

THE UNIVERSITY OF CHICAGO

KNOWING THYSELF:

ESSAYS ON THE ROLE OF SELF-AWARENESS IN INTERPERSONAL CONTEXTS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY

KRISTINA ANN WALD

CHICAGO, ILLINOIS

DECEMBER 2023

Copyright © 2023

Kristina Ann Wald

All rights reserved

To my family and friends

Table of Contents

| | |
|---|-----|
| List of Tables..... | vi |
| List of Figures..... | vii |
| Acknowledgements..... | ix |
| Overview..... | 1 |
| Chapter 1: The Credibility Dilemma: When Acknowledging a (Perceived) Lack of Credibility Can Make a Boast More Believable..... | 4 |
| Abstract..... | 5 |
| Introduction..... | 6 |
| Overview of Current Research..... | 21 |
| Study 1: Disclaimers Help Low-Credibility Speakers..... | 22 |
| Studies 2a-b: Disclaimers Do Not Help High-Credibility Speakers..... | 29 |
| Studies 3a-d: Generalizing to Multiple Traits..... | 36 |
| Study 4: Credibility Disclaimers in Realistic Interactions..... | 42 |
| Studies 5a-b: Disclaimers in Hiring Settings..... | 52 |
| General Discussion..... | 60 |
| Chapter 2: Ignorance can be Trustworthy: The Effect of Social Self-Awareness on Trust..... | 67 |
| Abstract..... | 68 |
| Introduction..... | 69 |
| Overview of Current Research..... | 78 |
| Study 1: Social Self-Awareness is Good, but Only if You're Nice to Me! | 79 |
| Study 2: The Price Rating is Right..... | 85 |
| Study 3: Recalling Non-Collegial Colleagues..... | 91 |

| | |
|--|-----|
| Study 4: If You're Not Going to Listen, At Least be Unaware..... | 97 |
| Study 5: It's Better if I Know You Can't Change..... | 110 |
| Study 6: It's Better if I Know You Can't Change, Take Two..... | 117 |
| Study 7: Do you really want to hurt me?..... | 122 |
| General Discussion..... | 127 |
| Chapter 3: When and Why People Discern Others' Degree of Social Self-Awareness..... | 132 |
| Abstract..... | 133 |
| Introduction..... | 134 |
| Pilot Study: Your (Lack of) Self-Awareness is Showing..... | 138 |
| Study 1: Self-Awareness—Online Dating Edition..... | 141 |
| Study 2: Negative Qualities Beget Thoughts on Self-Awareness..... | 149 |
| Study 3: Why are you behaving that way?..... | 155 |
| General Discussion..... | 159 |
| Appendix A: Supplemental Material, Additional Tables, and Stimuli for Chapter 1..... | 162 |
| Appendix B: Supplemental Material for Chapter 2..... | 177 |
| Appendix C: Study Materials and Additional Results in Chapter 3..... | 186 |
| References..... | 197 |

List of Tables

| | |
|--|-----|
| Table 1.1: Condition means and standard deviations for all studies in Chapter 1..... | 28 |
| Table 1.2: Summary of manipulations and stimuli in Studies 3a-d, Chapter 1..... | 38 |
| Table 2.1: Summary of studies and main findings in Chapter 2..... | 80 |
| Table 2.2: Results on manipulation check measures in Study 4, Chapter 2..... | 105 |
| Table 3.1: Selected responses from Pilot Study, Chapter 3..... | 140 |
| Table A.1.1: Additional results from Studies 2a-b, Chapter 1..... | 162 |
| Table A.1.2: Additional results from Studies 3a-d, Chapter 1..... | 163 |
| Table A.1.3: Additional results from Study 4, Chapter 1..... | 164 |

List of Figures

| | |
|---|-----|
| Figure 1.1: Proposed theoretical model in Chapter 1..... | 17 |
| Figure 1.2: Results from Study 1, Chapter 1..... | 27 |
| Figure 1.3: Results from Studies 2a-b, Chapter 1..... | 34 |
| Figure 1.4: Results from Studies 3a-d, Chapter 1..... | 40 |
| Figure 1.5: Overview of study procedure in Study 4, Chapter 1..... | 44 |
| Figure 1.6: Results from Study 4, Chapter 1..... | 50 |
| Figure 1.7: Stimulus 1 in Studies 5a-b, Chapter 1..... | 55 |
| Figure 1.8: Stimulus 2 in Studies 5a-b, Chapter 1..... | 56 |
| Figure 1.9: Results from Studies 5a-b, Chapter 1..... | 58 |
| Figure 2.1: Results from Study 1, Chapter 2..... | 85 |
| Figure 2.2: Example stimulus in Study 2, Chapter 2..... | 88 |
| Figure 2.3: Results from Study 2, Chapter 2..... | 90 |
| Figure 2.4: Overview of study procedures in Study 4, Chapter 2..... | 99 |
| Figure 2.5: Results from Study 4, Chapter 2..... | 107 |
| Figure 2.6: Results from Study 5, Chapter 2..... | 116 |
| Figure 2.7: Results from Study 7, Chapter 2..... | 126 |
| Figure 3.1: Results from Study 1, Chapter 3..... | 147 |
| Figure 3.2: Results from Study 2, Chapter 3..... | 152 |
| Figure 3.3: Additional results from Study 2, Chapter 3..... | 154 |
| Figure 3.4: Results from Study 3, Chapter 3..... | 158 |
| Figure A.1.1: Stimuli in Study 4, Chapter 1..... | 170 |
| Figure A.1.2: Additional stimuli in Study 4, Chapter 1..... | 171 |

| | |
|---|-----|
| Figure A.1.3: First half of job posting for Study 5b, Chapter 1..... | 172 |
| Figure A.1.4: Second half of job posting for Study 5b, Chapter 1..... | 173 |
| Figure A.2.1: Additional results from Study 1, Chapter 2..... | 178 |
| Figure A.2.2: Additional results from Study 1, Chapter 2..... | 178 |
| Figure A.2.3: Additional results from Study 5, Chapter 2..... | 180 |
| Figure A.2.4: Additional results from Study 5, Chapter 2..... | 181 |
| Figure A.3.1: Stimuli in Pilot Study, Chapter 3..... | 186 |
| Figure A.3.2: Instructions in Study 1, Chapter 3..... | 187 |
| Figure A.3.3: Additional results from Study 1, Chapter 3..... | 189 |
| Figure A.3.4: Additional results from Study 1, Chapter 3..... | 190 |
| Figure A.3.5: Additional results from Study 2, Chapter 3..... | 191 |
| Figure A.3.6: Additional results from Study 2, Chapter 3..... | 192 |
| Figure A.3.7: Research assistant coding for “social anxiety” from Study 3, Chapter 3..... | 193 |
| Figure A.3.8: Research assistant coding for “matching intentions” from Study 3, Chapter 3...194 | 194 |
| Figure A.3.9: Research assistant coding for “good reason” from Study 3, Chapter 3..... | 194 |
| Figure A.3.10: Research assistant coding for “ulterior motive” from Study 3, Chapter 3..... | 195 |
| Figure A.3.11: Research assistant coding for “impression incorrect” from Study 3, Chapter 3..... | 195 |
| Figure A.3.12: Research assistant coding for “other explanations” from Study 3, Chapter 3...196 | 196 |

Acknowledgements

I owe a great deal of gratitude to many people for supporting me both intellectually and emotionally throughout my time in the PhD (and, of course, for helping me to develop my own self-awareness ☺). I am deeply indebted to my advisor and Committee Chair, Shereen Chaudhry, for providing the most wonderful combination of mentorship, guidance, support, and congeniality I could have asked for throughout my time in the PhD. I am extremely proud to have been your first student. I am also incredibly grateful to the rest of my committee members and collaborators: Jane Risen, for additional excellent mentorship, support, advice, and encouragement; Emma Levine, for superb wisdom; and Nick Epley, for insightful guidance and feedback. Much gratitude as well to Ed O'Brien, for being another wonderful collaborator and teaching me a lot about the publication process, and to Adam Galinsky and Mabel Abraham, for being incredible collaborators from afar.

I also extend many, many thanks to everyone else at Booth who made my experience as wonderful, supportive, productive, and fun (!) as it was—too many to name, but I will try: Diag Davenport, Alex Moore, David Munguia Gomez, Yuji Winet, Annabelle Roberts, Danny Katz, Dan Medvedev, Akshina Banerjee, MK Jang, Lin Fei, Rusty Roberts, Nicholas Herzog, Rafael Batista, Soaham Bharti, Jiaqi Yu, Yena Kim, Nicholas Owsley, Umy Yasar, Stephanie Hong, Roman Gallardo, Tong Su, Jiabi Wang, Felicia Joy, Radhika Santhanagopalan, Sam Hirschman, Kariyushi Rao, Quinn Hirshi, Michael Kardas, Stav Atir, Rui Sun, Alex Koch, Erika Kirgios, Ayelet Fishbach, Bernd Wittenbrink, Becky White, Amy Boonstra, and many, many others.

I am grateful to my New York relatives—Duncan, Jenna, and Bob MacVicar, for providing an excellent New York crash pad throughout my PhD along with superb and hilarious

conversations, as well as Jane Weiss, for her boundless warmth and kindness right through the end of her life.

I want to express tremendous gratitude to my dad and my sister Sarah, for providing endless love and support through the highs and lows. I cannot express how thankful I am to have you as my family, and I don't know what I would have done without you.

And finally, I owe more than I can say to my mom, whom I am incredibly grateful to have had for the first 20 years of my life, who I so wish could have lived to see me write this, and who always—without fail—encouraged me to live life “with every single lasting ounce of passion that you have.” I have tried to do just that, and this dissertation is one outcome of that passion.

Overview

Think back to the last person you met or interacted with. What characteristics did you notice about them? In particular, did they seem *self-aware*, or not? In my dissertation, I explore the role of self-awareness in interpersonal contexts. While much existing research has examined how being self-aware affects one's own subsequent experiences and behaviors (Diener & Wallbom, 1976; Duval & Wicklund, 1972; Hass, 1984; Heatherton & Baumeister, 1991; Wicklund, 1975), I examine how people *perceive self-awareness in others*, and how the perception that another person is self-aware (or not) affects observers' subsequent judgments of that person. I focus in particular on *social* self-awareness, which I define as an accurate awareness of what others think of oneself (in contrast to other types of self-awareness, such as an awareness of internal experiences like thoughts or emotions).

In Chapter 1, I examine how expressing self-awareness can resolve a common impression management dilemma. Often, people who are judged negatively by others (e.g., as low in competence) face a dilemma: They may want to self-promote (to improve this negative impression), but may simultaneously worry that their claims may not seem *believable*. I dub this type of situation the credibility dilemma, and find that explicitly expressing self-awareness about one's perceived shortcoming helps to resolve this dilemma. In particular, prefacing one's self-promotional statement with a "credibility disclaimer" (e.g., "I'm not that smart, but..." or "I know this may seem hard to believe, but...") actually makes the statement seem *more* believable, and leads the speaker be perceived *more* positively, relative to self-promoting without any disclaimer. This occurs, at least in part, because these disclaimers increase the perception that the speaker is self-aware, which in turn yields more positive overall impressions.

In Chapter 2, I provide a more general framework for how perceiving self-awareness in another person affects observers' subsequent judgments of that person, and in particular, how trustworthy that person seems. I find that although self-awareness can signal positive qualities to others—like in Chapter 1—it does not universally enhance others' trust. This is because self-awareness also signals greater *intentionality* behind a target person's behaviors. In other words, when a target person appears to be high in social self-awareness, observers infer that the target's actions are more diagnostic of the target's true character and future behavior, which thus affects trust differently depending on the target's specific behaviors. When the target behaves in ways that positively impact others (e.g., being kind and friendly to others), exhibiting self-awareness increases trust, as the positive behaviors are interpreted as more intentional and diagnostic, but for behaviors that negatively impact others (e.g., being rude and unfriendly), exhibiting self-awareness *decreases* trust, as negative behaviors are seen as worse when more intentional.

Finally, in Chapter 3, I examine when and why observers are most likely to spontaneously evaluate a target person's degree of self-awareness in the first place, in the absence of specific cues. I propose that when observers are surprised by a target person's behavior, or evaluate the target negatively on other attributes or behaviors, the observer undergoes a more thoughtful attribution process in order to make sense of the behavior. In doing so, one explanation that sometimes comes to mind is the person's degree of self-awareness (e.g., that the person lacks self-awareness or is highly self-aware).

Overall, my findings suggest that self-awareness has important *interpersonal* consequences, not just *intrapersonal* consequences. Further, my findings suggest that the effect of self-awareness on interpersonal judgment is nuanced: While self-awareness is often considered a positive and desirable quality in others—and does indeed lead to more positive

judgments for those in particular circumstances (e.g., the credibility dilemma)—it does not universally enhance others' trust. Finally, my findings also suggest that self-awareness is a quality that people spontaneously evaluate in others even in the absence of specific expressions of self-awareness, suggesting that it may be worth considering alongside other commonly-researched traits such as competence and trustworthiness. Taken together, my research highlights the importance of examining self-awareness within the interpersonal domain.

Chapter 1:

The Credibility Dilemma: When Acknowledging a (Perceived) Lack of Credibility Can

Make a Boast More Believable

Abstract

People who are judged negatively by others (e.g., as low in competence) often face a dilemma: They may want to self-promote (to improve this negative impression), but worry their claims may not seem *believable*. We term this type of situation the “credibility dilemma,” and investigate how people can self-promote most effectively in such cases. In particular, we examine the impact of explicitly acknowledging one’s perceived lack of credibility while self-promoting (e.g., “I’m not that smart, but...” or “I know this may seem hard to believe, but...”). Across ten studies, we find that credibility disclaimers *improve* perceptions of the self-promoter (compared to self-promoting without them) by increasing perceptions of the speaker’s self-awareness and sincerity. In contrast, credibility disclaimers are ineffective (and sometimes backfire) when the speaker is *already* perceived as credible. Our findings suggest that common advice to avoid drawing attention to one’s flaws may sometimes be unwarranted.

Introduction

The 2001 film *Legally Blonde* depicts a young woman, Elle Woods, whose intelligence is constantly underestimated by others. On account of her penchant for pink and her bubbly personality, almost everyone who encounters her assumes she must be lacking in intellect—forcing her to prove her abilities to others time and again. The situation Elle faces is not uncommon. For a variety of reasons, people are often initially judged negatively by others (in many cases, mistakenly so), and thus need to find ways to correct others' unduly negative impressions of them. In particular, this type of situation is common in the workplace: We ran a pilot study among current full-time and part-time employees ($N = 202$), and found that 86% of people reported having felt underestimated, or unfairly negatively judged, by their workplace colleagues at some point. Some examples of the experiences people described included times when they made mistakes due to momentary anxiety or situational factors (rather than true inability); times when they were assumed to be inept due to newness at a position, youth, or being a non-native speaker; and times when they did not receive an adequate attribution of credit for their work or idea. Given that creating positive impressions on others is a crucial aspect of navigating the social world (Aloise-Young, 1993; Baumeister, 1982; Baumeister & Jones, 1978; Fiske, Cuddy, & Glick, 2007; Leary & Kowalski, 1990; Schlenker, 1980; Tedeschi, 2013), these types of misjudgments pose a challenge: How can such people correct others' negative pre-existing impressions?

There are many strategies people might use to try to improve others' perceptions of them, such as performing well on subsequent tasks (Borman, White, & Dorsey, 1995), improving others' perceptions of them on ancillary characteristics (Landy & Sigall, 1974; Stellar & Willer, 2018), or engaging in ingratiation (Gordon, 1996; Edward E. Jones, Gergen, Gumpert, &

Thibaut, 1965). The current research focuses on one ubiquitous, low-cost strategy: bragging (Chaudhry & Loewenstein, 2019; Giacalone & Rosenfeld, 1986; Heck & Krueger, 2016; Edward E. Jones & Pittman, 1982; Schlenker & Leary, 1982; Vonk, 1999). We define bragging broadly as communicating any kind of self-promotional information (Berman, Levine, Barasch, & Small, 2015; Chaudhry & Loewenstein, 2019), and we focus in particular on brags that are specific and verifiable. Note that we do not limit the term “bragging” to statements that are excessive or undeserved.

Our research examines the particular bragging dilemma described above: We investigate how people who are *already* perceived negatively on a given trait (e.g., low in competence, or as otherwise lacking on any given trait) can communicate self-promotional information about their abilities on that trait most effectively, given that their statements are less likely to be believed in the first place. We refer to this type of situation as the “credibility dilemma”: When an audience already has a negative impression of a target person on a particular dimension (e.g., math ability, athletic skills, generosity, etc.), this person is more likely to want or need to self-promote on this dimension (to improve this impression), but is also less likely to be *believed* when they do so because the audience’s prior impression will influence how they perceive the target’s subsequent statements. We refer to such speakers as “low-credibility,” where the lack of credibility can be with respect to any trait (including both competence- and warmth-related traits). In such cases, common advice might be to avoid drawing attention to one’s perceived flaws at all costs, in order to steer the audience away from their pre-existing negative perceptions and instead focus their attention on one’s positive attributes.

