The Overblown Implications Effect

Date Submitted: 11/6/2017

Abstract

People frequently engage in behaviors that put their competencies on display. But do actors understand how others view them in light of these performances? Seven studies support an *overblown implications effect* (OIE): Actors overestimate how much observers think an actor's one-off success or failure offers clear insight about a relevant competency. Such effects were observed on judgments of intelligence (Study 1) and likeability following a trivia quiz or getting-acquainted interview, respectively. To explain the OIE, we introduce the construct of *working trait definitions*—accessible beliefs about what specific skills define a general competency. When actors—those under the threat of evaluation—try to adopt observers' perspective, the narrow performance domain seems disproportionately important in defining the general trait. As this account anticipates, actors overblow performances' implications even in prospect, before there are successes or failures on which to ruminate (Study 3). Furthermore, such errors emerge when one considers being the object of evaluation, not merely when one considers how another would make social evaluations (Study 4). A novel intervention that broadened actors' working trait definitions to include other (unobserved) trait-relevant behaviors eliminated the OIE (Studies 5-6). A final study (Study 7) more precisely localized the meta-inferential error. Although meta-perceivers and observers agreed on what a single success or failure (e.g., the quality of a single batch of cookies) could reveal about actors' narrow competence (e.g., skill at baking cookies), meta-perceivers erred in thinking observers would feel this performance would reveal a considerable amount about the broader skill (e.g., cooking ability).

Keywords: meta-perceptions, social judgment, working trait definitions, behavioral diagnosticity

The Overblown Implications Effect

"Is that your *final* answer?" In the well-known game show *Who Wants to Be a Millionaire?*, the contestant sits in the hot seat answering trivia questions for a shot at riches. For the contestant, each question is high stakes. Most obviously, the monetary stakes are high: Correct answers are necessary to advance toward the shot at the million dollars. But less focally, the evaluative stakes are high as well: Contestants' every move is being closely watched by a couple hundred studio audience members and thousands more at home.

Social psychologists have long appreciated the importance of others' presence on behavior. Such evaluative pressure can at times facilitate effort and thus, performance (Harkins, 2006; Zajonc, 1965; see Seitchik, Brown, & Harkins, 2017, for a review), but at other times can lead performers to choke (Baumeister, 1984). The contrasting findings both reinforce that evaluative stakes matter, but do actors appreciate the true evaluative stakes when in the spotlight? In this paper, we argue that a concern about imminent evaluation leads people to show an *overblown implications effect* (OIE): Actors exaggerate how much their performance speaks to their broader competencies in observers' eyes. We ultimately explain this phenomenon by introducing and demonstrating the role of a novel construct, *working trait definitions*—what behaviors form one's momentary definition of a trait or competency. The OIE arises because observers' actual working trait definitions are broader than actors assume. In introducing this new construct (working trait definitions) and developing our theoretical account, we will give particular attention to how this novel psychological mechanism distinguishes itself from previously identified reasons why meta-perceptions—people's guesses of how others view them—err.

Meta-perceptions

To date, much of the research on meta-insight has examined whether people's meta-perceptions correlate with how others actually view them, or are merely an egocentric product of self-perceptions (Kenny, 1994; Kenny & DePaulo, 1993). And indeed, correlations between self- and meta-perceptions emerge. For example, those who view themselves as sociable believe others are more likely to view them as sociable. People believe their self-perceptions are accurate, and thus, leaning on them to form meta-perceptions would seem reasonable (e.g., Kenny & DePaulo, 1993; Albright, Forest, & Reiseter, 2001; Albright & Malloy, 1999; Malloy, Albright, Kenny, Agatstein, & Winquist, 1997). But beyond this, people also possess special meta-insight—an understanding of how they are viewed by others that does not merely stem from self-perceptions (Carlson, Vazire, & Furr, 2011; Carlson & Kenny, 2012; Vazire & Carlson, 2010). For example, people understand that they make different impressions on different groups of people, such as on their friends versus their parents (Carlson & Furr, 2009), suggesting that people's meta-perceptions do not merely rely on their global, stable self-perceptions.

In this paper, we examine *errors* in meta-insight, but depart from much of the just-reviewed approaches in two key ways. First, instead of examining whether people possess meta-insight for broad personality characteristics, we examine how impressions of competencies are formed or moved by one-off successes or failures. Second, we focus not on correlations between observers' social perceptions and actors' meta-perceptions, but instead at systematic mean-level miscalibration between them. After all, two drivers who complete a simple parallel parking job in 3 versus 23 turns clearly possess better and worse driving ability, respectively. In other words, we expect that meta- and social perceptions of these two drivers would likely correlate. But by

examining mean-level biases in impressions, we can know whether observers' impressions are influenced by such performances to a greater or lesser extent than actors assume.

Why would actors fail to appreciate how they are judged in light of their performance? To begin answering this question, it is important to consider what is being judged when evaluating a trait or competency. For example, what does it mean to evaluate someone as intelligent? Most obviously, intelligent people do intelligent things. They may use more ornate words, remember obscure facts from childhood, or even know the order of all 118 elements in the Periodic Table.

But cues to a trait or quality are rarely observed all at once and are not perfectly correlated. Trait-relevant behaviors show variability across situations (Fleeson, 2004; Pervin, 1994; Ross & Nisbett, 1991). In part, different situational factors either overwhelm or selectively activate different aspects of one's personality (Cramer et al., 2012). But also, any single behavior gives only so much information about a particular quality in question. Social perceivers appreciate this notion: When determining another's intelligence, one wants to know more than whether they make it all the way from hydrogen to ununoctium.

From this perspective, one potential pitfall to meta-insight is actors may fail to understand how observers characterize or define the trait or competency in question. Psychologists acknowledge that representations of complex constructs may be based on only a limited amount of information at any one time. Consider the self, a target about which we have an almost overwhelming amount of information. Despite these vast stores, people's *working self-concepts*—their accessible self-knowledge (Markus & Wurf, 1987)—show moment-to-moment variability (Markus & Kunda, 1986; DeSteno & Salovey, 1997; Cervone & Shoda, 1999;

McConnell, 2011) with predictable consequences for judgment and behavior (e.g., Showers & Zeigler-Hill, 2003).

Much as people have working self-concepts, we propose that people hold *working trait definitions*. That is, at any given moment, people define a trait or competency by drawing on only a subset of potentially relevant behavioral dimensions. Previous researchers have recognized that different people define traits or competencies in different (often self-serving) ways (Critcher, Helzer, & Dunning, 2011; Dunning, Meyerowitz, & Holzberg, 1989). We propose that even the same person will show variability in their trait definitions—more specifically, their working trait definitions—across time.

Recent research has argued that under threat, the working self-concept constricts around the threatened domain, leading this damaged identity to occupy a larger portion of the active self-concept (Critcher & Dunning, 2015). For example, after failing an exam, one's academic self looms large in the working self-concept, thereby exerting a disproportionately large effect on one's feelings of self-worth. By analogy, we suggest that working trait definitions may show similar properties. As actors consider the impressions observers form of them, this evaluative threat may cause the performance domain to loom large in actors' working trait definitions. For example, while parallel parking, drivers may assume that an observer's definition of good driving is heavily dominated by parallel parking ability. As a result, actors see their own performance as likely to exert considerable sway on observers' impressions. But observers' actual impressions do not have the same evaluative stakes for observers themselves. Thus, observers' working trait definitions may not be as dominated by the performance behavior as actors anticipate. That is, for the observer on the sidewalk, driving skills encompass not merely a particular type of parking maneuver, but also awareness of blind spots, attention to road signs,

and maintaining safe distances from other drivers, among other such indicators. We propose that this asymmetry is a primary cause of the overblown implications effect. We summarize this account in Figure 1.

[INSERT FIGURE 1 HERE]

Empirical and Conceptual Similarities To and Differences From Previous Work

To our knowledge, we are the first to posit the existence of working trait definitions and use them to explain why meta-perceptions may err. But we are of course not the first to explore whether people show systematic biases in understanding the impressions their performances leave on others. The most similar and influential work in this tradition has shown that people fail to recognize how charitably others respond to their own blunders (Epley, Savitsky, & Gilovich, 2002; Savitsky, Epley, & Gilovich, 2001). Just as decades of research have identified reasons why self and social perceptions display biases (e.g., Critcher, Dunning, & Rom, 2015; Dunning, 2005; Kruger & Dunning, 1999; Ross & Sicoly, 1979), we have no doubt that failures of meta-insight are themselves multiply determined. That said, we enumerate five ways that our account distinguishes itself conceptually and empirically from past research. We will repeatedly reference these differentiators by number and title as we present our results both in order to clarify our contribution and keep the focus on the larger issues that a cumulative science of meta-insight will ultimately have to address. In the General Discussion, we return to the question of whether the present account complements past research as a qualitatively new direction in research on meta-insight, or if instead it offers a theoretical reinterpretation of previous findings.

Issue 1: Scope (Overblowing successes as well as failures). We argue that when actors consider their behavior through the evaluative eye of observers, their working trait definitions constrict. This implies actors should overblow the social judgment implications of both failures

and successes. In contrast, Savitsky et al. (2001) and Epley et al. (2002) demonstrated their effects on fear of negative evaluations after failures exclusively. Savitsky et al. (2001) noted that although their primary mechanism anticipates that those who perform especially well would expect even more adulation than observers would actually grant them, two additional mechanisms they speculate are in play would have the reverse effect. They appeal to Kruger and Gilovich (1999), who coin and support the concept *naïve cynicism*—that people assume others are more self-serving and other-disparaging than they are. Savitsky et al. (2001) argue observers' empathy inspires "judgmental charity," which allows observers "to empathize with [actors'] misfortune and withhold harsh judgment" (p. 54). Epley et al. (2002) call actors' tendency to overlook observers' charity *empathy neglect*. This combination of mechanisms suggests that the overblown implications effect will either be reduced, eliminated, or even reversed for successes.

Issue 2: Actors' distortion (Trait definitions vs. performance focalism). We argue that meta-perceivers assume observers *define* traits or competencies more narrowly than they do. In contrast, previous work has attributed errors in meta-insight to a set of related mechanisms all placed under the umbrella term "focalism." One form of focalism occurs when meta-perceivers become focused on a small portion of their *performance.* Savitsky et al. (2001) suggest actors ruminate on their blunders, prompting those unrepresentative moments to loom large in mind. For example, participants (misleadingly and sometimes inaccurately) outed as bedwetters failed to appreciate how additional information would water down observers' negative impressions (Savitsky et al., 2001). As Epley et al. (2002) noted, "musicians who miss one key note in a concert still hit countless others" (p. 310). Whereas these accounts focus on *performance focalism*—disproportionate attention to a blunder while ignoring the largely competent moments—we focus on what could be called *definitional focalism*. In particular, our account

applies even in contexts in which a blunder or success occurs *in isolation*–-that is, without the context of opposing information (e.g., the rest of the concert). In such cases, observers' judgments are tempered because they recognize that they lack information about actors' relative competencies ("I only know he can bake cookies, not whether he is a good cook"), not because they focus on the actor's blunder at the expense of other information (which is unknown).

Issue 3: Observer's perspective (Reduced diagnosticity vs. insight into difficulty). In evaluating performance, one must translate an objective outcome (e.g., answering a trivia question right or wrong) into a subjective evaluation. By our account, observers' evaluations are muted (compared to actors' expectations) because they see limited informational value in any one performance domain in what it signals about a broader competency. This means that when an actor fails or succeeds, observers' subjective impressions will not shift considerably. In contrast, previous research has emphasized that observers are in privileged positions that help them recognize the difficulty of a performance context (Epley et al., 2002). That is, observers often have one additional data point compared to actors: Observers witnessing failure often know that *they too* would have failed in a context. This allows observers to understand—for example—that a trivia question is especially difficult, but actors lack the same privileged perspective. This too has been identified as a form of focalism: Actors focus on the superficial characterization of their performance as a "failure," not appreciating how observers will be able to empathize with that failure as understandable. This argues not that observers will see less diagnosticity in tasks, but that their own privileged perspective allows them to recognize a performance context as objectively difficult or easy. This alternative has empirically distinct implications: Observers' personal insight that a task is actually quite difficult [easy] means they should judge targets—

regardless of whether they succeed or fail—more positively [negatively] than actors would expect.

Issue 4: Timing (Prospective vs. retrospective). We have argued that constricted working trait definitions come from the evaluative apprehension of considering an observer's evaluative eye, not from rumination on a previous blunder or success. Previous research, on the other hand, argues that errors in meta-prediction occur because "people's mishaps are often highly salient to them" (p. 49, Savitsky et al., 2001) and become a focus of undue attention. This reasoning would not anticipate an overblown implications effect *before* there is a failure (or, if one extends their account, a success) to focus on. Furthermore, before an event plays out, there is not additional non-focal information that could temper meta-perceivers' hyperfocused perspectives. But by our account, as actors consider engaging in a future behavior, the perceived evaluative stakes grow in their mind due to the constricted working trait definitions. That is, such behaviors seem to carry evaluative weight not because actors continue to play out the highlights and lowlights of their previous performance, but because they imagine observers seeing those performance domains as relatively all-defining of the broader competencies in question.

Issue 5: Error specificity (OIE on global but not narrow competencies). By our account, meta-insight fails when translating an evaluation of a performance task (e.g., parallel parking) to an evaluation of a broader competency (e.g., driving). This contrasts with previous work, which has focused on an earlier translational step: deciding what a single performance, especially when undertaken under suboptimal conditions, reflects about the narrow competence in question. For example, Epley et al. (2002) had some participants sing the "Star Spangled Banner" while chewing gum. Because people overestimate how much observers display the correspondence bias (Van Boven, Kamada, & Gilovich, 1999), meta-perceivers fail to appreciate

how much observers actually do realize the gum plays an interfering role. We instead argue that there is only so much observers feel they learn about another's musical talents by hearing the performance of one song. That is, we expect the OIE to hold even when there are no situational explanations to appeal to (i.e., singing while chewing gum). For example, we predict that meta-perceivers will be quite accurate in guessing how much diagnosticity an observer will see in a one-off (unimpeded) performance of the "Star Spangled Banner" in diagnosing skill with singing the national anthem. But despite such accuracy, our account predicts meta-perceivers will fail to realize that observers will see little information about their musical ability more generally based on this one-off performance.

Overview of the Present Research

The present research examined if actors overweigh how much a specific behavioral performance factors into observers' perceptions of them. In an effort both to clarify our contribution and to help build the multiple strands of research on performance meta-insight into a more cumulative science, we repeatedly return to the five issues just summarized when considering the broader meaning of our results. We begin by testing the overblown implications effect in two evaluative domains. Unbeknownst to participants, we randomly assigned actors to fail or succeed (Issue 1: Scope) at performance tasks that could have implications for their intelligence (Study 1) and likeability (Study 2). We distinguished whether meta-perceivers' errors reflected their tendency to focus on a small part of their performance when commenting on their performance more generally (as past research might anticipate), or whether meta-perceivers misunderstood even what meaning would be seen in their specific performance behavior (Issue 2: Actor's distortion). Also, findings that observers' responses to successes versus failures are blunted—as opposed to generally more charitable or harsher than expected—

would reinforce our characterization that observers' see less diagnosticity in actions, not that they have greater insight into task difficulty (Issue 3: Observers' perspective).

Next, we proceeded to determine whether the OIE emerges in *prospective* judgments of behavior's diagnosticity (Issue 4: Timing), even before rumination on past performance or mitigating situational information could occur (Study 3). We also tested whether the OIE specifically characterizes actors'—as opposed to an uninvolved bystander's—guesses of observers' judgments (Study 4). Studies 5 and 6 directly manipulated actors' and observers' working trait definitions to determine whether expanding them would debias meta-perceptions while leaving observers' judgments unaffected. Study 7 tried to localize the OIE to the translation from evaluations of specific competencies (about which we expected actors and observers would agree) to the assessment of what these say about broader competencies (where we expected the error in meta-insight to emerge; Issue 5: Error Specificity).

Compliant with Simmons, Nelson, & Simonsohn (2011), we report all manipulations below. The only exclusions were participants for whom there was a problem collecting data (e.g., a video of them did not record). For the multi-stage lab studies (Studies 1, 2, and 6), we collected as many participants as we could in a semester for the subject pools that we used. For the survey studies (Studies 3-5), we aimed for at least 100 participants per cell. For Study 7, for which we expected smaller effect sizes, we aimed for 200 participants per cell.[1] We report exploratory or unanalyzed measures in the Supplemental Materials.

[1] To get a sense for whether such sample sizes are appropriate, the average sample sizes reported in the work we saw as most similar were approximately 23.6 per cell (Savitsky et al., 2001) and

Study 1

Study 1 tested the *overblown implications effect* in the domain of intelligence. Actors (contestants) took part in a mock game show and guessed how observers (audience members) would judge them. To provide information upon which baseline impressions could be based, actors began by answering a set of trivia questions. Based on this information, participants offered their baseline impressions of actors' intelligence: Observers offered social judgments of contestants, and actors offered meta-judgments of how observers likely viewed them. We then engineered a success or failure. We predicted an *overblown implications effect*, that meta-judgments would shift more in light of this performance than social judgments actually do.

Only because many readers may be familiar with Ross, Amabile, and Steimetz's (1977) classic quiz bowl paradigm, it may be useful to consider three ways in which the present paradigm intentionally differs. First, Ross et al. (1977) had participants generate questions for each other, whereas we used the same experimenter-provided questions across participants. Whereas Ross et al. (1977) focused on how observers fail to appreciate the structural advantage that (the question-generating) observers possess, we wanted to avoid such contexts given their relevance to other mechanisms (Van Boven et al., 1999) instead of our proposal. Second, we wished to standardize the performance task and feedback. Third, we wanted to remove the observer from the live context in an effort to be more confident that any observed effects reflect our newly proposed (instead of a previously studied) mechanism. That is, when observers are focused on their own performance and the impressions they are making, they can fail to notice

---

26.5 per cell (Epley et al., 2002). Of course, our goals and paradigms differ, so we felt it was good to err toward using that precedent merely as an approximate minimum standard.

variability in others' performance (Gilovich, Kruger, & Medvec, 2002). Because actors do not understand that observers are distracted, they experience a failure of meta-insight. But crucially, such differences were found to disappear when observers were completely removed from the performance context. We wanted to make sure that differences between meta-perceptions and social perceptions could be attributed to differences in the perceived evaluative implications of actors' performance, not a failure to notice them.

