neighbor asked you to pick up an Amazon.com package from their porch while they were away for a few nights. You forgot to pick up the package and the package was stolen. Measures 1.To what extent would your [friend/colleague/neighbor] form a negative impression of you? 2. To what extent would your [friend/colleague/neighbor] judge you negatively? 3, How much would your [friend/colleague/neighbor] dislike you?	Scenarios *Scenario 1:* You text a friend about a movie they recommended. Your friend forgets to text you back. *Scenario 2:* Your colleague gave you critical, negative feedback on your presentation. *Scenario 3:* Your asked your neighbor to pick up an Amazon.com package from your porch while you were away for a few nights. They forgot to pick up the package and the package was stolen. Measures 1. To what extent would you form a negative impression of your [friend/colleague/neighbor]? 2. To what extent would you judge your [friend/colleague/neighbor] negatively? 3. How much would you dislike your [friend/colleague/neighbor]?

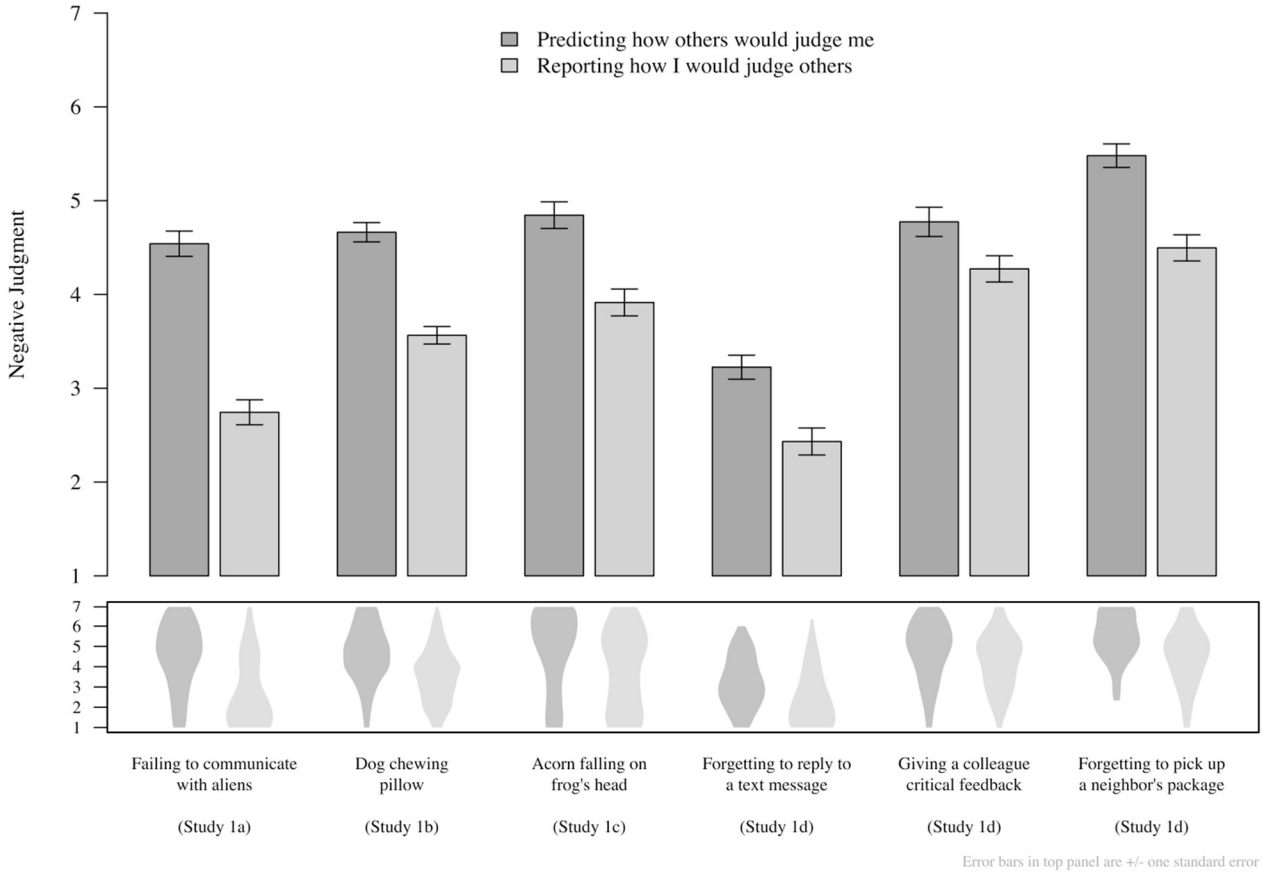
Table 2

Descriptive statistics and results from independent samples t-tests for Studies 1a – 1d

Study	Behavior	Condition	*N*	M (SD)	t(df)	*p*	Cohen's *d*
1a	Failing to communicate with aliens	Predict	158	4.54 (1.69)	t(312) = 9.50	< .001	1.07
		Report	156	2.74 (1.66)			
1b	Dog chewing pillow	Predict	202	4.66 (1.46)	t(409) = 7.92	< .001	0.78
		Report	209	3.56 (1.35)			
1c	Acorn falling on a frog's head	Predict	194	4.85 (1.98)	t(387) = 4.62	< .001	0.47
		Report	195	3.91 (2.00)			
1d	Forgetting to reply to a text message	Predict	98	3.22 (1.27)	t(187) = 4.13	< 001	0.60
		Report	91	2.43 (1.37)			
1d	Giving a colleague critical feedback	Predict	93	4.77 (1.50)	t(190) = 2.40	.02	0.35
		Report	99	4.27 (1.40)			
1d	Forgetting to pick up a neighbor's package	Predict	91	5.48 (1.20)	t(193) = 5.18	< .001	0.74
		Report	104	4.50 (1.42)			

Figure 2

Predicted versus reported negative judgments in Studies 1a – 1d



Discussion

Across Studies 1a–1d, the results followed the same pattern. Participants consistently predicted that others would judge them more negatively for a transgression than others said they would. This pattern emerged regardless of the content of the act – whether it involved being an acorn that fell on a frog's head or failing to text a friend back about a movie recommendation. This consistency is striking, and supports the hypothesis that condition differences in the questions people are asked, and thus the response pressures they face, may be sufficient to generate patterns of results that look like biases in metaperceptions.

Much of the existing literature interprets predictor–reporter differences as evidence of behavior-specific biases. For example, it suggests that people think they'll be judged more harshly for asking sensitive questions or giving critical feedback. But if the same pattern emerges across a wide range of negative acts, including some fantastical ones, that interpretation becomes more difficult to defend. The findings from Studies 1a–1d suggest that predictor–reporter differences do not hinge on the dynamics of any particular situation. Instead, they point to something more general at play, something that is not tightly linked to the specific act being evaluated.

What might that "something" be? We propose that it is, at least in part, a methodological artifact: that the predictor–reporter difference arises because the two roles carry different, and opposing, response pressures. These pressures can generate a pattern of results that looks like a bias in metaperception but is, in fact, produced by the nature of the paradigm itself.

Study 2a and 2b: The Glorpatch Biases

In Studies 1a-1d, we observed predictor-reporter differences across a wide range of behaviors, but they were all readily identifiable as negative acts. In Studies 2a and 2b, we go further by examining a made-up behavior that lacks an inherent valence, with no clear positive or negative meaning. This

design allows us to hold constant the act itself while varying – across studies – whether participants are asked about the act's positive impact or its negative impact.

According to our account, when you ask about the positive impact of a behavior, response pressures will push predictors' answers lower (not wanting to seem immodest) and reporters' answers higher (not wanting to seem unappreciative). Conversely, when you ask about the negative impact of a behavior, response pressures will push predictors' answers higher (not wanting to seem unaccountable) and reporters' answers lower (not wanting to seem unforgiving).

Studies 2a and 2b test this logic using a single, made-up behavior. Participants imagined participating in an alien "glorpatch" forum in which they either mentioned someone else's work or had their own work mentioned. This could be interpreted as a positive act, like a public accolade that evokes gratitude, or it could be seen as a negative act, like a public admonishment, that makes someone upset. We did not expect participants to spontaneously imbue the act with meaning, but that the meaning of the act would turn on the kinds of questions they were asked. In Study 2a, we asked about positive impact – predictions and reports of gratitude – whereas in Study 2b, we asked about negative impact – predictions and reports of negative judgment. Because the act itself is identical across studies, observing the standard predictor-reporter difference in both cases would demonstrate that the apparent biases can be produced solely by altering the evaluative question (i.e., "How grateful…?" vs. "How negatively…?"). Not necessarily because the biases are real, but because the questions themselves elicit response pressures that push predictors and reporters in opposite directions.

Method

Studies 2a and 2b followed the same structure. In both studies, adult participants recruited from Mechanical Turk read the following scenario: ″You are an alien living on the planet Zorplon. You speak

the alien language Glorvish and reside in a zebbth, which is like a community or neighborhood. Your zebbth has a glorpatch forum, which functions as a community meeting.″

Participants were then randomly assigned to one of two conditions. In the *prediction* condition, participants learned that they had mentioned another alien's work in the previous glorpatch forum. In the *report* condition, participants learned that another alien had mentioned their work in the forum.

In Study 2a (N = 393 after preregistered exclusions), participants evaluated gratitude. Predictors were asked, "How grateful would you be that the other alien mentioned your work during the forum?". Reporters were asked, "How grateful would the other alien be for you mentioning their work during the forum?" Participants answered on a 7-point scale (1 = *Not at all*; 7 = *Extremely*).

In Study 2b (N = 598 after preregistered exclusions), the focus shifted to negative judgment. Predictors were asked how negatively they thought the other alien would judge them for mentioning that alien's work in the forum. Reporters were asked how negatively they themselves would judge the other alien for mentioning their work. Negative judgment was measured as the average of the same three items used in Studies 1a, 1c, and 1d (α = .96), each using a 7-point scale.

Results

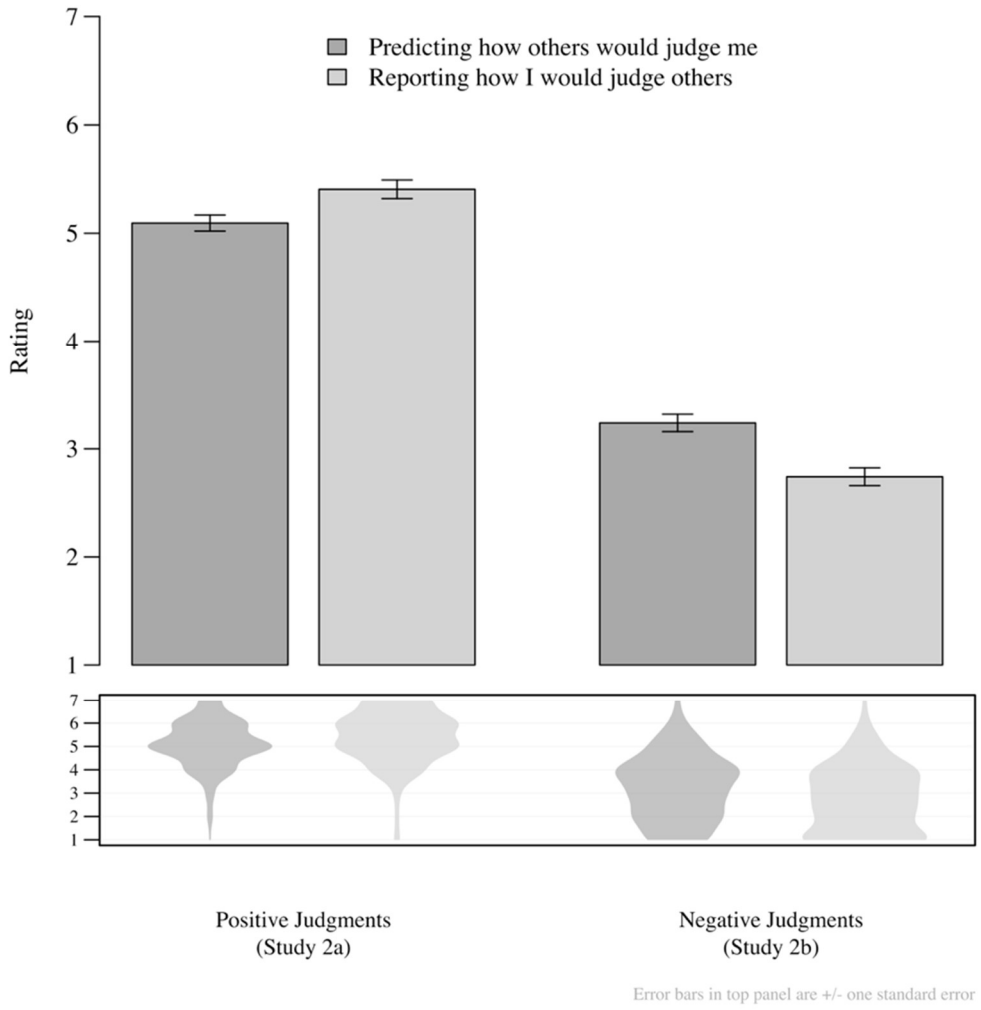
As shown in Figure 3, both studies revealed a glorpatch bias. In Study 2a, reporters said they would feel more grateful for having their work mentioned in the glorpatch forum (M = 5.41, SD = 1.19) than predictors said that they would (M = 5.09, SD = 1.05), $t(391) = 2.75$, p = .006, Cohen's d = 0.28. In Study 2b, reporters said that they would judge others less negatively for having their work mentioned in the glorpatch forum (M = 2.74, SD = 1.42) than predictors said that they would (M = 3.24, SD = 1.40), $t(596) = 4.31$, $p < .001$, Cohen's d = 0.35.