Yet we find that when such low-credibility speakers explicitly acknowledge their perceived shortcomings through a credibility disclaimer (e.g., “I’m not that good at writing,

but...” or “I know this may seem hard to believe, but...”), they are perceived *more* positively than when they brag without one. We find that the image boost from the credibility disclaimer occurs for (at least) two main reasons: First, the disclaimer demonstrates the speaker’s self-awareness and (accurate) perspective-taking. Second, the disclaimer signals the speaker’s sincerity in making the statement. Yet we also find that these same credibility disclaimers are not effective for all speakers: By the same logic, these disclaimers are ineffective (and sometimes backfire) for high-credibility speakers (i.e., those who were *not* already perceived negatively on a particular trait by their audience). Such speakers appear *less* self-aware and *less* sincere when offering credibility disclaimers.

Our findings make three primary contributions to prior literature. First, we contribute to research on person perception and impression management by highlighting two crucial mechanisms that influence impressions: the speaker’s perceived self-awareness and the speaker’s perceived sincerity. While some research has already examined the role of perceived sincerity in impression formation (Berman et al., 2015; Crant, 1996; Ohtsubo & Watanabe, 2009; Sezer, Gino, & Norton, 2018), less research (to our knowledge) has studied the role of perceived self-awareness. Yet, as our findings suggest, perceiving a target as self-aware is an important determinant of how one perceives the target, particularly with regard to the target’s warmth. In our case, a target who demonstrates self-awareness is also perceived as warmer overall.

Second, we contribute to research on impression management by demonstrating and explaining at least one situation in which the success of impression management strategies depends on *who* attempts them. That is, the extent to which a strategy is effective can depend on the specific characteristics of the actor (Berman et al., 2015). In our case, it depends on whether the listener’s prior impression of the actor seems to match the strategy itself—i.e., the benefits of

acknowledging one's shortcomings depend on whether the listener actually perceives the speaker as having these shortcomings. Understanding these moderators to impression management strategies is important because—outside of very first impressions—people constantly update their judgments of others after having already formed some impression. Clarifying how judgments of the same act may be differentially influenced by those prior impressions adds crucial nuance to our understanding of optimal impression management strategies. Moreover, if observers' pre-existing impressions matter in determining the optimal strategy, then the actor's success will in part be determined by their ability to accurately assess what others think of them. Given extensive research demonstrating that people have difficulty accurately assessing what others think of them (Boothby, Cooney, Sandstrom, & Clark, 2018; Bruk, Scholl, & Bless, 2018; Gilovich, Medvec, & Savitsky, 2000; Gilovich, Savitsky, & Medvec, 1998; Kenny & DePaulo, 1993; Moore-Berg, Ankori-Karlinsky, Hameiri, & Bruneau, 2020; Savitsky, Epley, & Gilovich, 2001; Zhao & Epley, 2021), our work underscores a significant challenge that impression managers face.

Third, we contribute to literature on establishing credibility. Some research in the realm of persuasion suggests that appearing confident can make one seem more credible to others (Anderson, Brion, Moore, & Kennedy, 2012; Price & Stone, 2004), while other research suggests that moderate—as opposed to high—levels of confidence are most persuasive (Cramer, Brodsky, & Decoster, 2009), and that high confidence is viewed negatively when unwarranted (Tenney, MacCoun, Spellman, & Hastie, 2007). We build on these findings by establishing when and why explicitly discounting one's own credibility—perhaps similar to displaying lower confidence—can make one seem *more* credible to others within the domain of self-promotion. In particular, we find support for the importance of calibration: Only speakers who are already

perceived as low in credibility benefit from discounting their own credibility while self-promoting.

The Benefits and Costs of Self-Promotion

As mentioned, it is well-established that people are motivated to create positive impressions on others (Aloise-Young, 1993; Baumeister, 1982; Baumeister & Jones, 1978; Fiske et al., 2007; Leary & Kowalski, 1990; Schlenker, 1980; Tedeschi, 2013). While there are many strategies people may use to improve others' impressions of them (Gordon, 1996; E.E. Jones, 1964), verbally conveying information about one's positive characteristics and accomplishments is a common way to do so (Scopelliti, Loewenstein, & Vosgerau, 2015; Sezer et al., 2018; Tice, Butler, Muraven, & Stillwell, 1995). Moreover, it is often successful: In the absence of any contradictory information, people generally tend to believe others' brags, and thus perceive the target more positively on the focal trait than if they did not brag (Berman et al., 2015; Vonk, 1999). Not surprisingly, then, self-promotion can have tangible benefits, such as aiding in professional advancement (Ellis, West, Ryan, & DeShon, 2002; Proost, Schreurs, De Witte, & Deros, 2010).

While past research has uncovered these benefits of bragging, much research on bragging has also focused on its costs. More specifically, a significant body of research has examined the so-called warmth-competence tradeoff that bragging entails (Berman et al., 2015; Chaudhry & Loewenstein, 2019; Pfeffer, Fong, Cialdini, & Portnoy, 2006; Rudman, 1998; Scopelliti et al., 2015). In other words, the act of bragging contradicts norms of humility and modesty (Baumeister & Ilko, 1995; Exline & Geyer, 2004; Forsyth, Berger, & Mitchell, 1981; Tice et al., 1995; Wosinska, Dabul, Whetstone-Dion, & Cialdini, 1996), which in turn can harm perceptions of the braggart's warmth (even while increasing perceptions of their competence). As a result,

much prior research has focused on strategies people can use to try to reap the benefits of bragging without incurring the costs (Pfeffer et al., 2006; Sezer et al., 2018; Tal-Or, 2010a, 2010b; VanEpps, Hart, & Schweitzer, 2023).

In the current research, however, we examine a different bragging dilemma than that of the warmth-competence tradeoff. Instead, we examine how people can brag most effectively when their claims are unlikely to be believed in the first place, i.e., when the listener has a prior impression of the speaker that contradicts his or her claim. We refer to this type of situation as a “credibility dilemma.” This dilemma occurs fairly often: In the same pilot study mentioned previously, we found that 33% of respondents reported wanting to tell a colleague something positive about themselves, but being worried that the other person might not believe the statement and/or might not attribute the accomplishment to the respondents’ own abilities. In such cases, we propose that explicitly acknowledging one’s perceived lack of credibility—via credibility disclaimers—may serve as one useful strategy for conveying self-promotional information more effectively.

Acknowledging Negative Impressions

One might expect that low-credibility speakers would be *worse off* using a disclaimer that acknowledges the brag’s lack of credibility, compared to bragging without one. Indeed, much popular advice warns against drawing too much attention to one’s shortcomings because it could make those shortcomings all the more focal for the listener. For example, job seekers are often encouraged to share weaknesses that actually communicate strengths when answering the question “What is your greatest weakness?” in a job interview (e.g., “I focus too much on the details” or “I get impatient when projects run beyond the deadline”) (“List of Weaknesses,” 2022). Beyond popular advice, existing research suggests that drawing attention to the negative

aspects of one's statement (via other types of disclaimers, e.g., "I don't mean to sound arrogant, but...") can lead the listener to perceive subsequent information as consistent with this negative interpretation (El-Alayli, Myers, Petersen, & Lystad, 2008).

However, other research suggests that acknowledging negative attributes may not always be a bad thing. For example, acknowledging a lack of clarity in communication or a poor performance can lead to more positive perceptions of those characteristics (E. Knowles & Linn, 2004; Ward & Brenner, 2006), and explicitly mentioning the negative aspects of a product, along with corresponding counterarguments, can result in more positive evaluations of that product (Crowley & Hoyer, 1994; Etgar & Goodwin, 1982; Rucker, Petty, & Briñol, 2008). More generally, leaders who reveal weaknesses come across as more authentic (Jiang, John, Boghrati, & Kouchaki, 2022); revealing negative information (as opposed to hiding it) can improve perceptions of trustworthiness (John, Barasz, & Norton, 2016); and sharing failures can mitigate envy from others (Brooks et al., 2019). Disclosing negative information can also improve perceptions of warmth (Hoffman-Graff, 1977), and revealing mistakes can be perceived positively when it humanizes the target (Aronson, Willerman, & Floyd, 1966).

Of course, the examples above are not in the context of self-promotion and are not focused on those trying to overcome a pre-existing negative impression. Inadvertently revealing a foible (like spilling coffee) improves perceptions of someone who is otherwise seen as highly competent by making them appear more relatable, but the same foible hurts impressions for those who are only seen as average in ability (Aronson et al., 1966). Thus, it may be particularly risky to acknowledge negative information when trying to overcome an initial negative impression.

To assess lay theories about what to do in these situations and to gauge whether people spontaneously acknowledge their audience’s negative impression when facing a credibility dilemma, we ran a second pilot study ($N = 100$). Participants imagined applying for a job, knowing that their qualifications were not very strong. They were encouraged to write an email to submit with their application that would describe an accomplishment of theirs that might help them get the job. After writing it, participants reported whether or not their email acknowledged the perceived lack of credibility of the brag (given their weak qualifications). We found a split among participants: 68% chose to use credibility disclaimers or similar language (e.g., “My GPA isn’t stellar but...” and “although my grades may be lacking slightly...”). Many of those who acknowledged the lack of credibility reported that it felt more honest or truthful to do so, or that they wanted to demonstrate an awareness of their apparent weaknesses. Among the 31%¹ of participants who chose not to acknowledge their perceived weakness in their email, many of them explained that they did not want to risk drawing attention to it.

In this paper, we test which of these two groups is correct—i.e., whether those who acknowledge their audience’s negative prior impression when they brag have the correct or incorrect lay theory for improving how people see them. Despite the potential risk, we predict that drawing attention to negative characteristics can be a good thing—even in the context of self-promotion. Further, we document the underlying psychological mechanisms that explain why this may be a helpful strategy for those facing the credibility dilemma.

Our specific paradigm involves testing the effect of prefacing a brag with a *disclaimer* that acknowledges one’s perceived lack of credibility. A disclaimer is formally defined as “a

¹ One additional participant selected that they neither included nor excluded an acknowledgement of their weaknesses. Upon reading their response, it appears that they did *not* include an acknowledgement (and perhaps misunderstood the self-coding question).

verbal device employed to ward off and defeat in advance doubts and negative typifications which may result from intended conduct” (Hewitt & Stokes, 1975, p. 3). In the context of the current research, we focus on disclaimers that acknowledge the apparent contradiction between one’s self-promotional claim and the listener’s (negative) prior impression of the speaker. This includes disclaimers that explicitly acknowledge the speaker’s perceived shortcoming on the focal trait, such as “I’m not that smart, but…” or “I know my record might not seem as impressive as other applicants’ records, but…”, as well as disclaimers that more generally acknowledge the speaker’s likely incredulous reaction, such as “This may sound strange to you, but…” or “I know this may seem hard to believe, but…”

Credibility Disclaimers Signal Self-Awareness and Sincerity for Low-Credibility Speakers

We propose two distinct reasons for why credibility disclaimers should *improve* perceptions of low-credibility speakers who self-promote. First, low-credibility speakers who use disclaimers with their brags could be perceived as having high self-awareness and as engaging in perspective-taking. Acknowledging a negative pre-existing impression via a disclaimer requires some degree of self-awareness; that is, the speaker must be directing their attention toward the self (Duval & Wicklund, 1972; Wicklund, 1975). More specifically, the act of acknowledging another person’s impression of oneself requires “social” self-awareness, which refers to an awareness of how one is being viewed by other people (Chon & Sitkin, 2021). In theory, people can be socially self-aware, but inaccurate; in other words, someone can acknowledge what others think of them, but be wrong about the impression they are making. However, self-awareness tends to improve perspective-taking, making individuals more accurate about others’ perspectives (Hass, 1984). Thus, we argue that when speakers display *accuracy* regarding others’ perspectives of them, observers will judge speakers to be self-aware. For this reason, we predict

that low-credibility speakers who preface their brags with an acknowledgement of their pre-existing negative impression will be perceived as more self-aware than if they make no acknowledgement.

The second reason that we expect disclaimers to help low-credibility speakers is because the disclaimer may change perceptions of the speaker's intentions in making the statement. Specifically, it may make the speaker appear more *sincere* in conveying the information in the statement, rather than appearing to manipulate the listener's impression of them. As discussed, low-credibility speakers' self-promotional claims (without a disclaimer) are especially unlikely to be believed given their incongruence with the listener's prior impression of the speaker. Indeed, the claim might be viewed as an attempt to get away with an unrealistically positive statement in the hopes that the listener will not remember their prior impression of the speaker's shortcomings. The perception of this motive, in turn, would lead the listener to perceive the speaker as insincere (E. Jones & Pittman, 1982).

The addition of a credibility disclaimer, however, could shift perceptions of the speaker's intentions. By definition, a disclaimer draws attention to something negative about one's forthcoming statement (i.e., the fact that the listener may have reason to discount the credibility of the speakers' statement). For low-credibility speakers, the act of highlighting this information is costly in that it risks drawing the listener's attention to the reasons they have to doubt the speaker's credibility. Yet it is precisely the costliness of this act that signals the speaker's sincerity in conveying honest information to the listener. In accordance with costly signaling theory, we expect people to infer that only honest speakers would be willing to incur such a cost (Chaudhry & Wald, 2022; Connelly, Certo, Ireland, & Reutzel, 2011; Gangestad & Thornhill, 2011; Zahavi & Zahavi, 1999). Thus, the claim itself may seem more believable. Similar

examples of the relationship between costly signaling and perceived sincerity have been illustrated in other forms of communication and interpersonal interactions, such as apologies (Ohtsubo & Watanabe, 2009), persuasive messaging (Crowley & Hoyer, 1994; Etgar & Goodwin, 1982; Rucker et al., 2008), and generosity (Swap, 1991; Tesser, Gatewood, & Driver, 1968). Overall, we expect that—relative to bragging with no disclaimer—including a credibility disclaimer increases the perceived sincerity of low-credibility speakers and, as a result, also the believability of the claim itself.

Credibility Disclaimers Improve Warmth Perceptions for Low-Credibility Speakers

Because of their predicted impact on perceived self-awareness and sincerity, we further expect disclaimers to have a positive impact on warmth-related impressions of low-credibility speakers. In other words, we predict that low-credibility speakers will be seen as warmer (more likable, honest, and trustworthy) when they brag with a disclaimer than without one. First, we expect that listeners will exhibit more liking and trust for individuals who demonstrate self-awareness and accurate perspective-taking because these traits are associated with beneficial social behaviors by those who possess them, such as reduced antinormative behavior and greater social coordination (Batson, Early, & Salvarani, 1997; Diener & Wallbom, 1976; Galinsky, Ku, & Wang, 2005; Galinsky, Maddux, Gilin, & White, 2008; Trötschel, Hüffmeier, Loschelder, Schwartz, & Gollwitzer, 2011; Wang, Kenneth, Ku, & Galinsky, 2014; Wang, Ku, Tai, & Galinsky, 2014; Wicklund, 1975). Self-awareness is also one component of humility (Chancellor & Lyubomirsky, 2013; Davis, Worthington, & Hook, 2010; Nielsen & Marrone, 2018; Tangney, 2000; Van Tongeren, Davis, Hook, & Witvliet, 2019), which can lead to interpersonal benefits (Owens, Johnson, & Mitchell, 2013) and may thus further lead to liking. Consistent with all of these findings, some of the literature on meta-perception—or one’s beliefs about what others

think of them—has found a correlation between accurately knowing how others view oneself and being perceived positively (Brion, Lount, & Doyle, 2015; Ohtsubo, Takezawa, & Fukuno, 2009).

Second, we also expect that perceptions of sincerity should lead to more positive warmth-related perceptions of the speaker. Perhaps not surprisingly, prior research indicates that perceptions of the speaker’s (in)sincerity influences the listener’s overall evaluations of them (Barasch, Levine, Berman, & Small, 2014; Crant, 1996; Ohtsubo & Watanabe, 2009; Sezer et al., 2018). In our case, if the speaker’s statement seems more sincere (when a credibility disclaimer is present), then we would expect observers to also infer that their statement is more likely to be honest and that they are a more honest person overall (compared to if the speaker did not use a disclaimer). We expect that greater perceived honesty will also lead observers to trust and like such targets more. For all of these reasons, we expect that disclaimers will increase warmth-related impressions of low-credibility speakers. See Figure 1.1 for our full proposed theoretical model.