Method

Participants and design. One hundred ninety-eight undergraduates from an American university completed a lab session for course credit. Actors were randomly assigned to a *success* or a *failure* condition. Each observer was yoked to a randomly selected actor.

Procedure and materials. We describe actors' experience first. Observers observed their yoked actor's complete experience—both by learning the instructions actors received and watching the actors' performance on video.

Actors. Actors took part in the study individually. Upon arrival, actors were seated in front of a laptop. The experimenter informed actors (accurately) that they would be videotaped throughout the entire study so that a future participant could observe their performance. Actors were told that they would be answering a series of multiple choice trivia questions. Along with providing each answer, actors explained their rationale behind their selection aloud. This made sure that actors would offer plenty of behavioral information, beyond their multiple-choice response, that could be interpreted to signal their competence and intelligence. Actors were told that for each of the ten questions they answered correctly, they would be given a ticket to enter into a lottery drawing for a $50 Amazon.com gift card. We included this ticket scheme so that—as we explain below—we could create a high-stakes moment for actors during the game.

Actors were presented 10 difficult trivia questions, one at a time, on the laptop. Each trivia question had two answer choices. As an example, one question asked: "Which city has the higher crime rate: Chicago or Detroit?" Actors read each question aloud, indicated their answer, and explained why they chose their answer. Actors completed all three steps out loud, so that their yoked observers would be able to observe the full process. After completing all 10 of the trivia questions and regardless of their actual performance, actors were informed that they had answered 7 out of the 10 questions correctly. Most likely because we did not indicate which of the questions were supposedly answered correctly or not, no actor or observer expressed suspicion during a funnel debriefing about this feedback's credibility.

At this point, the experimenter gave actors seven lottery tickets for the seven questions they had supposedly answered correctly. Following this standardized initial feedback, actors completed the *baseline perception* measures of intelligence. Actors provided both meta-perceptions (guesses of how the observers would rate them) and self-perceptions (evaluations of their own performance on the task), in a counterbalanced order. (Because we found that our effect was uninfluenced by order, we did not vary order in subsequent studies). These measures are described in more detail below.

Next, the experimenter returned to explain the performance task. Actors were told, "Now, you will answer one Trivial Pursuit question for *double or nothing*. That is, if you answer this question correctly, you will double your chances of winning with a total of 14 lottery tickets for the $50 Amazon.com gift card. But, if you answer incorrectly, you will lose all your tickets and be left with nothing. Again, please explain your reasoning *out loud* for the following question." We standardized participants' feedback on the first round so that the implications of this final

question would be equivalent for all participants. The final question was: "Which novel was published first: *To Kill a Mockingbird* or *The Catcher in the Rye*?"

Our goal was to make this final question feel especially high-stakes as a performance event. In addition to raising the stakes on this question (making it double or nothing), we upped the evaluative stakes by having the experimenter read the question, listen to the actor's reasoning and answer, and then provide verbal feedback. (In the initial round, these steps had been taken by the computer.) We were also careful to choose a question related to a topic with which our participants would be familiar (both books are staples on required reading lists), but not so familiar that they would clearly know the precise question being asked (their publication dates). This allowed us—unbeknownst to participants—to randomly assign participants to learn they had (supposedly) answered this question correctly or incorrectly. Based on this randomly assigned feedback, participants either received seven more lottery tickets (*success* condition) or had their seven tickets taken away (*failure* condition). Again, no actor or observer expressed suspicion about this feedback's credibility.

Finally, actors completed the *final perception* measures. These took the same form as the baseline perception measure—the meta-perception and self-perception measures of intelligence completed before the performance event. After completing these final measures, actors were debriefed and apologized to for the mild deception. They were informed that all participants had an equal chance of receiving the $50 prize.

Observers. Each observer was yoked to one actor. Observers had the same experience as actors, but as onlookers to the situation instead of as active participants. That is, they learned what instructions had been given to actors, but the observers then watched the actors perform on

video instead of answering the questions themselves. Prior to the experimental session, research assistants clipped the full-length footage of the actors into two shorter videos to show observers.

The first video showed the actor answering the first 10 trivia questions and ended with the experimenter coming in to give the actor the seven tickets for the supposed seven trivia questions that they answered correctly. After watching this video, observers rated their baseline *social perceptions* of the actor's intelligence. The second video showed the experimenter reading the final question to the actor, the actor answering the question, and the experimenter providing the final feedback. To make sure that observers were not less reactive to the feedback than actors were merely because observers failed to notice these performance details (Gilovich et al., 2002), the computer instructions reiterated the performance outcome to observers before they made their final judgments. At that point, observers completed their final social perceptions of the actors' intelligence.

Trait perceptions. The perception measure comprised five items that asked about the actor's intelligence. The *social perceptions* asked observers to judge the actors in light of their performance. The *meta-perceptions* instructed actors to guess how observers would judge them in light of their performance. More specifically, actors saw the same prompt given to the observers, and they were asked to guess the observers' responses. The *self-perceptions* asked actors to judge themselves in light of their performance.

The first three questions asked participants to rate the actor as competent, intelligent, and knowledgeable on 9-point scales anchored at 1 (*not at all*) and 9 (*extremely*). The fourth item asked what score the actor would likely get on an IQ test. Although participants supplied their own numerical score, they were given the following guide in case they were unfamiliar with the standard IQ scale: "80-89 = below average; 90-109 = average; 110-119 = above average; 120-

139 = gifted; >140 = genius." Finally, participants were asked into what percentile the actor's IQ fell in comparison to other undergraduates at their university.

Because our five items used different response scales, we first standardized responses. We calculated the grand mean and standard deviation of each item across both baseline and final meta, social, and self-perception judgments. By relying on these sample statistics when standardizing each measure, we preserved all effects of perception type (meta, social, and self) and time. The meta-perception ($\alpha = .77$), social perception ($\alpha = .75$), and self-perception ($\alpha = .77$) of intelligence composites all had good internal reliability.

Results and Discussion

We wanted to test whether responses to the performance feedback depended on the nature of the perception (meta, social, or self). Toward this end, we submitted the perception composites to a 2 (feedback: success or failure) X 3 (perception: meta, social, or self) X 2 (time: baseline or final) mixed-model ANOVA, with only the first factor manipulated between subjects. We found a significant three-way interaction, $F(2, 194) = 3.26$, $p = .04$, $\eta^2_p = .03$, demonstrating that different perceivers responded differently to actors' success versus failure. We proceeded to conduct a series of 2 (feedback) X 2 (perception) X 2 (time) ANOVAs to understand whose perceptions were out of step with whose.

[INSERT FIGURE 2 HERE]

First, did observers respond to the high-stakes final round like actors thought they would? A significant 2 (feedback) X 2 (perception: meta or social) X 2 (time) mixed-model ANOVA

suggested that that they did not, $F(1, 97) = 4.25$, $p = .04$, $\eta^2_p = .04$[2]. Decomposing this interaction provided evidence of the overblown implications effect: Observers responded less extremely to the actors' final performance than actors thought they would. As depicted in Figure 2, observers did shift their social impressions in response to actors' success versus failure, $F(1, 97) = 17.93$, $p < .001$, $\eta^2_p = .16$. But this shift was less pronounced than actors thought it would be, $F(1, 97) = 60.61$, $p < .001$, $\eta^2_p = .38$.

Second, we examined whether actors' meta-perceptions merely reflected their own self-perceptions of their performance. That is, although we argued that actors overblow the implications of their own performance when estimating how they are being evaluated, perhaps they more generally overblow the meaning of their own behavior in their own mind as well. Contradicting this possibility, we found that actors' self-perceptions and their meta-perceptions diverged. That is, a significant 2 (feedback) X 2 (perception: self or meta) X 2 (time) mixed-model ANOVA showed these judgments diverged, $F(1, 97) = 7.81$, $p = .006$, $\eta^2_p = .07$. Although self-perceptions were reactive to the feedback, $F(1, 97) = 37.81$, $p < .001$, $\eta^2_p = .28$, they were less so than meta-perceptions.

Third, and related to the last analysis, we asked whether actors would have had a more accurate understanding of how observers actually responded to their performance if actors had merely leaned on their own self-perceptions. As foreshadowed by these two previous analyses, a final 2 (feedback) X 2 (perception: self or social) X 2 (time) mixed-model ANOVA returned a

[2] We caution readers against comparing effect sizes of the 2(feedback) X 2(perception) X 2(time) ANOVAs because they vary in whether the two perception conditions are measured within-participants (e.g., self and meta) or manipulated between-participants (e.g., meta and social).

non-significant three-way interaction, $F < 1$. Thus, both actors and observers responded to the actors' high-stakes performance in a similar way—i.e, less reactively than actors thought observers would.

The divergence between meta-perceptions and social perceptions provides straightforward support for the overblown implications effect. But—as might be foreshadowed by the naïve cynicism and empathy neglect accounts—was this effect driven merely by divergent interpretations of failure (Issue 1: Scope)? To the contrary, we unexpectedly found that the effect was driven by different responses to success: Actors thought observers would be more responsive to actors' successes than observers were. On the one hand, the lesson for Issue 1 (Scope) is clear: The OIE extends to, and does not encounter a boundary condition at, the success domain. Before speculating why Study 1's result is asymmetric, we should foreshadow that this asymmetry is not robust. In fact, another study using this same paradigm (Study 6) found a symmetric OIE. Furthermore, an internal meta-analysis reported in the General Discussion will reinforce the idea that the OIE is, in general, symmetric between successes and failures.

Although our hypotheses focus on how different perceptions are differentially responsive to performance outcomes (Feedback X Perception X Time interactions), we did also find a main effect of Perception that is reminiscent of the alternative accounts. That is, we found that actors expected observers to judge actors more harshly than observers actually did (see General Discussion for a cross-study meta-analysis). In other words, this naïve cynicism may operate in the background (depressing all meta-perceptions), even as the orthogonal *overblown implications effect* may characterize actors' failure to understand dynamic shifts (or, more accurately, the lack of such shifts) in observers' perceptions.

We have argued that meta-perceivers think that observers define traits, such as intelligence, more narrowly than they do. That is, we argue that meta-perceivers assume the performance domain—in this case, trivia—would loom large in observers' working trait definitions of intelligence. For observers, their broader working trait definitions meant there was only so much they could learn about actors' intelligence from their performance on this task. But Study 1's results are also consistent with an alternative account—one rooted in performance focalism (Issue 2: Actor's distortion). By this alternative, actors were disproportionately focused on the high-stakes final round to the neglect of all of the cues to their intelligence they had already given. By this alternative account, actors and observers actually agreed on the implications of the failure or success, but actors simply failed to appreciate that for observers that was merely one question among eleven that they answered. Study 2 distinguishes these possibilities.

Study 2

Study 2 aimed to expand on Study 1 in two ways. First, we wanted to test the overblown implications effect in a new context and for a new trait. More specifically, we placed some participants (actors) in a social context in which they were socially accepted or rejected. Observers watched the social dynamic unfold. We then examined how this social success or failure changed perceptions of actors' likeability. According to the overblown implications effect, actors' meta-perceptions of how observers view them should be more reactive to the success or failure than observers' perceptions actually are.

Second, we wanted to begin to disentangle our explanation for why actors' impressions are distorted compared to a mechanism identified in previous research (Issue 2: Actors' distortion). By an alternative performance focalism account, actors fail to understand observers

not because they disagree on the implications of being rejected, but because actors forget all of the other cues they have passed on to the observers about their own sociability. That is, actors focus on their one sour note (or impressive high note) and forget about the rest of the song they performed. In contrast, our working trait definition account suggests that actors should fail to appreciate how little observers feel like they learn from this social performance context itself. That is, observers' broader working trait definitions of likeability should be much broader than what can be gleaned from the present context. We disentangle these possibilities by separating actors' performance into two halves. Their success or failure was based only on the second half. This allowed us to see if actors' meta-perceptions erred because they gave too much weight to the half on which they performed well or poorly (as a performance focalism account might anticipate), or because actors failed to understand the broader implications of the success or failure (even when everyone was more narrowly focused on the portion in which the success or failure occurred).

After the acceptance or rejection, participants offered impressions of two types. Like in Study 1, participants offered a *global* final impression. But new to Study 2, participants offered *feedback-informed* judgments. These were commentaries only on the performance event—i.e., the second-round interview that resulted in acceptance or rejection. By our account, we should observe the OIE even on this more specific, feedback informed perception measure. By a *performance focalism* account, actors should understand what observers see as the significance of this event, but simply fail to appreciate how much that impression is watered down when considering the global impression of the overall interaction.

Method

Participants. Two hundred forty-eight undergraduates completed a lab session for course credit. All actors were randomly assigned to the social *success* or *failure* condition. Observers were yoked to a randomly selected actor.

Procedure and materials. We describe actors' experience first. Observers learned the instructions actors received and read all of their specific actor's responses. Figure 3 summarizes the general procedure for actors and observers.

[INSERT FIGURE 3 HERE]

Actors. Actors were seated in a private room in front of a laptop. They were told that they would be completing a study examining: 1) whether people work better with those they like, and 2) whether outside observers can anticipate which groups will work best. Actors were led to believe that there were three participants present—two who would be interviewees and one who would be the interviewer. Actors were also informed (truthfully) that in a later session, an observer would watch the situation unfold from the vantage point of the actor, but would not participate in the groups themselves. A graphic was shown along with these instructions to more clearly illustrate the different roles involved.

Actors were told that they would be randomly assigned via a card-choosing task to either the interviewer role or one of two interviewee roles. Actors were asked to select one of three cards. Although participants did indicate a choice, all actors were told that their card selection assigned them to a role of "Interviewee." The interviewer and second interviewee were actually preprogrammed, fictitious participants.

Actors—as Interviewee #1—were told that they would answer two rounds of questions: (1) a practice round question, not shown to the interviewer, to familiarize them with the format of the question, and (2) an interview round question, based on which the interviewer would

choose whether they would like to work with one, both, or neither of the interviewees on the final "fun" task. A diagram (see Figure 4A) reiterated the role the actor would take as well as the roles of the other participants. To enhance the believability that there were other participants in the current session (playing the role of the interviewer and the other interviewee), the next screen displayed a spinning "loading" icon for five seconds and explained that the survey would automatically continue once everyone had finished learning about their roles.

[INSERT FIGURE 4 HERE]

First, actors completed the practice round of questions, which only the observer would see. This was included so that observers would have some grounds on which to offer baseline perceptions. In the practice round, actors listed three values that were important to them, then described why each of the values they listed had importance to them. They described each value for one minute, sequentially, until they described all three values. At this point, actors completed the baseline perception measures of likeability. Actors provided self-perceptions (evaluations of their own performance) followed by meta-perceptions (guesses of how the observers would rate them).

Next, actors completed the interview round of questions, which would be seen by both the interviewer and the observer. In the interview round, actors listed three of their best qualities, then described an instance in which they exhibited each. They described each quality for 90 seconds, sequentially, until they described all three qualities. After completing the interview round, actors waited for the interviewer to make a decision. Actors saw a spinning "loading" icon for 25 seconds and were informed that the interviewer was reading their responses (as well as those of the other interviewee) to decide whether they would like to work with one, both, or neither of the interviewees.

Actors were randomly assigned to learn that the interviewer had chosen to work with them but not the other interviewee (social *success*), or that the interviewer had chosen to work with the other interviewee but not with them (social *failure*). At this point, actors completed the *final* meta-perception and self-perception measures of likeability—both *global* (based on the entire study) and *feedback-informed* (focusing only on the act of social acceptance or rejection). Finally, actors were debriefed and apologized to for the mild deception. No actor or observer expressed suspicion about the legitimacy of the interaction or the feedback's credibility.

Observers. Each observer was yoked to one actor. We conducted the study until all actors were yoked to at least one observer. We averaged the responses of observers who were yoked to the same actor.

Like actors, observers were also seated in private rooms. But in observers' cases, they were told they would be *observing* a previous participant who completed a study on whether people work better with those they like, and whether outside observers can predict how well a group can work together. Observers saw everything that actors saw, but read actors' responses instead of providing responses of their own.

Observers were shown a diagram (see Figure 4B) that summarized the role they and others were playing. Observers completed *social* perception measures that were parallel to—in timing and in form—the actors' meta-perception and self-perception measures. (As before, observers did not see actors' social or meta-perception responses.) More specifically, observers completed the baseline measures of likeability after reading the actor's responses to the practice round. They completed final social perceptions measures after reading the actor's responses to the interview round then witnessing the social acceptance or rejection.

Trait perceptions. The perception measure comprised six items that assessed the actor's likeability. The *meta-perception* items asked actors to guess how observers would judge them in light of their social performance during the study. Actors answered these questions with the understanding that they were guessing observers' responses to those exact items. The *social perceptions* asked observers to judge the actors in light of their social performance. Similarly, the *self-perceptions* asked actors to judge themselves in light of their performance as well. The first four questions asked participants to rate the actor as "engaging", "likable", "warm", and "charming" on 9-point scales anchored at 1 (*not at all*) to 9 (*extremely*). The final two questions asked participants how much the actor would "make a good impression" and would be "able to get along with others" on 9-points scales anchored at 1 (*not at all*) to 9 (*extremely*). We averaged these items to create likeability perception composites.

Critically, Study 2 differed from Study 1 in that participants completed two sets of final ratings. The first set asked participants to rate the actors' performance in light of the entire study—i.e., based on the *global* set of information to which participants had been exposed. The second set (the *feedback-informed* final ratings) asked participants to rate the actor only based on the performance task in particular. The self-perception ($\alpha = .96$), meta-perception ($\alpha = .96$), and social perception ($\alpha = .97$) of likeability composites all had good reliability.