Discussion

The behavior described in Studies 2a and 2b had no inherent positive or negative meaning, yet we nevertheless observed the same predictor-reporter differences that permeate the metaperception literature. When we asked about gratitude, it looked like predictors underestimated the impact of their behavior; when we asked about negative judgment, it looked like predictors overestimated the impact of their behavior. As in Studies 1a–1d, these results argue against act-specific metaperception biases and are fully consistent with their arising from question-induced response pressures, from participants' tendency to adhere to social norms when answering survey questions.

Figure 3

Predicted versus reported gratitude (Study 2a) and negative judgments (Study 2b)



Study 3: LXK versus KXL

Study 3 builds on Studies 2a and 2b by examining a real behavior that participants would clearly recognize as neutral. In Study 3, participants assigned another participant to work with one of two nonsense letter strings that were functionally identical. Because the task was straightforward and clearly devoid of any inherent valence, participants knew exactly what the behavior was and that it had no obvious positive or negative implications for the other person. This allowed us to test whether the same predictor-reporter difference would emerge even when participants had full information about the neutrality of the behavior. If the standard predictor–reporter differences emerge here as well, this would provide further support for the hypothesis that apparent metaperception biases are driven by question-induced response pressures rather than by the intrinsic meaning of the act.

Method

Participants and Decision Task

We recruited adult participants from Mechanical Turk to complete a word-generation task for a monetary bonus (N = 580 after preregistered exclusions). Each participant was assigned a letter string and given 20 seconds to generate three or more words from it in order to earn a $0.10 bonus.

Participants were randomly assigned to one of two roles: Decider or Receiver. Deciders chose between two letter strings - PESQALTUHEPN and ATLHENPQSPEU – and assigned one to themselves and the other to the Receiver. These strings contained the same letters in different orders, ensuring that neither offered an inherent advantage.

Receivers did not make a choice but were informed of the letter string assigned to them. To maintain experimental control, Receivers' strings were randomized rather than yoked to actual Decider choices. This preserved the perception that another participant determined their outcome while ensuring

balanced conditions. After learning about the assignment, both Deciders and Receivers completed the word-generation task to qualify for the bonus.

Measuring Predicted and Actual Reactions

As in prior studies, we randomly assigned participants to different roles: predictor or reporter. Deciders, who were in the *prediction* condition, answered three questions predicting how they thought the Receiver would react to their decision. Receivers, in the *report* condition, answered the same questions but reported their actual reactions to the Decider's choice.

We again varied the valence of the questions used to assess predicted and reported reactions. Participants answered questions framed either positively (as in Study 2a) or negatively (as in Study 2b) regarding the Receiver's reaction to the Decider's decision.

Specifically, Deciders predicted the Receiver's reaction by answering the following questions, with negatively framed items in brackets: "How much will the other participant like [dislike] your decision?"; "How happy [unhappy] will the other participant be with your decision?"; and "To what extent will the other participant think you are a good [bad] person for the decision you made?"

Receivers answered analogous questions about their own reactions: "How much do you like [dislike] the other participant for their decision?"; "How happy [unhappy] are you with the other participant's decision?"; and "To what extent do you think the other participant is a good [bad] person for the decision they made?"

All responses were made on 7-point scales ranging from 1 = *not at all* to 7 = *extremely*, and were averaged together. Reliability was high for both the positively framed items ($\alpha = .92$) and the negatively framed items ($\alpha = .85$).

Analytic Approach

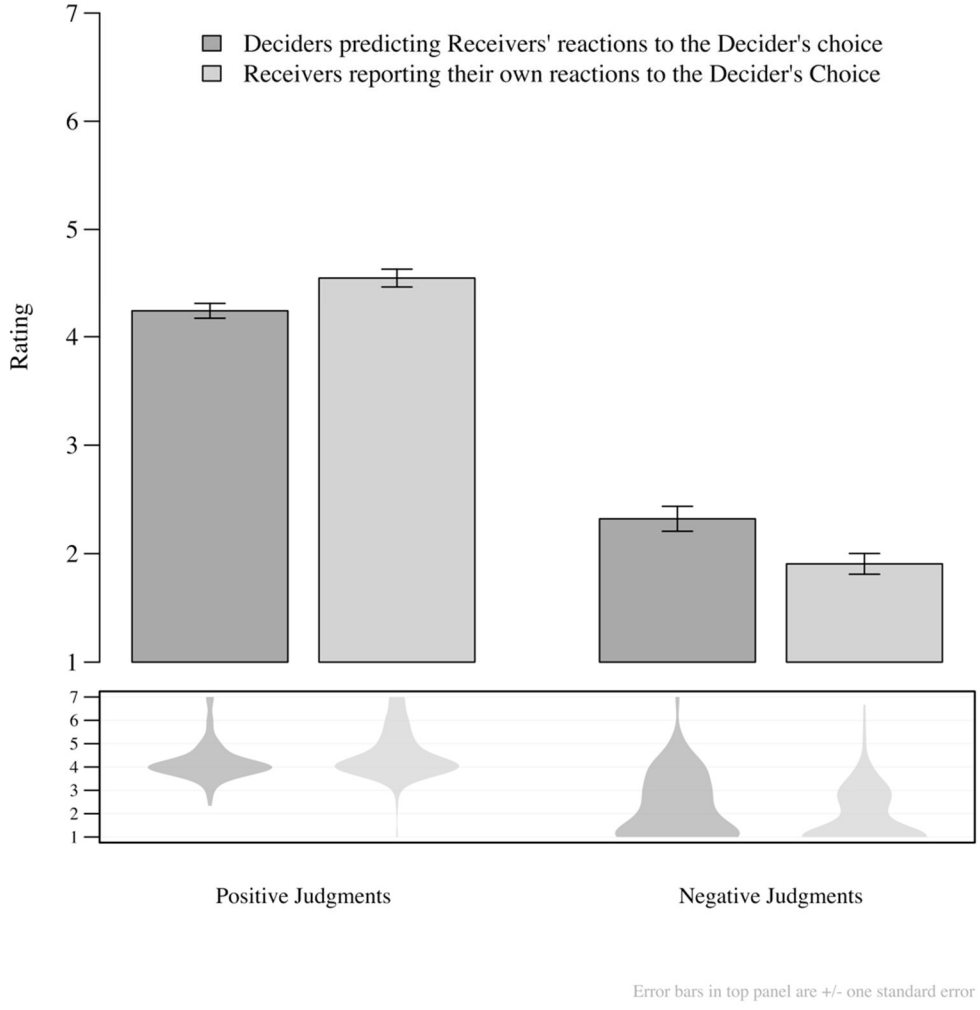
Our preregistered plan was to first recode the negatively framed items so that all responses were on the same scale, then collapse across the framing conditions and test the difference between the Deciders' predictions and the Receivers' reports. Instead, for clarity and for consistency with Studies 2a and 2b - where the positive and negative frames were run as separate studies – we analyze the framing conditions separately in Study 3. For transparency, we report the preregistered analysis in the SOM. This decision was entirely motivated by presentational considerations; the preregistered analyses fully support our hypothesis.

Results

As shown in Figure 4, Study 3 produced another set of apparent metaperception biases. In the positive frame condition, Deciders (n = 141) predicted that their choice would have less of a positive impact on the Receiver (M = 4.24, SD = 0.81) than the Receivers (n = 149) reported experiencing (M = 4.55, SD = 1.00), $t(288) = 2.90$, $p = .004$, Cohen's $d = 0.34$. In the negative frame condition, Deciders (n = 141) predicted that their choice would have a more negative impact on the Receiver (M = 2.32, SD = 1.37) than the Receivers (n = 149) reported experiencing (M = 1.91, SD = 1.16), $t(288) = 2.79$, $p = .006$, Cohen's $d = 0.33$.

Figure 4

Predicted versus reported positive reaction and negative reaction to the Decider's choice.



Discussion

Our key claim is that the questions used to assess metaperception biases may themselves be responsible for producing the observed effects. Because these questions are confounded with experimental condition, differences in how participants interpret, use, or respond to them can, on their own, generate predictor-reporter differences. Since metaperception accuracy is measured by comparing responses across these different questions, discrepancies arising from interpretation or response

tendencies may be misread as psychological bias. We argue that a critical contributor to these discrepancies is response pressure, the pressures that shape how some participants answer questions, independent of their true perceptions.

Studies 1a–3 provide consistent evidence for this claim. We observed the same biased pattern across a wide range of scenarios, from the fanciful (acorns falling on frogs), to the ambiguous (aliens discussing each other's work), to the transparently neutral (assigning a letter string). Across these contexts, predictors appeared to underestimate the positive impact of their good actions and overestimate the negative impact of their bad ones. These results very much support the possibility that apparent biases in metaperceptions are driven by response pressure confounds rather than by genuine miscalibration.

Still, although these studies represent an important test of the response pressure account, they do not, on their own, definitively establish that response pressure is responsible for the observed effects. Even if a response pressure confound is present, it may not be the primary driver of the predictor-reporter difference. For example, people may hold a generalized negativity bias, one not tied to any particular behavior, that leads them to assume they will be judged more harshly than they actually are, regardless of what they have done (see Elsaadawy & Carlson, 2022). Participants may have drawn on familiar experiences of harm, embarrassment, or failure when responding to unfamiliar or unusual scenarios, applying a generalized belief that others tend to judge them harshly.

In the next set of studies (Studies 4a-6b), we more directly test whether differential response pressures operate in this setting, and whether they alone can produce apparent biases in metaperceptions. First, we manipulate response pressures by asking participants to respond as particularly polite or impolite individuals, allowing us to examine whether these pressures alone can generate or eliminate the apparent bias (Studies 4a and 4b). Next, we investigate how third-party

observers evaluate predictors and reporters based on their responses, examining whether what counts as a "likable" response differs for predictors and reporters (Studies 5a and 5b). Finally, we examine whether metaperception biases are larger when predictors and reporters face divergent response pressures and smaller (or eliminated) when their pressures are more aligned (Studies 6a and 6b).

Studies 4a and 4b: Answering as a Saint or as a Jerk

In Studies 4a and 4b, we again assigned participants to the role of a predictor or a reporter, but this time we also manipulated whether they were instructed to respond in more or less socially appropriate ways. Specifically, we asked participants to answer either as themselves (*self* condition), as a very selfless, considerate, and polite person (*saint* condition), or as a very selfish, inconsiderate, and rude person (*jerk* condition). Participants evaluated positive behaviors in Study 4a and negative behaviors in Study 4b.

If the metaperception biases observed in prior research are devoid of response pressure confounds, such that what constitutes a "good" answer for a predictor is the same as what constitutes a "good" answer for a reporter, then these instructions should not alter responses. However, if, as we contend, a "good" answer for a predictor differs from a "good" answer for a reporter, then we should see systematic differences between these conditions. That is, if response pressures influence these effects, then altering those pressures should alter the effects.

In the *jerk* condition, in which participants were instructed to answer as an impolite and selfish person, we expected response pressures to be minimized or reversed. Without the expectation to appear humble or grateful, predictors should be more willing to assert the impact of their own good deeds, and reporters should be less inclined to overstate their appreciation. This should eliminate – or even reverse – the standard metaperception pattern.