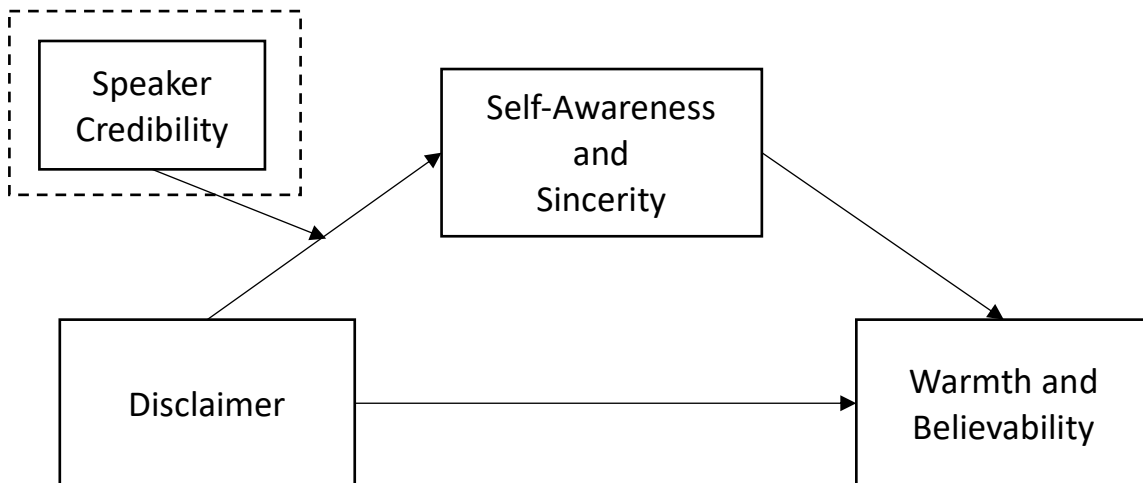


Figure 1.1. Proposed theoretical model in Chapter 1. Note that Studies 1 and 5 focus on low-credibility speakers, where we expect that including a disclaimer when bragging will increase warmth and believability by increasing perceptions of self-awareness and sincerity. Studies 2-4 test our prediction that this process will only hold for low-credibility speakers. When the brag already seems credible, we predict that using a disclaimer will not increase, and may even decrease, warmth perceptions by decreasing perceptions of self-awareness and sincerity.

We make less strong predictions about how disclaimers will affect perceived competence. On the one hand, disclaimers could *increase* perceived competence for low-credibility speakers. Demonstrating self-awareness may showcase cognitive capacities that in turn make one appear more competent. Further, as described above, we expect disclaimers to increase the believability of the self-promotional claim. If and when this self-promotional claim is about the speaker's competence, then finding the statement more believable ought to in turn increase the speaker's perceived competence. On the other hand, given that one of our key mechanisms is perceived sincerity, which relates most strongly to warmth-related characteristics, it could be that the effect of disclaimers on perceived competence will be *less strong* than that on warmth perceptions. At minimum, however, we do not expect to find a cost to using a disclaimer on perceived competence. Thus, we expect that low-credibility speakers who self-promote with a disclaimer will be seen as warmer and potentially more (but no less) competent compared to those who do not use a disclaimer.

Differentiating Credibility Disclaimers from Self-Deprecation, Modesty, and Attempts at Humor

We suggest that low-credibility speakers are seen more positively when they brag with a disclaimer because the disclaimer makes them seem more self-aware and sincere. It is possible, however, that disclaimers have their effect through a different process. For example, because disclaimers focus attention on negative information about oneself, disclaimers could be seen as self-deprecating statements, which can sometimes have interpersonal benefits for the speaker (Speer, 2019). Similarly, disclaimers could be viewed as attempts at humor, which can also increase positive perceptions of the speaker in some cases (Bitterly, Brooks, & Schweitzer,

2017). Or, they may be seen as a sign of modesty, which is generally considered a positive trait (Wosinska et al., 1996).

Previous research leads us to be skeptical that disclaimers are working through any of these avenues in this context. First, self-deprecating humor is perceived more positively from high-status than low-status speakers (Greengross & Miller, 2008). A self-deprecation account would therefore suggest that disclaimers should be, if anything, *more* effective when coming from someone perceived to be high in competence and status on a given dimension. We predict the opposite. Second, humor has been found to decrease the perception that the speaker has the goal of conveying accurate information (Bitterly & Schweitzer, 2019). If disclaimers are seen as attempts at humor, we would expect the subsequent bragging statement to be perceived as *less* believable. Again, we predict the opposite. Finally, modesty has been defined as “the underrepresentation of one’s positive traits, contributions, expectations, or accomplishments” (Wosinska et al., 1996, p. 626). Using disclaimers when facing a credibility dilemma does not seem to fit this definition because the disclaimer is accurately (rather than under-) representing the audience’s prior impression and because the disclaimer is paired with a self-promotional claim, which conflicts with downplaying achievements. Furthermore, as with self-deprecation or humor, if disclaimers worked simply by showing modesty, then they should benefit high-credibility speakers too. We predict the opposite.

To empirically address these alternatives, we test whether the positive effect of disclaimers holds for all speakers or whether it is limited to speakers who seem more self-aware and more sincere when using them. If disclaimers improve warmth perceptions either because they are seen as self-deprecatory, humorous, or modest (rather than signaling self-awareness and sincerity), then they should improve perceived warmth even among those who do not face a

credibility dilemma—that is, they should also increase the perceived warmth of high-credibility speakers (i.e., speakers for whom the listener’s prior impression does *not* contradict the content of the self-promotional claim).

However, we propose that the impact of credibility disclaimers differs from the impact of self-deprecation, humor, and modesty. In particular, we expect that disclaimers will not help, and may even harm, perceptions of high-credibility speakers—in terms of both warmth and believability—because of what disclaimers’ signal about self-awareness and sincerity. First, high-credibility speakers who use credibility disclaimers are likely to be perceived as *lacking* in self-awareness and *failing* to engage in perspective-taking, relative to those who brag without a disclaimer, given that the disclaimer contradicts the listener’s impression of the speaker. That is, referring to one’s own shortcomings on the relevant trait (e.g., “I’m not that smart”) or the statement’s lack of believability (e.g., “I know this may seem hard to believe”) would signal that the high-credibility speaker is out of touch with the listener’s true perceptions of them, thus signaling a lack of (accurate) self-awareness. Because we expect high-credibility speakers to be perceived as *less* self-aware when they brag with a disclaimer than without one, we expect high-credibility speakers to be perceived as less warm when they brag with a disclaimer than without one.

Second, high-credibility speakers who use credibility disclaimers are likely to be perceived as insincere relative to those who brag without a disclaimer. When a speaker is already perceived as credible, their claim by itself will already seem believable (Vonk, 1999), and thus is more likely to be seen as a sincere attempt to convey true information. As a result, the addition of a disclaimer may actually disrupt this perception of sincerity. In this case, the disclaimer is more likely to be perceived as an attempt at false modesty—rather than a genuine acknowledgement of

one's perceived shortcoming—and thus insincere. This prediction could be seen as analogous to the ingratiation's dilemma (i.e., that ingratiation must prevent their audience from *realizing* they are engaging in ingratiation) (E.E. Jones, 1964). In our case, disclaimers from high-credibility speakers might be perceived as signaling an intention to appear modest and likable, and may thus seem manipulative rather than sincere. This prediction is also consistent with the (negative) effects of humblebragging (which has been studied in cases where the speaker does *not* seem to lack credibility) (Sezer et al., 2018), and with the finding that downplaying the importance of an accomplishment (when one's claim is credible) backfires by reducing perceived modesty (Schlenker & Leary, 1982). Therefore, we expect high-credibility speakers to be perceived as *less* sincere when they brag with a disclaimer than without one, and as a result, less warm and less believable.

Overview of Current Research

Altogether, we propose that prefacing brags with credibility disclaimers can improve warmth perceptions of low-credibility speakers (i.e., those facing a credibility dilemma). We also propose that disclaimers have this impact because they signal self-awareness and sincerity, not because they involve self-deprecation, humor, or modesty. Thus, rather than expecting disclaimers to improve warmth perceptions for all speakers, we expect that disclaimers will not help, and may hurt, warmth perceptions of high-credibility speakers.

We tested these hypotheses across ten pre-registered studies. In Study 1, we tested whether credibility disclaimers improve warmth-based perceptions and behavioral trust toward low-credibility speakers. In Studies 2a-b, we varied speaker credibility to test whether such disclaimers help warmth perceptions of *only* low-credibility—not high-credibility—speakers, thus helping to rule out alternative accounts of the benefit of disclaimers. Because our theory of

credibility suggests that disclaimers should be useful for addressing negative perceptions on any trait, not just competence, Studies 3a-d tested the generality of our effects with a variety of traits, including those unrelated to workplace competence. In Study 4, we replicated our key findings with richer stimuli that simulated live interactions. Finally, in Studies 5a-b, we tested whether credibility disclaimers have material consequences for low-credibility speakers in an important real-world setting: hiring decisions. All of our study pre-registrations, materials, data, and code (including for the two pilot studies described in the Introduction) can be found on OSF:

https://osf.io/y7bj8/?view_only=1f71e792fafa452687f5e1439cb5e5aa.

Study 1: Disclaimers Help Low-Credibility Speakers

In Study 1, we conducted an initial test of whether credibility disclaimers improve perceptions of low-credibility speakers when they self-promote. We showed participants another person’s performance on a math quiz, which was always poor, in order to create an initial negative impression of the person’s competence (in reality, this quiz performance was created by the research team). We then showed participants an “interesting fact” that this person had shared about themselves, which consisted of a self-promotional statement related to quantitative abilities, and we varied whether this statement was prefaced by a disclaimer acknowledging the person’s poor algebra skills or not. We predicted that when the speaker used a disclaimer, they would be perceived as more self-aware and sincere, as making a more believable statement, and as warmer than when the speaker did not use a disclaimer. Further, we predicted that these perceptions would affect a real behavioral choice toward the target—whether to trust the target’s advice or not.

Participants. Participants were recruited online via Prolific Academic and completed the study in exchange for \$1.40, with an additional \$0.60 bonus paid out as described below. We

used Prolific’s prescreen filters to select only those who indicated having a college degree, as we wanted those who would be best positioned to evaluate the target’s math skills. We pre-registered that we would collect 400 participants after excluding those who failed the attention check, did not finish the survey, and/or failed the comprehension check. Of the 481 who started the study, we ended up with a final sample of 393 participants (47.84% female; 26.97% non-White; $M_{\text{age}} = 39.87$) who fit these criteria.

Procedure. Participants read that we had previously run a study in our research center’s laboratory, and that we would show them some information about a randomly-selected participant from this prior study—we will refer to this person as the “target” here. In reality, we never ran such a study, and instead showed participants materials created by the research team.

Participants read that the target they would be evaluating had been asked to complete a short math quiz, and that we had video recorded the target’s screen while they had taken it (thus showing the target typing out their solutions as well as the multiple-choice answers they ultimately selected). We further told participants that we had modified the video such that we had highlighted which answer was correct for ease of evaluation while they watched, and we noted that the target had of course not seen the correct answers while they took the quiz. We also asked participants to try to solve each problem in their head as they watched, to help keep them engaged throughout the video.

Participants then watched a nearly three-and-a-half-minute video showing the target answering four math questions (e.g., solving equations). See Supplement for the link to this video. At the end of the video, participants saw a screen (that the target had also supposedly seen) telling them the target had answered one out of the four questions correctly, and that the target had scored lower than 80% of others who had taken this same quiz.

On the next page, participants read that we had asked the target to share one interesting fact about themselves, and that we had told the target that we would share both their math quiz performance and their interesting fact with another participant who would evaluate them (i.e., the real participant). We made this clear so that participants would know that the target was aware— at the time they provided the interesting fact—that the (real) participant would have formed an impression of them by the time they read the fact.

Participants were randomly assigned to one of two conditions: no disclaimer or disclaimer. All participants read the same interesting fact, which consisted of a bragging statement: “I recently got a job in a top software engineering position.” We manipulated whether this statement was made by itself (no disclaimer condition) or was prefaced by a disclaimer acknowledging the target’s apparently poor math skills: “I may not be that good at algebra, but [rest of statement]” (disclaimer condition). We chose this brag because it may seem somewhat implausible for someone with poor algebra skills, but also does not seem completely impossible (based on a pretest, $N = 50$).

To help ensure that participants had paid attention to all of the information we had given them, we showed them a reminder of both the final math quiz results page and the interesting fact, below which we asked them to write one sentence describing their overall impressions of the target given the information they had seen.

Next, participants responded to our dependent measures in randomized order (all on a slider scale, $-30 = \textit{extremely [opposite of trait]}$, $30 = \textit{extremely [trait]}$): “Based on what you know about them, how trustworthy do you think this participant is?”; “Based on what you know about them, how honest do you think this participant is?”; “Based on what you know about them, how likable do you think this participant is?”; “Based on what you know about them, how

competent do you think this participant is?"; "Based on what you know about them, how believable do you think this participant's interesting fact is?"; "Based on what you know about them, how self-aware do you think this participant is?"; and "Based on what you know about them, how sincere do you think this participant's interesting fact was?" The last two questions were intended to capture our proposed mechanisms.

On the next page, participants saw instructions for a task that served as our behavioral dependent measure: trusting the target's advice in an incentivized game. We designed this game to capture integrity-based trust in particular, based on the sender-receiver game (Gneezy, 2005), as we expected our manipulation to have the strongest effect on perceptions of integrity (Mayer, Davis, & Schoorman, 1995). Participants read that they would play a short "investment game," in which their choice would affect both their own bonus payment and the bonus payment of the target they had just evaluated. They were told that they would have to choose between two stocks, Stock A or Stock B, but that they would not know the payouts of each stock before selecting one. They read that one of the stocks would pay them \$0.60 and the target \$0.40, while the other would pay them \$0.20 and the target \$0.80. They further read that the target had been told which stock (A or B) provided which set of payouts, and had been given the opportunity to send the participant a suggestion about which stock to choose; the target had specifically been asked to suggest the stock that would give the participant the higher bonus payment. The participant read that they would then have the choice to either pick the suggested stock or to pick the other stock.

On the next page, participants were required to answer three comprehension check questions about the game correctly in order to proceed; they were given three tries to answer these correctly, and if they failed after the third try, the study automatically ended.

Once they correctly answered these questions, they were shown which stock the target had suggested the participant pick. We randomized whether the recommended stock was Stock A or Stock B. We then asked participants: “Which stock would you like to pick?” (*I will take the other participant’s recommendation and choose Stock [A/B] vs. I will NOT take the other participant’s recommendation and instead choose Stock [A/B]*). Regardless of which choice participants made, the next page told participants that they would receive a \$0.60 bonus based on their choice; thus, all participants received the higher bonus payment. Finally, participants reported demographic information (gender, age, race) and were given an optional space to provide feedback. We obtained IRB approval not to debrief participants about the deception in the study, so as not to create undue distress or suspicion in the participant pool.

Results

In this and our subsequent studies, we combined our measures of trustworthiness, honesty, and likability into a composite of speaker warmth ($\alpha = 0.92$ in this study), and analyzed the rest of our measures individually. We conducted independent t-tests between disclaimer conditions for each of our perception measures, and a chi-square test of proportions for our behavioral choice measure. Means and standard deviations for this and all subsequent studies can be found in Table 1.1.

Self-awareness and sincerity. Consistent with hypotheses, participants found the target to be more self-aware, $t(391) = 8.14, p < .001, d = 0.82$, and to be more sincere in their bragging statement, $t(391) = 5.49, p < .001, d = 0.55$, when they used a disclaimer compared to no disclaimer.

Believability. Similarly, participants found the bragging statement more believable when it was prefaced by a disclaimer compared to no disclaimer, $t(391) = 6.17, p < .001, d = 0.62$.

Warmth. As hypothesized, participants found the target warmer overall when they used a disclaimer than when they used no disclaimer, $t(391) = 5.17, p < .001, d = 0.52$.

Competence. Participants also found the target more competent when they used a disclaimer than when they used no disclaimer, $t(391) = 5.94, p < .001, d = 0.60$.

Behavioral choice: Integrity-based trust. Finally, participants were more willing to trust the target’s advice—by selecting the stock that the target had suggested—when the target used a disclaimer (60.31%) than when they used no disclaimer (41.71%), $\chi^2(1, N = 393) = 13.60, p < .001$ —see Figure 1.2.

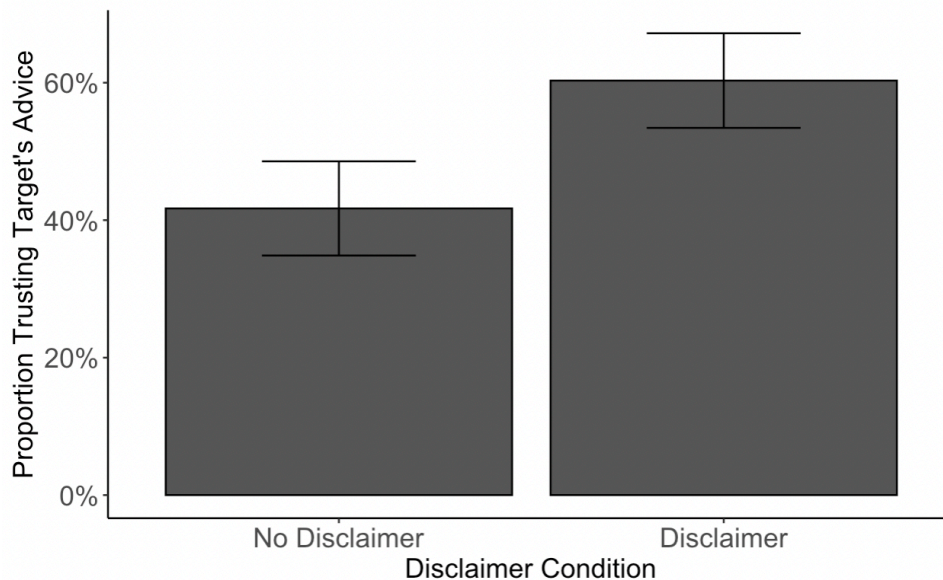


Figure 1.2. Proportion of participants who trusted the target’s advice in Study 1, Chapter 1. Error bars represent 95% confidence intervals.