Results and Discussion

The two competing accounts—*overblown implications* versus *performance focalism* — make the same prediction about how meta-perceptions will err on their global impressions. The two accounts differ in whether meta-perceptions will err on the more specific feedback-informed impressions. We conducted a 2 (feedback: success or failure) X 3 (perception: meta, social, or self) X 2 (time: baseline or final) mixed-model ANOVA. For our first set of analyses, the final

measure was the global impression (parallel to that in Study 1). For the second set of analyses, the final measure was the feedback-informed impression (based on the performance that led to the success or failure).

Global final impression. Conceptually replicating Study 1, we found the same pattern of results on the global impressions. That is, the predicted 2 (feedback: success or failure) X 3 (perception: meta, social, or self) X 2 (time: baseline or final) interaction emerged, $F(2, 226) = 6.67$, $p = .002$, $\eta^2_p = .06$. To determine whether this interaction reflected the predicted pattern, we proceeded to test all three 2 (feedback) X 2 (perception) X 2 (time) interactions.

First, we tested whether social perceivers were less reactive to the actors' social success or failure than observers assumed they would be. Consistent with Study 1, a 2 (feedback) X 2 (perception: meta or social) X 2 (time) interaction suggested that this was the case, $F(1, 113) = 10.50$, $p = .002$, $\eta^2_p = .08$. As depicted in Figure 5A, although observers did shift their global social impressions in response to actors' success versus failure, $F(1, 113) = 4.23$, $p = .04$, $\eta^2_p = .04$, this shift was less pronounced than actors thought it would be, $F(1, 113) = 46.86$, $p < .001$, $\eta^2_p = .29$.

[INSERT FIGURE 5 HERE]

Second, we once again found that meta-perceivers were not merely using their own personal interpretations of their performance when making meta-judgments about others. That is, we also found a significant 2 (feedback) X 2 (perception: meta or self) X 2 (time) interaction, $F(1, 113) = 9.71$, $p = .002$, $\eta^2_p = .08$. That is, actors' self-perceptions were not as reactive to their own success or failure, $F(1, 113) = 33.03$, $p < .001$, $\eta^2_p = .23$, as they assumed observers' perceptions would be.

Third, we observed that actors would have been more accurate in forecasting observers' shift if they had just leaned on their own shift in self-perception. That is, the 2 (feedback) X 2 (perception: social or self) X 2 (time) interaction failed to reach significance, $F < 1$.

Feedback-informed final impression. We tested whether perceptions differed in their interpretation of the final performance that—the one that produced the social acceptance or rejection—and what it signified about the actors' likeability. Suggesting they did, we found a significant 2 (feedback) X 3 (perception: meta, social, or self) X 2 (time) interaction, $F(2, 226) = 3.70$, $p = .03$, $\eta^2_p = .03$. As before we decompose this interaction by comparing each type of perception using 2 (feedback) X 2 (perception) X 2 (time) mixed-model ANOVAs.

First, we tested whether actors assumed that observers would see actors' social success or failure as more informative about actors' likeability than observers themselves actually did. As expected, the 2 (feedback) X 2 (perception: meta or social) X 2 (time) interaction was significant, $F(1, 113) = 5.46$, $p = .02$, $\eta^2_p = .05$. As depicted in Figure 5B, observers did shift their social impressions in response to actors' success versus failure, $F(1, 113) = 4.90$, $p = .03$, $\eta^2_p = .04$, but actors thought this shift would be more pronounced, $F(1, 113) = 46.86$, $p < .001$, $\eta^2_p = .29$.

Second, we tested whether actors merely leaned on their own self-perceptions of what the social acceptance or rejection implied in guessing observers' reactions, or whether actors thought observers would be particularly reactive. Supporting the latter interpretation, we observed a 2 (feedback) X 2 (perception: meta or self) X 2 (time) interaction, $F(1, 113) = 7.21$, $p = .01$, $\eta^2_p = .06$. Actors saw fewer implications in their own social acceptance or failure, $F(1, 113) = 33.03$, $p < .001$, $\eta^2_p = .23$, than they thought that observers would.

Third, we asked whether actors would have been better off leaning on their own self-perceptions in judging how observers viewed them. Consistent with this possibility (and with

Study 1), we failed to observe a significant 2 (feedback) X 2 (perception: social or self) X 2 (time) interaction, $F < 1$. That is, social observers and actors themselves saw similar, but relatively smaller, implications in the episode that produced social acceptance or rejection.

In short, not only did actors fail to understand how few implications observers would see in observers' final global impressions of them, but they were similarly wrong even when everyone was more narrowly focused on the social interaction that got them included or excluded. (See the Supplemental Materials for a conceptual replication of this latter finding.) Although we found that actors and observers disagreed on the implications of the portion of the performance upon which the social acceptance or rejection was based, it is possible that meta-perceivers still focused on a narrower portion of their performance than observers did. In part, this was the tradeoff that came from employing a behaviorally rich context like those used in the first two studies; such contexts provide so much information that even by constraining the target of judgments, we may not be completely confident that actors and observers are not focusing on somewhat different performance information. Study 3 moves away from live performance contexts in order to address this and another issue more conclusively.

Study 3

We have twice demonstrated that actors fail to understand how observers' impressions shift in light of their successes and failures. Study 2 offered initial evidence that meta-perceivers were inaccurate about observers' perceptions of the broader implications of actors' failure or success. The evidence was less consistent with an alternative account whereby meta-perceivers simply failed to give sufficient weight to their baseline performance, that which was not met with signs of social success or failure. But in Study 3, we wanted to go further in showing that actors' working trait definitions constricted around the performance domain (what one might call

definitional focalism) instead of that actors became fixated on their failures or successes while ignoring other relevant behavioral (*performance focalism*) or contextual information.

In Study 3, we asked participants to simulate the perspective of either an actor or an observer to an upcoming situation (Issue 4: Timing). In each vignette, a simple performance context was described. There was no additional information that meta-perceivers, but not observers, might neglect (Issue 2: Actor's distortion). Furthermore, we provided no information about whether the performance was a success or a failure. Observers indicated how much they would learn about another person's trait based on their performance on a trait-relevant behavior. Actors were asked to consider that they were about to perform those behaviors, but guessed how those who would be observing (and judging) them would answer the same questions.

By our reasoning, when actors adopt the perspective of an observer, they become focused on an evaluative threat. That is, the observer is someone who is, or who will be, watching and judging actors based on their performance. This leads the performance context to loom disproportionately large in actors' meta-perceptions. But because social perceptions do not have personal stakes for observers, such constriction does not actually occur.

By an alternative *self-rumination* hypothesis, actors overblow the implications of their performance because they ruminate on their recent failures and successes. That is, actors may replay in their minds the specifics of their embarrassing, demotivating defeats or the glorious details of their energizing victories. If social observers do not have the same penchant for ruminating on others' performance, then this could offer an alternative explanation for meta-perceivers' relatively constricted working trait definitions. On the one hand, this unpacks the psychology of performance focalism (Issue 2: Actor's distortion). But also, it illustrates how this alternative account should apply to retrospective instead of prospective judgments (Issue 4:

Timing). That is, if the OIE comes from a constricted working trait definition, then we should still observe it in prospect, when there is no past performance to focus on.

We predicted that those taking actors' perspective should guess that their evaluators would see more diagnosticity in upcoming behaviors than those taking observers' perspective actually see. If instead the OIE stems from performance focalism (neglecting additional aspects of one's performance) or ruminating on previous failures or successes, then there should be no disagreement about the behavior's implications.

Method

Participants. Two hundred fifteen undergraduates from an American university completed a lab session for course credit. Participants were randomly assigned to one of two perspective conditions: *actor* or *observer.*

Procedure. Participants were asked to simulate being in different situations. We emphasized that they should try to fully place themselves in the context, to vividly visualize being there, and to be attuned to what they would be thinking and feeling. Each scenario described an interaction in which an actor's skills or abilities would be on display to an observer. The wording was varied such that the exact same situation was described from the vantage point of the actor or the observer. Crucially, all scenarios asked participants to consider these behavioral contexts in prospect. That is, no information was provided about whether the behavior was performed successfully or not. A picture also accompanied each scenario to reinforce the perspective manipulation (see Figure 6).

[INSERT FIGURE 6 HERE]

As an example, one scenario described a person who had baked and brought cookies to a party. Actors were told, "You baked cookies to take to a party. You overhear someone mention

to a friend that you baked the cookies. You watch as the person picks up a cookie to try one." Observers learned this same information, but from the vantage point of the person about to try the cookies, "A person baked cookies to take to a party. Someone mentions to you which person baked the cookies. You pick up a cookie to try one."

For this item, observers were asked, "After sampling their cookies, how much do you feel like you would have learned about whether or not the other person is a good cook?" Actors were asked to predict observers' responses: "After sampling your cookies, how much do you think the person would feel like they have learned about whether or not you are a good cook?" Each judgment was made on 11-point scales anchored at 0 (*not at all*) and 10 (*a great deal*). Each scenario described a different behavior that spoke to a different trait (see Table 1 for a summary of the behaviors and traits associated with each scenario). The 10 scenarios were presented in a random order.

Results and Discussion

To determine whether those simulating the perspective of observers would see less diagnosticity in actors' upcoming behavior than those considering the situations as actors guessed, we submitted participants' diagnosticity ratings to a 2 (perspective: actor or observer) X 10 (scenario) mixed-model ANOVA. Only the second factor was measured within-subjects. As hypothesized, there was a strong main effect of perspective, $F(1, 213) = 13.29$, $p < .001$, $\eta^2_p = .06$. Those adopting the perspective of observers saw significantly less diagnosticity in actors' upcoming behavior ($M = 5.93$, $SD = 1.42$) than actors thought observers would ($M = 6.61$, $SD = 1.33$). Table 1 presents these results by scenario.

That actors and observers have prospective disagreement about behaviors' implications cannot be explained by actors' ruminating on their recent performance (Issue 4: Timing), nor can

it be explained by actors ignoring or discounting other relevant performances (again, because there were none to consider; Issue 2: Actor's distortion). Instead, these findings are consistent with our account that considering how others are evaluating the self causes working trait definitions to constrict around the source of evaluative apprehension. That said, this story emphasizes that constricted working trait definitions stem not merely from imagining how another evaluates just anyone, but from considering how another evaluates the self. Study 4 tests this boundary condition.

Study 4

In Study 4, we aimed to test whether the overblown implications effect stems from people estimating how *they themselves* will be evaluated. By an alternative account, the OIE does not reflect the narrowed working trait definitions that come from considering being evaluated, but instead reflects a general property of how people attempt to read the minds of another person. That is, do people simply guess that others see more diagnosticity in anyone's actions, not the self's own actions in particular? This alternative possibility is supported by research showing that people think others make more extreme dispositional attributions—not just about the self, but about anyone—than they actually do (Van Boven, White, Kamada, & Gilovich, 2003; Pronin, Lin, & Ross, 2002).

To address this alternative explanation, Study 4 added a third perspective condition. We retained our actor and observer perspective conditions, but added a third group who considered the situations from the perspective of uninvolved bystanders. Bystanders considered the interactions of actors and observers from afar, but had to estimate observers' perceptions of actors. In this way, bystanders estimated the perceptions of someone else (just like actors), but

not someone else who was judging the self (unlike actors; see Epley et al., 2002, Study 2, for similar experimental reasoning).

Comparing actors' and bystanders' estimates of observers' perceptions is particularly informative. By our reasoning, the constricted working trait definitions come from actors considering being evaluated. Imagining how another will judge the self focuses actors on that performance judgment in observers' minds. But if such constriction emerges merely from trying to adopt someone else's perspective (instead of the perspective of someone else who is evaluating the self), actors and bystanders should agree on their guesses of observers' judgments.

As in Study 3, we measured the perceived diagnosticity of behaviors. But unlike Study 3, we did not probe it directly. Instead, we allowed diagnosticity to be revealed through a pair of judgments. That is, we asked participants to report how they would judge the actor twice—if (hypothetically) the actor were to perform well, and then again if the actor were to perform poorly. When behaviors are particularly diagnostic of traits, they should prompt more divergent judgments when considering success versus failure. This measure of diagnosticity returned us to a method in which we could test whether the OIE emerges due to meta-perceives overblowing actions' implications instead of thinking observers will be especially uncharitable judges (Issue 1: Scope), but also in a judgment context in which no additional behavioral information was provided (Issue 2: Actor's distortion).

Method

Participants. Three hundred one participants recruited from Amazon's Mechanical Turk (MTurk) completed the study for nominal payment. Participants were randomly assigned to one of three perspective conditions: actor, observer, or bystander.

Procedure. Like in Study 3, we told participants they should fully throw themselves into every simulation, to visualize the scene unfolding and to be attuned to what they would be thinking or feeling. The scenarios were the same as those used in Study 3. The instructions for actors and observers were nearly equivalent to those used in Study 3. Participants in the new bystander condition were informed that they would consider each situation as an outside onlooker. Their task would be to predict how a person (the observer) would form judgments about another (the actor). As before, an image accompanied every scenario to reinforce the perspective manipulation.

Next, participants indicated how the observer would view the actor if the actor were successful as well as if the actor were unsuccessful. As one example, consider again the scenario in which an actor brings cookies to a party. Those in the observer condition made two judgments: "After sampling [the actor's] cookies, if you thought the cookies tasted *good [bad]*, how good of a cook would you think the other person is?" Those taking the actor's perspective tried to guess how observers would respond to this question. Bystanders made similar judgments, but they did not consider that they themselves had baked the cookies; they considered how another person (the observer; Person Y) judged someone else (the actor; Person X). For the cookie scenario, bystanders saw questions of this form: "After sampling their cookies, if Person Y thinks the cookies taste *good* [*bad*], how good of a cook do you think Person Y would think Person X is?" All judgments were made on 11-point scales, anchored at 0 (*not at all…*) and 10 (*extremely…*).

Results and Discussion

For each scenario, we took the trait judgment for a successful performance and subtracted off the trait judgment for a failed performance. Greater numbers imply greater perceived

diagnosticity of the behavior for the trait.[3] We submitted these inferred diagnosticity scores to a 3 (perspective: actor, observer, or bystander) X 10 (scenario) mixed-model ANOVA. Only the first factor was varied between-subjects. The predicted main effect of perspective was significant, $F(2, 298) = 7.31$, $p < .001$, $\eta^2_p = .05$. (See Table 2 for results by scenario).

We conducted a series of 2 (perspective) X 10 (scenario) repeated-measures ANOVAs to better understand the main effect of perspective. Providing evidence of the overblown implications effect, actors guessed that observers would be more reactive to performance events ($M = 4.07$, $SD = 1.93$) than those in the observer perspective condition were ($M = 3.09$, $SD = 1.86$), $F(1, 299) = 13.76$, $p < .001$, $\eta^2_p = .04$.

Did actors' meta-perceptions see greater diagnosticity in these behaviors because actors were considering being personally evaluated (as we have argued), or merely because they were making judgments about someone else's inferences? Providing support for the predicted account, bystanders did not think that observers would be particularly reactive ($M = 3.36$, $SD = 1.87$). That is, their own guesses about observers' inferences showed less evidence of an overblown implications effect than did actors', $F(1, 299) = 6.88$, $p = .009$, $\eta^2_p = .02$. Instead, bystanders' guesses were fairly accurate, statistically indistinguishable from the observers', $F(1, 299) = 1.30$, $p = .26$, $\eta^2_p < .01$.

Although we have pointed out how our findings thus far shed light on the scope (Issue 1), actor's distortion (Issue 2), and timing (Issue 4), we also wish to draw attention to how we have

[3] This analysis is equivalent to one in which we include feedback (success or failure) as a third factor, but the present approach will simplify the description of the results. Table 2 also provides descriptive statistics for successes and failures separately in addition to difference scores.

also been able to speak to the observer's perspective issue (Issue 3). The idea that observers may have special insight into the difficulty of performance challenges offers another previously identified reason why meta-perceptions may err. But our support for the overblown implications effect is not empirically consistent with this account. That is, we find evidence of *interactions*: Meta-perceivers think successes will be met with more adulation and failures will be met with more harsh judgment. If instead observers simply realized that performance contexts were more challenging or simple than actors realized, this would predict only *main effects:* It would lead observers to be more charitable or less kind, respectively, in their characterization of actors than meta-perceivers would realize. Of course, these possibilities are empirically distinguishable but not mutually exclusive.

Study 5

Studies 3 and 4 helped to pinpoint what it is about actors' perspective and experience that is (i.e., being the object of evaluation) and is not (i.e., having past performance to selectively ruminate on or additional contextual cues to ignore) necessary for *the overblown implications effect* to emerge. Although we repeatedly offered tests that differentiate our account from previously documented ones, we have yet to examine directly the role of working trait definitions. In Study 5, we aimed to manipulate actors' and observers' working trait definitions. More specifically, we asked some actors and observers to list other behaviors—beyond those demonstrated in the performance context (cf. Savitsky et al., 2001, Study 3)—that could speak to the trait in question.

If actors' meta-perceptions, compared to observers' social perceptions, operate under a constricted working trait definition, then this *broadening* manipulation should debias actors' subsequent meta-perceptions. That is, by encouraging them to appreciate that the behaviors that

could be demonstrated in this context constitute only a narrow sliver of what defines the broader trait, meta-perceivers may come to appreciate what observers do spontaneously. If the intervention corrects for a mismatch in working trait definitions, then it should attenuate the overblown implications effect.

This intervention should work if actors' distortion involves constricted working trait definitions, but not if it merely involves performance focalism or actors' initial failure to realize the mitigating circumstances that would explain away their poor performance. By way of contrast, consider a "defocusing" manipulation used by Savitsky et al. (2001, Study 3). That intervention asked actors to consider what other factors could influence observers' judgments. Actors came to realize that observers would be less harsh on them once actors began to focus on factors like "the difficulty of the questions." In our intervention, we neither focus participants on other aspects of their performance nor on situational factors. Instead, we keep people from adopting a narrow definition of the broader quality they are judging.

Given that Studies 3 and 4 found that merely simulating the role of actor or observer was sufficient to produce the overblown implications effect, we begin by testing this intervention in a simulation scenario. Study 6 will return to the lab to test whether the intervention can debias judgment following actual performance behavior. In this way, Study 5 has the advantage of testing these ideas in a helpfully impoverished context in which there are no situational factors to notice or not. But Study 6 will have the potential to offer assurance that the intervention's effects are robust in more behaviorally and informationally rich contexts.