In the *saint* condition, the anticipated effects were less straightforward. On one hand, instructing participants to respond as extremely polite and considerate individuals could increase response pressures, potentially amplifying the standard metaperception pattern. On the other hand, because many people already strive to be good and considerate (Aquino & Reed II, 2002; Monin & Jordan, 2009; Wojciszke, 2005), responding as a 'saint' may not meaningfully change their answers compared to responding as themselves. As a result, we expected the *self* and *saint* conditions to be more similar to each other than to the *jerk* condition.

Study 4a

Method

We recruited adult online participants (N = 507 after exclusions[5]) to complete a study examining the perceived positive impact of good deeds.[6] Participants either predicted how positively another person would feel about a good deed they performed (*prediction* condition) or reported how positively they would feel about a good deed someone performed for them (*report* condition). Additionally, we asked participants to adopt different roles when answering these questions. Specifically, they responded either as themselves (*self* condition), as a very considerate and polite person (*saint* condition), or as a very rude and impolite person (*jerk* condition).[7] The exact instructions and questions corresponding to each condition are detailed in Table 3.

---

[5] We do not know why, but a very high percentage of Study 4a participants (44.2%) had duplicate IP Addresses, resulting in a smaller final sample than intended. In the SOM, we report analyses that include these participants. Doing so yields the same overall conclusions, with the only meaningful difference being that the predictor-reporter difference in the *self* condition becomes statistically significant (p = .001). This issue of excessive duplicate IP Addresses occurred only in Study 4a.

[6] Study 4a is the only study in which we collected demographic information. Participants were asked to answer, "What is your age?" (open-ended) and "What is your gender?", with the response options: "Male", "Female", "Non-binary", "Different gender identity" (along with a box to specify their gender identity), and "Prefer not to state." The average age was 43.6, and 238 participants selected "Male", 257 selected "Female", 6 selected "Non-binary", 1 selected "Different gender identity", and 5 selected "Prefer not to state".

[7] Rather than instructing participants to imagine themselves as either a very considerate and polite ("saint") or rude and impolite ("jerk") person, we asked them to respond from the perspective of a character, LN. We chose to do this because asking participants to imagine that they themselves are particularly considerate or inconsiderate could be perceived as insulting if it contradicts their self-view. Similarly, instructing participants to adopt a rude or selfish persona could trigger

The study employed a 2 (Perspective: Predict other's response vs. Report own response) × 3 (Role: Self, Saint, Jerk) between-subjects design. For stimulus-sampling purposes, we varied the type of good deed – small favor, compliment, or constructive feedback – as well as how the saint and jerk roles were described. As preregistered, we collapsed across these variables in the primary analyses.

Results

We conducted a 2 (Perspective: Predict other person's response vs. Report own response) × 3 (Role: Self, Saint, Jerk) between-subjects ANOVA to analyze positive impact ratings of a good deed. There were significant main effects of both perspective and role, which were further qualified by a significant interaction, $F(2, 501) = 161.28$, $p < .001$, $\eta_p^2 = 0.39$.

As illustrated in Figure 5, a simple effects analysis showed that participants answering as themselves exhibited the standard effect, though it was not quite statistically significant: predictors rated their good deed as having less positive impact than reporters said it would, $b = 0.33$, $SE = 0.17$, $p = .059$, $\eta_p^2 = 0.02$. This difference was directionally larger when participants answered as saints, $b = 0.72$, $SE = 0.16$, $p < .001$, $\eta_p^2 = 0.12$,[8] and fully reversed in the jerk condition. Specifically, predictors in the jerk condition rated their good deed as having more positive impact than reporters said it would, $b = -3.12$, $SE = 0.17$, $p < .001$, $\eta_p^2 = 0.62$.

reactance, leading to reactance or disengagement. By having participants answer as someone with these qualities rather than embodying the traits themselves, we sought to create the necessary psychological distance while maintaining engagement and minimizing potential biases in responding.

[8] To test the amplification effect, we ran a two-way Perspective x Role ANOVA that excluded the *jerk* condition. The interaction fell short of statistical significance, $F(1, 330) = 3.07$, $p = .081$, $\eta_p^2 = 0.01$.

Figure 5

The Effect of Perspective Condition (Prediction vs. Report) on Positive Impact Ratings Across Role Conditions (Self, Saint, Jerk) in Study 4a.

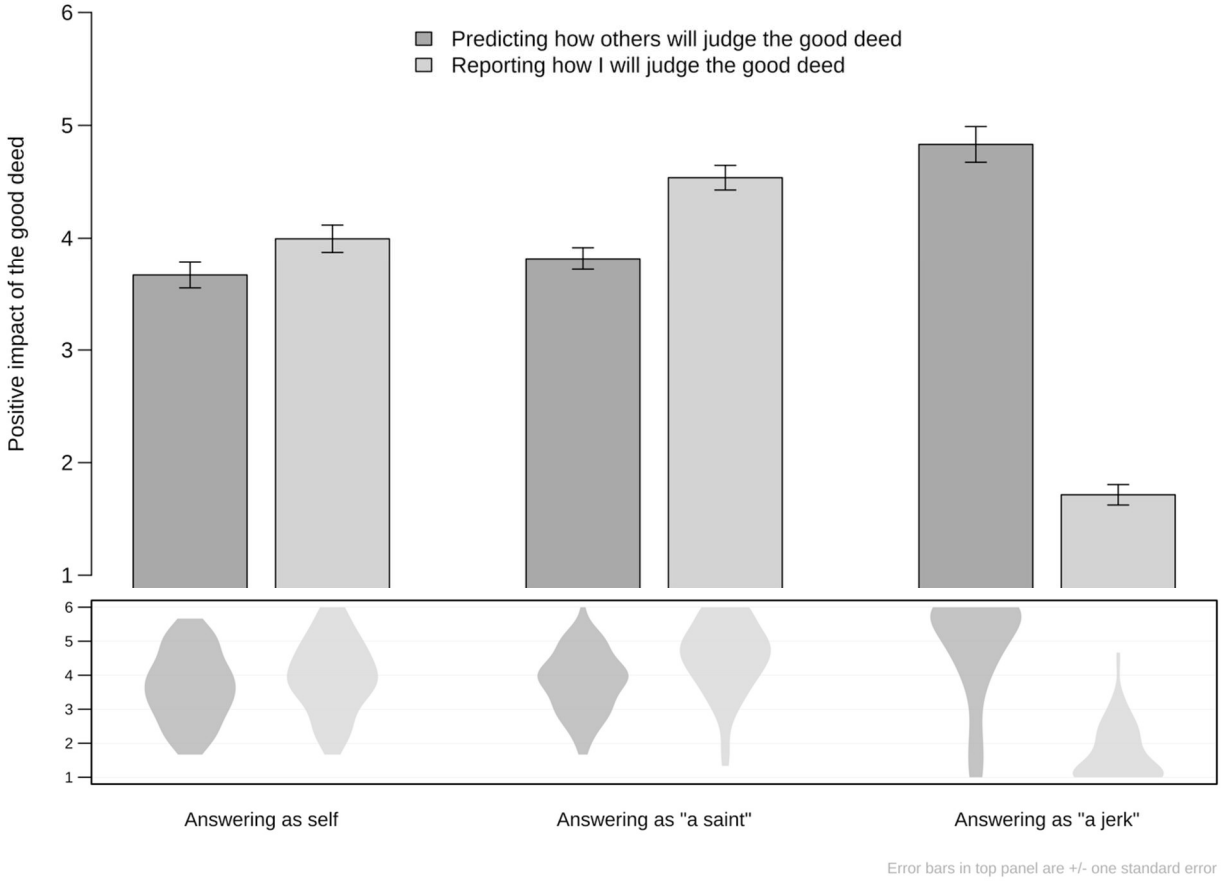


Table 3. Manipulations and measures in Study 4a

Manipulation of Role	**Answer as self**		**Answer as a saint**		**Answer as a jerk**	
	In this study, you will imagine yourself in a scenario. Think about how you would feel and respond in the following situation.		In this study, you will imagine how a certain type of person would respond to a scenario. Imagine the person LN. LN is someone who is humble, sensitive to other people's feelings, and unselfish. LN is respectful and considerate of others.		In this study, you will imagine how a certain type of person would respond to a scenario. Imagine the person LN. LN is someone who is arrogant, insensitive, selfish, and full of themselves. LN is manipulative and inconsiderate of other people.	
Manipulation of perspective	**Predict**	**Report**	**Predict**	**Report**	**Predict**	**Report**
	Imagine you do the following. You do a small favor for a neighbor. Think about how you would feel about the small favor you did.	Imagine the following happens to you. A neighbor does a small favor for you. Think about how you would feel about this small favor.	Imagine LN does the following. LN does a small favor for a neighbor. Think about how LN would feel about the small favor they did.	Imagine the following happens to LN. A neighbor does a small favor for LN. Think about how LN would feel about this small favor.	Imagine LN does the following. LN does a small favor for a neighbor. Think about how LN would feel about the small favor they did.	Imagine the following happens to LN. A neighbor does a small favor for LN. Think about how LN would feel about this small favor.
Measure of positive impact	How grateful will your neighbor be for the small favor you did? How happy will this small favor make your neighbor feel? How positive of an impact will this small favor have on your neighbor's life?	How grateful are you for this small favor? How happy does this small favor make you feel? How positive of an impact does this small favor have on your life?	How grateful does LN think their neighbor will be for the small favor LN did? How happy does LN think their small favor will make their neighbor feel? How positive of an impact does LN think this small favor will have on their neighbor's life?	How grateful is LN for this small favor? How happy does this small favor make LN feel? How positive of an impact does this small favor have on LN's life?	How grateful does LN think their neighbor will be for the small favor LN did? How happy does LN think their small favor will make their neighbor feel? How positive of an impact does LN think this small favor will have on their neighbor's life?	How grateful is LN for this small favor? How happy does this small favor make LN feel? How positive of an impact does this small favor have on LN's life?

Note. We stimulus sampled the way we described the saint and jerk. The other "saint" version was: "LN is someone who is considerate, kind, respectful, and empathetic towards others. LN is generous, grateful, and humble." The other "jerk" version was: "LN is someone who is selfish and boastful. LN is inconsiderate and thinks they are better than other people." We also stimulus sampled the type of positive act. The other acts were giving a compliment to a friend and providing constructive feedback to a colleague. All questions were answered on 6-point scales (1 = *Not at all*; 6 = *Extremely*).