Mediation (Exploratory). In a non-preregistered analysis, we tested whether self-awareness and sincerity (independently) mediated the effect of disclaimer condition on warmth perceptions and (separately) behavioral choices. We conducted this and all other mediation analyses in R with the “mediation” package, using 5,000 bootstrapped samples (or in some cases

| Study | Variable | Low Credibility | | | High Credibility | | |
|-------|-----------------------------|-------------------------|----------------------|-----|-------------------------|----------------------|------|
| | | No Disclaimer Mean (SD) | Disclaimer Mean (SD) | Sig | No Disclaimer Mean (SD) | Disclaimer Mean (SD) | Sig |
| 1 | Self-awareness | -5.27 (15.47) | 9.27 (17.43) | *** | - | - | |
| | Sincerity | -4.16 (17.64) | 5.58 (17.54) | *** | - | - | |
| | Believability | -9.07 (18.38) | 2.84 (19.82) | *** | - | - | |
| | Warmth | -3.23 (13.91) | 4.41 (15.40) | *** | - | - | |
| | Competence | -6.45 (15.98) | 3.40 (16.91) | *** | - | - | |
| | Advice choice | 41.71% | 60.31% | *** | - | - | |
| 2a | Self-awareness | -7.14 (15.48) | 3.00 (18.74) | *** | 14.46 (13.04) | 10.62 (15.44) | |
| | Warmth | -6.70 (11.54) | -2.09 (13.34) | ** | 11.21 (10.53) | 9.81 (11.02) | |
| | Competence | -8.22 (13.36) | -5.46 (15.77) | | 21.51 (8.90) | 18.90 (11.49) | |
| 2b | Self-awareness | -9.04 (16.07) | 1.92 (18.53) | *** | 13.89 (13.06) | 11.52 (15.60) | |
| | Warmth | -7.79 (13.39) | -0.61 (11.71) | *** | 13.62 (9.06) | 11.29 (11.39) | |
| | Competence | -8.10 (16.31) | -4.87 (15.01) | | 21.27 (8.08) | 16.64 (11.59) | (*) |
| 3a | Self-awareness | -1.36 (16.34) | 5.05 (15.48) | ** | 5.37 (14.53) | -0.61 (16.34) | (**) |
| | Believability | -11.70 (14.62) | -4.46 (15.92) | ** | 6.77 (15.56) | 7.37 (16.65) | |
| | Warmth | -2.20 (12.33) | 3.88 (11.03) | *** | 7.81 (9.16) | 4.45 (11.50) | (*) |
| | Competence | 3.35 (12.78) | 7.98 (11.81) | ** | 8.99 (10.28) | 7.80 (10.26) | |
| 3b | Self-awareness | -6.19 (18.04) | -2.13 (16.45) | | 8.68 (14.20) | 7.94 (15.55) | |
| | Believability | -13.76 (13.74) | -8.53 (14.78) | ** | 8.44 (13.57) | 5.62 (15.49) | |
| | Warmth | -11.36 (10.85) | -6.95 (11.45) | ** | 12.40 (10.51) | 10.46 (11.56) | |
| | Competence | -0.33 (11.62) | 4.01 (10.76) | ** | 12.81 (10.88) | 12.39 (9.86) | |
| 3c | Self-awareness | -14.70 (14.65) | -7.22 (18.28) | *** | 13.41 (12.15) | 8.35 (14.48) | (*) |
| | Believability | -16.71 (14.00) | -14.26 (15.06) | | 17.55 (9.43) | 13.86 (13.85) | (*) |
| | Warmth | -6.81 (13.39) | -3.50 (12.87) | * | 13.44 (7.90) | 10.28 (10.98) | |
| | Competence | -5.61 (13.60) | -2.32 (12.90) | * | 17.80 (8.54) | 15.73 (10.83) | |
| 3d | Self-awareness | -14.77 (13.28) | -0.24 (17.71) | *** | 10.62 (14.21) | 7.85 (15.16) | |
| | Believability | -19.72 (11.09) | -7.97 (15.57) | *** | 13.75 (12.13) | 15.84 (11.19) | |
| | Warmth | -10.97 (11.50) | -1.96 (12.56) | *** | 8.47 (10.37) | 8.45 (10.28) | |
| | Competence | -7.82 (12.89) | 1.01 (13.16) | *** | 14.92 (10.50) | 14.93 (10.09) | |
| 4 | Awareness of how others see | 4.07 (14.44) | 9.76 (11.86) | ** | 12.23 (11.47) | 8.54 (11.93) | (*) |
| | Sincerity | 2.01 (16.53) | 9.32 (14.43) | *** | 13.93 (10.67) | 8.45 (15.78) | (**) |
| | Warmth | 5.45 (11.30) | 9.51 (9.21) | ** | 14.02 (9.43) | 10.95 (10.41) | (*) |
| | Competence | 7.79 (14.92) | 9.76 (11.97) | | 22.57 (7.86) | 17.61 (10.26) | (**) |
| | Task choice (continuous) | -10.46 (20.30) | -10.23 (19.95) | | 5.45 (21.72) | 2.66 (20.15) | |
| | Task choice (binary) | 33.67% | 34.69% | | 59.00% | 58.59% | |
| | Social hour choice | 3.77 (12.96) | 4.46 (13.41) | | 11.80 (12.89) | 8.69 (13.15) | |
| 5a | Hiring recommendation | -1.15 (1.43) | -0.76 (1.49) | ** | - | - | |
| | Self-awareness | 2.79 (15.06) | 11.28 (14.40) | *** | - | - | |
| | Sincerity | 7.15 (14.51) | 11.24 (13.58) | ** | - | - | |
| | Believability | 3.27 (16.01) | 7.68 (14.75) | ** | - | - | |
| | Warmth | 7.16 (11.38) | 10.45 (11.39) | ** | - | - | |
| 5b | Competence | 0.85 (13.68) | 3.12 (14.41) | | - | - | |
| | Hiring recommendation | -1.14 (1.42) | -0.92 (1.45) | | - | - | |

Note. "Sig" refers to the significance of a pairwise contrast (or t-test) between the no disclaimer and disclaimer conditions, within each credibility condition. For the high-credibility condition, parentheticals around the asterisks indicate that significance is in the *opposite* direction from the low-credibility condition (there were no cases in which the high-credibility condition was significant and in the same direction as the low-credibility condition). Significance of overall interactions is not shown. * $p < .05$, ** $p < .01$, *** $p < .001$. Control condition in Studies 2a-b is not shown.

Table 1.1. Condition means and standard deviations for all studies in Chapter 1.

10,000 bootstrapped samples when testing moderation specifically). Using separate models for each mediator, we observed that self-awareness and sincerity each mediated the effect on warmth perceptions (indirect effect of self-awareness: $b = 8.28$, 95% CI = [6.10, 10.73], $p < .001$; sincerity: $b = 6.94$, 95% CI = [4.35, 9.45], $p < .001$) and behavioral choices (self-awareness: $b = 0.14$, 95% CI = [0.09, 0.19], $p < .001$; sincerity: $b = 0.10$, 95% CI = [0.06, 0.14], $p < .001$), respectively.

Discussion

Study 1 provided initial support for our hypothesis that low-credibility speakers are perceived to be *warmer* when they acknowledge a prior impression of a perceived shortcoming before a bragging statement than when they do not. People were also more likely to trust the speaker's advice when they bragged with a disclaimer, suggesting that these impressions have consequences for real behavior. Finally, in an exploratory analysis, we found support for our hypothesis that self-awareness and sincerity mediate the relationship between disclaimers and perceived warmth. We even found that disclaimers boosted perceptions of competence as well in this study.

If credibility disclaimers increase warmth perceptions because of our proposed mechanisms of self-awareness and sincerity, then we would expect disclaimers to *only* increase perceptions of low-credibility speakers, and not high-credibility speakers. If, however, an alternative explanation accounts for our effects—such as people simply liking self-deprecation, modesty, or statements that they perceive as an attempt at humor—then our findings should generalize across all types of speakers, regardless of credibility. Thus, in our next study, we tested whether the effect of disclaimers is moderated by the speaker's credibility.

Studies 2a-b: Disclaimers Do Not Help High-Credibility Speakers

In Studies 2a-b, we tested whether the effect of credibility disclaimers would be moderated by the speaker's credibility, thus suggesting that our effects are not simply due to an appreciation for self-deprecatory statements, modesty, or humor. Specifically, we hypothesized that—as before—low-credibility speakers would be perceived as warmer when they bragged with a disclaimer than without one, but that such disclaimers would *not* help high-credibility speakers.

For this study, we stimulus sampled different types of disclaimers in order to test generalizability across different ways of acknowledging a perceived lack of credibility. In Study 2a, we tested disclaimers that acknowledge that the listener may find the forthcoming statement to be contradictory to their prior impression of the speaker (e.g., “This may sound strange to you, but...”). In Study 2b, we tested disclaimers that specifically refer to the speaker's apparent shortcoming (e.g., “I'm not that smart, but...”; similar to that in Study 1).

Participants. Participants were recruited online via Amazon Mechanical Turk and completed the study in exchange for \$0.10, with an additional \$0.15 bonus if they correctly answered the comprehension questions. We pre-registered that we would collect 600 participants in each study after excluding those who failed the attention check, did not finish the survey, and/or failed the comprehension check. Of the 752 who started the survey in Study 2a and 734 who started the survey in Study 2b, we ended up with the following samples that fit this criteria for each study: in Study 2a, we ended up with a final sample of 600 participants (51.33% female; 25.17% non-White; $M_{\text{age}} = 37.04$); in Study 2b, we ended up with a final sample of 599 participants (54.76% female; 23.71% non-White; $M_{\text{age}} = 38.21$).

Procedure. Participants read a scenario that told them to imagine they worked at a mid-size corporation and were at work, having a conversation with one of their colleagues in the

office break room (exact wording for this scenario is provided in Appendix A). Each participant was randomly assigned to one of six conditions in a 2 (Disclaimer condition: no disclaimer vs. disclaimer) x 3 (Credibility condition: low vs. high vs. control) between-subjects design.

In the high-credibility conditions, participants read: “You know that this colleague has a high-ranking position at the company, is very well-respected, and is very competent at their job.” In the low-credibility conditions, we changed this to “low-ranking position,” “not very well-respected,” and “not very competent.” In the control conditions, participants were told that they did *not* know the colleague’s ranking at the company, how well-respected they were, or how competent they were at their job. Thus, the control condition allowed us to test the effect of disclaimers when the listener’s prior impression of the speaker is relatively neutral (rather than positive), but is still not directly contradictory to the information conveyed in the brag; we therefore expected this condition to follow a similar pattern as the high-credibility condition.

Next, participants were told to imagine that the colleague said something to them during the conversation. Participants then read the colleague’s bragging statement, either without a disclaimer (no disclaimer conditions) or with a disclaimer (disclaimer conditions). For stimulus sampling, participants saw one of two brags (either about having won an award at their previous job or about having graduated from college with honors). These brags were taken from a pilot test ($N = 50$) in which we asked participants to report something they wanted to share with their colleagues, but were worried it would sound like a brag.

In addition, we randomly assigned participants in the disclaimer conditions to view one of two phrasings for each type of disclaimer (within each study), to ensure that our effects were not due to any one specific phrasing. For Study 2a, our disclaimers were: “I know this may seem hard to believe, but...” and “This may sound strange to you, but...”. For Study 2b, our

disclaimers were: “I’m not that smart, but...” and “I’m no genius, but...”. Again, we expected all of these credibility disclaimers to yield similar effects.

Participants then responded to nearly the same perception measures as in Study 1 (in randomized order), except for three changes. First, we did not include the believability and sincerity measures. Second, we slightly modified the wording of the questions by taking out “Based on what you know about them...” at the beginning of each question and by referring to “this person” rather than “this participant.” Third, we included two additional questions: “How powerful do you think this person is?” and “How high status do you think this person is?” (again anchored at -30 = *extremely [opposite of trait]*, 30 = *extremely [trait]*).

Results

For each study version, we conducted two-way ANOVAs with disclaimer condition, credibility condition, and their interaction as factors on each of our dependent measures. We collapsed across brags and across disclaimer phrasings within each study version. As in Study 1, we combined our measures of trustworthiness, honesty, and likability into a warmth composite (Study 2a: $\alpha = 0.88$; Study 2b: $\alpha = 0.90$).² We examined competence on its own, rather than combining it with power and status to create a composite (as is sometimes done), because the three did not consistently load together across the studies where we measured them, and we wanted to maintain consistency in how we reported competence across studies. For conciseness, we report results on the power and status items for this study and subsequent studies in the Supplement: Across the studies where we measured power and status, we observed a null difference between disclaimer conditions within the low-credibility condition, which was our

² For Studies 2a-b only, we did not pre-register to combine these items into a composite, but analyzing each measure individually yields the same results—see Supplement.

main comparison of interest. That is, disclaimers neither help nor hurt low-credibility speakers' power and status.

We have summarized all main effects in Appendix A (Table A.1.1), and report the results of the interactions in-text below, as this was our primary result of interest. We report means and standard deviations for the low- and high-credibility conditions in Table 1.1.

Self-awareness. In both Studies 2a and 2b, there was a significant interaction between disclaimer condition and credibility condition (Study 2a: $F(2,594) = 11.87, p < .001, \eta_p^2 = .04$; Study 2b: $F(2,593) = 11.51, p < .001, \eta_p^2 = .04$). Low-credibility speakers were perceived to be more self-aware when they used a disclaimer than when they did not (p 's $< .001$), but there was no difference ($p = .296$ in Study 2b) or a marginal difference in the *opposite* direction ($p = .097$ in Study 2a) for high-credibility speakers, with no significant difference for speakers in the control condition (p 's $\geq .184$ for both studies).

Warmth composite. All condition means for the warmth composite are visualized in Figure 1.3. As hypothesized, in both studies, there was a significant interaction between disclaimer condition and credibility condition (Study 2a: $F(2,594) = 6.69, p = .001, \eta_p^2 = .02$; Study 2b: $F(2,593) = 11.79, p < .001, \eta_p^2 = .04$). Low-credibility speakers were perceived as warmer when they used a disclaimer than when they did not (p 's $< .004$), but there was no difference for high-credibility speakers (p 's $> .162$), and in the control condition, speakers were perceived as marginally or significantly *less* warm with a disclaimer than without one (p 's $< .068$).

Competence. Our results on perceptions of speaker competence followed the same pattern as those of speaker warmth, albeit with weaker effects. In both studies, there was a significant interaction between disclaimer condition and credibility condition (Study 2a: $F(2,594) = 4.49, p$

= .012, $\eta_p^2 = .01$; Study 2b: $F(2,593) = 5.86, p = .003, \eta_p^2 = .02$). Low-credibility speakers were perceived as directionally ($p = .118$ in Study 2a) or marginally ($p = .081$ in Study 2b) more competent when the speaker used a disclaimer than when they did not, but in the high-credibility condition, speakers were directionally ($p = .147$ in Study 2a) or significantly ($p = .013$ in Study 2b) perceived as *less* competent with a disclaimer than without one, and speakers were also perceived as significantly *less* competent with a disclaimer than without one in the control condition (p 's < .017).