Method

Participants and design. Eight hundred seventeen participants took part in the study. Participants were randomly assigned to one of 8 conditions in a 2 (perspective: actor or observer)

X 2 (feedback: success or failure) X 2 (trait definition: broadened or control) full-factorial design. In order to achieve a large sample size, we recruited participants from two sources simultaneously: an American university subject pool and Amazon's Mechanical Turk.

Procedure. Participants were asked to read a short passage about a work meeting. For those in the *actor* condition, the passage read as follows:

> "Imagine that you are in a meeting at work with several co-workers. You have previously worked with each of these co-workers, and you often have meetings during which you generate ideas for new projects. You are in your second month at this job. As with most people's, some of the ideas you have come up with have gone on to be successful projects and some have never taken off the ground. At today's meeting, everyone is brainstorming how best to recruit a new client who could potentially bring in a lot of work. You feel like you have identified an area that this new client could work on to improve their business, and you are very excited about making a pitch with this in mind."

At this point, it was said that the idea was either accepted (*success* condition) or rejected (*failure* condition). In both conditions, the description began, "After you have shared your idea with your team of co-workers, there is…" At that point, what happened varied by condition. The success condition continued, "…excited chatter among everyone, and it is clear that everyone is interested in following up on your idea. The rest of the meeting focuses on how to best execute your idea." The failure condition concluded with, "…a long moment of silence among everyone, and it is clear that no one is interested in following up on your idea. The rest of the meeting shifts to other people's suggestions."

The *observer* condition was parallel to the actor condition, but all descriptions were written from the perspective of another co-worker who was present at the meeting. The actor—

i.e., the co-worker who shared the idea—was referred to as "Person X." All participants were provided with a picture to help them to both: (1) visualize the scenario, and (2) understand the perspective that they were supposed to take in the scenario (see Figure 7).

[INSERT FIGURE 7 HERE]

Definitional broadening manipulation. At this point, participants assigned to the broadened trait definition condition completed the definitional broadening manipulation. Calling their attention to the fact that competence is defined by more behaviors than "offering ideas in a meeting," the manipulation made certain that participants had an expanded working trait definition of competence. The wording for the actor [observer] condition was as follows:

> "Before answering questions about this situation, we would like you to think about things that could happen outside of this situation. In a workplace, like the one described, there are many ways that employees could display their competence or incompetence. Please think of and list 5 ways that an employee like you ["Person X"] in this scenario could display competence or incompetence."

Crucially, this manipulation asked participants to focus on "things that could happen outside of this situation." This permitted us to distinguish our account from performance focalism (Issue 2) or a focus on mitigating circumstances that could lead actors to identify why this described context was uniquely challenging (Issue 3).

Trait perceptions. The perception measures comprised three items that assessed the actor's competence. *Observers*—who considered the situation from the vantage point of a coworker—were asked to judge Person X in light of the meeting. *Actors* were asked to guess how another co-worker would judge them (i.e., "Person X") in light of the meeting. Three items measured perceptions of workplace competence: *competent*, *intelligent*, *creative*. Responses

were made on 9-point scales anchored at 1 (*not at all*) and 9 (*extremely*). We averaged these measures to create a competence perception composite ($\alpha = .92$).

Results and Discussion

Given our predictions that control actors (i.e., those who did not complete the broadening intervention) would be unique in having a constricted working trait definition, we began by defining a *contrast* that differentiates control actors (+3) from those in the other three conditions: broadened actors (-1), control observers (-1), broadened observers (-1). This differentiated those participants predicted to have a working trait definition of competence that focused on the specific type of behavior described in the scenario (control actors) from those who—due to the intervention or their baseline expanded perspective—should have a broader definition of the trait (those in the three other conditions). We then submitted the competence perception composite to a two-way 2 (contrast) X 2 (feedback: success or failure) between-subjects ANOVA.

As predicted, there was a significant Contrast X Feedback interaction, such that participants in the control actor condition were more reactive to the performance event than were those in the other three conditions, $F(1, 813) = 8.43$, $p = .004$, $\eta^2_p = .01$ (see Figure 8). To make certain that the predicted contrast fit the data well, we conducted a series of six pairwise comparisons to understand the relative reactivity of the four groups. As described next, every pattern emerged as expected.

[INSERT FIGURE 8 HERE]

Control actors were significantly more reactive to the performance event compared to participants in the broadened actor condition, $F(1, 813) = 5.40$, $p = .02$, $\eta^2_p = .01$; participants in the control observer condition, $F(1, 813) = 4.82$, $p = .03$, $\eta^2_p = .01$; and participants in the broadened observer condition, $F(1, 813) = 7.56$, $p = .01$, $\eta^2_p = .01$. The three other

comparisons—broadened actor vs. control observer, broadened actor vs. broadened observer, control observer vs. broadened observer—were all non-significant, Fs < 1, ps > .57. This suggests that the OIE can be eliminated when actors' working trait definition is broadened to encompass behaviors outside of the performance context. This highlights the key role of how meta-perceivers define the trait or competency in question in producing the overblown implications effect (Issue 2: Actor's distortion).

Study 6

Study 5 tested our explanation for the overblown implications effect by showing that a definitional broadening intervention—one that expanded some participants' working trait definitions to include performance behaviors outside of those observed in the performance context—eliminated the bias. On the one hand, Study 5 might be considered a particularly conservative test of our hypotheses. After all, when under actual (instead of merely simulated) evaluative threat, there may be more of a constricted working trait definition for the broadening manipulation to undo. On the other hand, one might worry that Study 5 was instead a liberal test of our hypotheses. That is, under actual evaluative threat, perhaps a minimal intervention like the definitional broadening intervention would not be sufficient to expand a more rigidly constricted working trait definition. Given our interest in examining a practical debiasing intervention in addition to testing our novel theoretical account, it is important to determine whether the broadening task has an influence in real situations as well.

In Study 6, we returned to the paradigm used in Study 1, the trivia contest. As before, audience members (observers) provided ratings of the contestants' (actors') intelligence both before and after a success or failure. Those randomly assigned to complete the broadening manipulation listed other ways that people like the actors could display whether or not they were

intelligent. We expected to conceptually replicate the effects of Study 5, that actors in the control condition would show evidence of the overblown implications effect. That is, control actors' meta-perceptions of how observers would view them should be more reactive to their success or failure than actors who completed the broadening manipulation. Broadened actors should be relatively accurate compared to observers (regardless of whether such observers completed the broadening manipulation).

Method

Participants and design. Two hundred thirty-six undergraduates from an American university completed a lab session for course credit. Actors were randomly assigned to the *success* or *failure* condition. Observers were yoked to a randomly selected actor. As with Study 2, when more than one observer was paired with an actor, we averaged the observers' responses for the purpose of analyses.

Procedure. The procedure and measures were similar to those used in Study 1. Actors began by answering ten trivia questions. Regardless of their actual performance, they were told (and their yoked observers also learned) that they answered seven correctly. Actors and observers then completed baseline perceptions (unlike Study 1, actors only offered meta-perceptions; observers provided social perceptions). We used the same intelligence perception measures as in Study 1: meta-perception ($\alpha = .93$) and social perception ($\alpha = .91$). On a subsequent high-stakes double-or-nothing round, actors received success or failure feedback. Unbeknownst to actors (or to the observers who saw this), this feedback was randomly determined.

Before completing the final perception measures, some participants were randomly assigned to complete the definitional broadening manipulation. The goal was to expand these

participants' working trait definition of intelligence to include additional behaviors than the one that defined the current performance context. The instructions were similar to those used in Study 5, but modified for the current context:

> "Before answering the next set of questions, we would like you to think beyond the tasks in this specific experiment and think about other contexts in which a student from [masked university name] like you could demonstrate that they are or are not intelligent. Please list 5 different ways that a student from [masked university name] could demonstrate intelligence or lack thereof."

Results and Discussion

We used a nearly identical analytic approach to that described in Study 5. That is, we first defined a variable *contrast* to differentiate those participants whose perceptions were expected to be more reactive to the performance feedback (control actors: +3) from those hypothesized to be less reactive (broadened actors, control observers, broadened observers: -1). We then submitted the intelligence perception measures to a 2 (contrast) X 2 (feedback: success or failure) X 2 (time: baseline or final) mixed-model ANOVA, with only the final factor measured within-subjects.

The predicted Contrast X Feedback X Time interaction emerged, $F(1, 210) = 11.36$, $p < .001$, $\eta^2_p = .05$ (see Figure 9). As expected, the meta-perceptions of those in the control actor condition were more reactive to the performance feedback than were the perceptions of those in the other three conditions. That is, control actors showed more of an overblown implications effect than did those in the broadened actor condition, $F(1, 210) = 6.31$, $p = .01$, $\eta^2_p = .03$. Furthermore, control actors' meta-perceptions were more reactive than the social perceptions of observers, regardless of whether observers were in the control condition, $F(1, 210) = 4.66$, $p =$

.03, η^2_p = .02, or the broadened condition, $F(1, 210) = 12.81$, $p < .001$, η^2_p = .06. The three other comparisons—broadened actor vs. control observer, broadened actor vs. broadened observer, control observer vs. broadened observer—were all non-significant, Fs < 1.61, ps > .20.

[INSERT FIGURE 9 HERE]

Study 7

Studies 5 and 6 directly tested the role of mismatches in working trait definitions as an explanation for the overblown implications effects. The studies reinforced our account that meta-perceivers err in translating impressions on the specific performance task in question (e.g., skill at answering trivia questions) to impressions on the more global competencies to which those specific skills might speak (e.g., intelligence). The present account differs from past research which instead focused on various reasons why the specific successes or failures themselves may be non-diagnostic—factors that make a task so difficult that it does not easily speak to even the narrow performance domain in question. Our proposal instead causes us to focus on performances that cannot be so easily dismissed. That is, whereas past research focuses on errors in diagnosing the significance of the specific performance, our account suggests that errors arise in translating impressions of specific skills into impressions on broader competencies (Issue 5: Error specificity).

Study 7 probes this distinction empirically. Much like in Study 3, those taking the perspective of actors or observers considered what actors would reveal in various prospective performance contexts. For example, actors considered taking homemade cookies to a party, whereas observers considered attending a party where they tried another's homemade cookies. But participants either estimated what would be learned about the specific competency in question (e.g., skill at baking cookies) or the broader one that this could reflect (e.g., skill at

being a cook). We expected to find that those considering these situations as meta-perceivers would diverge more from observers when considering the broad as opposed to the specific competencies. This would further implicate working trait definitions, which identify how people lean on specific skill perceptions when evaluating broad competencies, in the overblown implications effect.

As an additional aim, Study 7 investigated the generality of the OIE. Typically in social judgment studies, people are judging strangers. For example, when Jones and Harris (1967) examined whether observers drew inferences about actors' true Castro attitudes even when actors had been forced to write a pro- or anti-Castro essay, the experimenters did not conduct this study among good friends. Such decisions tend to be made both in order to standardize the targets being judged and to make sure that everything else one knows about a target does not overwhelm the influence of the specific causal effect being studied. Despite this typical rationale, Study 7 probed the robustness of the OIE by having people consider a specific friend as an actor or an observer. If the OIE emerged in this context, we would be more confident that it reflects a failure of meta-insight that emerges not merely during initial impressions but even among existing relationships.

Method

Participants. Eight hundred one participants recruited from Amazon's Mechanical Turk (MTurk) completed the study for nominal payment. Participants were randomly assigned to one of 4 conditions in a 2 (perspective: actor or observer) X 2 (competency: general or specific) full-factorial design.

Procedure. Those in the actor conditions were told, "To start, think of someone specific who would consider you a friend. This person should be someone who thinks of you as more

than an acquaintance but not quite a best friend. Please write the initials of the person who thinks of you as a friend." Those in the observer condition were given symmetric instructions: "To start, think of a specific friend. This person should be someone who is more than an acquaintance but not quite a best friend. Please write the initials of the friend below."

Participants received instructions similar to those used in Study 3. They learned they would consider seven briefly described scenarios. Participants were asked to fully place themselves in those situations: "How would the situation look to you from the perspective described? What would you be feeling? What would you be thinking?"

We used the scenarios from Study 3 except with three changes. First, actors considered being observed by the friend they identified, and observers considered observing the friend. Second, we did not include three scenarios, some of which could not be easily adapted to a context involving a good friend: conversing with a stranger, leaving work at an unusual time, and splitting the bill. Third, although the general competencies judged in Study 3 were again used in our general condition, participants in our specific conditions considered narrower competencies that corresponded more closely to the specific behaviors described. Consider the situation in which observers imagine trying a friend's cookies. Those in the observer general condition indicated, "After sampling their cookies, how much do you feel like you would have learned about whether or not [FRIEND'S INITIALS] is a good cook?" Those in the observer specific condition would instead answer, "…how much do you feel like you would have learned about whether or not [FRIEND'S INITIALS] is good at baking cookies?"

As in Study 3, each judgment was made on 11-point scales anchored at 0 (*not at all*) and 10 (*a great deal*). Each scenario described a different behavior that spoke to a different general

and specific trait or competency (see Table 3). The seven scenarios were presented in a random order.

Results and Discussion

In order to probe the error specificity of the OIE (Issue 5), we submitted participants' diagnosticity ratings to a 2(perspective: actor or observer) X 2 (competency: general or specific) X 7 (scenario) mixed-model ANOVA. Only the final factor was measured within-subjects. As hypothesized, there was a significant Perspective X Competency interaction, $F(1, 797) = 4.43$, $p = .04$, $\eta^2_p = .01$. Table 3 summarizes the results by scenario.

When considering the general trait, meta-perceivers and observers had different perspectives. Namely, observers saw less diagnosticity in actors' upcoming behavior ($M = 6.37$, $SD = 1.87$) than actors thought observers would ($M = 6.86$, $SD = 1.51$), $F(1, 797) = 10.63$, $p = .001$, $\eta^2_p = .01$. This replicates Study 3, but in the context of existing relationships. But illustrating that the OIE reflects a difference in how meta-perceivers and observers and observers translate specific impressions into more general ones, this difference went away in the specific condition. More specifically, actors knew about how much diagnosticity ($M = 7.87$, $SD = 1.34$) observers would report ($M = 7.84$, $SD = 1.29$) when considering what the one-off performance would reveal about the specific competency, $F < 1$.

Of course, this is not to say that actors and observers will always agree on what a single performance says about the matching, specific competencies. After all, some of the past research on meta-insight has shown that observers are surprisingly (at least surprising in the eyes of meta-perceivers) quick to recognize when actors' performance is limited not by actors' own competence but by features of the performance context. The overblown implications effect identifies a different reason why meta-perceptions and social perceptions may diverge. Even

when actors know how much observers think actions are diagnostic of an actor's skill, actors think those implications are more general than they actually are. That is, everyone can agree that a single instance of a person baking cookies does speak to how well that person bakes cookies. But actors mistakenly think that observers will go further by taking that single and assuming it speaks to how good of a cook the actor is (Issue 5: Error specificity). In short, actors overblow the implications of their own performance.

General Discussion

People care how others view them. But without direct access to others' perceptions, understanding how we are perceived entails guesswork. Across seven studies, we provided evidence for an *overblown implications effect.* Actors see their own performance as having more evaluative impact on observers than it actually does. By introducing the heretofore unidentified construct of working trait definitions, we were able to localize this error to a difference in how meta-perceivers and observers were defining the competencies in question.

Issue 1: Scope. We found that meta-perceivers overblew the implications not merely of their failures, but also of their successes. In some studies, effects looked asymmetric. But in a cross-study meta-analysis of the relevant studies (Studies 1-2, 4-6), we found that meta-perceivers clearly exaggerated the effects of both. That is, actors overestimated the negative implications observers would see in failures, $2.78 <$ Stouffer's Zs < 3.66, $.0003 < p$s $< .005$ (Study 1: $Z = -1.21$; Study 2: $Z_{\text{Global}} = 2.75$, $Z_{\text{Feedback-Informed}} = 1.26$; Study 4: $Z = 2.72$, Study 5: $Z = 1.81$, Study 6: $Z = 1.62$). But they also overblew successes, $3.75 <$ Stouffer's Zs < 3.96, $.00008 < p$s $< .0002$ (Study 1: $Z = 4.23$; Study 2: $Z_{\text{Global}} = 0.86$, $Z_{\text{Feedback-Informed}} = 1.33$; Study 4: $Z = 2.58$,

Study 5: $Z = .16$, Study 6: $Z = .55$).[4]

Issue 2: Actor's distortion. Actors erred because they assumed observers had different working trait definitions than they actually did, not because they assumed observers focused on a narrower portion of their own behavior. Evidence for this conclusion came in several forms. First, meta-perceivers were just as inaccurate when considering how observers would interpret the portion on which they performed well or poorly as when they considered how observers would judge their performance as a whole (Study 2). Second, even when performance occurred "in a vacuum," meaning there were not highlights or lowlights to selectively focus or situational factors to neglect, we still observed the OIE (Studies 3-5 & 7). Third, by experimentally manipulating working trait definitions to broaden them, we debiased meta-perceptions while leaving social perceptions untouched (Studies 5-6).

Issue 3: Observer's perspective. To understand how impressive it is that a basketball player makes a shot, one would want to know how far from the goal she was standing and how closely she was being defended. Past research has demonstrated that actors either do not appreciate or are not aware of observers' attunement to just how challenging a performance context is. By this account, observers may be generally more charitable or generally harsher than actors would think. That is, observers' special insight should allow them to apply a directional correction. Instead, we argued and demonstrated that observers appreciate the limited

[4] Study 2 had both global and feedback-informed final trait measures. This means that for the purpose of any cross-study meta-analysis including this study, there are three ways to select which measure to use from this study (global, feedback-informed, or both). The range of meta-analytic results reflects that the results are robust to all specifications.

diagnosticity of behaviors (Studies 3 & 7), which prompts more muted (instead of merely directional) inferences (Studies 1-2 & 4-6). That said, we certainly are not saying that meta-perceivers are not *also* directionally biased. In fact, a cross-study meta-analysis across the relevant studies (Studies 1-2 & 4-6) reveals that meta-perceptions are generally harsher than observers' actual perceptions, 3.08 < Stouffer's Zs < 3.44, .0006 < ps < .002 (Study 1: $Z = 3.31$; Study 2: $Z_{\text{Global}} = 0.45$, $Z_{\text{Feedback-Informed}} = 1.26$; Study 4: $Z = 0.52$, Study 5: $Z = 1.00$, Study 6: $Z = 1.60$). In other words, the overblown implications effect and this previously identified result are not mutually exclusive.