Study 4b

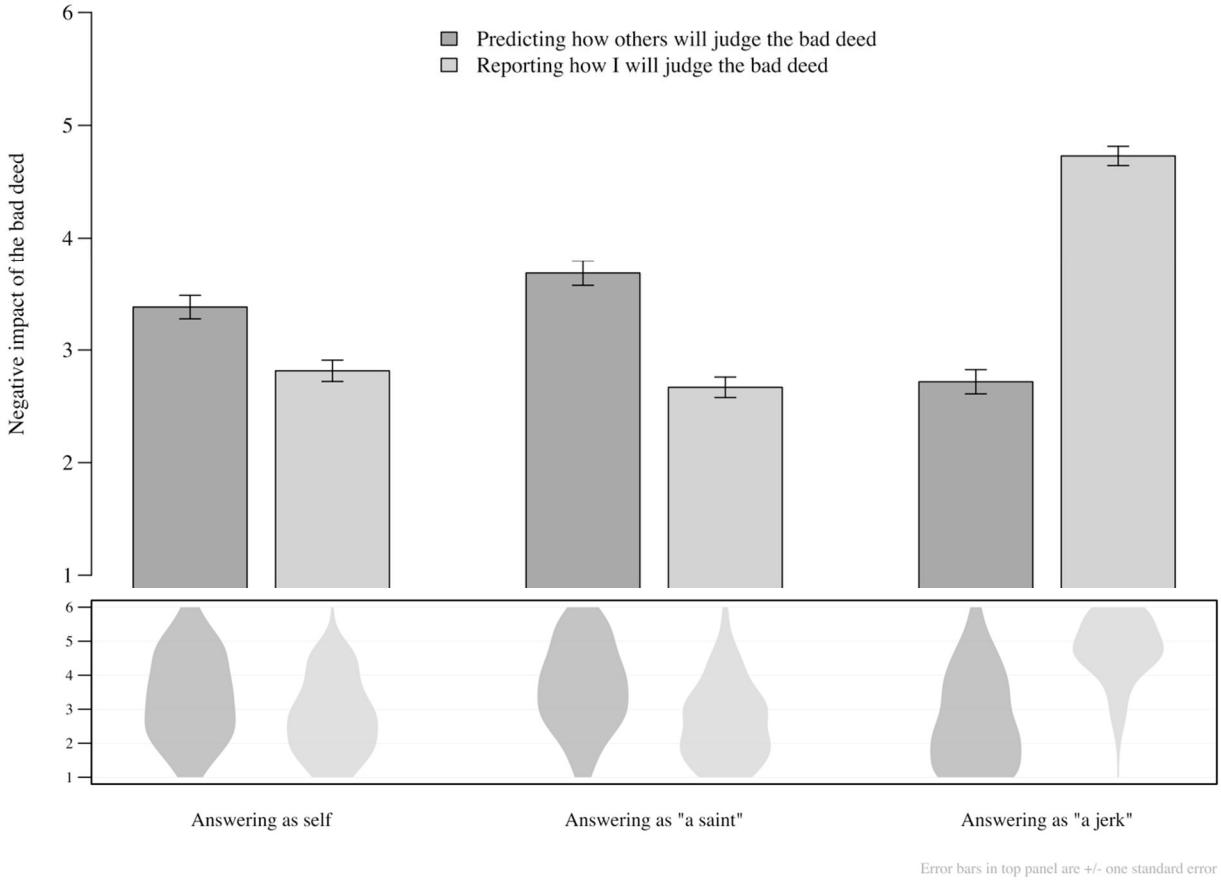
Method

The study was identical to Study 4a except that participants ($N = 884$ after preregistered exclusions) evaluated the negative impact of bad actions rather than the positive impact of good deeds. Overall, the study had a 2 (Perspective: Predict other's response vs. Report own response) × 3 (Role: Self, Saint, Jerk) between-subjects design. As in Study 4a, for stimulus-sampling purposes we varied both the type of bad action - not returning a friend's text, criticizing a colleague's presentation, forgetting to pick up a neighbor's package - and our descriptions of the saint and jerk roles (see Table 3), and we preregistered that we would collapse across these factors in the main analysis.

We assessed the rated negative impact of the bad action with three items, each measured on a 6-point scale (1 = *not at all*; 6 = *extremely*). In the *prediction* condition, participants estimated: (1) how upset the other person would be, (2) how unhappy the other person would feel, and (3) how much of a negative impact their actions would have on the other person. In the *report* condition, participants reported: (1) how upset they would be, (2) how unhappy they would feel, and (3) how much of a negative impact the other person's actions would have on them. The exact wording of these questions varied depending on the Role condition and the specific type of bad action, as detailed in the SOM.

Figure 6

The Effect of Perspective Condition (Prediction vs. Report) on Negative Impact Ratings Across Role Conditions (Self, Saint, Jerk) in Study 4b.



Results

We conducted a 2 (Perspective: Predict other person's response vs. Report own response) × 3 (Role: Self, Saint, Jerk) between-subjects ANOVA to analyze negative impact ratings of a bad action. There was a significant main effect of Role, but not of Perspective. Most importantly, as in Study 4a, there was a significant Perspective x Role interaction, $F(2, 878) = 133.67$, $p < .001$, $\eta_p^2 = 0.23$.

As shown in Figure 6, participants answering as themselves exhibited the standard metaperception effect: predictors rated their bad action as having a greater negative impact than reporters did, $b = 0.57$, $SE = 0.14$, $p < .001$, $\eta_p^2 = 0.05$. This difference was amplified when participants answered as saints, *b*

= 1.02, *SE* = 0.14, $p < .001$, $\eta_p^2 = 0.15$,[9] and fully reversed in the jerk condition. Specifically, predictors in the jerk condition rated their bad deed as having less negative impact than reporters said it would, *b* = -2.01, *SE* = 0.14, $p < .001$, $\eta_p^2 = 0.41$.

Discussion

In both Studies 4a and 4b, participants instructed to answer politely and considerately reproduced the standard predictor-reporter differences, whereas those instructed to be rude and impolite showed the opposite pattern. These results demonstrate that predictors and reporters face different response pressures and that what counts as a socially appropriate response for a predictor is not the same as what counts as a socially appropriate response for a reporter. They also show that simply responding the "appropriate" way will produce the very pattern typically interpreted as evidence for biases in metaperceptions.

The violin plots in the bottom panels of Figures 5 and 6 vividly illustrate these divergent pressures. In the positive realm, reporters avoid appearing ungrateful by choosing the very high end of the scale, whereas predictors avoid appearing immodest by choosing lower numbers. In the negative realm, reporters avoid appearing harsh or unforgiving by selecting the low end of the scale, while predictors avoid appearing irresponsible or unaccountable by selecting higher numbers.[10] Saints tend to amplify these tendencies, and jerks reverse them. The resulting distributions diverge not necessarily because

[9] The amplification effect was significant. A two-way ANOVA excluding the *jerk* condition showed that role (answering as a self vs. answering as a saint) significantly moderated the effect of the perspective condition (predicting vs. reporting) on negative judgments, $F(1, 584) = 5.07$ $p = .025$, $\eta_p^2 = 0.01$.

[10] Interestingly, Figure 5 shows that predictors in the positive domain also seem to avoid extremely low responses, probably because it feels inappropriate to suggest that someone would be entirely ungrateful for a good deed they performed. In this context, the "right" response is a balanced one: indicating that the recipient will feel grateful without implying that the deed was worthy of extreme praise. Consistent with this, the modal "saint" predictor responded with a 4 out of 6. Similarly, Figure 6 shows that predictors in the negative domain also seem to avoid extremely high responses, probably because it feels inappropriate to suggest that someone would be overly harsh in reacting to their small mistake. In this context, the "right" response is also a balanced one: indicating the recipient will react sensibly without implying that the mistake was too minor to justify a negative reaction. Consistent with this, the modal "saint" predictor responded with a 3 out of 6. In general, it appears as though predictors are pressured to avoid both extreme ends of the scale and thus to provide more moderate responses, whereas reporters are pressured to select responses at one extreme end of the scale while avoiding the other.

predictors are miscalibrated, but because predictors and reporters are navigating different social expectations embedded in the questions they are asked.

Studies 5a and 5b: Judging Predictors and Reporters Based on Their Responses

If response pressures can generate apparent biases in metaperceptions, then observers should judge the same response differently depending on whether it represents a prediction of someone else's reaction or a report of one's own reaction. Studies 5a and 5b tested this prediction by examining whether identical scale responses are evaluated differently when framed as coming from a predictor versus a reporter.

In our earlier studies, participants either predicted how their actions would be perceived by someone else or reported how they would respond to someone else's actions. In Studies 5a and 5b, we shifted the focus. Instead of having participants answer these questions themselves, we asked them to evaluate someone else based on that person's answers. This allowed us to test whether predictors and reporters face different response pressures by measuring how others perceive them based on their responses.

We predicted that people would form different impressions of predictors and reporters in ways consistent with the response pressures we have described. In Study 5a, which focused on positive actions, we expected that reporting high gratitude for a favor (e.g., "I am extremely grateful") would make a person more likable than predicting that someone else would feel the same way (e.g., "They would be extremely grateful"). Conversely, we expected that reporting low gratitude (e.g., "I am slightly grateful") would make a person less likable than predicting that someone else would feel the same way (e.g., "They would be slightly grateful").

In Study 5b, which focused on negative actions, we expected the mirror-image pattern. Predicting that a bad action would have a severe negative impact (e.g., "They will be extremely upset") should make a person more likable than reporting that they themselves are extremely upset. Conversely,

predicting that a bad action would have only a mild impact (e.g., "They won't at all be upset") should make a person less likable than reporting that they themselves wouldn't be upset.

This design provides a direct test of response pressure differences between predictors and reporters. By examining whether identical responses lead to different social judgments depending on whether they come from predictors or reporters, Studies 5a and 5b provide a critical test of whether metaperception biases necessarily reflect true miscalibration or whether they may instead reflect socially shaped response tendencies.[11]

Study 5a

Method

We recruited adult online participants (N = 454 after preregistered exclusions) to evaluate their impressions of someone based on how that person answered questions in a study. The study featured a 2 (Perspective: Predict other person's gratitude vs. Report own gratitude) × 6 (Gratitude Rating: Not at all, Slightly, Somewhat, Quite, Very, Extremely) between-subjects design. We varied the type of good deed – small favor vs. positive feedback – for stimulus-sampling purposes, and we preregistered to collapse across this factor in the main analysis.

Participants were asked to consider a person named L.N. and were shown how L.N. responded to a gratitude question in a hypothetical study. They were randomly assigned to read a scenario in which L.N. either (1) predicted how grateful someone else would feel for a good deed L.N. performed (*prediction* condition) or (2) reported how grateful they would feel for a good deed someone performed for them (*report* condition). L.N.'s response was displayed on a 6-point scale – Not at all grateful,

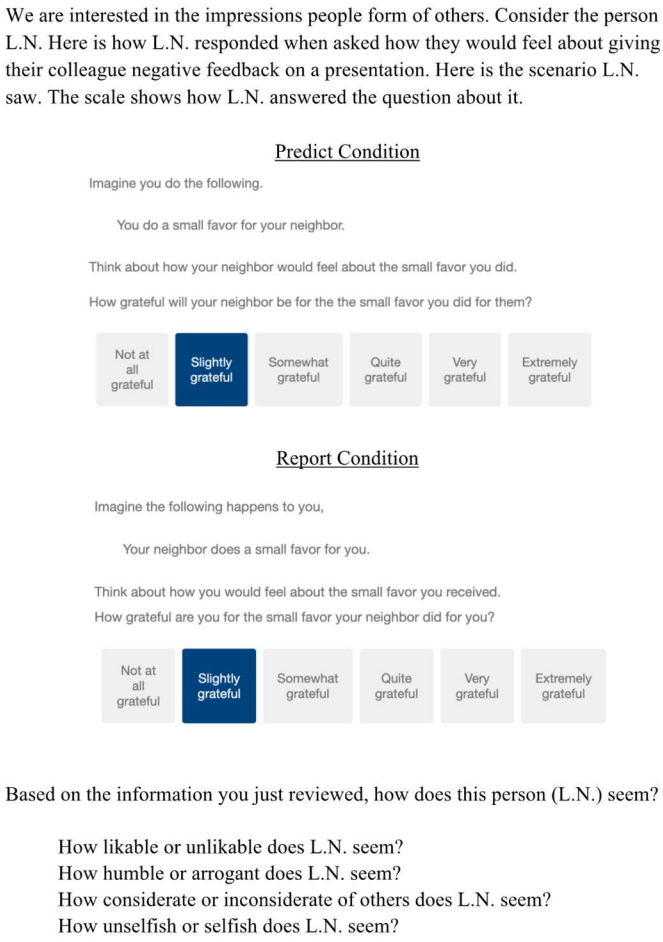
[11] Studies 5a and 5b were run in a single session using a single survey, and participants were randomly assigned to complete either Study 5a or 5b. We have a single preregistration for both studies, and we preregistered to treat the studies as separate, despite the fact that the data were collected concurrently.

Slightly grateful, Somewhat grateful, Quite grateful, Very grateful, Extremely grateful – and was experimentally varied to be one of these six responses (see Figure 7 for an example).

Participants then answered four questions about L.N.: “How likable or unlikable does L.N. seem?”; “How humble or arrogant does L.N. seem?; “How considerate or inconsiderate of others does L.N. seem?”; and “How unselfish or selfish does L.N. seem?” Responses were recorded on 7-point scales (e.g., 1 = *very unlikable*; 7 = *extremely likable*) and averaged to create an overall positive impression score (Cronbach’s $\alpha = .95$).

Figure 7

Example of the manipulation and stimuli used in Study 5a.



Note. Participants were randomly assigned to the *prediction* or *report* condition, and L.N.'s scale response was experimentally manipulated. The example shown depicts the *slightly grateful* condition. The type of good deed was stimulus sampled; the other good deed involved giving positive feedback to a colleague.

Results

We preregistered to treat L.N.'s gratitude rating as a continuous variable. To analyze the data, we regressed positive impression ratings on perspective (prediction vs. report), gratitude rating, and their interaction. Consistent with the response-pressure account, there was a significant interaction between perspective and gratitude rating, $b = 0.26$, $SE = 0.06$, $p < .001$.

The left panel of Figure 8 shows this interaction. Although the pattern can be described in several ways, all interpretations point to the same conclusion: the social implications of a given gratitude rating depend on whether it represents a prediction of someone else's reaction or a report of one's own.