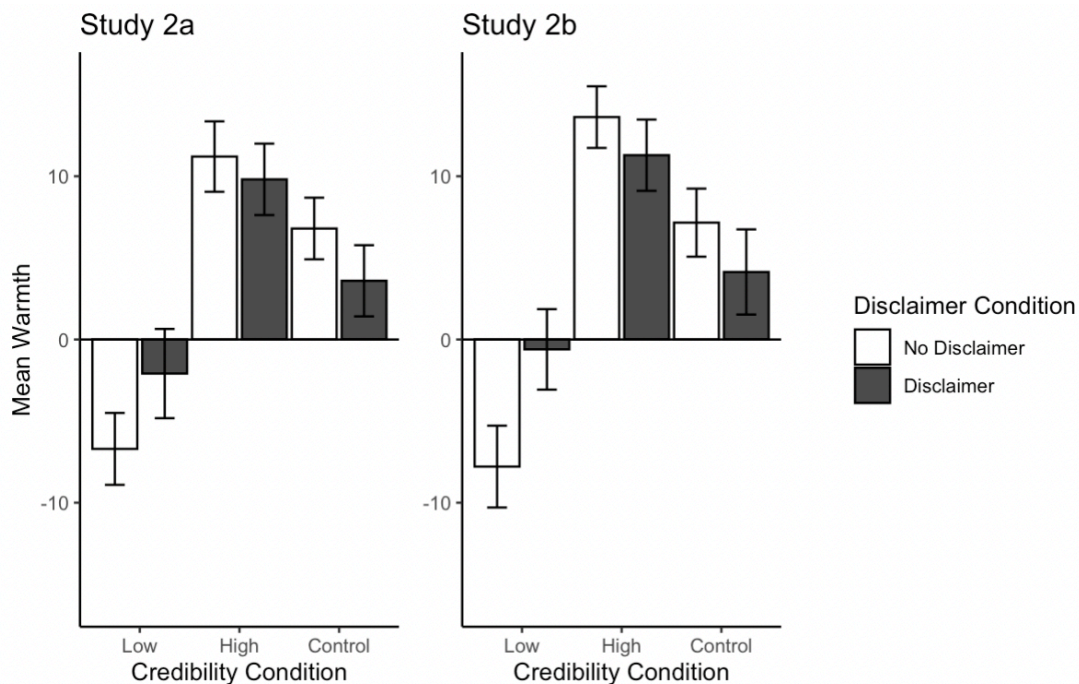


Figure 1.3. Mean warmth perceptions in Studies 2a-b, Chapter 1. Error bars represent 95% confidence intervals.

Moderated Mediation (Exploratory). We again conducted a non-preregistered analysis in which we tested whether self-awareness mediated the effect of disclaimer condition on warmth perceptions, and whether this was moderated by credibility condition. For simplicity and ease of interpretation, we limited this mediation to comparing the high and low credibility conditions

(without the control condition). Consistent with our framework, we observed significant moderated mediation in both studies (Study 2a indirect effect: $b = 5.04$, 95% CI = [2.65, 7.53], $p < .001$; Study 2b indirect effect: $b = 5.88$, 95% CI = [3.08, 8.82], $p < .001$). Specifically, when focusing on the low-credibility condition, we see that self-awareness *positively* mediated the effect of disclaimer condition on warmth (Study 2a indirect effect: $b = 4.02$, 95% CI = [1.96, 6.35], $p < .001$; Study 2b indirect effect: $b = 5.16$, 95% CI = [2.80, 7.92], $p < .001$). In contrast, when focusing on the high-credibility condition, we see that self-awareness marginally or non-significantly *negatively* mediated the effect (Study 2a indirect effect: $b = -1.18$, 95% CI = [-2.66, 0.07], $p = .064$; Study 2b indirect effect: $b = -0.94$, 95% CI = [-2.62, 0.59], $p = .232$).

Discussion

The results of Studies 2a-b showed that disclaimers help low-credibility, but not high-credibility, speakers—suggesting that our effects are driven by the disclaimer’s signal of self-awareness and sincerity rather than a more general effect of self-deprecation, modesty, or humor. Our results on perceived competence were weaker than those of warmth in this study, but we still do not find any evidence that the disclaimer *harms* perceptions of low-credibility speakers’ competence, thus creating no cost to using the disclaimer for low-credibility speakers (likewise, as mentioned, we find no cost to perceptions of low-credibility speakers’ power and status: see Supplement).

One limitation of this study is that it manipulated the power, status, and competence of the speaker in tandem (as a means of manipulating speaker credibility). If indeed our effects are driven by speaker credibility, then we would expect the results to be driven primarily by the trait most directly relevant to the speaker’s brag (in this case, competence), as that would have the strongest influence on perceptions of the speaker’s credibility. We find support for this

proposition in Supplemental Study B, where we manipulate speaker power and competence orthogonally. In our subsequent studies, we manipulate only the trait most relevant to the brag.

Studies 3a-d: Generalizing to Multiple Traits

To show that our findings apply to any speaker perceived as low-*credibility* on a trait—not just those perceived as low in *competence* specifically—Studies 3a-d each tested our effects for a different trait (Study 3a: sociability, 3b: generosity, 3c: musical talent, and 3d: athletic skill). Our design was very similar to that of Studies 2a-b: For each trait, we first manipulated speaker credibility by varying the description of the speaker’s abilities with regard to that particular trait (as either high or low on that trait). Next, the speaker bragged about something related to that particular trait, and we manipulated whether the speaker used a disclaimer or not when making this self-promotional claim. As in our previous studies, we hypothesized that the disclaimer would improve warmth perceptions of low-credibility speakers, but not high-credibility speakers.

In these studies, we also included a condition that tested other types of disclaimers that did *not* specifically address the brag’s credibility. We report the design and results for this condition in the Supplement, as they are incidental to our main hypotheses for this paper. (In short, including the additional condition does not alter any of the conclusions that can be drawn from these results: Consistent with our theory, disclaimers that do *not* address a brag’s credibility—i.e., “non-credibility disclaimers”—and thus do not signal self-awareness and sincerity, have a similar effect to not using a disclaimer at all.) In the text below, we report only the sample, methods, and results from the conditions that tested our primary hypotheses.

Participants. Participants were recruited online via Prolific Academic and completed the study in exchange for \$0.50, with an additional \$0.45 bonus if they correctly answered the

comprehension questions. We pre-registered that we would collect 100 participants per condition (totaling 400 for the conditions reported here) in each study after excluding those who failed the attention check, did not finish the survey, and/or failed the comprehension check. We ended up with the following samples that fit this criteria in the focal conditions for each study: In Study 3a (sociability), 418 started the survey and we ended up with a final sample of 404 participants (48.27% female; 34.65% non-White; $M_{\text{age}} = 31.00$); in Study 3b (generosity), 438 started the survey and we ended up with 401 participants (50.87% female; 28.43% non-White; $M_{\text{age}} = 32.92$); in Study 3c (musical talent), 423 started the survey and we ended up with 400 participants (47.75% female; 27.75% non-White; $M_{\text{age}} = 34.37$); and in Study 3d (athletic skill), 422 started the survey and we ended up with 396 participants (47.47% female; 36.11% non-White; $M_{\text{age}} = 31.52$).

Procedure. We used the exact same scenario as in Studies 2a-b that described having a conversation with a colleague. Participants were randomly assigned to one of four conditions in a 2 (Disclaimer condition: no disclaimer vs. disclaimer) x 2 (Credibility condition: low vs. high) between-subjects design. In Table 1.2 below, we outline our key manipulations and stimuli in each study version. Exact wording for these stimuli can be found in Appendix A.

In all study versions, the brag was either made by itself (no disclaimer conditions) or with a disclaimer (disclaimer conditions). We used the same disclaimers as in Studies 2a-b with some slight wording adaptations as needed for the particular context (see Appendix A). For instance, two of our disclaimers in Study 3a (sociability) were “I’m no social butterfly, but...” and “I’m not that outgoing, but...” We collapsed across disclaimer phrasings, and across the brags in each study version, in our analyses.

| Study | Low-Credibility Condition | High-Credibility Condition | Brag Version 1 | Brag Version 2 |
|---------------------|---|---|---|--|
| Sociability (3a) | Colleague is shy and socially awkward; does not tend to make friends easily or go to many social gatherings | Colleague is outgoing and charming; tends to make friends easily and go to many social gatherings | Being voted most popular in high school | Being invited to seven different parties that weekend |
| Generosity (3b) | Colleague is selfish and inconsiderate | Colleague is generous and considerate | Volunteering | Donating to charity |
| Musical Talent (3c) | Colleague is not musically talented at playing the guitar | Colleague is known to be musically talented at playing the guitar | Winning an award for their guitar playing | Being the lead guitar player in a band that signed with a major record label |
| Athletic Skill (3d) | Colleague is not athletically skilled and is uncoordinated | Colleague is athletically skilled and well-coordinated | Having a fast mile running time | Being a recruited athlete in college |

Table 1.2. Summary of manipulations and brags in Studies 3a-d, Chapter 1.

Participants then responded to the exact same set of questions as in Studies 2a-b, plus a question about the brag’s believability (similar to in Study 1) and several other exploratory measures that we report in the Supplement.³ As in the previous studies, participants answered two comprehension check questions (which determined whether they received the bonus payment for answering them correctly), and reported demographic information.

Results

For each study version, we conducted a two-way ANOVA with disclaimer condition, credibility condition, and their interaction as factors on each of our dependent measures. We

³ Once again, power and status were pre-registered as main dependent measures rather than exploratory, but we report them in the Supplement for the same reasons described earlier. We also included a measure of arrogance, but because arrogance does not test the primary hypotheses of this paper, we report it in the Supplement as well.

conducted our analyses on the same composite of speaker warmth as in the previous studies (α 's > 0.81) and the rest of the measures individually. We report all main effects in Appendix A (Table A.1.2) and interactions in-text below. Means and standard deviations for each condition can be found in Table 1.1.

Self-awareness. Disclaimers improved perceptions of the speaker's self-awareness for low-credibility speakers, but not high-credibility speakers. For three of the four study versions, there were significant interactions between disclaimer condition and credibility condition (sociability: $F(1,400) = 15.56, p < .001, \eta_p^2 = .04$, musical talent: $F(1,396) = 17.36, p < .001, \eta_p^2 = .04$, athletic skill: $F(1,392) = 32.06, p < .001, \eta_p^2 = .08$). In one case, the interaction was non-significant (generosity: $F(1,397) = 2.20, p = .139, \eta_p^2 < .01$). In all cases, low-credibility speakers seemed significantly or marginally more self-aware when using a disclaimer than without one (p 's $< .069$). However, for high-credibility speakers, there was either no significant difference between disclaimer conditions (p 's $> .199$ for generosity and athletic skill), or self-awareness was *lower* with a disclaimer than without one (p 's $< .004$ for sociability and musical talent).

Believability. As expected, disclaimers improved the believability of the brag for low-credibility speakers, but not for high-credibility speakers. In all study versions, there were significant interactions between disclaimer condition and credibility condition (sociability: $F(1,400) = 4.49, p = .035, \eta_p^2 = .01$, generosity: $F(1,397) = 7.78, p = .006, \eta_p^2 = .02$, musical talent: $F(1,396) = 5.33, p = .021, \eta_p^2 = .01$, athletic skill: $F(1,392) = 14.42, p < .001, \eta_p^2 = .04$). In three of the four studies, low-credibility speakers' brags were significantly more believable with a disclaimer than without one (p 's $< .009$). In one study, there was no difference (musical talent: $p = .195$). However, for high-credibility speakers, there was either no significant

difference between disclaimer conditions (p 's $> .180$ for sociability, generosity, and athletic skill) or believability was *lower* with a disclaimer than without one ($p = .050$ for musical talent).

Warmth composite. All condition means for warmth are visualized in Figure 1.4. Once again, disclaimers improved perceptions of the speaker's warmth for low-credibility speakers, but not high-credibility speakers. In all study versions, there were significant interactions between disclaimer condition and credibility condition (sociability: $F(1,400) = 18.07, p < .001, \eta_p^2 = .04$, generosity: $F(1,397) = 8.14, p = .005, \eta_p^2 = .02$, musical talent: $F(1,396) = 7.89, p = .005, \eta_p^2 = .02$, athletic skill: $F(1,392) = 16.04, p < .001, \eta_p^2 = .04$). In all cases, low-credibility speakers seemed warmer when using a disclaimer than without one (p 's $< .043$). However, for high-credibility speakers, there was either no significant difference between disclaimer conditions (p 's $> .053$ for generosity, musical talent, and athletic skill), or warmth was *lower* with a disclaimer than without one ($p = .034$ for sociability).

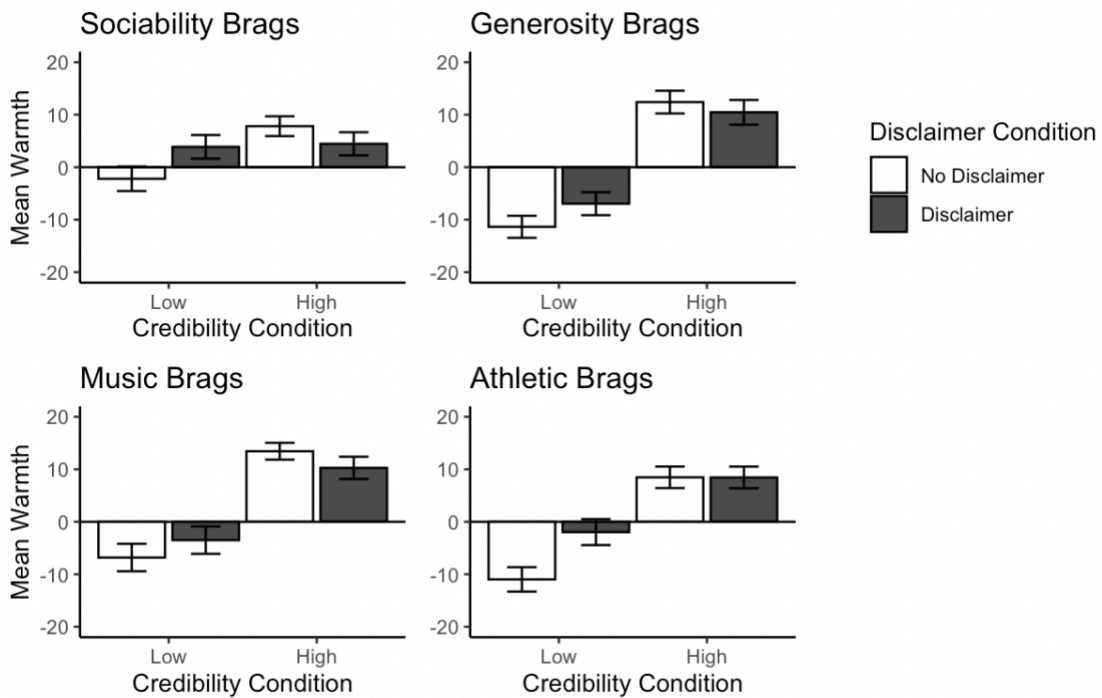


Figure 1.4. Mean warmth perceptions in Studies 3a-d, Chapter 1. Error bars represent 95% confidence intervals.

Competence. Similarly, disclaimers improved perceptions of the speaker's competence for low-credibility speakers, but not high-credibility speakers. In all study versions, there were significant interactions between disclaimer condition and credibility condition (sociability: $F(1,400) = 6.59, p = .011, \eta_p^2 = .02$, generosity: $F(1,397) = 4.84, p = .028, \eta_p^2 = .01$, musical talent: $F(1,396) = 5.27, p = .022, \eta_p^2 = .01$, athletic skill: $F(1,392) = 13.98, p < .001, \eta_p^2 = .03$). In all cases, low-credibility speakers seemed more competent when using a disclaimer than without one (p 's $< .047$). However, for high-credibility speakers, there was no significant difference between disclaimer conditions (p 's $> .210$).

Moderated Mediation (Exploratory). Again, we conducted a non-preregistered analysis in which we tested whether self-awareness mediated the effect of disclaimer condition on warmth perceptions, and whether this was moderated by credibility condition. We observed significant moderated mediation in three studies (Study 3a indirect effect: $b = 4.46, 95\% \text{ CI} = [2.22, 6.86], p < .001$; Study 3c indirect effect: $b = 5.03, 95\% \text{ CI} = [2.60, 7.55], p < .001$; Study 3d indirect effect: $b = 7.22, 95\% \text{ CI} = [4.51, 10.02], p < .001$). In Study 3b, the indirect effect was non-significant ($b = 1.71, 95\% \text{ CI} = [-0.54, 3.99], p = .138$). In the three studies in which the effect was significant, self-awareness *positively* (in one case, marginally) mediated the effect of disclaimer condition in the low-credibility condition (p 's $< .070$), but directionally or significantly *negatively* mediated in the high-credibility conditions ($p < .001$ for Study 3a; p 's $> .130$ for Studies 3c-d).

Discussion

Study 3 provides additional support for our hypothesized framework by demonstrating generality to a variety of traits besides workplace-related competence. In each of these cases, speakers who are perceived to have little ability on a given dimension (whether warmth-related

traits or particular types of skills) are not perceived as very credible when bragging about their accomplishments on that trait. As such, acknowledging the perceived lack of credibility on that particular dimension with a disclaimer improves the speaker's perceived warmth (relative to bragging without a disclaimer). Yet for high-credibility speakers, the disclaimer has no beneficial effect, and even sometimes backfires.

Study 4: Credibility Disclaimers in Realistic Interactions

Study 4 sought to replicate our findings with richer stimuli—including spoken, rather than written, disclaimers—and in a non-hypothetical setting. While Study 1 also used richer stimuli and measured real behavior, it focused exclusively on low-credibility speakers. Studies 2 and 3 used simple scenarios to compare the effect of disclaimers for low and high-credibility speakers. Thus, Study 4 tests the predicted interaction on warmth-perceptions from Studies 2 and 3 with more realistic stimuli. To do so, we showed participants videos of a target, in which we manipulated whether the target spoke about an accomplishment with or without a disclaimer. To create greater realism, we conducted this experiment over Zoom (a video-calling platform) and led participants to believe that they were evaluating another participant in the study; in reality, they simply watched pre-recorded videos of the other alleged participant.