Issue 4: Timing. The OIE did not merely reflect actors retrospectively ruminating on their own recent successes and failure, thereby blowing up their implications in their own mind. Instead, the OIE occurred even in prospective judgments (Studies 3 & 7), even when there was no additional non-focal performance information that actors might ignore. What actors ignored was the tenuous connection between the specific performance and the broader competency in observers' minds. To use our theoretical formulation, as people considered being evaluated for a behavior they had yet to engage in, their working trait definitions constricted. As a result, actors prospectively approached—and did not merely retrospectively interpret—their performances with the idea that they have more on the line than they actually do.

Issue 5: Error specificity. Finally, we found that the OIE emerges even when actors understand how observers will view the *narrow* diagnosticity of their behavior. That is, the OIE need not reflect that observers excuse a one-off performance as non-diagnostic of skills on that specific performance task. Instead, the OIE is rooted in the distorting influence of working trait definitions that help people understanding the broader implications of narrow competencies.

Actors think that observers will move draw inferences from a specific skill (e.g., baking cookies) to a general one (e.g., being a good cook) more than they actually do (Study 7).

Connecting constricted working trait definitions to evaluative threat. By our theoretical reasoning, meta-perceivers' working trait definitions should constrict around the performance domain due to the threat of considering how one would be evaluated. But if this logic is correct, it suggests that there should be individual variability in the OIE that is tied to individual variability in how concerned people are with evaluation. Public self-consciousness (PSC) has been shown to predict anxiety around being evaluated (e.g., Hope & Heimberg, 1988; Turner, Carver, Scheier, & Ickes, 1978). Given the OIE was often measured using higher-order interactions, we knew that we would not have the power to consistently detect even higher-order interactions with PSC. Instead, we measured PSC—using Fenigstein, Scheier, and Buss's (1975) public self-consciousness scale—in every relevant study with the intention of conducting a cross-study meta-analysis. Consistent with our logic, we found that the overblown implications effect grew stronger as actors' public self-consciousness increased, 2.66 < Stouffer's Zs < 3.55, .0004 < ps < .008 (Study 1: $Z = 1.57$; Study 2: $Z_{\text{Global}} = 2.63$, $Z_{\text{Feedback-Informed}} = 2.29$; Study 5: $Z = -0.02$, Study 6: $Z = 1.47$). This result provides another piece of convergent evidence for our experimental logic.

A reinterpretation of or a complement to previous research? Psychologists have spent decades studying many psychological mechanisms that lead to accuracy and error in self and social judgment (Dunning, 2005; Funder, 1987; Vazire, 2010). Although meta-judgments are a natural extension of work on self and social judgment, research in this area is in an earlier stage of development. Just as there are many reasons why people's self and social judgments are accurate or inaccurate, the same is no doubt true of meta-judgment. In studying whether actors

know what impressions they convey through their performance successes and failures, we identified five issues on which our theoretical account could be differentiated from the focus of past research.

Although our results consistently confirmed our take on these questions, this does not mean that the explanations and accounts articulated in previous research do not sometimes contribute to errors in meta-insight. That is, we believe the overblown implications effect is one of several effects that explain when meta-judgments will be accurate or err. We do not call into question that musicians may become fixated on their one beautiful (or perhaps more likely, one sour) note, or that observers do sometimes have a privileged perspective to realize that a task is more challenging than actors realize. What does distinguish the overblown implications effect is it looks elsewhere—not at people's consideration of their specific performance, but instead at the gap between evaluation of a narrow skill and judgments about the broader competency that skill speaks to—to identify that error. Future theoretical work may wish to more precisely specify the conditions under which these various mechanisms are likely to lead meta-insight astray. Such a cumulative science could build toward a checklist that would allow one to identify *a priori* whether a particular context is particularly ripe for inaccurate meta-judgments.

Reconciling the Overblown Implications Effect with Other Related Research

At first glance, the overblown implications effect might seem inconsistent with the actor-observer effect (Jones & Nisbett, 1971; but see Malle, 2006)—the tendency for observers to make more dispositional inferences than actors. But crucially, we do not test whether actors and observers make different attributions for actors' actions. Instead, we examine the accuracy of actors' guesses about how observers view them. But might the OIE reflect actors' sense that

observers commit the fundamental attribution error more strongly than they actually do (Van Boven et al., 1999)?

For three reasons, we would not characterize the OIE as a false belief that observers embrace dispositional instead of situational explanations for others' behavior. First, actors' misestimates of observers' perceptions stemmed from the narrowness with which meta-perceivers thought about the trait category, not actors' explanations for the behavior. Only our preferred account can explain the influence of definitional broadening intervention (Studies 5-6) or why actors and observers agreed on the narrow implications of a behavior (Study 7). Disagreement about whether a behavior actually just reflected situational influence would have produced differences on the narrow impressions as well.

Second, and relatedly, in some of our studies—particularly our scenario studies (Studies 3-5 & 7)—the behavioral contexts were described in a vacuum—i.e., without information about how the situation may affect performance success. It is thus hard to imagine what situational contexts observers would have been relying upon that actors would have neglected. Such hypothetical scenarios were sufficient (or perhaps even ideal) for localizing effects to definitional misunderstandings. Third, if the overblown implications effect were merely another example of people exaggerating how much others display the fundamental attribution error (Pronin, Lin, & Ross, 2002; Van Boven et al., 2003), then participants in the bystander condition (Study 4) should also have overestimated the extent to which observers would draw inferences from actors' behavior, but in fact, such bystanders did not.

Although we did not give much attention to actors' self-perceptions in Studies 1 and 2, some might be surprised that they did not show the same evidence of the overblown implications effect that actors' meta-perceptions did. This might seem inconsistent with the literature on the

looking glass self, the idea that self-views derive from how (they believe) others view them (Cooley, 1902; Tice, 1992; Tice & Wallace, 2003). Although it is the case that self-perceptions did not move to the same extent as meta-perceptions, they still did show sizable shifts in light of recent performance. Furthermore, the correlations between the change in meta- and self-perception were strong (r = .63 in Study 1 and r = .44 in Study 2). We, of course, cannot say whether this relationship between meta- and self-perceptions was causal; nonetheless, these findings do illustrate how the overblown implications effect should not be interpreted to exist instead of, but rather on top of, that which results from the looking glass self.

Questions For Future Research

Can observers' reactions mitigate the OIE? In an effort to isolate the hypothesized overblown implications effects, we did not have observers directly interact with actors. Previous research has found that participants in an interaction become sufficiently focused on their own behavior that they can fail to notice variability in their interaction partners' performance (Gilovich et al., 2002). For this reason, our in-lab paradigms preserved the subtleties of actors' behavior (by presenting them on video or computer-mediated communication) but did not place the observer into the live context (in which self-presentational concerns could distract).

But had the observer been present, then the actor would have an additional source of information—observers' verbal and non-verbal reactions—when forming meta-perceptions. Though even when observers are present, observers are notoriously hard to read. For example, in the classic spotlight effect studies (Gilovich et al., 2000), the live presence of observers did not help actors realize they were not the clear focus of attention. Furthermore, many performance behaviors occur not in dyadic contexts (in which only one observer's reactions could be monitored), but in front of an audience. As the negativity dominance literature and many a

professor's experience teaching suggests, that one scornful audience member can loom large in our attentional field (e.g., Hansen & Hansen, 1988; Pinkham, Griffin, Baron, Sasson, & Gur, 2010). Such attentional biases could distort meta-perceptions of the audience as a whole.

Furthermore, there is one hint that by leaning on uninvolved observers, our tests may have been especially conservative. Campbell and Fehr (1990) had actors guess how interaction partners and uninvolved observers viewed them. Actors didn't distinguish between these two groups in their meta-perceptions. But uninvolved observers were actually harsher in their assessments of the actors than were the interaction partners. At least when it comes to anticipating how others will respond to one's poor performance, this suggests that the overblown implications effect may be even stronger when live observers are present.

Does the OIE extend to naturalistic performances? In some of our studies, the fact that participants had succeeded or failed was made explicit. For example, trivia contestants were told that they had gotten the double-or-nothing question right or wrong (Studies 1 and 6). Interviewees learned that their interview responses caused them to be selected or excluded (Study 2). On the one hand, this raises a worry that constricted working trait definitions stemmed not from actors considering being personally evaluated but instead from the unusualness of receiving blunt feedback. But such a characterization would not be consistent with our studies that included no feedback whatsoever (Studies 3 & 7), that merely asked people to consider the conclusions that would follow from things going well or poorly (Study 4), or that described a situation in which direct feedback was not offered (Study 5).

In our everyday lives, sometimes performance can be murkier, in part because the feedback we get is circumspect. That is, when a student makes a comment, and the professor fills the silence in the classroom with a slowly delivered "That's interesting," there can be ambiguity

in what was meant. But the existence of the overblown implications effect does not suggest that performance feedback is always clear. It simply means that the perceived implications of the performance—even when those implications are misassessed—get overblown. This in itself can be a cause for additional error. If, for example, the student didn't realize his observers saw his comment as (mildly) embarrassing, he may assume they took it as strongly revealing of his superior intellect.

Another property of these engineered successes and failures is that they could have created a demand effect of some sort: that is, perhaps meta-perceivers felt that these performances *should* be incorporated into their judgments. However, it is unclear whether or why such a demand effect would apply only to meta-perceivers and not observers asked to judge the *same* performances. Moreover, a demand effect explanation should not apply in Studies 3 and 7 in which we examine diagnosticity rather than judgments of tasks in which participants were provided feedback.

Can invested bystanders also exhibit the OIE? We argued that the constricted nature of actors' working trait definitions stems from the evaluative threat posed by the performance situation. Bolstering this logic, our cross-study meta-analysis found that those most disposed to performance anxiety showed the strongest overblown implications effect. But this also suggests that there may be times in which bystanders show the overblown implications effect as well. For example, parents who feel highly invested in their child's soccer game or spelling bee performance may feel empathic evaluative apprehension as their child is on stage. As such, it may feel that their child's image as athletic or intelligent is on clear display to observers. Future research should examine whether such invested bystanders exhibit the OIE as well.

How does the OIE change over time? Finally, our studies investigated how actors' and observers' perceptions respond to a single performance event. What would happen if actors' skills—both successes and failures—are on display over multiple rounds? Do actors feel most under evaluation when they know observers do not know them well, meaning that the OIE may diminish across time? Or instead will actors' meta-perceptions respond to what is evaluatively focal, that which has just occurred (or is about to occur)? On one hand, in Study 7, we found that the OIE also extends to well-acquainted others: Meta-perceivers still overblow the implications of their performances when considering evaluation by a friend. This suggests that the OIE may continue to persist over time.

On the other hand, with repetition, actors may forget just how much their talents will surprise and thus impress others. Consider a professional singer who mindlessly sings along to the radio. She may fail to realize just how impressed her taxi driver will be. Or a party guest who brings his tried and true recipe, though one with which he has become somewhat bored, may fail to appreciate how much his cookies will be encoded as a success that reflects his superior cooking abilities. In such contexts, actors' performance may actually loom larger in the eyes of observers than actors realize.

Finally, there may be real differences between qualities in how much information observers feel they need before updating their impression of another. For example, Kammrath, Ames, and Scholer (2007) found that observers update their impression of others' agreeableness more quickly than impressions of others' conscientiousness. If actors fail to anticipate that observers are quick versus slow to update their impressions in certain domains, then this itself could reduce or enhance the overblown implications effect. More generally, although we think

that working trait definitions are a key but overlooked construct in making sense of meta-insight, a more complete understanding will require additional empirical work as well.

Conclusions

As people navigate through their personal and professional lives, they aim not merely to passively estimate but also to actively manage others' impressions (e.g., Jones & Pittman, 1982; Leary & Kowalski, 1990; Schlenker & Weigold, 1992). This means meta-perceptions are important barometers of whether people (think they) are doing so effectively. And thus, when people's meta-perceptions are inaccurate, they may make suboptimal decisions about how best to invest in further impression management. Those who make a single inane comment during a work meeting may go to unnecessary lengths to redeem themselves in the eyes of their colleagues, and those who offer a single stroke of genius may be mistaken about how much they can rest on these (thin) laurels (see Anderson, Ames, & Gosling, 2008; Elfenbein, Eisenkraft, & Ding, 2009). We may do well to keep in mind that although our specific competencies are sometimes on full display, our broader abilities almost never are.

References

Albright, L., Forest, C., & Reiseter, K. (2001). Acting, behaving, and the selfless basis of metaperception. *Journal of Personality and Social Psychology*, *81*, 910–921.

Albright, L., & Malloy, T. E. (1999). Self-observation of social behavior in meta-perception. *Journal of Personality and Social Psychology, 77,* 726–734.

Anderson, C., Ames, D. R., & Gosling, S. D. (2008). Punishing hubris: The perils of overestimating one's status in a group. *Personality and Social Psychology Bulletin*, *34*, 90–101.

Baumeister, R.F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology, 46,* 610-620.

Campbell, J.D., & Fehr, B. (1990). Self-esteem and perceptions of conveyed impressions: Is negative affectivity associated with greater realism? *Journal of Personality and Social Psychology, 58,* 122-133.

Carlson, E. N., & Furr, R. M. (2009). Evidence of differential meta-accuracy: People understand the different impressions they make. *Psychological Science*, *20*, 1033–1039.

Carlson, E. N., & Kenny, D. A. (in press). Do we know how others see us? In S. Vazire & T. D. Wilson (Eds.), *Handbook of self-knowledge*. New York, NY: Guilford Press.

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology*, *101*, 831–846.

Cervone, D. & Shoda, Y. (1999). Beyond traits in the study of personality coherence. *Current Directions in Psychological Science*, *8*, 27-32.

Cooley, C. H. (1902). Human nature and the social order (Rev. ed.). New York: Scribner's.

Cramer, A., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., & Aggen, S. et al. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality, 26*, 414-431.

Critcher, C. & Dunning, D. (2015). Self-affirmations provide a broader perspective on self-threat. *Personality and Social Psychology Bulletin*, *41*, 3-18.

Critcher, C. R., Dunning, D., & Rom, S. C. (2015). Causal trait theories: A new form of person knowledge that explains egocentric pattern projection. *Journal of Personality and Social Psychology, 108,* 400-416.

Critcher, C. R., Helzer, E. G., & Dunning, D. (2011). Self-enhancement via redefinition: Defining social concepts to ensure positive views of self. In M. D. Alicke, & C. Sedikides (Eds.), Handbook of self-enhancement and self-protection (pp. 69-91). New York, NY: The Guilford Press.

DeSteno, D. & Salovey, P. (1997). Structural dynamism in the concept of self: A flexible model for a malleable concept. *Review of General Psychology*, *1*, 389-409.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself*. New York: Psychology Press.

Dunning, D., Meyerowitz, J.A., & Holzberg, A.D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57,* 1082-1090.

Elfenbein, H. A., Eisenkraft, N., & Ding, W. W. (2009). Do we know who values us? Dyadic meta-accuracy in the perception of professional relationships. *Psychological Science*, *20*, 1081–1083.

Epley, N., Savitsky, K., & Gilovich, T. (2002). Empathy neglect: Reconciling the spotlight effect and the correspondence bias. *Journal of Personality and Social Psychology*, *83*, 300-312.

Fenigstein, A., Scheier, M., & Buss, A. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, *43*, 522-527.

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science, 13,* 83– 87.

Funder, D.C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75-90.

Gilovich, T., Kruger, J., & Medvec, V.H. (2002). The spotlight effect revisited: Overestimating the manifest variability of our actions and appearances. *Journal of Experimental Social Psychology, 38,* 93-99.

Gilovich, T., Medvec, V., & Savitsky, K. (2000). The spotlight effect in social judgment: An egocentric bias in estimates of the salience of one's own actions and appearance. *Journal of Personality and Social Psychology*, *78*, 211-222.

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology*, *54*, 917–924.

Harkins, S. (2006). Mere effort as the mediator of the evaluation-performance relationship. *Journal of Personality and Social Psychology, 91,* 436-455.

Hope, D. & Heimberg, R. (1988). Public and private self-consciousness and social phobia. *Journal of Personality Assessment*, *52*, 626-639.

Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. Morristown, NJ: General Learning Press.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. In J. Suls (Ed.), *Psychological perspectives on the self* (pp. 231-261). Hillsdale, NJ: Lawrence Erlbaum.

Kammrath, L.K., Ames, D.R., & Scholer, A.A. (2007). Keeping up impressions: Inferential rules for impression change across the Big Five. *Journal of Experimental Social Psychology, 43,* 450-457.

Kenny, D.A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kenny, D.A., & DePaulo, B.M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin, 114,* 145–161.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134

Kruger, J., & Gilovich, T. (1999). "Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology, 76,* 743-753.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, *107*, 34–47.

Malle, B. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, *132*, 895-919.

Malloy, T.E., Albright, L., Kenny, D.A., Agatstein, F., & Winquist, L. (1997). Interpersonal perception and metaperception in nonoverlapping social groups. *Journal of Personality and Social Psychology, 72,* 390–398.

Markus, H. & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology*, *51*, 858-866.

Markus, H. & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology*, *38*, 299-337.

McConnell, A. (2011). The multiple self-aspects framework: Self-concept representation and its implications. *Personality And Social Psychology Review*, *15*, 3-27.

Pervin, L. (1994). A critical analysis of current trait theory. *Psychological Inquiry*, *5*, 103-113.

Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & Gur, R. C. (2010). The face in the crowd effect: Anger superiority when using real faces and multiple identities, *Emotion, 10,* 141-146.

Pronin, E., Lin, D., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*, 369-381.

Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology, 35,* 485-494.