We first examined the simple effects of gratitude rating within each perspective condition. Gratitude ratings shaped observers' impressions more strongly when L.N. reported their own gratitude ($b = 0.69$, $SE = 0.04$, $p < .001$) than when L.N. predicted someone else's gratitude ($b = 0.43$, $SE = 0.04$, $p < .001$).

Next, we examined the simple effect of perspective at different levels of gratitude. For this exploratory analysis, we first created a variable that captured whether L.N. gave low gratitude ratings (i.e., "Not at all," "Slightly," or "Somewhat") or high gratitude ratings (i.e., "Quite," "Very," and "Extremely"). We then regressed participants' positive impression ratings on perspective (coded as *prediction* = 0; *report* = 1) within low vs. high levels of gratitude. The results are clear. People judged L.N. less positively when they reported rather than predicted low levels of gratitude ($b = -0.58$, $SE = 0.16$, $p < .001$), but more positively when they reported rather than predicted high levels of gratitude ($b = 0.44$, $SE = 0.15$, $p = .004$).[12]

Study 5b

Method

The study used the same methodology as Study 5a, but instead of evaluating someone based on their responses to a good deed, participants (N = 481 after preregistered exclusions) evaluated someone based on their responses to a bad action (e.g., causing small harm to a neighbor's yard). The design was

[12] For transparency, we report the differences in positive impressions between predicting and reporting gratitude at each level of the gratitude scale. The *prediction* condition is coded as 0 and the *report* condition is coded as 1. Higher values of the dependent variable signify more positive judgments. The results at each scale point are: Not at all ($b = -0.09$, $SE = 0.24$, $p = .72$); Slightly grateful ($b = -1.08$, $SE = 0.25$, $p < .001$); Somewhat grateful ($b = -0.60$, $SE = 0.25$, $p = .015$); Quite grateful ($b = 0.11$, $SE = 0.24$, $p = .471$); Very grateful ($b = 0.61$, $SE = 0.24$, $p = .013$); Extremely grateful ($b = 0.64$, $SE = 0.24$, $p = .007$).

a 2 (Question Perspective: Predict how upset someone else would be vs. Report how upset you would be) × 6 (Upset Rating: Not at all, Slightly, Somewhat, Quite, Very, Extremely) between-subjects design. The type of bad action was varied for stimulus-sampling purposes – minor yard harm vs. negative feedback – and we preregistered to collapse across this factor in the main analysis. Participants rated their impressions of L.N. using the same four items as in Study 5a (Cronbach's $\alpha = .94$).[13]

Results

We conducted the same analyses as described in Study 5a. As preregistered, we treated the upset rating condition as a continuous variable and regressed positive impression ratings on perspective, upset rating, and their interaction. Consistent with the response-pressure account, there was a significant interaction between perspective and upset rating, $b = -0.76$, $SE = 0.06$, $p < .001$.

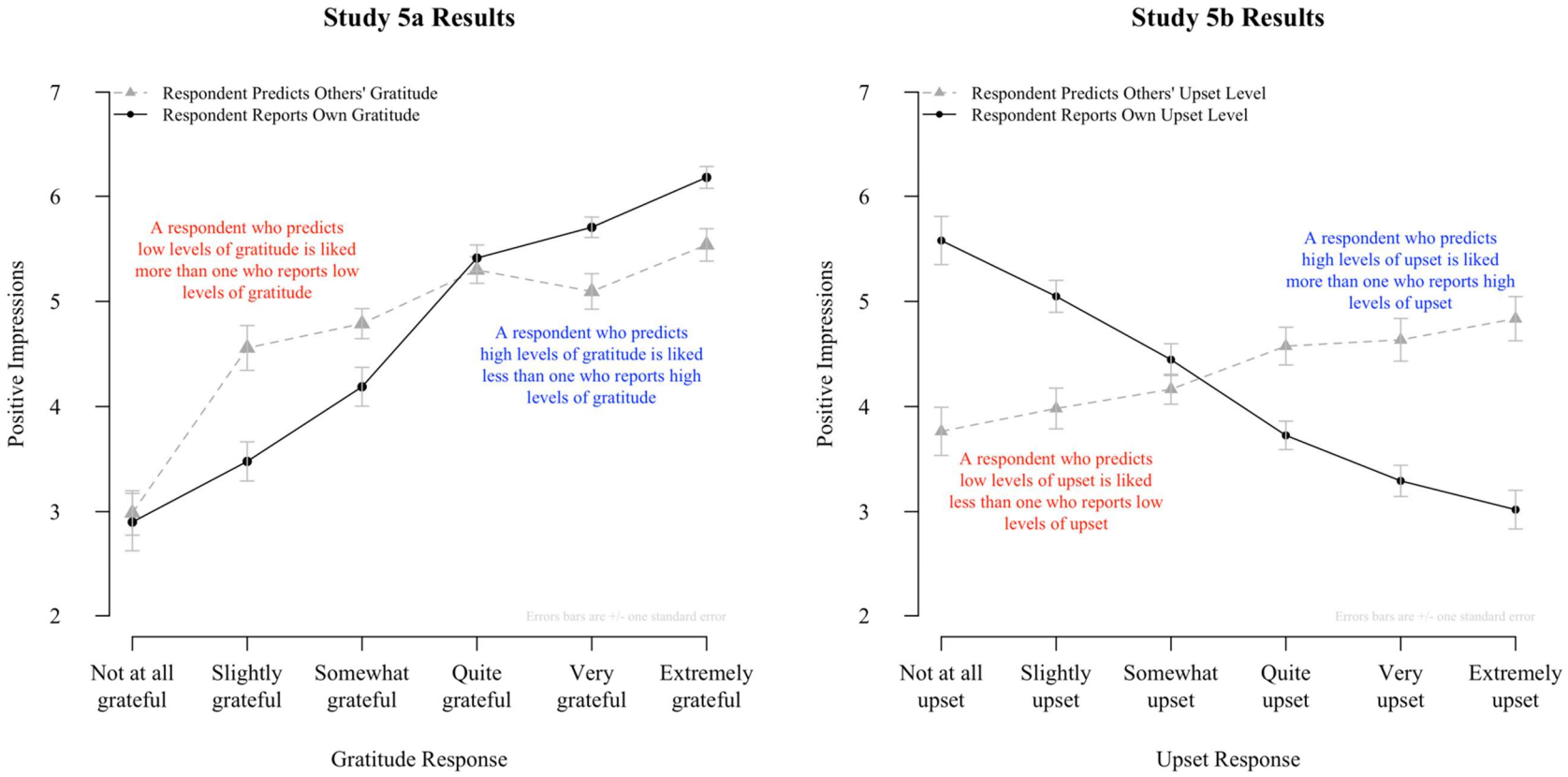
As shown in the right panel of Figure 8, impressions of the respondent differed substantially depending on whether the respondent predicted how upset others would be or reported how upset they would be themselves.

We first examined the simple effects of upset ratings within each perspective condition. When L.N. reported how upset they would be, higher upset ratings were associated with less positive third-party judgments ($b = -0.54$, $SE = 0.04$, $p < .001$). In contrast, when L.N. predicted how upset others would be, higher upset ratings were associated with more positive third-party judgments ($b = 0.22$, $SE = 0.04$, $p < .001$).

[13] We varied the scale anchors to match the valence of the action being evaluated. In Study 5a (positive actions), the impression scales placed the positive pole at the high end (e.g., extremely unlikable → extremely likable). In Study 5b (negative actions), the scales were reversed so that the negative pole appeared at the high end (e.g., extremely likable → extremely unlikable). For our analyses, we reverse-scored the Study 5b items so that, in both studies, higher numbers reflected more positive impressions of the respondent.

Figure 8

Effect of Predicting vs. Reporting Gratitude (Study 5a) or Upset Level (Study 5b) on Impressions of the Respondent



Note. How the respondent supposedly answered the question and whether the respondent predicted someone else's gratitude or reported their own were manipulated.

We next analyzed the simple effect of perspective (prediction vs. report) at different levels of upset rating. For this exploratory analysis, we first created a variable that captured whether L.N. gave low upset ratings (i.e., "Not at all," "Slightly," or "Somewhat") or high upset ratings (i.e., "Quite," "Very," or "Extremely"). We then regressed participants' positive impression ratings on perspective (coded as *prediction* = 0; *report* = 1) within low vs. high levels of upset. The results are again clear. Although L.N. was judged more positively when they reported rather than predicted low levels of upset (b = 1.03, SE = 0.15, $p < .001$), they were judged more negatively when they reported rather than predicted high levels of upset (b = -1.33, SE = 0.15, $p < .001$).[14]

Discussion

The results of Studies 5a and 5b are not subtle. Observers consistently judged the same response more or less favorably depending on whether it came from a predictor or a reporter. In Study 5a, people formed more positive impressions of someone who predicted low levels of gratitude than of those who reported the same low levels. Conversely, someone who predicted high levels of gratitude were judged more negatively than those who reported the same high levels.

In Study 5b, this pattern was even more striking. The responses that made predictors seem most likable made reporters seem least likable. Predictors were evaluated most positively when they predicted that their small bad action would make someone extremely upset, whereas reporters were evaluated most positively when they stated that the same bad action would not make them upset at all.

Taken together, these findings reveal a fundamental problem: identical scale responses carry different social meanings depending on whether they come from a predictor or a reporter. This is

[14] For transparency, we report the differences in positive impressions between predicting and reporting gratitude at each level of the upset scale. The *prediction* condition is coded as 0 and the *report* condition is coded as 1. Higher values of the dependent variable signify more positive judgment. The results at each scale point are: Not at all (b = 1.82, SE = 0.26, $p < .001$); Slightly upset (b = 1.07, SE = 0.26, $p < .001$); Somewhat upset (b = 0.28, SE = 0.25, $p = .267$); Quite upset (b = -0.85, SE = 0.26, $p = .001$); Very upset (b = -1.34, SE = 0.26, $p < .001$); Extremely upset (b = -1.81, SE = 0.26, $p < .001$).

precisely what one would expect if the two roles elicit different response pressures. And if predictors and reporters face different response pressures, then any observed differences between conditions could be an artifact of these pressures rather than evidence of genuine miscalibration. What looks like a bias – such as underpredicting how grateful someone will be or overpredicting how upset they will be – may instead arise because some participants simply say what a more likable person would say.

Studies 6a and 6b: When Response Pressures Are More or Less Aligned

Thus far, our investigation has focused on a central question in the metaperception literature: Do people accurately forecast the impact that their actions have on other people? We have shown that response pressures differentially influence the responses of predictors and reporters, and that those pressures alone can generate apparent biases in metaperceptions, making it seem like people underestimate the impact of their good actions and overestimate the impact of their bad ones.

But not all metaperception questions may elicit such divergent pressures. Consider the difference between asking people to forecast *how much impact* their positive action will have on another person versus asking them to forecast *how they will behave* toward that person. In the previous studies, questions about positive impact (e.g., gratitude or appreciation for an act) consistently produced opposite pressures for predictors and reporters.

By contrast, we hypothesized that questions about positive behavior (e.g., how nice or polite someone will be during an interaction) would evoke more aligned pressures. Predictors may feel compelled to say they will be nice during the interaction (and thus expect others to see them as nice), while reporters may feel compelled to say they will find someone else to be nice (since calling someone unkind can itself seem unkind). In both cases, the social pressure encourages higher ratings.

If this is correct, then predictors and reporters should be pulled in more similar directions when judging positive behavior than when judging positive impact. And if response pressures shape how

people answer metaperception questions, as we suggest they do, then the classic predictor-reporter gap should be larger for judgments of positive impact than for judgments of positive behavior. We test this prediction in Study 6b.

Before turning to this specific test, it is important to verify that response pressures between predictors and reporters are indeed more aligned for judgments of behavior (niceness or politeness) than for judgments of impact (gratitude or appreciation). Study 6a tests this by using the same methodology as Studies 5a and 5b. Specifically, Study 6a assessed how observers judge predictors and reporters based on their answers to positive impact or positive behavior questions, and Study 6b assessed how predictors and reporters answer questions about positive impact or positive behavior.

Study 6a

Method

We used the same basic design as Studies 5a and 5b to examine how people evaluate someone based on their answers to different types of questions. Participants (N = 1,444 after preregistered exclusions) evaluated a target person, S.N., after seeing how S.N. responded to a question from Study 6b. The study employed a 2 (Question Perspective: Predict other person's response vs. Report own response) × 2 (Question Category: Positive Impact vs. Positive Behavior) x 5 (Question rating: Not at all, Slightly, Somewhat, Quite, Very) between-subjects design. As preregistered, we treated question rating as a continuous variable, expecting its relationship to positive impressions of S.N. to interact with both question perspective and question category.