Participants. Participants were recruited online via the virtual laboratory participant pool of a large Midwestern University, and completed the study in exchange for a \$12 Amazon gift card.⁴ The study was advertised as “Pair Task Study” in order to make it convincing that the study would involve interaction with another participant; in reality, we only scheduled one participant for each study session. We included three comprehension check questions in the

⁴ We originally started paying participants a \$10 gift card for the study, but due to relatively slow recruitment, we increased the payment to a \$12 gift card after data collection had begun. We observed no differences on our main dependent measures between payment amounts, p 's > .239.

middle of the study; participants were given three tries to answer these questions correctly, and if they failed on the third try, they were allowed to proceed with the study, but we pre-registered to exclude their data. We pre-registered to collect 400 participants after excluding those who did not finish the survey, failed the comprehension checks, encountered procedural errors (as noted by the research assistant, e.g., technical difficulties that interfered with important aspects of the study procedure), or requested that their data be withdrawn after they were fully debriefed about the study's deception. We excluded two participants based on procedural errors noted by the research assistant: one who saw the wrong manipulation video for their condition, and one who was unable to hear the audio in our manipulation video. We ended up with a final sample of 395 participants (75.19% female; 48.61% non-White; $M_{age} = 29.67$).⁵

Procedure. We randomly assigned participants to one of four conditions in a 2 (Disclaimer condition: disclaimer vs. no disclaimer) x 2 (Credibility condition: low vs. high) between-subjects design. Participants joined the study session on Zoom, at which point they were greeted by the research assistant. Unbeknownst to the participant, the research assistant had used a second device to join the Zoom call under the name “Jamie Thompson” (before letting the real participant enter the Zoom call), in order to make it look like there was another participant on the call. The research assistant kept the account's video off and sound muted for the entire duration of the study. The research assistant began by sending the (real) participant the link to the study survey in a private Zoom chat message, at which point the participant filled out the

⁵ Our data file includes a large number of individuals who mistakenly got access to the survey link and thus started the survey, but did not get very far. They are excluded from all counts and analyses here. The data file also includes 8 pilot participants (with ID numbers starting at 1001) whom we ran to test our study procedures prior to pre-registering our study (with no intention of including their data in our analyses). Again, they are excluded from all counts and analyses here.

consent form and then completed the first part of the study silently on their own (while remaining on the Zoom call).

In the first part of the study, the participant read that they and the other participant would each be randomly assigned a role—either “writer” or “judge”—that would determine their first task. They read that the writer would spend the first 15 minutes of the study writing an essay, during which time the judge would work on a separate task. Then, they read that the judge would read the writer’s essay and answer some questions about it, during which time the writer would work on a separate task.

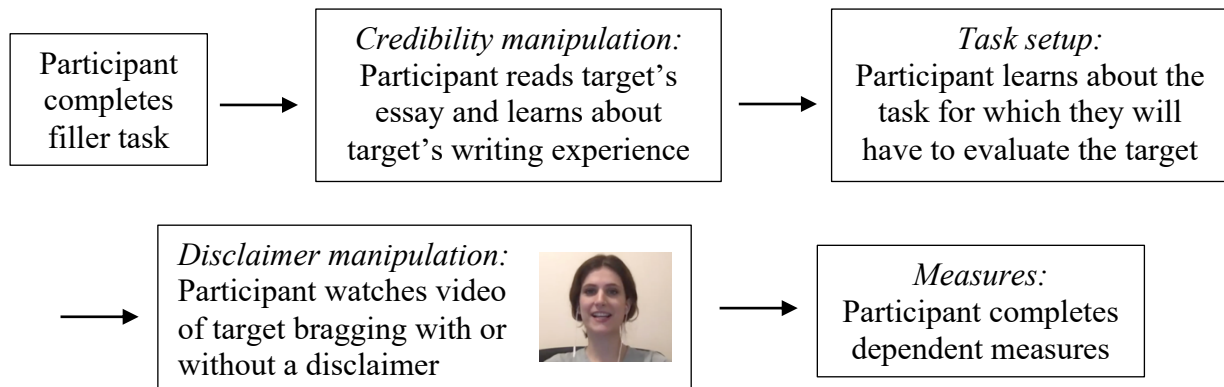


Figure 1.5. Overview of study procedure in Study 4, Chapter 1.

Credibility manipulation. All participants then found out their role: In reality, all participants were assigned to the role of judge. While (allegedly) waiting for the other participant to write their essay, the participant completed a filler task (a number logic task). Once they finished this task, the research assistant created a separate “breakout room” on Zoom and put the fake participant into it, in order to increase believability for some of the next steps of the study and for greater ease of interaction with the (real) participant. The research assistant then shared his or her screen over Zoom in order to show the google document with the essay that was supposedly written by the other participant (a movie review). Participants in the high-credibility

condition saw a high-quality essay, while participants in the low-credibility condition saw a low-quality essay. In reality, the essays were written in advance by the researchers, and we confirmed the difference in essay quality with a pre-test ($N = 58, p's < .001$). See Appendix A for the exact essays. While viewing the essay, participants answered three manipulation check questions (e.g., “Overall, how would you rate the quality of this essay?”) on a scale from 1 = *extremely low quality* to 7 = *extremely high quality*.

Next, the research assistant explained that the other participant had also filled out some additional survey questions, and had given permission for the participant to view these survey responses. The research assistant then shared these survey responses on his or her screen, which served to further strengthen our credibility manipulation. Specifically, the survey questions asked about Jamie’s past experiences and training in writing. In the low-credibility conditions, Jamie’s responses indicated that (s)he had had little past writing experience and no formal instruction in writing. In the high-credibility conditions, Jamie’s responses indicated that (s)he was studying English in college and had additional experiences and instruction in writing. In both conditions, participants also saw that Jamie had answered “yes” to a question asking whether (s)he gave permission for his/her responses to be shared with the other participant in the study. This allowed us to make it clear to participants that Jamie would be aware of the participant’s prior impression of him/her (before they heard his/her bragging statement). See Appendix A for exact stimuli. While viewing these responses, the participant responded to additional manipulation checks (e.g., “Based on all of the information you have seen, how good of a writer do you think Jamie is in general?”; 1 = *extremely poor*, 7 = *extremely good*).

Task setup. Next, the research assistant put the participant into their own “breakout room” on Zoom and told them that they could proceed with the rest of the survey on their own.

The next part of the survey explained that both participants would have to complete another writing task. This task helped us create a context within which we could ask participants to evaluate the other participant, and within which it would make sense for the other participant to brag.

Participants learned that they would again each be randomly assigned to a role—either “first writer” or “second writer.” In reality, all participants learned that they were assigned to the role of “first writer.” Participants then read several pages of detailed instructions explaining the writing task and the decision they would have to make. We designed the task to roughly correspond to an enhanced trust game, in which judgments of both trustworthiness and competence would influence participants’ decisions. See the Supplement and our OSF page for the full set of task instructions.

Disclaimer manipulation. Next, participants were told that the other participant had recorded a video introducing themselves, and that this other participant had known the (real) participant would be making a choice about them regarding a writing task after watching the video. The research assistant sent participants the link to the recording of the “other participant’s” video. For stimulus sampling, we recorded two versions of each video using one male and one female confederate, and showed each participant the video corresponding to the gender of Jamie they had been randomly assigned to. In all conditions, Jamie started by saying “Hi, my name is Jamie, and I’m a third-year student in the college. I hope you pick me for the writing task.” In the no disclaimer condition, Jamie then said: “I won a college essay contest last year.” In the disclaimer condition, Jamie then said: “I’m not that good at writing, but I won a college essay contest last year.” Links to the videos are available in the Supplement. We asked participants to type the last sentence they heard in the video into the survey, to help ensure that

they were paying attention and would actually hear the brag (and disclaimer). We also asked participants to message the experimenter if they experienced any technical difficulties viewing the video.

Exploratory task choice measures. Next, we asked participants a series of questions to help them make their choice for the writing task, and then asked them to report their ultimate choice for the task. The task told participants that they would have to write an essay, and asked them whether they would want to submit this essay independently from the other participant or jointly with the other participant. In the latter case, the participant would have the potential to earn a higher bonus but would also be vulnerable to the other participant “stealing” their essay and taking the entire bonus payment. We pre-registered the results on all of these task choice measures as exploratory. Our first measure was continuous to capture the strength of their preference: “What are you thinking about what choice you’d like to make for the writing task?” (-30 = *strongly prefer to submit independently*, 30 = *strongly prefer to submit jointly*). Our second measure was binary to capture their actual choice: “Make your final (binding) choice below” (*I choose to submit independently* vs. *I choose to submit jointly*).

Exploratory warmth behavioral intentions measure. We also included a measure to capture participants’ behavioral intentions toward the other participant using a simpler choice that relies solely on perceptions of warmth (rather than both warmth and competence). We pre-registered this measure as exploratory. Specifically, we asked participants to re-watch the other participant’s video (in order to ensure that it was fresh in their minds), and then told them that our research center was considering piloting a “social hour” initiative to allow participants to socialize with each other, in which they might be paired with another participant to chat one-on-one. Participants then responded to the following question: “If you were to participate in such a

‘social hour,’ how much would you want to be paired up with the other participant in this study (as opposed to someone else) in order to have the chance to socialize with them?” (-30 = *strongly prefer NOT to be paired with this participant*, 30 = *strongly prefer to be PAIRED with this participant*). We informed participants that the other participant would *not* know their answer to this question.

Perception measures. Next, participants responded to a similar set of measures as in the prior studies that asked about participants’ perceptions of Jamie’s trustworthiness, honesty, likability, competence, self-awareness, and the sincerity of Jamie’s statement in the video. We did not include the measure of brag believability in this study. We also included another measure in this study (-30 = *extremely unaware*, 30 = *extremely aware*): “Based on what you know about them, to what extent do you think the other participant is aware of how others see them?” We included this question in order to assess more precisely what we refer to as “self-awareness.”

Then, participants were asked to actually write their essay for the writing task. This essay doubled as a suspicion check: We asked participants to write a short essay on what they thought the purpose of the experiment was. Finally, participants answered demographic questions, read a short debrief (that did not reveal the true nature of the study), and filled out a separate form to receive their payment.

After data collection finished, we emailed all participants and informed them of the deception in the study, giving them the option to retract their data if they chose to do so once they learned the true nature of the study. Seven participants requested that their data be withdrawn and have been excluded from our dataset. We also gave all participants the highest possible bonus payment based on their choice in the study (\$1.70 if they chose to submit independently and \$1.85 if they chose to submit jointly).

Results

As indicated in our preregistration, we ran our main analyses with and without participants who indicated suspicion about the nature of the study in their essay (e.g., suspecting the other participant was not a real participant; 18.23% of participants), and since results remain meaningfully unchanged with and without them, we report analyses on the full sample.

We conducted a two-way ANOVA with disclaimer condition, credibility condition, and their interaction as factors on each of our dependent measures. We conducted our analyses on the same composite of speaker warmth as in the previous studies ($\alpha = 0.82$) and the rest of the measures individually. We report all main effects in Appendix A (Table A.1.3) and interactions in-text below. We report means and standard deviations for our key measures in Table 1.1.

Manipulation checks. Our manipulations of the target's credibility worked as intended: Participants perceived Jamie's essay as higher quality in content, $t(386.78) = 26.51, p < .001$, technical features, $t(388.57) = 29.58, p < .001$, and overall, $t(389.82) = 27.28, p < .001$, in the high credibility condition compared to the low credibility condition; they also perceived Jamie as a better writer overall, $t(344.77) = 24.90, p < .001$, with high confidence (above the midpoint of the scale), $t(394) = 23.12, p < .001$.

Awareness of how others see him/her. On our measure of how aware Jamie seemed of what others think of him/her (intended to more precisely capture what we mean by self-awareness), we observed the predicted Disclaimer condition x Credibility condition interaction, $F(1,391) = 13.95, p < .001, \eta_p^2 = .03$: Using a disclaimer made the low-credibility speaker seem more aware of how others view them ($p = .002$), but made the high-credibility speaker seem *less* aware of how others view them ($p = .037$). On our other measure of self-awareness, we found this exact same pattern of results (see Supplement).

Sincerity. On our measure of sincerity, we also observed the predicted Disclaimer condition x Credibility condition interaction, $F(1,391) = 19.15, p < .001, \eta_p^2 = .05$: Using a disclaimer made the low-credibility speaker seem more sincere ($p < .001$), but made the high-credibility speaker seem less sincere ($p = .008$).

Warmth composite. Similarly, on our composite of speaker warmth, we observed the predicted Disclaimer condition x Credibility condition interaction, $F(1,391) = 12.29, p < .001, \eta_p^2 = .03$. For the low-credibility speaker, using a disclaimer made the speaker seem warmer ($p = .005$), but for the high-credibility speaker, using a disclaimer made them seem less warm ($p = .033$; see Figure 1.6).

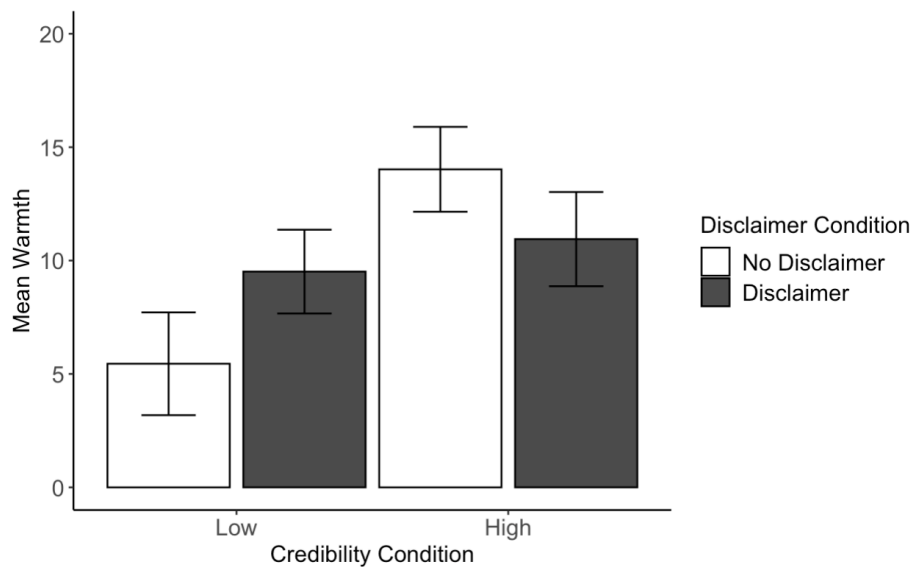


Figure 1.6. Mean warmth perceptions in Study 4, Chapter 1. Error bars represent 95% confidence intervals.

Competence. We pre-registered our measure of competence as exploratory given that we observed weaker results on this measure in our previous studies, but we nevertheless observed a Disclaimer condition x Credibility condition interaction, $F(1,391) = 8.94, p = .003, \eta_p^2 = .02$,

such that disclaimers did not significantly affect perceived competence for low-credibility speakers ($p = .232$), but lowered perceptions of competence for high-credibility speakers ($p = .003$).

Exploratory task choice measures. Contrary to predictions, we did not observe a significant interaction between disclaimer condition and credibility condition on our measures of task choice in our enhanced trust game setup, either for the continuous preference measure, $F(1,391) = 0.53, p = .466, \eta_p^2 < .01$, or for the binary choice measure, $b = 0.06, p = 0.881$. However, we did find (in an analysis that was not pre-registered) that responses on both measures were predicted by warmth perceptions, even when controlling for perceived competence (continuous choice measure: $b = 0.61, p < .001$; binary choice measure: $b = 0.06, p < .001$). These results suggest that—even though our manipulation was not strong enough to yield a significant overall interaction on these measures—the perception measures do seem to predict behavioral choices in the expected directions.

Exploratory social hour measure. We also did not observe a significant interaction between disclaimer condition and credibility condition on our measure of desire for social interaction with Jamie, $F(1,391) = 2.08, p = .150, \eta_p^2 < .01$. However, as with our task choice measures, we did find (again, not pre-registered) that perceptions of warmth predicted desire for social interaction, even when controlling for perceived competence, $b = 0.52, p < .001$.

Moderated Mediation (Exploratory). Again, we conducted a non-preregistered analysis in which we tested whether the target's perceived awareness of how others see them and perceived sincerity (separately) mediated the effect of disclaimer condition on warmth perceptions, and whether this was moderated by credibility condition. Using separate models for each mediator, we observed significant moderated mediation for both awareness of how others see them

(indirect effect: $b = 3.82$, 95% CI = [1.80, 5.89], $p < .001$) and sincerity (indirect effect: $b = 6.64$, 95% CI = [3.65, 9.67], $p < .001$). Specifically, when focusing on the low-credibility condition, we see that awareness and sincerity each *positively* mediated the effect of disclaimer condition (indirect effect of awareness: $b = 1.83$, 95% CI = [0.59, 3.22], $p = .002$; indirect effect of sincerity: $b = 3.57$, 95% CI = [1.47, 5.77], $p < .001$), but when focusing on the high-credibility condition, awareness and sincerity each *negatively* mediated the effect (indirect effect of awareness: $b = -1.90$, 95% CI = [-3.56, -0.24], $p = .027$; indirect effect of sincerity: $b = -3.06$, 95% CI = [-5.15, -1.02], $p = .003$).