Ross, L. & Nisbett, R. (1991). *The person and the situation*. Philadelphia: Temple University Press.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology, 37,* 322-336.

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*, 44-56.

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

Seitchik, A.E., Brown, A.J., & Harkins, S.G. (2017). Social facilitation: Using the molecular to inform the molar. In S.G. Harkins, K.D. Williams, & J.M. Burger (Eds.), *The Oxford Handbook of Social Influence* (pp. 183-203). New York: Oxford University Press.

Showers, C. J., & Zeigler-Hill, V. (2003). Organization of self-knowledge: Features, functions, and flexibility. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 47-67). New York: Guilford Press.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359-1366.

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

Tice, D. M., & Wallace, H. M. (2003). The reflected self: Creating yourself as (you think) others see you. In M. R. Leary, & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 91-105). New York: Guilford Press.

Turner, R., Carver, C., Scheier, M., & Ickes, W. (1978). Correlates of self-consciousness. *Journal of Personality Assessment*, *42*, 285-289.

van Boven, L., Kamada, A., & Gilovich, T. (1999). The perceiver as perceived: Everyday intuitions about the correspondence bias. *Journal of Personality and Social Psychology, 77,* 1188-1199

van Boven, L., White, K., Kamada, A., & Gilovich, T. (2003). Intuitions about situational correction in self and others. *Journal of Personality and Social Psychology*, *85*, 249-258.

Vazire, S. (2010). Who knows what about a person? The Self-Other Knowledge Assymetry (SOKA) model. *Journal of Personality and Social Psychology, 28,* 281-300.

Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves? *Social and Personality Psychology Compass, 4,* 605– 620.

Zajonc, R.B. (1965). Social facilitation. *Science, 146,* 269-274.

Table 1
Diagnosticity Ratings by Perspective for Each Trait-Relevant Behavior (Study 3).

Context	Behavior	Trait	Observer	Actor Rating	Observer Rating	*t*
Restaurant with a group	Splitting the bill	Mathematical ability	Person looking over actor's shoulder	6.17 (2.28)	6.96 (2.23)	-2.56*
Cocktail party	Conversing with stranger	Social skills	Person overhearing actor's conversation	6.74 (2.20)	7.08 (2.16)	-1.13
Game night at a friend's house	Answering a trivia question	Intelligence	Person reading actor the question	6.02 (2.45)	5.69 (2.31)	0.99
On a flight	Playing chess on your computer	Analytical thinking ability	Person next to actor on the flight	6.22 (2.24)	5.68 (2.25)	1.75†
Cash-only restaurant	Remembering to pay back the $20 you borrowed	Exploitativeness	Acquaintance that lent actor the money	6.60 (2.62)	5.66 (2.55)	2.64**
Party	Baking cookies	Cooking ability	Person sampling actor's cookie	7.43 (2.01)	6.58 (2.27)	2.90**
Near work right before a meeting	Parallel parking	Driving ability	Coworker waiting for actor to walk in together	7.85 (2.36)	6.86 (2.58)	2.96**
Restaurant with a group	Accepting/ rejecting a fork for dessert	Self-control	Person that knows actor is on a diet and asks if actor wants a fork	6.75 (2.64)	5.54 (2.30)	3.54***
Party	Introducing a new person to your friend	Inconsiderateness	New person actor is introducing, whom actor just met and had a conversation with	5.87 (2.68)	4.41 (2.54)	4.09***
Office	Leaving work at an unusual time	Work ethic	Coworker asking actor for a ride home at usual time	6.48 (2.50)	4.82 (2.67)	4.71***
Total:				6.61 (1.33)	5.93 (1.42)	13.29***

Note. Ratings indicate means and standard deviations (in parentheses); † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$,

Table 2
Trait Inferences by Perspective For Hypothetical Success or Failure on Each Trait-Relevant Behavior (Study 4).

	Diagnosticity (Positive – Negative)			Successful Outcome			Failed Outcome		
Trait	Actor	Observer	Bystander	Actor	Observer	Bystander	Actor	Observer	Bystander
Mathematical ability	4.22 (2.52)	3.78 (2.85)	4.05 (2.60)	8.08 (1.58)	8.08 (1.61)	8.12 (1.43)	3.57 (2.94)	4.31 (2.65)	4.05 (3.16)
Social skills	3.51 (2.65)	3.90 (2.78)	3.45 (3.07)	7.84_{a} (1.54)	8.32_{b} (1.61)	7.84_{a} (1.63)	4.33 (1.70)	4.43 (1.90)	4.39 (1.83)
Intelligence	3.18_{a} (2.18)	1.82_{b} (2.15)	2.36_{b} (2.43)	8.37_{a} (1.20)	7.83_{b} (1.51)	7.96_{b} (1.59)	5.19_{a} (1.52)	6.01_{b} (1.40)	$5.60_{a,b}$ (1.62)
Analytical thinking ability	3.30_{a} (2.41)	2.30_{b} (2.64)	$2.77_{a,b}$ (2.07)	8.42 (1.51)	8.21 (1.61)	8.24 (1.64)	5.12_{a} (1.62)	5.92_{b} (1.70)	$5.48_{a,b}$ (1.51)
Exploitativeness	5.21_{a} (3.82)	4.02_{b} (3.72)	$4.41_{a,b}$ (3.46)	7.89_{a} (2.33)	7.05_{b} (2.23)	7.18_{b} (2.35)	2.68 (2.40)	3.03 (2.76)	2.77 (2.35)
Cooking ability	5.17 (2.67)	4.79 (3.24)	5.13 (2.72)	8.67 (1.16)	8.66 (1.44)	8.64 (1.31)	3.49 (2.19)	3.87 (2.42)	3.51 (1.81)
Driving ability	4.43 (2.85)	3.80 (2.72)	3.95 (2.67)	9.01 (1.61)	8.96 (1.53)	8.87 (1.44)	4.58_{a} (1.92)	5.17_{b} (2.04)	$4.93_{a,b}$ (1.83)
Self-control	3.89_{a} (3.24)	3.03_{b} (3.02)	2.09_{c} (3.14)	8.70_{a} (1.84)	8.56_{a} (1.76)	7.32_{b} (1.85)	4.81_{a} (2.23)	5.53_{b} (1.86)	$5.23_{a,b}$ (1.71)
Inconsiderateness	3.05_{a} (4.14)	$.09_{b}$ (3.32)	1.64_{c} (3.92)	6.62_{a} (2.41)	4.92_{b} (2.38)	5.69_{c} (2.26)	3.57 (2.94)	4.31 (2.65)	4.05 (3.16)
Work ethic	4.72_{a} (2.80)	3.37_{b} (2.93)	$3.76_{a,b}$ (3.13)	9.34_{a} (1.49)	8.67_{b} (1.71)	8.69_{b} (1.77)	4.63a (2.03)	5.30_{b} (1.91)	$4.94_{a,b}$ (1.94)
Total	4.07_{a} (1.93)	3.09_{b} (1.86)	3.36_{b} (1.87)	8.29 (.95)	7.93 (.92)	7.86 (.93)	4.23_{a} (1.29)	4.79_{b} (1.41)	$4.50_{a,b}$ (1.25)

Note. Ratings indicate means and standard deviations (in parentheses). Means in the same row that reflect the same type of score (diagnosticity, successful outcome, failed outcome) but that do not share the same subscripted letter differ at the $p < .05$ level.

Table 3
Diagnosticity Ratings by Perspective and Specificity for Each Trait-Relevant Behavior (Study 7).

		Broad			Specific		
General Competency	Specific Competency	Actor	Observer	*t*	Actor	Observer	*t*
Intelligence	Trivial Pursuit ability	6.94 (2.34)	6.47 (2.56)	1.96*	7.32 (2.15)	7.32 (2.17)	-.04
Analytical thinking ability	Chess-playing ability	6.64 (2.44)	6.54 (2.61)	.39	7.34 (2.23)	7.69 (2.16)	.39
Exploitativeness	Ability to remember to pay people back	6.81 (3.16)	5.88 (3.14)	3.01**	8.65 (2.42)	7.97 (2.38)	2.78**
Cooking ability	Ability to bake cookies	8.00 (1.96)	7.39 (2.40)	2.84**	8.79 (1.84)	8.70 (1.83)	.53
Driving ability	Parallel parking ability	7.38 (2.60)	6.59 (2.87)	2.91**	8.56 (2.27)	8.51 (2.09)	.20
Self-control	Ability to refuse complimentary desserts	7.07 (2.42)	6.40 (2.73)	2.64**	7.24 (2.39)	7.23 (2.61)	.03
Inconsiderateness	Ability to remember new people's names	5.20 (2.78)	5.34 (2.98)	-.52	7.21 (2.23)	7.43 (2.27)	-.96
Total		6.86 (1.51)	6.37 (1.87)	2.93**	7.87 (1.34)	7.84 (1.29)	.26

Note. Ratings indicate means and standard deviations (in parentheses). * $p < .05$, ** $p < .01$.

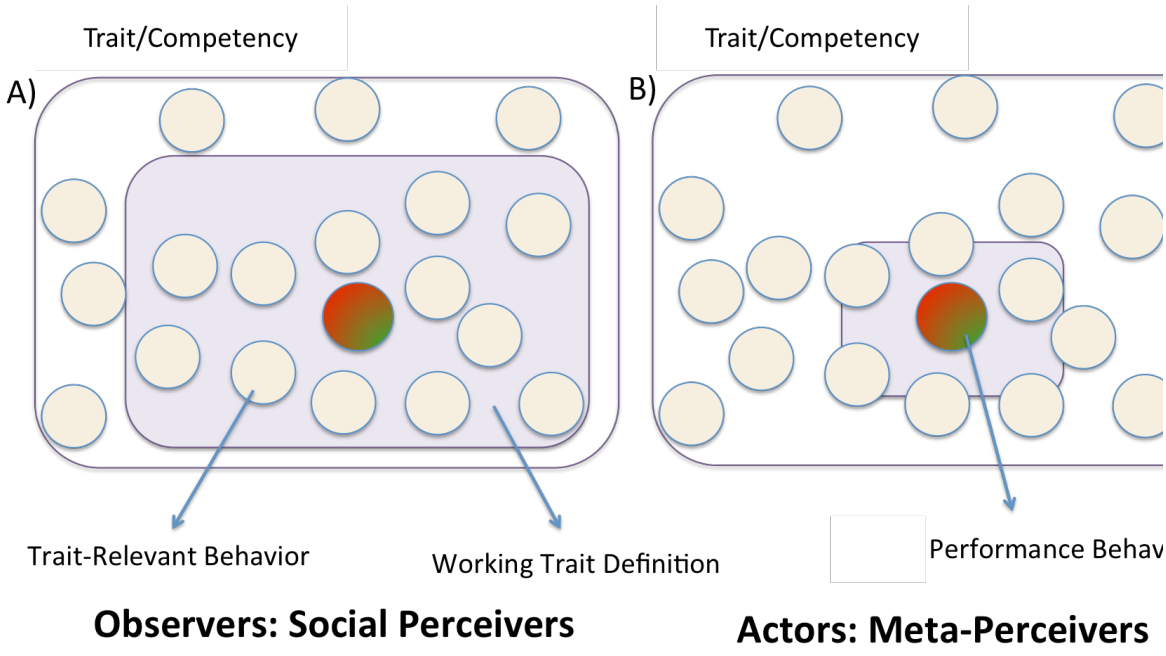


Figure 1. Why social and meta-perceptions are hypothesized to diverge and produce the overblown implications effect. At baseline, observers' working trait definitions (the purple may not account for all trait-relevant behavior (explaining why some trait-relevant behavio always outside of the working trait definition; panel A), but actors' meta-perceptions constr around their performance behavior under evaluative threat (panel B). This predicts that acto meta-perceptions will be more reactive to their own successes or failures than observers' so perceptions, but that expanding actors' working trait definitions should debias them.

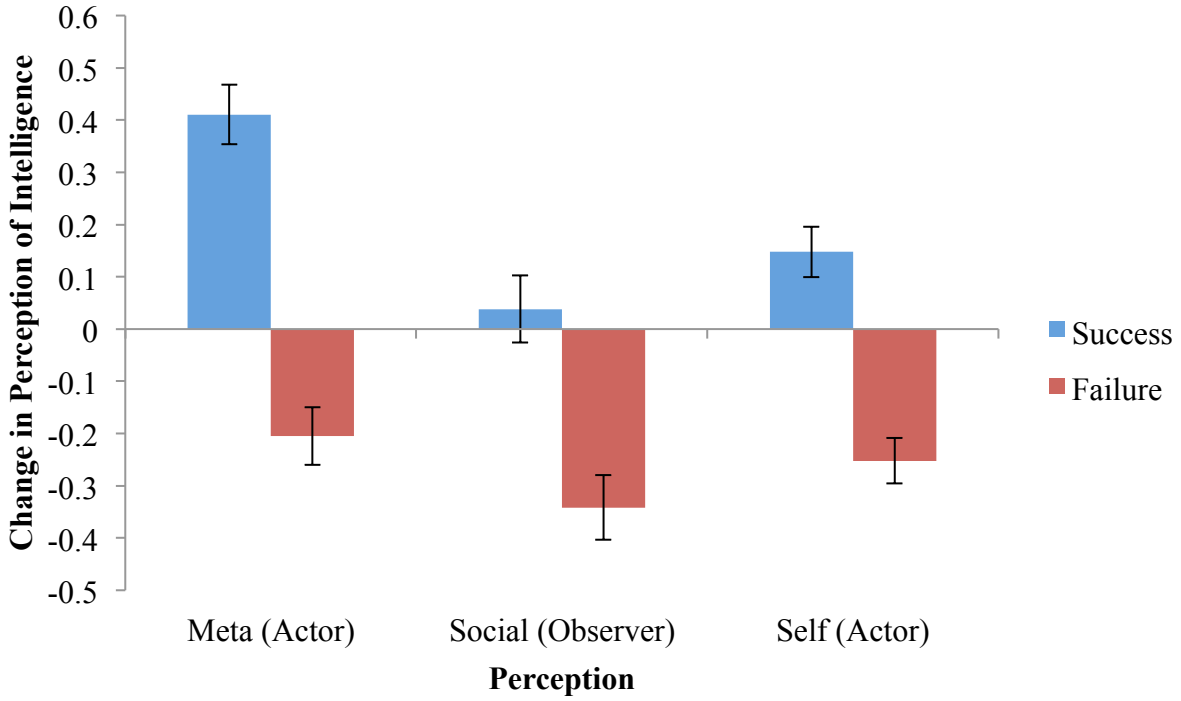


Figure 2. The change in perception (final – baseline) of actors' intelligence by feedback condition and type of perception (Study 1). Which participant offered each perception is in parentheses. The overblown implications effect is reflected by the larger gap between the two meta-perception bars compared to the two social perception bars.

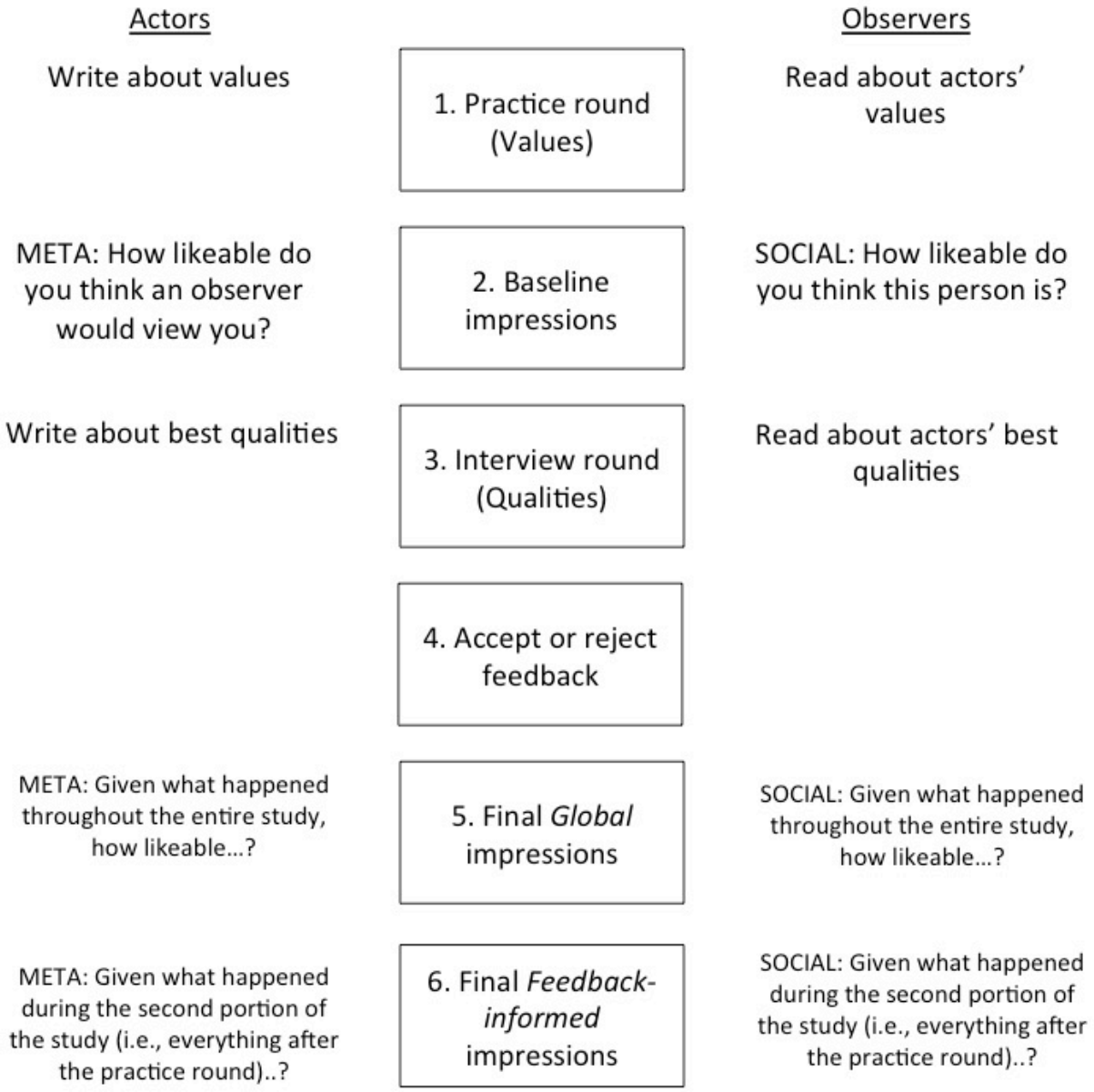
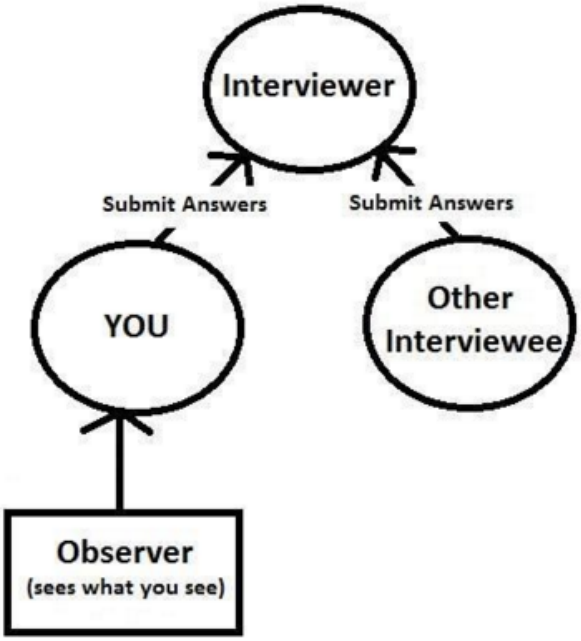
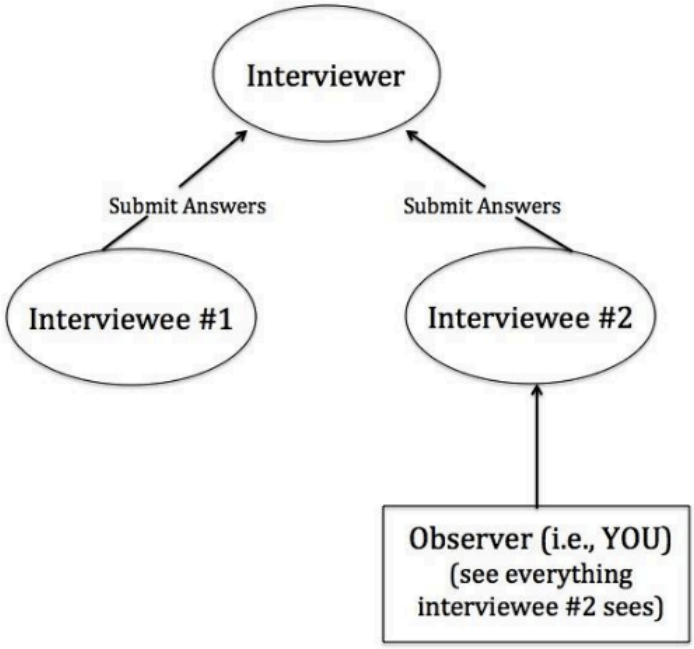