Participants read: "We are interested in the impressions people form of others. Consider the person S.N. Here is how S.N. responded when asked how they would feel about [giving their colleague advice about an upcoming presentation at work | their colleague giving them advice about their upcoming

presentation at work]. Please read the scenario and how S.N. responded to it. Then answer four questions about your thoughts about S.N."

The scenario text varied by question perspective. In the *prediction* condition, S.N. was described as answering a question about the following scenario: "You have a conversation with a colleague about their upcoming presentation at work. You give your colleague advice about their presentation." In the *report* condition, S.N. answered a question about the complementary situation: "You have a conversation with a colleague about your upcoming presentation at work. They give you advice about your presentation."

The question category manipulation determined the construct that S.N. was shown to be responding to. In the *positive impact* condition, participants saw that S.N. predicted the other person's gratitude or appreciation for the advice (in the *prediction* condition) or reported their own gratitude or appreciation for the advice (in the *report* condition). In the *positive behavior* condition, participants saw that S.N. either predicted how nice or polite their colleague would find them during the conversation (*prediction* condition) or reported how nice or polite they would find their colleague during the conversation (*report* condition). Participants than saw S.N.'s response on a 5-point scale (*not at all, slightly, somewhat, quite, very*). Figure S1 provides an example of the response format shown to participants.

For stimulus-sampling purposes, we varied the specific traits within the *positive impact* and *positive behavior* conditions (i.e., grateful vs. appreciative; and nice vs. polite). As preregistered, we collapsed across traits within each category.

After reviewing this information, participants answered four questions assessing their impressions of S.N. The questions were answered on 7-point bipolar scales (e.g., 1 = *very unlikable*; 7 = *very likable*). We averaged the responses to create a positive-impression score ($\alpha = 0.95$). The exact materials and questions appear in Figure S1 of the SOM.

Results

We predicted a three-way interaction among Question Perspective, Question Category, and Question Rating on positive impressions of S.N. To test this prediction, we regressed positive impressions of the target on question perspective, question category, question rating (treated as a continuous variable), and all two- and three-way interactions. The predicted three-way interaction was significant, $b = 0.32$, $SE = 0.08$, $p < .001$, indicating that the relationship between S.N.'s rating and impressions of S.N. differed across perspective and category conditions.

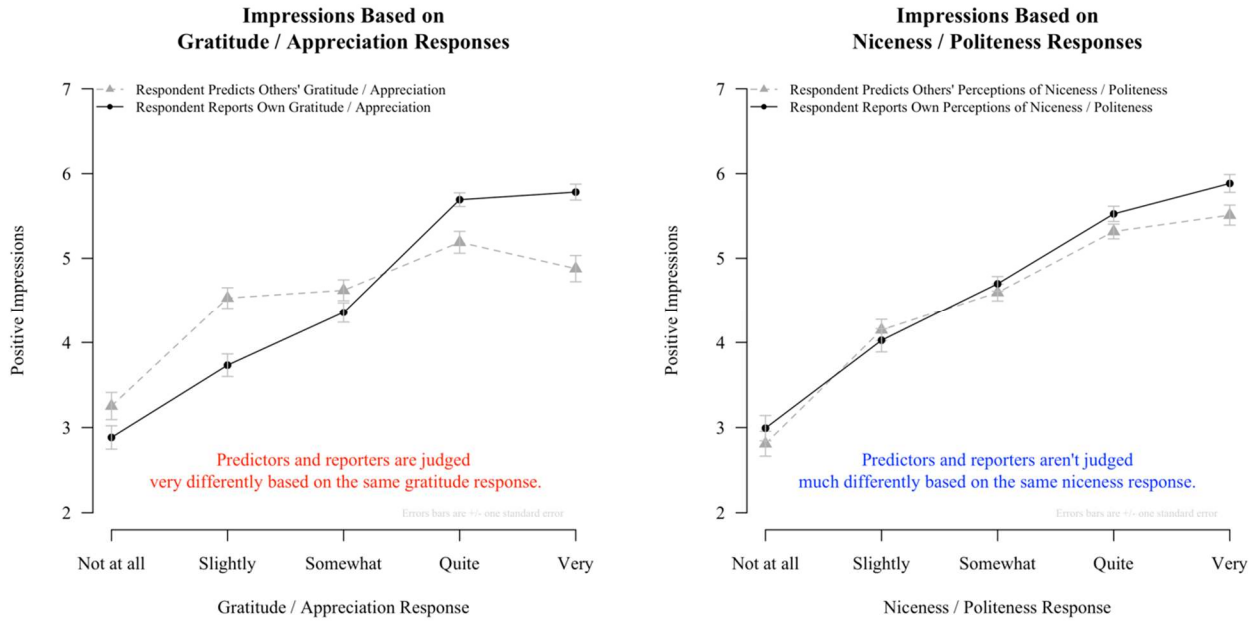
We first examined the pattern within the *positive impact* condition. The two-way interaction between question perspective and question rating was significant, $b = 0.39$, $SE = .06$, $p < .001$. As shown in Figure 9, the relationship between the magnitude of gratitude that S.N. expressed and observers' positive impressions of S.N. was stronger in the *report* condition, $b = 0.78$, $SE = 0.04$, $p < .001$, than in the *prediction* condition, $b = 0.39$, $SE = 0.04$, $p < .001$. These findings replicate Study 5a, and provide further evidence that predictors and reporters face systematically different response pressures when evaluating the positive impact of advice (e.g., gratitude, appreciation).

In contrast to the results observed in the *positive impact* condition, the interaction between question perspective and question rating was not significant within the *positive behavior* condition, $b = 0.07$, $SE = .05$, $p = .194$. This indicates that the relationship between S.N.'s behavior rating (i.e., niceness, politeness) and participants' impressions of S.N. did not differ significantly between the *prediction* and *report* conditions. In both perspective conditions, higher ratings of niceness/politeness produced more favorable impressions of S.N., and the magnitude of this relationship was very similar for predictors and reporters. These results suggest that the response pressures associated with positive behavior

judgments are more aligned for predictors and reporters, in contrast to the divergent pressures observed for positive impact judgments.[15] Additional exploratory analyses are reported in the SOM.

Figure 9

Effect of Predicting vs. Reporting Gratitude/Appreciation Level or Niceness/Politeness Level on Impressions of the Respondent in Study 6a



Discussion

Study 6a showed that response pressures operate when people respond to both types of questions – those about gratitude/appreciation (positive impact questions) and those about niceness/politeness (positive behavior questions) – but that the degree to which predictors and reporters face *divergent* pressures depends on the question type. While predictors and reporters face markedly different response

[15] This is not to say that response pressures between predictors and reporters are perfectly aligned. Indeed, it is extremely difficult to know whether one group experiences exactly the same response pressures as another. For example, perhaps both predictors and reporters are pressured to give higher responses, but maybe this pressure influences 40% of predictors and only 30% of reporters (or vice versa). And maybe the magnitude of that influence is greater in one condition vs. another. We can say that the pressures faced by those asked about niceness or politeness are *more* aligned than those asked about gratitude or politeness, but we cannot say that they are perfectly equivalent.

pressures when answering questions about positive impact, they face very similar response pressures when answering questions about positive behavior.

If apparent metaperception biases arise from such response pressures, then the predictor-reporter difference should be largest when these pressures push predictors and reporters in opposite directions, as they do for judgments of gratitude and appreciation, and smallest when the pressures push them in the same direction, as they do for judgments of niceness and politeness. This logic led us to expect – and to test in Study 6b – whether the traditional predictor-reporter difference observed for positive impact judgments would weaken or disappear when participants evaluated positive behavior.

Study 6b

Method

Online participants (N = 752 after preregistered exclusions) considered a situation in which they either gave or received advice from a colleague about a presentation. They then answered questions about the advice. We varied whether they evaluated the positive impact of the advice on the advisee (*positive impact* condition) or the positive behavior of the advisor during the conversation (*positive behavior* condition). This resulted in a 2 (Perspective: Predict [give advice] vs. Report [receive advice]) × 2 (Rating Type: Positive Impact vs. Positive Behavior) between-subjects design.

We operationalized *positive impact* as ratings of how grateful or appreciative the advisee would feel (e.g., How grateful [appreciative] will your colleague be for your advice? | How grateful [appreciative] will you be for the advice?). *Positive behavior* was operationalized as ratings of how nice or polite the advisor would be during the conversation (e.g., How nice [polite] will your colleague find you during the conversation? | How nice [polite] will you find your colleague during the conversation?). Participants answered on a 5-point scale (1 = *not at all*; 5 = *very*).

For stimulus-sampling purposes, we manipulated the specific adjective used in each condition (i.e., grateful vs. appreciative; nice vs. polite) and preregistered that our main analyses would collapse across this factor.

We predicted that rating type would moderate the effect of perspective. Specifically, in the *positive impact* condition, we expected advice givers (predictors) to estimate a smaller positive impact of their advice than advice receivers (reporters) would report. In contrast, we expected this predictor-reporter difference to be attenuated or eliminated in the *positive behavior* condition, where response pressures are more aligned.

Results

We ran a 2 (Perspective: Predict [give advice] vs. Report [receive advice]) × 2 (Rating Type: Positive Impact vs. Positive Behavior) between-subjects ANOVA on positive ratings. As predicted, there was a significant interaction between perspective and rating type, $F(1, 748) = 23.28$, $p < .001$, $\eta_p^2 = 0.03$.

As shown in Figure 10, in the *positive impact* condition predictors estimated that their advice would be judged less positively than reporters said it would, $b = 0.39$, $SE = 0.08$, $p < .001$, $\eta_p^2 = 0.06$. In contrast, in the *positive* behavior condition, a different pattern emerged. Here, predictors estimated that the advisee would find them to be nicer than reporters said they would, $b = -0.16$, $SE = 0.08$, $p = .042$, $\eta_p^2 = 0.01$.

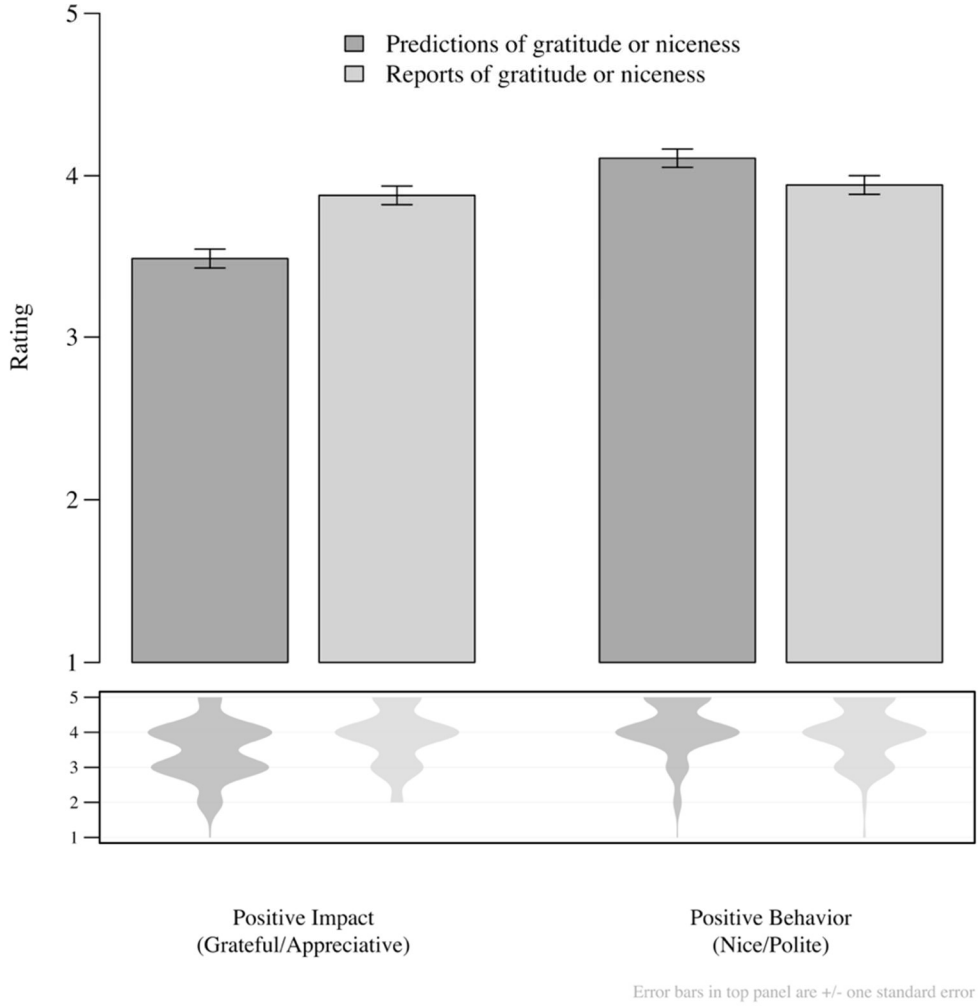
Discussion

If (apparent) biases in metaperceptions are at least partially driven by predictors and reporters facing different response pressures, then such (apparent) biases should be reduced when those pressures are more aligned. And this is what we found in Study 6b. The large predictor-reporter differences observed for impact judgments (i.e., gratitude or appreciation) disappeared – and even slightly reversed

– when participants judged behavior (i.e., niceness or politeness) instead. These results provide further evidence that apparent biases in metaperceptions may stem from differences in the response pressures faced by predictors and reporters.