Discussion

Study 4 replicated our findings in a more realistic context with richer stimuli. All told, our disclaimer manipulation was relatively subtle compared to all the other information participants were able to observe about the target in this study (e.g., the target's essay and writing history, as well as the target's face, voice, and other information from the videos)—yet we nevertheless observed that the disclaimer influenced perceptions of the target. Our results thus further underscore the influence of disclaimers in everyday live interactions. The lack of significant results on our behavioral measures in this study may perhaps support our proposition in Study 1 that disclaimers most strongly affect integrity-based trust, rather than benevolence-based trust (which is more closely related to the measures captured here).

Studies 5a-b: Disclaimers in Hiring Settings

In our final set of studies, we sought to further test the benefits of disclaimers for low-credibility speakers in a consequential real-world context: hiring decisions. We reasoned that, because people often prefer likable, honest, and trustworthy work colleagues, the impact of disclaimers on perceived warmth might trickle down to the behavioral decision to recommend a

person for a job. We conducted both an online study and a field study with the same design. In both studies, we presented participants with a brief resume of a (fictitious) job candidate, who always had poor qualifications. We then varied whether the candidate used a disclaimer or not (between-participants) in a separate message in which they self-promoted. Participants then indicated to what extent they would recommend hiring that candidate.

Study 5a: Participants. Participants were recruited online via Prolific Academic and completed the study in exchange for \$1.20. We used Prolific's prescreen filters to select only individuals with hiring experience, in order to ensure that they would be best suited to evaluate candidates for a job. We pre-registered that we would collect 400 participants after excluding those who failed the attention check, did not finish the survey, and/or provided gibberish or bot-like responses to a free-response question. Of the 422 who started the survey, we ended up with a final sample of 402 participants (45.02% female; 20.15% non-White; $M_{\text{age}} = 45.58$).

Study 5a: Procedure. Participants read that they would be shown the profile of a candidate for a job, and would be asked how much they would recommend we hire that candidate. We described the job as an entry-level research assistant position in the social sciences, and told participants that we were interested in candidates who had the most potential in terms of both general intelligence and interpersonal skills. We further told them that we particularly valued candidates with integrity and self-awareness, as those are key to successful interpersonal interactions and the ability to adapt and improve. Participants also read that we would conduct a second round of review after this initial round, during which we could collect more information from the candidates, but that we wanted their recommendations now based on the information we had so far.

Participants read that we would first show them the profile of the previous person who held this job (when they applied) to help give them a sense of the kind of candidate we wanted. On the next page, they saw the educational record and work experience of the previous candidate. As shown in Figure 1.7, we made this profile seem high-quality (high grades, relevant work experience), in order to ensure that participants viewed the target profile as not very high-quality by contrast.

On the next page, participants read that they were about to see the current candidate's profile, and we explained that all candidates had filled out a standardized form with their information that we would show participants. We also noted that this particular candidate had sent us an email message with some additional information, which would be presented below their profile. Participants then clicked to the next page and saw the target profile. As shown in Figure 1.8, the target candidate generally had low qualifications for the job (low grades, a lack of relevant work experience), so that their subsequent brag would seem low-credibility. Below the profile, we showed the candidate's email message. It was in this email message that we placed the brag, either by itself or prefaced by a disclaimer. In the no disclaimer condition, the email message read: "Hi, I wanted to add something that wasn't included on the form I submitted. I did an honors thesis project in college that won a top award. Thank you, [name redacted]." In the disclaimer condition, the email text was exactly the same, except we added a disclaimer to the beginning of the brag about the honors thesis project: "I know my record might not seem as impressive as other applicants' records, but..."

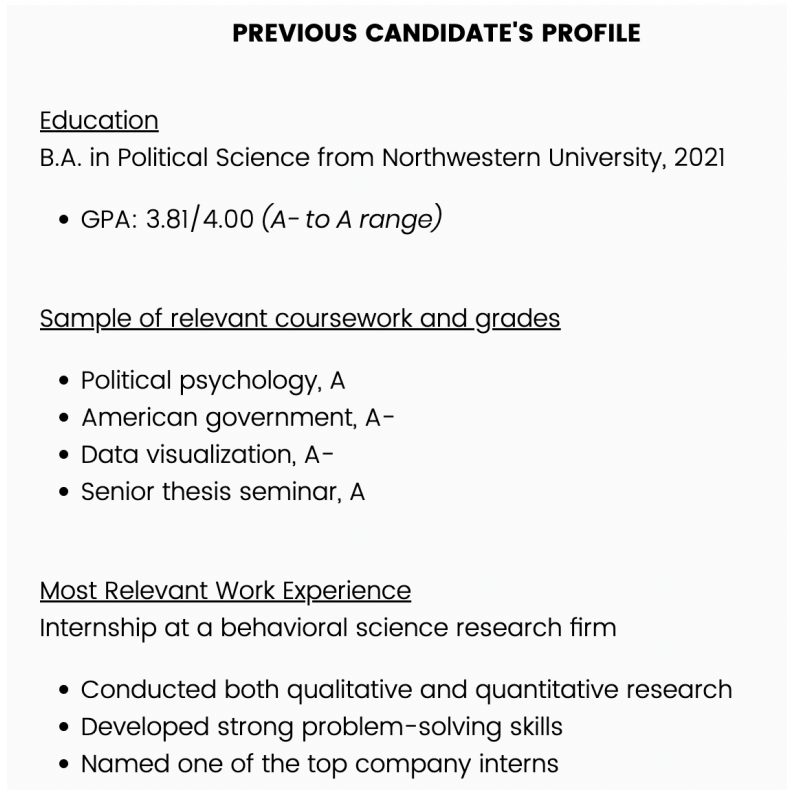


Figure 1.7. Previous candidate's profile in Studies 5a-b, Chapter 1.

On the next page, participants saw a reminder of the target candidate's profile, as well as a reminder of the nature of the job and what we were looking for in a candidate. We then asked participants for their hiring recommendation: "With all of this in mind, to what extent would you recommend we hire this candidate?" (-3 = *strongly recommend you NOT hire this candidate*, 3 = *strongly recommend you HIRE this candidate*). Below that, we asked participants: "Please write ~1 sentence about why you made your choice above" (free-response).

On the next page, participants again saw a reminder of the target candidate's profile, followed by the same set of perception questions used in Study 1 (modified as needed to refer to "this candidate" rather than "this participant" and to refer to the candidate's specific brag in the sincerity and believability questions). Finally, participants reported demographic information, and were given an optional space to provide feedback on the study.

CANDIDATE PROFILE

Education

B.A. in Sociology from Bradley University, 2023

- GPA: 2.65/4.00 (C+ to B- range)

Sample of relevant coursework and grades

- Research methods, C-
- Pop culture and mass media, B
- Algebra, C
- Writing seminar, B

Most Relevant Work Experience

Summer job as an administrative assistant at a mid-size company

- Experience with standard administrative tasks

Email message from the candidate:

**"Hi, I wanted to add something that wasn't included on the form I submitted. I know my record might not seem as impressive as other applicants' records, but I did an honors thesis project in college that won a top award.
Thank you,
[name redacted]"**

Figure 1.8. Target candidate's profile, including their email message (in the disclaimer condition), in Studies 5a-b, Chapter 1.

Study 5b: Participants. Participants were recruited via Upwork, a platform on which individuals can be hired to complete freelance tasks for pay. Participants responded to a job posting that asked them to help evaluate candidates for a job, unaware that they were participating in an experiment. We paid them \$10 for an estimated 10-minute task.⁶ We ended up

⁶ Four participants were paid a different amount—see details in Appendix A.

with our preregistered target of exactly 300 participants (we did not collect demographic information in this field study).

Study 5b: Procedure. We created a hiring account on Upwork and posted a job listing—see Appendix A for this job advertisement. A research assistant managed the account and handled all interactions with potential participants. Initially, we accepted only individuals who indicated having hiring-related experiences and specialties on their profile (e.g., recruiters, HR workers); however, due to limitations on available participants, we later expanded to those with other relevant experiences, such as those with their own research experience (since the job they were evaluating candidates for was a research assistant position). For each participant, we sent them a word document with the exact same information provided in Study 5a, and asked them to fill out their hiring recommendation in this word document and send it back to us. Unlike in Study 5a, we did not ask the perception measures. We report the full set of procedures in Appendix A, and have posted additional study materials on OSF.

Study 5a: Results

For our Prolific study, we conducted independent t-tests between disclaimer conditions for each of our dependent measures. We combined our measures of trustworthiness, honesty, and likability into a composite of warmth ($\alpha = 0.89$), and analyzed the rest of our measures individually. We report means and standard deviations for the key variables in Table 1.1.

Hiring recommendation. On our hiring recommendation measure, participants were significantly more likely to recommend hiring the candidate when the candidate used a disclaimer than when they used no disclaimer, $t(400) = -2.69$, $p = .007$, $d = -0.27$. See Figure 1.9, left panel.

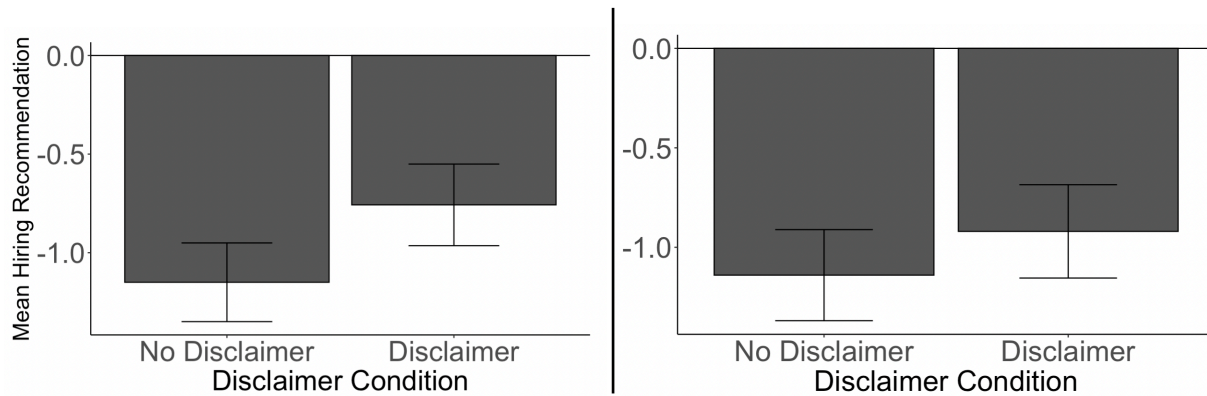


Figure 1.9. Mean hiring recommendations in Study 5a (left panel) and 5b (right panel), Chapter 1. Error bars represent 95% confidence intervals.

Self-awareness and sincerity. Once again, participants found the target to be more self-aware, $t(400) = -5.78, p < .001, d = -0.58$, and to be more sincere in their bragging statement, $t(400) = -2.92, p = .004, d = -0.29$, when they used a disclaimer compared to no disclaimer.

Believability. Similarly, participants found the bragging statement more believable when it was prefaced by a disclaimer compared to no disclaimer, $t(400) = -2.87, p = .004, d = -0.29$.

Warmth. Participants also found the target warmer overall when they used a disclaimer than when they used no disclaimer, $t(400) = -2.90, p = .004, d = -0.29$.

Competence. Finally, there was no significant difference in perceived competence when the target used a disclaimer compared to no disclaimer, $t(400) = -1.62, p = .106, d = -0.16$. Once again, the direction of means suggests that disclaimers at the very least do not *harm* perceptions of competence.

Mediation (Exploratory). As in all our other studies, we conducted a non-preregistered analysis in which we tested whether self-awareness and sincerity (independently) mediated the effect of disclaimer condition on hiring recommendations and (separately) warmth perceptions. Using separate models for each mediator, we observed that self-awareness and sincerity each mediated the effect on hiring recommendations (indirect effect of self-awareness: $b = 0.37, 95%$

CI = [0.23, 0.51], $p < .001$; sincerity: $b = 0.17$, 95% CI = [0.05, 0.29], $p = .005$) as well as the effect on warmth perceptions (indirect effect of self-awareness: $b = 4.45$, 95% CI = [2.86, 6.18], $p < .001$; sincerity: $b = 2.57$, 95% CI = [0.81, 4.29], $p = .004$).

Study 5b: Results

In our field study, contrary to predictions, we observed a non-significant effect of disclaimer condition on hiring recommendations, $t(298) = -1.33$, $p = .186$, $d = -0.15$. We expect that this non-significant effect may have been due to insufficient power given the added noise inherent in a field setting. As shown in Figure 1.9, right panel, the direction of means was consistent with our predictions and with Study 5a.

To better understand these results, we conducted two non-preregistered analyses using participants' free-response explanations of why they made the recommendation they did. We had a research assistant who was blind to hypotheses read each response and code whether the response included a reference to 1) the applicant's thesis/award, and/or 2) the disclaimer in the email (the latter of which, of course, should have been mentioned only by participants in the disclaimer condition). We did so in order to test whether those who paid enough attention to the content of the email—including the disclaimer—to write about it in their response were especially likely to favor the applicant who used the disclaimer.

With regard to mentions of the thesis/award, 161 participants across both conditions mentioned the thesis/award somewhere in their free-response. Analyzing only these participants still yielded a non-significant effect of condition, $t(159) = -0.64$, $p = .523$, $d = -0.10$.

With regard to mentions of the disclaimer specifically, 42 participants in the disclaimer condition mentioned the disclaimer somewhere in their free-response.⁷ Within the disclaimer

⁷ The research assistant, who was blind to condition, coded 5 additional participants in the no disclaimer condition as having mentioned the disclaimer; we have excluded those from the total of 42.

condition, we compared the 42 who mentioned the disclaimer to the 108 who failed to mention it. Those who mentioned the disclaimer were more likely to recommend hiring the candidate than those who did not $t(62.10) = -2.26, p = .027, d = -0.45$. We also compared the 42 who mentioned the disclaimer to all participants in the no disclaimer condition. Participants who mentioned the disclaimer recommended the candidate significantly more than participants in the no disclaimer condition, $t(58.59) = -2.43, p = .018, d = -0.47$. Thus, this finding—though, again, non-preregistered—may provide some suggestive evidence that the disclaimer increased hiring recommendations among those who most considered it in their evaluations.

Discussion

Study 5a demonstrated that disclaimers can impact consequential behavioral decisions: the decision to recommend a person for a job. Attempting to replicate this in a field setting where participants did not know they were participants, we found a weaker, non-significant impact of disclaimers on hiring recommendation. However, post-hoc analysis of free responses provided some confidence that disclaimers were having the expected impact even in the field: When people gave the disclaimer enough consideration to write about it in their responses, their evaluations of the candidate were higher than those who did not write about it or those in the no disclaimer condition.

General Discussion

People often face the dual dilemma of needing to correct others' negative impressions of them while simultaneously knowing that their attempts to self-promote may seem especially unbelievable in such cases—a situation we call the “credibility dilemma.” Our pilot data suggests that when facing such a dilemma, people are split in terms of whether they choose to explicitly acknowledge that their claim seems unbelievable. Many do not, explaining that they do not want

to draw attention to their weakness. Many others, however, choose to acknowledge their weakness when self-promoting. Our paper tests the effectiveness of this strategy and finds support for it.

Across ten pre-registered studies, we find that acknowledging the perceived lack of credibility of a self-promotional claim—with a disclaimer—is beneficial: Such disclaimers make the claim seem more believable, and lead the speaker to be perceived as warmer (relative to making the claim without a disclaimer). In particular, these disclaimers increase the perceived self-awareness (Studies 1-5a) and sincerity (Studies 1 and 4-5a) of the speaker, which in turn increase the perceived warmth, and in some cases competence, of the speaker (Studies 1-5a). Further, these perceptions translate to consequential behavioral decisions, such as trusting someone’s advice (Study 1) or deciding whether to hire them (Study 5a). In addition, the effect of disclaimers generalizes to a variety of different types of self-promotional claims (Studies 3a-d), and replicates in the context of spoken (rather than written) interactions (Study 4). Yet we also find that these credibility disclaimers are not universally beneficial: When the speaker is already perceived as credible in their claim, adding a disclaimer does not help, and can even backfire (Studies 2-4). This finding suggests that the benefit of disclaimers is not simply due to a general effect of self-deprecation, perceived modesty, or humor, but rather, is contingent on the match between the audience’s prior impression of the speaker and the speaker’s disclaimer.