Figure 3. Summary of the procedure in Study 2 for actors (left) and observers (right). In addition to meta-perception ratings, actors also completed self-perception ratings.



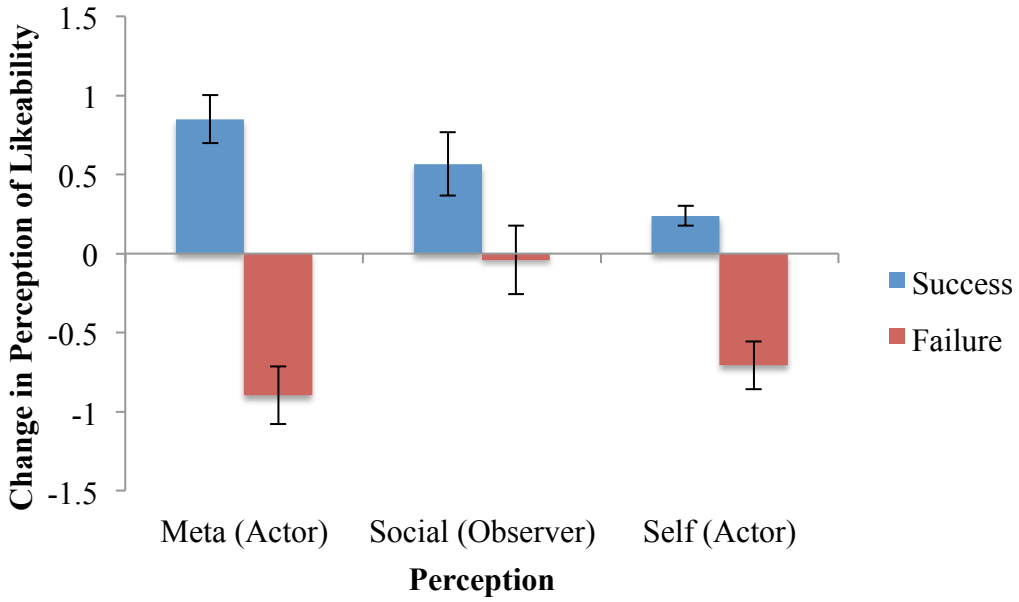
A)



B)

Figure 4. Diagrams explaining to actors (A) and observers (B) their role. Actors knew all of their behaviors and experience would be observed by an outside observer. Observers saw the experiment through the perspective of an actor (whom we called “Interviewee #2”).

A)



B)

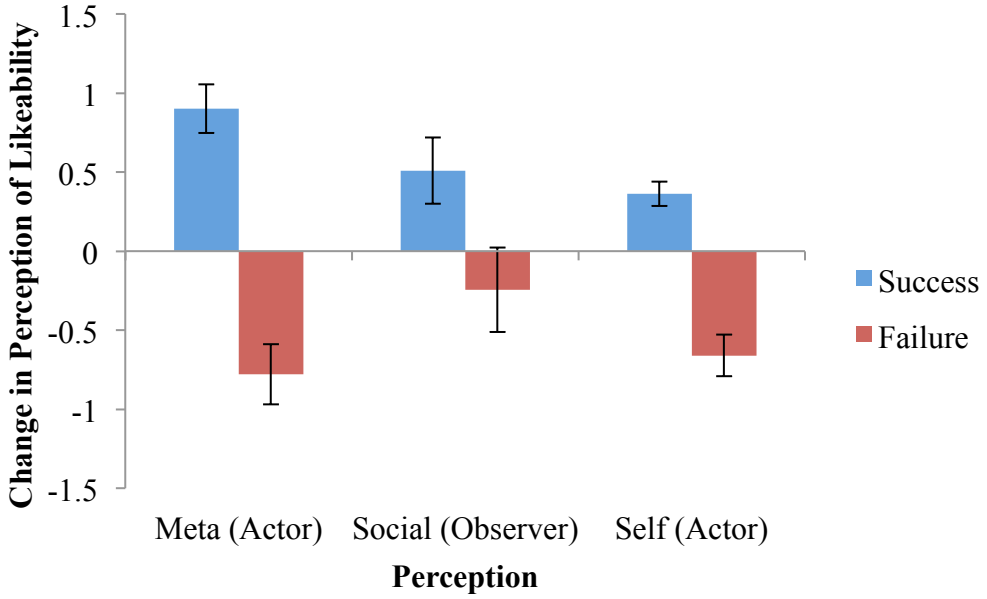


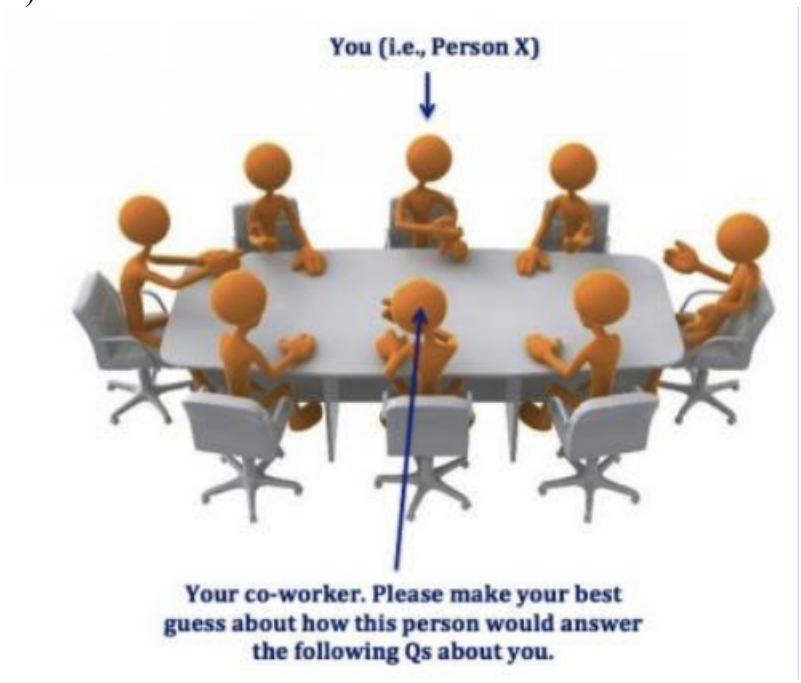
Figure 5. The change in perception (final – baseline) of actors' likeability by feedback condition and type of perception (Study 2). Panel A uses the global final perceptions. Panel B uses the feedback-informed final perceptions. Within each panel, the overblown implications effect is reflected by the larger gap between the two meta-perception bars compared to the two social perception bars. Which participant offered each perception is in parentheses.

A)

B)

Figure 6. Picture accompanying the baking cookies scenario (Studies 3, 4, and 7). Actors saw Panel A. Observers (and bystanders in Study 4) saw Panel B.

A)



B)

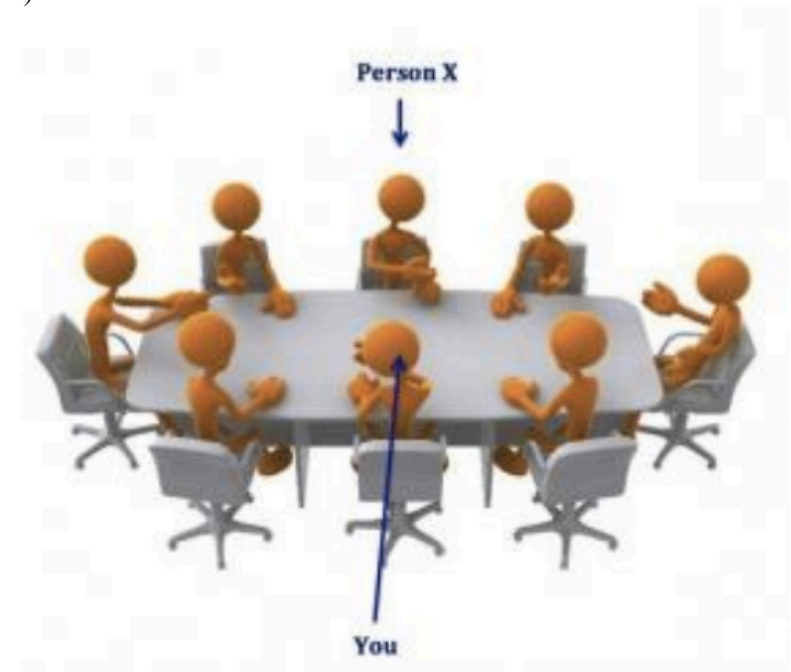


Figure 7. Graphic seen by observers (A) and actors (B) in Study 5 to clarify the trait perception task. Observers made judgments of the described actors (Person X). Actors estimated how an observer would view them.

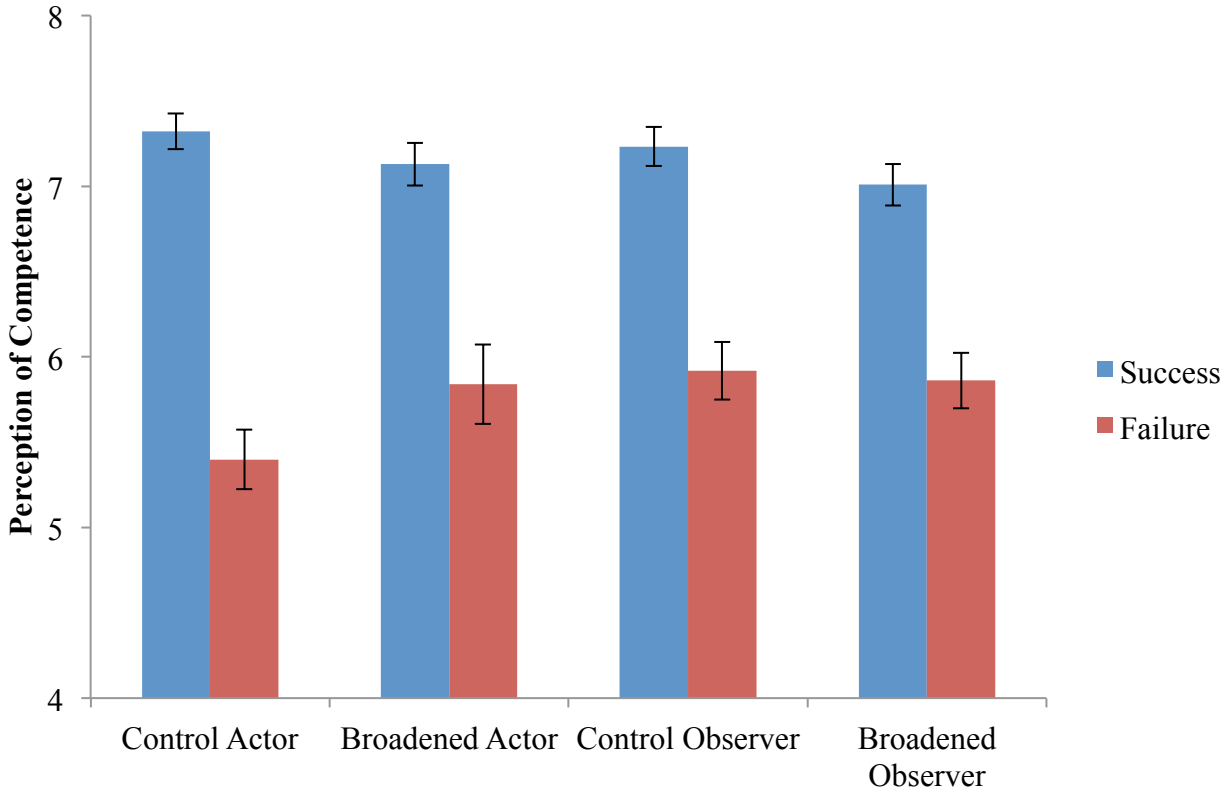


Figure 8. Perceptions of actors' competence by feedback condition, perspective, and working trait definition intervention (Study 5). The broadening intervention eliminated the overblown implications effect, as seen by the greater gap between the two bars for control actors compared to gap between the two bars for the other three perspective-intervention combinations.

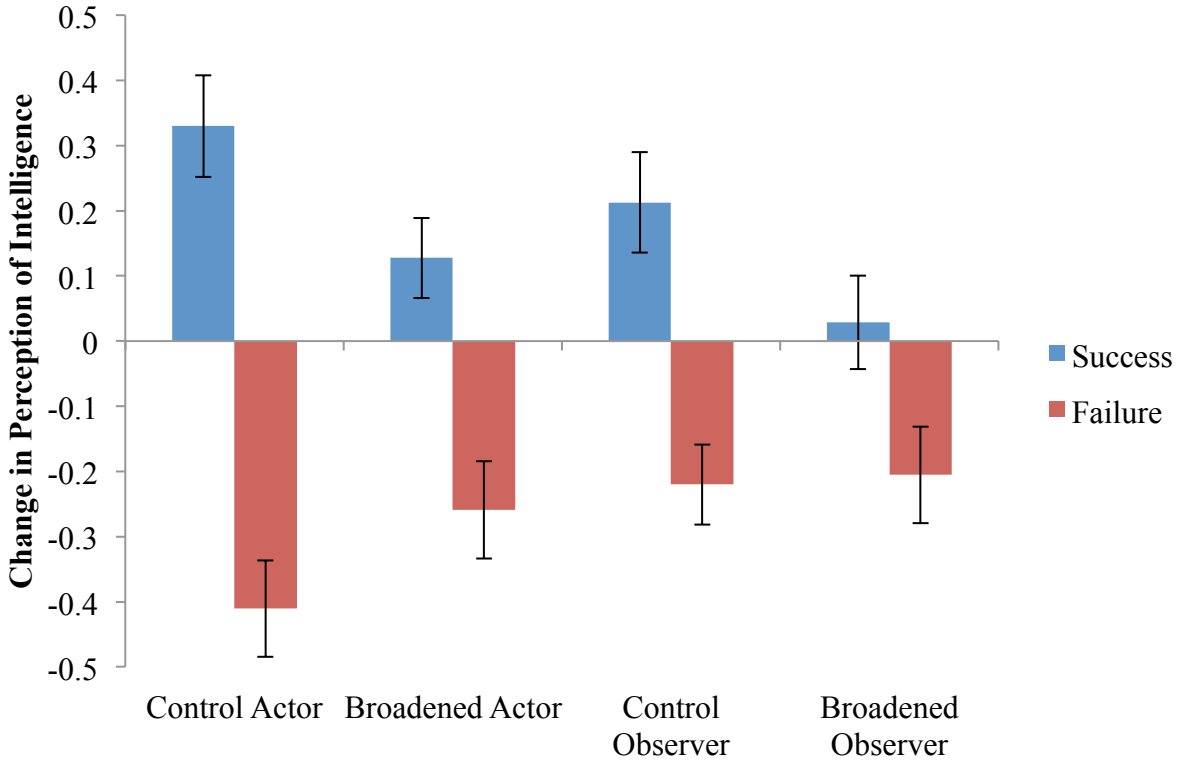


Figure 9. The change in perception (final – baseline) of actors' intelligence by feedback condition, perspective, and working trait definition intervention (Study 6). The broadening intervention eliminated the overblown implications effect, as seen by the greater gap between the two bars for control actors compared to gap between the two bars for the other three perspective-intervention combinations.