Figure 10

The Effect of Perspective (Prediction vs. Report) on Judgments of Positive Impact (Gratitude / Appreciation) vs. Positive Behavior (Niceness / Politeness) in Study 6b.



General Discussion

Metaperception research portrays people as systematically misjudging how others respond to their actions: overestimating how harshly others will judge their missteps and underestimating how

positively they will judge their prosocial acts. These findings have been used to explain social withdrawal and undersocialization (e.g., Epley et al., 2022), and invoked to justify a range of behavioral recommendations, such as giving more advice, reaching out more, or bargaining more assertively. We do not know whether this advice is good or bad. But we believe that researchers should not offer such recommendations solely based on evidence that people's metaperceptions are biased, because the methods used to reach these conclusions are inherently confounded.

Across 13 studies, we demonstrate that the apparent biases in metaperception research may not reflect genuine psychological miscalibration. Instead, they may be artifacts of how these studies are designed, arising because predictors and reporters are asked different questions that carry different social pressures.

In our first set of studies (Studies 1a–3), we tested whether the standard pattern of predictors misperceiving their impact on others would emerge across a wide range of situations, even those that were fanciful, ambiguous, or neutral. And, indeed, we observed the same apparent biases when people predicted how upset someone would be at them for mentioning their work in an alien forum or how grateful someone would be for receiving a nonsense letter string. That the same effects emerge in both fanciful and mundane contexts shows that they cannot be taken as evidence of something meaningful about or specific to any one context. Moreover, these effects, particularly the findings from Studies 2a-3, are difficult to explain in terms of genuine beliefs about gratitude or blame. Rather, we believe that the findings point to an endemic methodological issue as a likely source of the effect, namely the confounding of the question with the condition.

The next set of studies (4a-6b) showed that these findings appear to emerge across these contexts because predictors and reporters answer different questions, and these questions carry opposing response pressures. In Studies 4a and 4b, we manipulated the response pressures participants faced by

asking them to respond either as themselves, as a very polite person, or as a very impolite one. We found that the size and direction of the predictor-reporter difference changed accordingly. It was magnified under instructions to be polite and reversed under instructions to be impolite. In Studies 5a and 5b, we found that third-party observers evaluated the same responses differently depending on whether they came from a predictor or a reporter. For example, someone who reported they would feel very grateful was rated as more likable than someone who predicted that someone else would feel very grateful. Conversely, someone who reported that they would feel very upset was rated as less likable than someone who predicted that someone else would feel very upset.

These findings suggest that the social meaning of the same response differs across conditions. In other words, the same response may make a predictor more (or less) likable than a reporter. This further supports the idea that the predictor–reporter gap can emerge simply from differences in what feels socially appropriate to say.

Finally, in Studies 6a and 6b, we demonstrated that biases are strongest when predictors and reporters face opposing response pressures (e.g., as they do when assessing gratitude) and much weaker when those pressures are better aligned (e.g., when assessing niceness).

Together, these findings do not prove that there are no biases in metaperceptions. But they do show that even in the absence of any real bias, the confound alone could produce effects that look like bias.

Rethinking the Accuracy Benchmark

You cannot say that something is a bias without comparing it to a ground truth. But what, exactly, is the ground truth in metaperception research? Metaperception researchers act as if they know the answer, as they accept reporters' ratings as the truth against which predictions are judged. When predictions do not match those reports, the predictions are assumed to be wrong. The assumption is that reporters' ratings offer a direct and unbiased window to their internal experience.

On its face, this might seem reasonable. But upon a moment's reflection, it is not. Psychologists know that an average scale response to a question about a person's internal state is unlikely to represent the true average. They know that you cannot ask people how grateful they would feel for receiving a small compliment, and then presume that the observed average is correct, just as you cannot ask a sample of people how much they like Donald Trump, or affirmative action, or sushi, and presume that those observed averages are correct. It is understood that some participants do not report their true feelings or beliefs, whether due to social desirability concerns, response biases, Gricean norms, confusion, scale (mis)interpretations, etc. Someone might say '5' when they are truly a '4'. And if *any* respondents do that, the observed average won't be right.[16] And if the observed average isn't right, then you cannot use it as a ground truth against which to judge the accuracy of people's predictions.

It's worth pausing to consider what it would mean to take this assumption seriously. If we truly believed that self-reports provide direct insight into internal states, we would be publishing single-cell studies claiming, for example, that the "true" level of gratitude for receiving a compliment is 4.2 on a 5-point scale. But we would never do that. Researchers understand that how a sample of people answers a question about their feelings may not reflect how those people actually feel.

Consider attitudes toward outgroup members. If participants report an average favorability rating of 4.4 out of 5 on a feeling thermometer, we don't accept that as the true level of warmth that people have toward outgroups. We recognize that these responses are shaped by moral and social pressures. As noted in the Introduction, researchers have developed indirect elicitation methods for the sole purpose of trying to overcome or sidestep these distortions (e.g., Fazio et al., 1995; Greenwald et al., 1998; Sigall & Page, 1971). Yet, for some reason, when it comes to comparing predictors and reporters,

[16] Strictly speaking, the average can be right if the biased responses all miraculously cancel out – if, for example, some participants overstated their beliefs by a given amount while an equal number understated theirs by the same amount. This is extremely unlikely, and there is no basis for expecting it to happen in any given circumstance.

we drop that caution. We ignore the same pressures and problems that we elsewhere worry about and work so hard to avoid. Reporters' answers are taken as neutral benchmarks. Predictors are judged against them. And any discrepancy is interpreted not as a methodological artifact, but as a failure of social perception.

Our findings reinforce this point. If reporters' responses were neutral reflections of internal states, then they should not be influenced by response pressure. But we find that reporters' responses do change when they are asked to respond politely vs. impolitely in Studies 4a and 4b, and that their own answers resemble those of extremely polite responders, while diverging from those instructed to respond like impolite or inconsiderate people.

These results indicate that there is a socially appropriate way to respond to questions about how someone else's actions affect you, and that response pressures may operate on reporters at least as strongly as they operate on predictors. Self-reports about how others' actions affected them are not socially neutral. They are shaped by social expectations, and people seem to know what kinds of answers those expectations favor.

Non-Solutions and Semi-Solutions

In this paper we have suggested that the study of metaperceptions has been plagued by a built-in confound: the question participants are asked differs across conditions. We appreciate that any paper that exposes a methodological problem is expected to offer a solution. And so we also appreciate how unsatisfying it is when we say that a perfect solution does not exist.

The question that motivates research on metaperceptions is deeply compelling. It speaks to something we all wonder: Is the way I see the world the way others see it? How do people feel about what I have done? What do other people truly think of me, feel about me, remember about me? We crave access to others' minds. It seems tempting to tackle these questions by comparing what people

think is in someone's head to what others report is in their head. But once we recognize that we don't actually have unbiased access to what is in people's heads, the limitations become obvious. Once you accept that the thing you want is someone's actual internal state, and that the thing you are measuring is a self-report of that state, the mismatch becomes clear. You cannot measure what you want unless what you want is just a potentially biased number on a scale.

And so what can we do?

Let's begin by emphasizing what we can't do. We cannot fix the problems identified in this article by the use of statistical controls, or by the simple methods of incentivizing responses or assuring anonymity. First, controlling for response pressures – by, for example, controlling for individual differences in social desirability – simply will not fix the problem. For it to work, you'd need to control for a variable that perfectly correlates with exactly how much each participant's answer on that specific item was distorted by response pressures. For example, you'd need a variable that correctly reflects the fact that Participant 1 lowered their answer by 1, Participant 2 increased their answer by 2, Participant 3 did not distort their answer, and so on. In real life, no such variable exists.

Second, because we can't observe reporters' true internal states, we can't incentivize reporters to accurately report those states or predictors to forecast them. You cannot incentivize accuracy without a verifiable measure of accuracy. Third, anonymity assurances are likely to fail for at least two reasons. One is that some participants may not believe those assurances, and perhaps reasonably so. Another is that some people don't simply turn off the desire or habit to be polite just because they feel anonymous (Cooley, 1922; Scheff, 1988). Social norms and expectations aren't carried in luggage that people leave at the door when they know they are anonymous and alone. Many people carry large parts of their social selves with them everywhere they go. They say "please" when they ask ChatGPT for help in

private,[17] they say "oops" when they make a mistake in private (Goffman, 1978), and they may express exaggerated appreciation for a gift when they are privately asked how they would feel upon receiving it. As noted in the Introduction, this problem does not require *all* participants to carry their social selves with them when responses are anonymous; it merely requires *some*.

These non-solutions are even less promising than this makes it seem. In this paper we have focused only on one source of the problem of question/condition confounds: response pressures caused by a desire to respond to questions in socially appropriate ways. But there are many potential sources of bias that may influence how people respond to different questions. For example, what if participants who are asked to predict how happy someone would be with a small gift use the rating scale differently than those who are asked to report how happy they are with that gift? Like, what if predictors are more likely to compare the small gift to a large gift, effectively answering the question, "Compared to a larger gift, how happy would the gift recipient be?", while reporters are more likely to compare the small gift to no gift at all, effectively answering the question, "Compared to no gift, how happy would I be with this small gift?" Even careless or inattentive responses can generate artifactual condition differences if some questions prompt more inattention than others, particularly since inattention can generate biased responding (e.g., Zorowitz et al., 2023). Whenever questions are confounded with condition, it will be very difficult to know whether any effect is driven by the confound or by the intended manipulation. And assurances of anonymity aren't going to solve this problem.

Of course, just because people are asked different questions in different conditions does not mean that this confound will necessarily be consequential. So how do you know if it matters? Our early studies (Studies 1a–3) offer a place to start. In these studies, we showed that the predictor–reporter gap

[17] https://web.archive.org/web/20250426080141/https://www.usatoday.com/story/tech/2025/04/22/please-thank-you-chatgpt-openai-energy-costs/83207447007/

appeared even in contexts where it arguably shouldn't, such as predicting or reporting reactions to receiving a nonsense letter string or being mentioned in an alien forum. The fact that they appeared in these contexts strongly suggests that these effects weren't indicative of a "real" mismatch between predictions and reports, but rather by the structure of the experiment, the confound of question and condition. Now these studies do not tell you exactly which aspect of the questions causes the gap, but they do represent a simple diagnostic test, an "acorn test:" If the effect shows up even where it shouldn't, then it's probably caused by a built-in confound.

How could you apply an "acorn test?" Suppose you are studying people's predictions and reports of how happy they would feel about winning an award, and you find that people predict being happier than they report. Now imagine running the same prediction–report study in a context that shouldn't exhibit the effect, such as receiving one meaningless letter string instead of another or receiving an acorn, and you find the same gap. That tells you that there likely is something about the questions that produces the difference, not necessarily the experience of receiving an award. On the other hand, if the gap appears for the award but not for the letter string or acorn, you can be more confident that the effect reflects something meaningful about the object of judgment, not just about the way the questions were asked. Putting your study through an "acorn test" doesn't tell you everything, particularly since "failing" the test is informative whereas "passing" it may be less so (for although it passed one acorn test, it doesn't mean that it would pass others). But it gives you a way to start figuring out whether your effect is caused by the intended manipulation or whether it is caused by your method.