Supplemental Materials

1. **Exploratory measures**
2. **Table of raw scores (Studies 1, 2, and 6)**
3. **Supplemental Study**

1. Exploratory measures

The following measures were added for exploratory purposes. Many of these measures were never analyzed. We thought they might be informative if our primary hypotheses had not been confirmed.

Study 1

In Study 1, actors were also asked to rate their general perceptions of their own intelligence prior to participating in the trivia contest – that is, prior to being asked any trivia questions and prior to any meta-judgments of intelligence. Although we had no *a priori* hypothesis about this measure, we included this measure in case the study did not produce the expected results in the off chance that this measure could help us to understand why our study had failed.

Study 5

In Study 5, we included three additional DVs. Actors guessed how observers would respond. Observers merely offered a judgment.

One item asked how likely the actor would be to make helpful comments in the next group project meeting (how *characteristic* the actor's performance was of their usual performances).

The second item asked how knowable the actors' level of competence seemed during the meeting using a revised version of the iceberg measure in Pronin, Kruger, Savitsky, and Ross (2001). The item read,

> Everyone has some part of them that others do not **know, understand, or "get."** In this way, people are like icebergs - part of us is visible and known to others, and part of us is hidden beneath the surface. Of course, exactly how much is *above* the surface and how much is *below* the surface varies from person to person, and how much can become visible or knowable varies from situation to situation. We would like you to think about how **the situation Person X was in** does or does not permit Person X's competence (or lack thereof) to even be knowable or visible.
>
> How much do you feel Person X's level of competence (or lack thereof) was *knowable* or visible to you during the meeting?

The item asked participants to choose an iceberg that corresponded with the knowability or visibility of Person X's level of competence.

The third item asked how many more meetings you felt that you would need to observe the actor in before making a meaningful decision on whether to keep the actor on the team or not.

For the *first* DV (as with our main DV), the interaction between the 3 (*broadened actor, original observer, broadened observer*) vs. 1 (*original actor*) contrast and valence was significant, $F(1, 813) = 6.39$, $p = .01$. To the extent that a stronger inference about the actor's

competence would encourage more confidence that successes or failures would be repeated, then this could be taken as additional evidence of the overblown implications effect.

With the latter two measures, the hope was we could determine whether the behavior was seen to be diagnostic of the broader trait of competence. Unexpectedly, we observed a massive main effect of valence on both items, Fs > 12, ps < .001. In other words, these measures were not capturing what we had hoped. Instead of assessing the extent to which the behavioral context reflected a smaller or larger part of participants' working trait definitions, these measures clearly became a commentary on the meaning of success vs. failure. This study was actually run before Study 3. Based on our experience with these measures, we knew to sidestep this issue in Study 3 by measuring the trait diagnosticity of behaviors before observing the traits.

2. Table of Raw Scores

Study 1:

	Meta Time 1	**Meta Time 2**	**Other Time 1**	**Other Time 2**
Success	-.33 (.81)	.082 (.88)	.20 (.68)	.24 (.79)
Failure	-.25 (.74)	-.45 (.80)	.19 (.70)	-.16 (.78)

Study 2:

	Meta Time 1	**Meta Final Global**	**Meta Feedback-Informed**	**Other Time 1**	**Other Final Global**	**Other Feedback-Informed**
Success	6.01 (1.31)	6.86 (1.06)	6.91 (1.11)	6.09 (1.58)	6.65 (1.50)	6.60 (1.44)
Failure	5.99 (1.15)	5.09 (1.42)	5.21 (1.45)	5.30 (1.82)	5.26 (1.66)	6.06 (1.79)

Study 6:

Control conditions:

	Meta Time 1	Meta Time 2	Other Time 1	Other Time 2
Success	-.08 (.85)	.25 (.81)	.22 (.71)	.43 (.68)
Failure	-.20 (.82)	-.61 (.91)	-.01 (.59)	-.23 (.56)

Broadened conditions:

	Meta Time 1	Meta Time 2	Other Time 1	Other Time 2
Success	-.14 (.84)	-.01 (.93)	.33 (.57)	.36 (.67)
Failure	-.23 (.92)	-.49 (1.03)	.28 (.51)	.08 (.56)

3. Supplemental Study: Romantic Partner Domain

This study, like Study 2 in the paper, tested whether the overblown implications effect occurred due to a *performance focalism hypothesis* or our preferred *overblown implications hypothesis*. This study also examined these hypotheses in a new context for a new trait: "dateability" based on a dating profile video. Participants were told they were taking part in a study on how people present themselves and communicate in dating contexts. For this reason, actors were prompted to record a video dating profile, in which they answered a series of interview questions about their dating style and preferences. After actors recorded this video or after observers had watched it, participants offered baseline impressions of actors' dateability: Observers made social judgments of actors' dateability, and actors made meta-judgments of how they thought observers saw their dateability. Once back on camera, actors were asked a "relationship IQ" question on which they were randomly assigned to succeed or fail. To see evidence of the overblown implications effect, we would expect that actors' meta-perceptions would be more reactive to this focal feedback than would observers' actual social perceptions.

After the focal success or failure, participants offered impressions of two types. Crucially, participants were asked to make judgments based only on the specific *feedback-*

informed event—whether the actor was dateable or not based only on the focal failure or success. According to the *performance focalism hypothesis*, actors and observers should agree on the meaning of the feedback-informed event; they merely incorporate that information into an overall impression differently. But according to the idea that the overblown implications effect stems from constricted working trait definitions, actors should believe that observers will see more meaning in that feedback-informed performance event than they actually do.

Method

Participants and design. One hundred eighty-two undergraduates at an American university completed a lab session for course credit. Participants were assigned to be an *actor* or an *observer*. Actors were randomly assigned to a performance *success* or *failure* condition.

Procedure and materials. As in Study 1 of the main text, each actor was yoked to an observer. We describe actors' experience first. Observers observed actors' complete experience – both by seeing the instructions actors received and watching the actors on video.

Actors. Actors took part in the study individually. Upon arrival, actors were seated in front of a laptop. They learned they would complete a study on "how people present themselves and the impressions they communicate." Actors were told that they would create a video dating profile that consisted of two parts. In the first part, participants answered questions about themselves—the "get to know you" questions. In the second part, actors answered a relationship IQ question. The experimenter informed actors (accurately) that they would be videotaped throughout the entire study so that a future participant could observe them and their performance.

Actors were given three minutes to prepare before the video dating profile was shot. Actors introduced themselves before reading aloud and answering seven questions. They had received a list of these questions three minutes before filming began. This gave actors some time

to consider responses to the prompts. Two of the questions were: “Describe the sorts of activities you would like to do on a first date” and “Think of your ideal dating partner. In what ways would you want the two of you to be similar? In what ways would you want the two of you to be different?” Actors were asked to spend around 30-60 seconds on each answer.

At this point, actors completed the baseline perception measures of their “dateability” (how good of a dating partner they would be). Actors provided self-perceptions (evaluations of their own performance at the task) followed by meta-perceptions (guesses of how the observers would rate them). These measures are described in more detail below.

Next, the experimenter returned to explain the focal task (i.e., the relationship IQ question). Actors were told, “For the last question, we will be testing your relationship IQ. This question has an objectively right or wrong answer as determined by previous research on the psychology of close relationships, and gives a sense of your ‘relationship IQ.’ Also, a fun fact: the producers of the TV shows *The Bachelor* and *The Bachelorette* use these questions to screen potential candidates for their show!” This final fib was meant to give a sense of legitimacy and interest value to the question. Actors were then presented with what was ostensibly a relationship IQ question: “Research has shown that five key qualities differentiate happy couples from unhappy couples. Which quality is MOST important for a couple’s relationship satisfaction? Rank the following from MOST important to LEAST important. (a) communication, (b) flexibility, (c) emotional closeness, (d) compatibility of personalities, and (e) conflict handling.” Experimenters read the question to the actor and provided the actor with a written version of the question to examine.

Experimenters first confirmed verbally with actors that the answer they provided was from most to least important. Then, based purely on random assignment, actors were told that

they had answered the question in the exact right order (*success*) or in the exact opposite of the right order (*failure*). In the *success* condition, experimenters said, "Wow, actually that's the exact order!" and repeated the items in the order that the actor had given, followed by, "Good job!" In the *failure* condition, experimenters said, "Wow, that's actually the exact opposite of what it should be," and repeated the items in the opposite order as the actor had given. This entire exchange happened on camera. Finally, actors completed the final measures of dateability: the same meta-perception and self-perception items they completed before the focal evaluative event. After completing these final measures, actors were debriefed and apologized to for the mild deception.

Observers. Each observer was yoked to one actor. We conducted the study until all actors were yoked to at least one observer. We collected additional observer data until the end of the semester. For the sake of both analytic consistency across studies and simplicity of presentation, we averaged the measures for observers who were yoked to the same actor.

Each actor video was clipped into two shorter videos. The first video showed the actor answering the "get to know you" questions. The second video showed the experimenter reading the relationship IQ question to the actor, the actor answering the question, and the experimenter providing the final feedback. A screen with text reiterated that the actor either answered the question correctly (*success* condition) or that the actor answered the question (*failure* condition).

Observers, seated in private rooms, had the same experience as actors, but as onlookers to the situation instead of as active participants. That is, they learned what instructions had been given to actors, but the observers then watched their yoked actor's video dating profile instead of creating one themselves.

Whenever actors completed measures of their meta-perceptions and self-perceptions, observers completed social perception measures. That is, after the first part of the actor's video dating profile but before the relationship IQ question, observers offered their baseline social perceptions. After observing the second part of the video dating profile, observers then offered their final social perceptions of their actor's dateability.

Trait perceptions. The perception measures comprised five items that assessed the actor's dateability. The *meta-perceptions* asked actors to guess how observers would judge them in light of their dating video. Actors knew they were guessing how observers would respond to those exact questions. The *social perceptions* asked observers to judge the actors in light of the dating video. The *self-perceptions* asked actors to judge themselves in light of the dating video they recorded.

The five questions asked participants to rate the actor compared to other undergraduates at the same university: "would make a better date than," "would make a better relationship partner than," "is more knowledgeable about dating than," "is likely to be in a happier relationship than," and "is more desirable as a dating partner than." Each of the questions was rated on a 101-point scale from 0% of fellow undergraduates at their university to 100% of fellow undergraduates at their university. For instance, an answer of 60% would indicate that the participant thought the actor would make a better date than 60% of fellow undergraduates at their university.

In this study, as in Study 2 in the main text, participants completed two sets of final ratings. One set asked participants to rate the actors' performance in light of the total dating video—i.e., based on the *global* set of information to which participants had been exposed. The other set asked participants to rate the actors' performance only based on the focal task, the

relationship IQ question—i.e., by responding to the *feedback-informed* behaviors for which actors had received success or failure feedback.

We averaged these items to create dateability perception composites. The self-perception ($\alpha = .97$), meta-perception ($\alpha = .97$), and social perception ($\alpha = .97$) of dateability composites all had good reliability.

Results and Discussion

Feedback-informed final impression. For the feedback-informed, we submitted the perception composites to a 2 (feedback: success or failure) X 3 (perception: self, meta, or social) X 2 (time: baseline or final feedback-informed) mixed-model ANOVA, with only the first factor manipulated between subjects. The Feedback X Perception X Time interaction was significant, $F(2, 176) = 7.93$, $p < .001$, $\eta^2_p = .08$, demonstrating that perceivers responded differently to success versus failure outcomes. We proceeded to conduct a series of 2 (feedback) X 2 (perception) X 2 (time) ANOVAs to understand whose perceptions were out of step with whose.

First, a significant 2 (feedback) X 2 (perception: meta or social) X 2 (time) mixed-model ANOVA revealed that actors believed observers would be more reactive to the final relationship IQ question than they actually were, $F(1, 88) = 11.51$, $p = .001$, $\eta^2_p = .12$. As depicted in Figure S1, observers did shift their social impressions in response to actors' success versus failure, $F(1, 88) = 22.41$, $p < .001$, $\eta^2_p = .20$. But this shift was less pronounced than actors thought it would be, $F(1, 88) = 80.20$, $p < .001$, $\eta^2_p = .48$.

Second, consistent with Studies 1 and 2 in the main text, a significant 2 (feedback) X 2 (perception: self or meta) X 2 (time) mixed-model ANOVA suggested that actors' meta-perceptions did not simply reflect their own self-perceptions, $F(1, 88) = 4.23$, $p = .04$, $\eta^2_p = .05$.

Although self-perceptions were reactive to the feedback, $F(1, 86) = 65.23$, $p < .001$, $\eta^2_p = .43$, they were less so than meta-perceptions.

Third, unlike Studies 1 and 2 in the main text, a significant 2 (feedback) X 2 (perception: self or social) X 2 (time) mixed-model ANOVA demonstrated that actors' self-perceptions were also out of sync with, and more extreme than, observers' social perceptions, $F(1, 88) = 5.37$, $p = .02$, $\eta^2_p = .06$. In short, actors' meta-perceptions were more reactive than were their self-perceptions, though observers' social perceptions were least reactive of all.

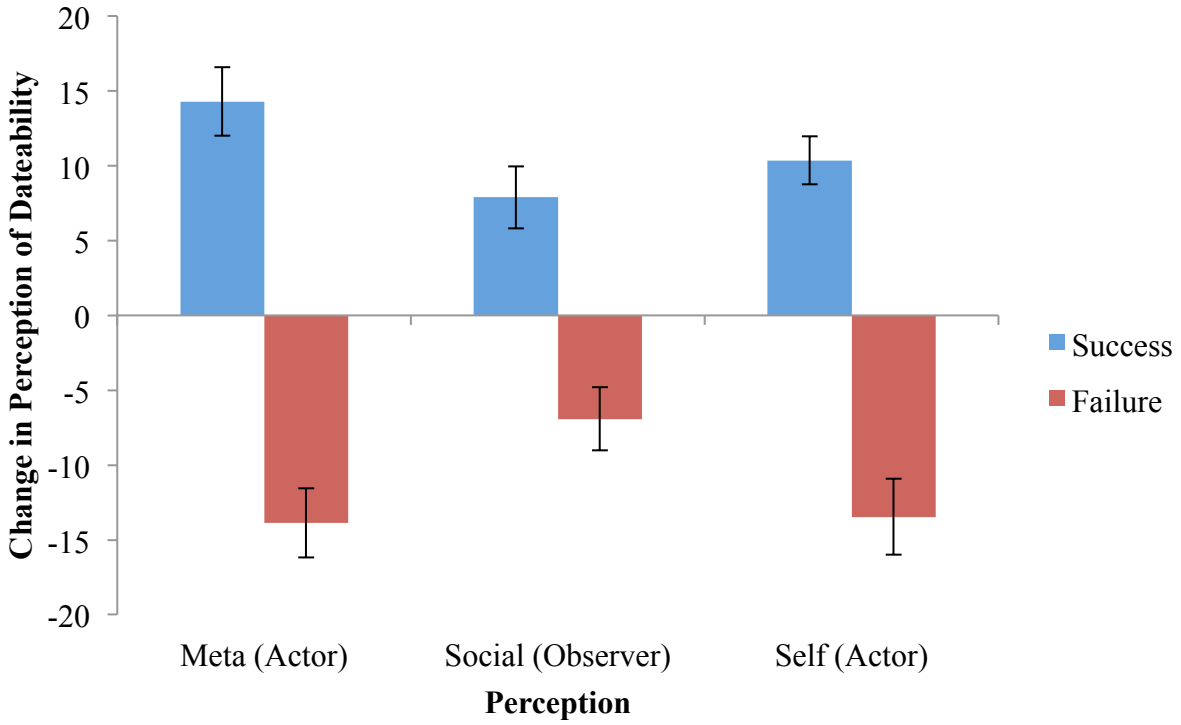


Figure S1. The change in perception (final – baseline) of actors' dateability by feedback condition and type of perception (Supplemental Study). Which participant offered each perception is in parentheses. The overblown implications effect is reflected by the larger gap between the two meta-perception bars compared to the two social perception bars.

Global final impression. For the global final impression, we submitted the perception composites to a 2 (feedback: success or failure) X 3 (perception: self, meta, or social) X 2 (time: baseline or final global) mixed-model ANOVA, with only the first factor manipulated between subjects. Surprisingly, this three-way interaction was not significant, $F < 1$.

We found this null effect somewhat baffling, especially given that the three perspectives differed in their interpretation of the focal event—the only event that occurred between the baseline and final perceptions. Regardless, we present this study and full results both in the interest of disclosure and because the effects on the feedback-informed impression measures provide clear support for the *overblown implications* account over the *performance focalism hypothesis*. But given the surprising null effect on the global perception measure, we felt it best to replicate our findings once more before moving to a more detailed examination of why the overblown implications effect emerges. Study 2 in the main text tests our hypotheses once again, but in a new context and with a new performance event.

Supplemental Study: Exploratory Measures and Additional Analyses

In this study, observers were also asked to report their sexual orientation. We had originally intended to match actors and observers based on their sexual orientations (e.g., heterosexual female with heterosexual male). But once it was clear we had mostly female participants, we dropped plans to match actors and observers in this way. Also, given we did not tell actors that we would select observers based on any particular demographic, we felt it best not to impose constraints when pairing actors and observers.

Furthermore, in addition to the five item dateability composite described in the text, we added a sixth item asking participants to rate the actor's physical attractiveness. Although physically attractive people may seem more dateable, it seems odd to say that someone score on

a relationship IQ question gives insight into that part of dateability. This is especially true when judges can see the actor. Nonetheless, had we added the physical attractiveness item to the composites, all results remain significant at the $p < .05$ level.

Including this study in the cross-study meta-analysis on public self-consciousness ($Z_{Global} = .25$, $Z_{Feedback\text{-}Informed} = -.35$) did not change our interpretation: the OIE grew stronger as actors' public self-consciousness increased, Stouffer's *Z*s between 2.22 and 2.96, *p*s between .03 to .003. (Like Study 2 in the paper, this study had both global and feedback-informed final measures. The range of meta-analytic results reflects the analyses depending on which measure(s) is chosen.)