In the context of research on metaperceptions, we believe that the best (though imperfect) solution is to change what we are studying. Instead of assessing how well people predict others' (unmeasurable) internal states, we can assess how well people predict others' behavior, because others' behavior has a verifiable ground truth. Will someone yell if I make a mistake? Will they return the favor? Will they

call to say thank you? Will they circle "5" on the rating scale? So instead of asking people to predict whether someone will like their gift, we can ask them to predict how much that person will *say* they like their gift. This prediction does not purely capture people's metaperceptions, but rather beliefs about someone's overt behavior, which may be driven by some combination of true appreciation for the gift and adherence to social norms. Thus, we can alter our research agenda so that we are asking predictors to forecast what people do, rather than what their true internal states are.

Note that even this is imperfect, because the *forecasts* may not be purely authentic, as they may be altered by response pressures, confusion, careless responding, etc. For example, a modesty norm might compel us to say that our gift recipient will provide a rating of 4 even if we think they will provide a rating of 5. But perhaps these sorts of response pressures are less potent when asked to predict behavior than when asked to predict internal states, and perhaps they can be attenuated by incentivizing people to make accurate forecasts, something you can do when you are asking people to forecast measurable *behavior* but not when you are asking them to forecast (unobservable) internal states. We do not think this is perfect, but it is an improvement.

Consider the study by Flynn and Lake (2008), where participants had to guess how many strangers they would need to ask on Columbia's campus before someone agreed to lend them a phone or accompany them to the gym. Participants underestimated others' willingness to help, predicting that they would have to ask more people than they did. Why? The evidence suggests that it's because they didn't fully account for the social pressure on others to say yes. That is, the very mechanism we highlight here (i.e., response pressure) may have operated on the participants in this study, compelling them to say "yes" to another person's request. Of course, predictors in that study may have felt response pressure to say the task was going to be harder than it was (e.g., it may be preferable to say it will take three asks than to more presumptuously assert that it'll take only one). But what's important about this

design is that it gives us something we can verify, and in so doing it changes the research question. Instead of asking whether people can guess how much others *want* to comply with a request, it asks whether they can predict whether they *will* comply with the request. That is not the same as understanding someone's thoughts and feelings. But it is a tractable alternative.

Implications Beyond Metaperception Research

These methodological issues have broader implications for experimental psychology, as the same problem can arise any time the assessment of the dependent variable differs across conditions. In such circumstances, it is difficult to know whether a difference between conditions is due to the intended manipulation or to the confound introduced by asking different questions in different conditions.

Any study with this structure is vulnerable to this concern. This includes any research comparing predictions and reports, because the way you ask a prediction question is not the same as how you ask a report question. Affective forecasting is a clear example. Participants are asked how they would feel after an event, and then later asked how they felt. The questions are inherently different. It could be that the questions carry different response pressures. It could also be that they are interpreted differently or evoke different reference points, leading people to use the response scale differently when they make predictions than when they report on their experiences (see Levine et al., 2012, for compelling evidence). If any of these differences push even some of the responses in different directions, then the questions alone could generate apparent errors in affective forecasts.

Conclusion

The question that motivates metaperception research is compelling. Are we biased perceivers? Do we misjudge how our actions affect others? Perhaps. Our findings cannot say one way or the other. But what they do show is that the current methods used to study these questions cannot answer them. And

unless we develop tools that allow us to elicit predictions and reports in ways that reveal purely authentic responses, the answers we seek will remain out of reach.

Constraints on Generality

In this research, we demonstrate that evidence for biases in metaperceptions comes from studies that confound experimental role – predictor vs. reporter – with the questions participants are asked and the response pressures those questions elicit. These differences in response pressures can generate results that look like biases in metaperceptions. It is important to acknowledge constraints on the generality of our investigation. First, our studies focus specifically on action-based metaperceptions, not on metaperceptions more broadly. We do not claim that all metaperceptual phenomena originate from the mechanisms we identify here, and future research should examine whether (and how) these concerns extend to other research on metaperceptions. Second, all studies were conducted online with U.S.-based samples. Although we have no reason to expect the observed effects to differ in laboratory or field settings, this possibility should be investigated. Moreover, cultural norms may shape both the nature and the magnitude of response pressures (e.g., Lalwani, Shrum, & Chiu, 2009), and so it is plausible that predictors' and reporters' answers would diverge differently in other cultural contexts. We hope future research will explore the boundary conditions of our claims and further clarify whether metaperception "biases" reflect genuine misunderstandings or artifacts of the questions people are asked to answer.

References

Abi-Esber, N., Abel, J. E., Schroeder, J., & Gino, F. (2022). "Just letting you know … " Underestimating others' desire for constructive feedback. *Journal of Personality and Social Psychology, 123*(6), 1362–1385.

Adams, G. S., Flynn, F. J., & Norton, M. I. (2012). The gifts we keep on giving: Documenting and destigmatizing the regifting taboo. *Psychological Science*, *23*(10), 1145-1150.

Aquino, K., & Reed II, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423.

Berry, Z., Lucas, B. J., & Jachimowicz, J. M. (2025). People overestimate how harshly they are evaluated for disengaging from passion pursuit. *Journal of Personality and Social Psychology*, 1102-1129.

Birnbaum, H. J., Wilson, D., & Waytz, A. (2024). Advantaged groups misperceive how allyship will be received. *Organizational Behavior and Human Decision Processes*, *181*, 104309.

Boothby, E. J., & Bohns, V. K. (2021). Why a simple act of kindness is not as simple as it seems: Underestimating the positive impact of our compliments on others. *Personality and Social Psychology Bulletin*, *47*(5), 826-840.

Cooley, C. (1922). *Human Nature and the Social Order*. New York: Scribners.

Crowne, D. P., & Marlowe, D. (1964), *The Approval Motive: Studies in Evaluative Dependence*. New York: Wiley.

Donnelly, K., Moon, A., & Critcher, C. R. (2022). Do people know how others view them? Two approaches for identifying the accuracy of metaperceptions. *Current Opinion in Psychology*, *43*, 119-124.

Dungan, J. A., & Epley, N. (2024). Surprisingly good talk: Misunderstanding others creates a barrier to constructive confrontation. *Journal of Experimental Psychology: General*, *153*(3), 779-797.

Dungan, J. A., Munguia Gomez, D. M., & Epley, N. (2022). Too reluctant to reach out: Receiving social support is more positive than expressers expect. *Psychological Science*, *33*(8), 1300-1312.

Elsaadawy, N., & Carlson, E. N. (2022). Do you make a better or worse impression than you think? *Journal of Personality and Social Psychology*, *123*(6), 1407-1420.

Elsaadawy, N., Carlson, E. N., Chung, J. M., & Connelly, B. S. (2023). How do people think about the impressions they make on others? The attitudes and substance of metaperceptions. *Journal of Personality and Social Psychology*, *124*(3), 640-658.

Epley, N., Kardas, M., Zhao, X., Atir, S., & Schroeder, J. (2022). Undersociality: Miscalibrated social cognition can inhibit social connection. *Trends in Cognitive Sciences*, *26*(5), 406-418.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013-1027.

Flynn, F. J., & Lake, V. K. (2008). If you need help, just ask: Underestimating compliance with direct requests for help. *Journal of Personality and Social Psychology*, *95*(1), 128-143.

Giurge, L. M., & Bohns, V. K. (2021). You don't need to answer right away! Receivers overestimate how quickly senders expect responses to non-urgent work emails. *Organizational Behavior and Human Decision Processes*, *167*, 114-128.

Givi, J., & Kirk, C. P. (2024). Saying no: The negative ramifications from invitation declines are less severe than we think. *Journal of Personality and Social Psychology*, *126*(6), 1103-1115.

Goffman, E. (1978). Response cries. *Language, 54*(4), 787-815.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464-1480.

Grice, H. P. (1975), Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts* (pp. 41-58). San Diego, CA: Academic Press.

Grutterink, H., & Meister, A. (2022). Thinking of you thinking of me: An integrative review of meta-perception in the workplace. *Journal of Organizational Behavior*, *43*(2), 327-341.

Haltman, C., Herziger, A., Donnelly, G. E., & Reczek, R. W. (2025). Better late than never? Gift givers overestimate the relationship harm from giving late gifts. *Journal of Consumer Psychology, 35*, 567-583.

Hart, E., Bear, J. B., & Ren, Z. B. (2024). But what if I lose the offer? Negotiators' inflated perception of their likelihood of jeopardizing a deal. *Organizational Behavior and Human Decision Processes*, *181*, 104319.

Hart, E., VanEpps, E. M., & Schweitzer, M. E. (2021). The (better than expected) consequences of asking sensitive questions. *Organizational Behavior and Human Decision Processes*, *162*, 136-154.

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin*, *76*(5), 349-364.

Kardas, M., Kumar, A., & Epley, N. (2024). Let it go: How exaggerating the reputational costs of revealing negative information encourages secrecy in relationships. *Journal of Personality and Social Psychology*, *126*(6), 1052-1083.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537-567.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29-44.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025-2047.

Kumar, A., & Epley, N. (2018). Undervaluing gratitude: Expressers misunderstand the consequences of showing appreciation. *Psychological Science*, *29*(9), 1423-1435.

Kumar, A., & Epley, N. (2023). A little good goes an unexpectedly long way: Underestimating the positive impact of kindness on recipients. *Journal of Experimental Psychology: General*, *152*(1), 236-252.

Lalwani, A. K., Shrum, L. J., & Chiu, C. Y. (2009). Motivated response styles: The role of cultural values, regulatory focus, and self-consciousness in socially desirable responding. *Journal of Personality and Social Psychology*, *96*(4), 870-882.

Levine, E. E., & Cohen, T. R. (2018). You can handle the truth: Mispredicting the consequences of honest communication. *Journal of Experimental Psychology: General*, *147*(9), 1400-1429.

Levine, L. J., Lench, H. C., Kaplan, R. L., & Safer, M. A. (2012). Accuracy and artifact: reexamining the intensity bias in affective forecasting. *Journal of Personality and Social Psychology*, *103*(4), 584-605.

Liu, P. J., Rim, S., Min, L., & Min, K. E. (2023). The surprise of reaching out: Appreciated more than we think. *Journal of Personality and Social Psychology*, *124*(4), 754-771.

Lu, J., Fang, Q., & Qiu, T. (2023). Rejecters overestimate the negative consequences they will face from refusal. *Journal of Experimental Psychology: Applied*, *29*(2), 280-291.

Monin, B., & Jordan, J. (2009). Dynamic moral identity: A social psychological perspective. In D. Narvaez & D. K. Lapsley (Eds.), *Personality, Identity, and Character: Explorations in Moral Psychology* (pp. 341–354). Cambridge University Press.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776-783.

Paulhus, D. L. (1991), Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes, Vol. 1* (pp. 17-59). San Diego, CA: Academic Press.

Pei, R., Grayson, S. J., Appel, R. E., Soh, S., Garcia, S. B., Bouwer, A., Huang, E., Jackson, M. O., Harari, G. M., & Zaki, J. (2025). Bridging the empathy perception gap fosters social connection. *Nature Human Behaviour*, *9*(10), 2121-2134.

Savitsky, K., Epley, N., & Gilovich, T. (2001). Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps. *Journal of Personality and Social Psychology*, *81*(1), 44-56.

Scheff, T. J. (1988). Shame and conformity: The deference-emotion system. *American Sociological Review*, *53*, 395-406.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93-105.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology*, *18*(2), 247-255.

Tedeschi, J. T., Schlenker, B. R., & Bonoma, T. V. (1971). Cognitive dissonance: Private ratiocination or public spectacle? *American Psychologist*, *26*(8), 685-695.

Thomas, M., & Kyung, E. J. (2019). Slider scale or text box: How response format shapes responses. *Journal of Consumer Research*, *45*(6), 1274-1293.

Vorauer, J. D., Hodges, S. D., & Hall, J. A. (2025). Thought-feeling accuracy in person perception and metaperception: An integrative perspective. *Annual Review of Psychology*, *76*(1), 413-441.

Whillans, A. V., Yoon, J., & Donnelly, G. (2022). People overestimate the self-presentation costs of deadline extension requests. *Journal of Experimental Social Psychology*, *98*, 104253.

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology*, *16*(1), 155-188.

Zhao, X., & Epley, N. (2021a). Insufficiently complimentary? Underestimating the positive impact of compliments creates a barrier to expressing them. *Journal of Personality and Social Psychology*, *121*(2), 239-256.

Zhao, X., & Epley, N. (2021b). Kind words do not become tired words: Undervaluing the positive impact of frequent compliments. *Self and Identity*, *20*(1), 25-46.

Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, *7*(10), 1667-1681.