Theoretical Implications

Theoretically, our research highlights at least one way in which impression management strategies may need to be adapted according to the actor’s characteristics. In particular, there are some cases in which the optimal use of impression management techniques requires one to accurately know what one’s audience thinks of oneself (in this case, whether the audience’s

initial perceptions would seem incongruent with a subsequent self-promotional claim). Much research, however, points to the fact that people often have systematically inaccurate (and in particular, more negative) perceptions of what others think of them (Boothby et al., 2018; Bruk et al., 2018; Gilovich et al., 2000, 1998; Kenny & DePaulo, 1993; Moore-Berg et al., 2020; Savitsky et al., 2001), leading people to make sub-optimal choices of how and what to communicate (Kumar & Epley, 2018; Zhao & Epley, 2021). In our case, we similarly expect these systematic inaccuracies to lead people to make sub-optimal choices of how to use credibility disclaimers in the real world (e.g., believing that others perceive them as less credible than they really do, and thus choosing to add a disclaimer when they should not).

Our findings also contribute more generally to existing literature on the efficacy of different types of disclaimers as impression management tools. Most prior research that has empirically tested the impact of disclaimers on perceptions of speakers has been outside of bragging contexts (Bell, Zahn, & Hopper, 1984; Bradley, 1981; Hamilton, Vohs, & McGill, 2014; Shapiro & Bies, 1994) or has tested different types of disclaimers that do not address brag believability (El-Alayli et al., 2008). Our research instead tests the efficacy of a particular set of disclaimer types—those that could be used to address a perceived lack of credibility in a self-promotional claim. The credibility dilemma we examine here showcases a common—but to our knowledge largely overlooked—challenge that potential braggarts face, and one that has not yet been linked to the disclaimers literature. Our findings also suggest that disclaimers may operate via additional mechanisms (perceived self-awareness and sincerity) than those explored in previous research on disclaimers, such as confirmation bias (El-Alayli et al., 2008). In fact, our results seem to point to the *opposite* of confirmation bias in this setting, since acknowledging the statement’s apparent lack of believability actually made the statement more—rather than less—

believable. We speculate that confirmation bias may not apply here, in contrast to previous research, because the disclaimer is referring to an existing impression of the person rather than a (yet unformed) impression of the statement itself (e.g., “I don’t mean to sound arrogant, but...”). Thus, in the latter cases, the listener has not already formed an impression of the statement before hearing it (e.g., how arrogant it will be), so the disclaimer has more of an influence than it does when the disclaimer refers to a pre-existing impression. However, additional research may help to further unpack when and why each of these mechanisms comes into play.

Finally, our research contributes to work that demonstrates ways in which highlighting negative information can actually lead others to form more *positive* impressions. Although previous research has highlighted various contexts and processes by which this can occur (Aronson et al., 1966; Brooks et al., 2019; Crowley & Hoyer, 1994; Etgar & Goodwin, 1982; Hoffman-Graff, 1977; John et al., 2016; Knowles & Linn, 2004; Rucker et al., 2008; Ward & Brenner, 2006), to our knowledge, this has not been previously tested when someone is trying to overcome an initial negative impression—a situation where highlighting negative information may seem even riskier. Despite the risk, we find that drawing attention to one’s lack of credibility actually *helps* low-credibility speakers. In contrast, we find that referring to a lack of credibility does not help (and can sometimes hurt) high-credibility speakers. Thus, our findings can be seen as a “flip” from previous work that has examined unintentional blunders (Aronson et al., 1966). Namely, unintentional blunders humanize those who are already seen as highly competent, but hurt impressions for those who are just average. Instead, our work shows that intentionally highlighting one’s shortcomings can actually help overcome negative impressions when doing so signals high self-awareness and sincerity.

Practical Implications

Practically, the current research suggests that people should take their audience's prior impression of them into account when deciding how to convey self-promotional information. If the audience already perceives one as having high skill on the dimension on which one would like to brag—or if the audience lacks information about one's skill on this dimension—then people may be advised to avoid credibility disclaimers when they want to brag. On the other hand, if the audience already perceives one as having low skill on the dimension on which one would like to brag, then people may be advised to address the brag's lack of credibility with a disclaimer—despite any common wisdom that drawing more attention to one's flaws will make things worse. Indeed, using a disclaimer in such cases is necessary, given that simply not bragging at all is better than bragging *without* a disclaimer (see Supplemental Studies C-D). Once again, our research cautions that the effectiveness of disclaimer usage relies on the speaker having an accurate perception of their audience's prior impression of them. Thus, our work underscores a (relatively under-examined) challenge that impression managers face in order to select the right self-presentation strategy: accurately knowing others' impressions of them. Impression managers may therefore be well-served to gather as much accurate information as possible about what others think of them prior to selecting the appropriate strategy.

Limitations and Future Directions

Our studies have several limitations that open up interesting directions for future research. First, future research should examine the intersection between credibility disclaimer usage and other communication strategies or intentions. For example, although our studies suggest that the benefit of disclaimers in these contexts is not due to perceived humor, there may be other contexts in which the use of a disclaimer may be perceived as an attempt at humor, or as otherwise intended to not be taken literally. Since humor can change the perceived motive of the

speaker (Bitterly & Schweitzer, 2019), additional research is needed to understand whether credibility disclaimers are perceived differently when used humorously or facetiously.

Second, future research could test the extent to which people's disclaimer usage in the real world is optimal. In our pilot study as well as Supplemental Study D, a majority of people seemed to make the "right" choice of disclaimer when they were fully informed of their audience's prior impression of them. However, as mentioned above, we expect that people's beliefs about others' impressions of them are often systematically mis-calibrated; thus, we expect that people may sometimes use disclaimers when they should not. Additional research could further test these types of speaker choices, using more naturalistic study designs. Relatedly, future research could examine a wider variety of contexts in which disclaimers might naturally occur. For instance, low-credibility speakers might sometimes use them (along with a brag) in response to negative feedback—which could perhaps backfire if it comes across as too defensive.

Third, future research should explore whether disclaimers can ever benefit speakers who are already seen positively. Imagine that a highly competent colleague wants to share a weakness or vulnerability (e.g., "I'm a terrible speller"). Would your impression be more positive if they prefaced the comment by acknowledging that this is discordant with their image (e.g., "I know this may be surprising, but...")? We suspect that speakers who explicitly acknowledge the audience's prior (positive) impression of them prior to making a self-deprecatory statement (thereby contradicting the audience's prior impression) may indeed be perceived more positively than those who make no acknowledgement, again by signaling self-awareness.

Finally, future research should continue to explore when and why expressing self-awareness is perceived positively versus negatively. While the studies presented here found that disclaimers help in part by signaling self-awareness, we also found in several supplemental

studies (and consistent with prior research) that not all types of disclaimers help, even while ostensibly indicating awareness of how the audience will perceive the statement. In particular, disclaimers that acknowledge other negative aspects of a brag besides its perceived lack of credibility—such as its obnoxiousness or its likelihood of irritating the listener—tend to backfire (regardless of speaker credibility). We speculate that there may be certain characteristics for which it is *worse* to show awareness of how one is perceived (compared to seeming unaware), such as characteristics for which the speaker might be expected to actually change their behavior if they are aware—see Chapter 2 of this dissertation.

While future research could help shed additional light on the effects of disclaimers, the current research nevertheless suggests that acknowledging one's (perceived) lack of credibility may be a useful tool for speakers who face a credibility dilemma.

Chapter 2:

Ignorance can be Trustworthy: The Effect of Social Self-Awareness on Trust

Abstract

Self-awareness is often thought of as a positive quality in others, yet there are cases in which self-awareness may send a negative signal. Specifically, we propose that when a target person appears to be high in social self-awareness—i.e., seems to accurately know what others think of them—observers infer that the target’s actions are more *intentional*, because the target is acting while seemingly knowing what others think of their actions. In turn, observers will perceive those actions as more diagnostic of the target’s true character and future behavior. Consequently, the target’s exhibited self-awareness should affect observers’ trust toward the target differently depending on whether the target behaves in ways that positively or negatively impact others. When the target behaves in positive ways, exhibiting self-awareness should increase trust, as the positive behaviors will be interpreted as more intentional and diagnostic. However, for negative behaviors, exhibiting self-awareness should *decrease* trust, as negative behaviors are seen as worse when more intentional. Across seven studies ($N = 6,164$) using online experiments, a recall study paradigm, and live interactions in a laboratory setting, we find support for this framework. We also show that when we constrain the extent to which people can infer intentions toward others from a target’s behaviors—by reducing the target’s control over changing their behavior or by reducing the impact of the target’s actions on others—the effect of self-awareness on trust attenuates. Our findings suggest that self-awareness, though often considered a desirable quality, does not universally increase others’ trust.

Introduction

Imagine meeting a new colleague at work. After interacting several times with this colleague, you form the impression that he seems to be an unpleasant sort; he often comes across as unfriendly, and sometimes even downright rude. Now imagine learning an additional piece of information about this colleague: He seems to be *aware* that you perceive him as rude. How might this information affect how much you trust this colleague? On the one hand, perhaps you would trust him *more* knowing that he is aware of your perception because it would show that he has the ability to take others' perspectives and to potentially improve his behavior in the future. On the other hand, perhaps you would trust him *less* because it signals that he was willing to be rude *despite* knowing you perceive him as such—and that he is, therefore, more likely to continue being rude in the future.

In the current research, we examine how people evaluate a target individual based on how socially self-aware the target appears to be—that is, whether the target appears to accurately know what others think of them. Much existing research on self-awareness has examined the construct from an *intrapersonal* lens, such as how an individual's degree of self-awareness affects their own subsequent cognitions and behaviors (Diener & Wallbom, 1976; Duval & Wicklund, 1972; Hass, 1984; Heatherton & Baumeister, 1991; Wicklund, 1975). By contrast, we examine self-awareness from an *interpersonal* lens, i.e., how observers perceive a target who appears to be high or low in self-awareness—and in particular, how much observers *trust* the target. In contrast to past work that has examined correlations between self-awareness and interpersonally relevant behaviors (Atwater & Yammarino, 1992; Church, 1997; Van Velsor et al., 1993), we use experiments to assess the causal impact of perceived self-awareness on observers' trust toward the target, controlling for the target's other behaviors. While there are

several different types of self-awareness, our theory centers on social self-awareness—an awareness of what others think of oneself—as we expect social self-awareness to be most directly tied to judgments of trustworthiness.

Intuition might suggest that self-awareness, including social self-awareness, should be perceived as a desirable quality in others and should, therefore, increase observers' trust toward those who exhibit it. However, we propose that the effect of social self-awareness on trust is more nuanced. Specifically, we propose that social self-awareness signals more than just the (often desirable) quality of being able to take others' perspectives and potentially regulate or improve one's behavior accordingly; it also serves as a signal of greater *intentionality* behind the target person's behavior. If a target individual appears to know how others perceive them, then observers will infer that the target more intentionally created these impressions, suggesting that their behavior is more diagnostic of the target's true intentions toward others and thus more predictive of the target's future behavior. As a result, a target's exhibited social self-awareness will have different effects on observers' trust of the target depending on the target's specific behaviors and will, in some cases, actually lead to a decrease in trust.

Our findings contribute theoretically to the literature and practically to impression management techniques. Theoretically, our work is among the first to causally demonstrate the role of perceived social self-awareness in trust formation. We show that a target's social self-awareness sends a complex signal to observers about the target's underlying character, leading to different effects on trust depending on the target's specific behaviors. Practically, our insights provide guidance for optimal impression management strategies. We suggest that it may increase others' trust to exhibit social self-awareness to others in some cases—but not in all cases.

Defining Social Self-Awareness

Self-awareness has been defined as attention that is directed toward the self, rather than features of the external environment (Duval & Wicklund, 1972; Wicklund, 1975). There are several types of self-awareness: internal self-awareness, defined as an awareness of one’s internal experiences, such as thoughts or emotions; external self-awareness, defined as an awareness of one’s external features, such as appearance and behaviors; and social self-awareness, defined as an awareness of how other people might perceive or interpret oneself (Chon & Sitkin, 2021). Given our focus on the relationship between self-awareness and social perception—namely, perceived trustworthiness—we examine social self-awareness in particular, and we explain in a forthcoming section why we do not expect our predictions to hold for other types of self-awareness.

Social self-awareness is closely related to perspective-taking, which has been defined as “the process of imagining the world from another’s vantage point or imagining oneself in another’s shoes” (Galinsky et al., 2005; Ku et al., 2015). Indeed, self-awareness has been shown to lead to greater perspective-taking (Hass, 1984). Social self-awareness can be thought of as perspective-taking that is specifically applied toward how others view *oneself*, as opposed to how they view other features of the environment.

A separate but closely related body of work studies meta-perception, which has been defined as one’s beliefs about what others think of oneself (Laing et al., 1966). Meta-accuracy, in turns, refers to being *accurate* in these beliefs (Donnelly et al., 2022; Kenny & DePaulo, 1993). Some research in this domain has examined how meta-accurate people tend to be on average (Boothby et al., 2018; Carlson et al., 2010; Carlson & Furr, 2009; Carlson & Kenny, 2012; Donnelly et al., 2022; Eisenkraft et al., 2017; Elsaadawy, 2018; Kenny & DePaulo, 1993; Moon et al., 2020), as well as what factors might make people more or less meta-accurate (Elsaadawy

et al., 2021). To our knowledge, none of this research has examined “meta-meta-perception,” or how people evaluate those who appear to have high or low meta-accuracy.

Drawing from research on both social self-awareness and meta-accuracy, we define social self-awareness as not only an *awareness* of what others think of oneself, but also an *accurate* inference of what others think. This definition is consistent with research that has operationalized self-awareness as agreement between others’ ratings of oneself and one’s prediction of those ratings (Taylor et al., 2012). We define being high in social self-awareness as accurately knowing what others think of oneself, and we define being low in social self-awareness as one of two possibilities. The first is being totally *unaware* of what others think of oneself because attention is not directed toward the self, thus not allowing one to form a judgment about what others think of oneself. The second is being *inaccurately* aware of what others think of oneself because one directs attention to the self but is incorrect in discerning what others think. In the current work, we focus on observers’ *perceptions* of a target person’s social self-awareness, regardless of whether the target truly is high or low in self-awareness. Thus, we compare perceptions of targets who seem accurately self-aware versus perceptions of targets who seem to be either of the other possible alternatives (i.e., un-self-aware or inaccurately self-aware).

We distinguish our construct from the related concept of self-monitoring in three main ways. Self-monitoring corresponds to “self-observation and self-control guided by situational cues to social appropriateness” (Snyder, 1975), and thus encompasses behavioral modifications in response to different situations. By contrast, our definition of social self-awareness does not include behavioral modifications but rather focuses on the awareness itself, regardless of whether the awareness may lead to behavioral changes. In addition, an individual could engage in self-

monitoring for reasons beyond others' perceptions of oneself, e.g., in response to general situational norms. Finally, and relatedly, self-monitoring does not have to include the accuracy component that we include in our definition of self-awareness. That is, one could change behavior in response to situational cues but be inaccurate in choosing the appropriate modification.

Social Self-Awareness as a Trustworthy Quality?

Trust can be defined as the willingness to be vulnerable due to positive expectations of the intentions or behavior of another person (Kramer, 1999; Kramer & Lewicki, 2010; Mayer et al., 1995; Rousseau et al., 1998). There are many reasons to expect that a target's degree of social self-awareness would lead observers to have more positive expectations of the target's future behavior, thereby increasing trust. For instance, both self-awareness (broadly defined) and perspective-taking (a necessary feature of social self-awareness) are associated with behaviors that should lead to desirable consequences for interpersonal contexts. Self-awareness can lead to motivation to improve the self (Wicklund, 1975), can reduce certain socially undesirable behaviors such as cheating (Diener & Wallbom, 1976), and is associated with greater helping behaviors (Wegner & Schaefer, 1978). Perspective-taking is associated with empathy (Batson et al., 1997), social bonding and social coordination (Galinsky et al., 2005), willingness to interact with outgroup members (Wang, Kenneth, et al., 2014), reduced stereotyping (Wang, Ku, et al., 2014), better joint outcomes in negotiations (Galinsky et al., 2008; Trötschel et al., 2011), and better marital adjustment (Long & Andrews, 1990). If people anticipate these behaviors from those who are more socially self-aware, then one would expect people to trust a target individual more when that target displays high (versus low) social self-awareness. Even if the target is behaving in undesirable ways, greater social self-awareness might still suggest a greater

likelihood that the target will improve their behavior in the future (Wicklund, 1975), again leading to more positive expectations of future behavior and thus higher trust.

In further support of this possibility, the literature on meta-perception has found in some correlational research that meta-accuracy is often related to positive interpersonal outcomes. People who are more accurate in knowing how much others trust them tend to be subsequently trusted more (Brion et al., 2015), and meta-accuracy in dyads of strangers is correlated with mutual liking (Ohtsubo et al., 2009). This work might again suggest that the causal effect of social self-awareness on trust is likely to be positive.

Taken together, there are many reasons to expect that social self-awareness signals to observers that the target may be more likely to engage in prosocial behaviors, and thus will lead to higher trust from observers. However, most existing research has examined only the *correlation* between accurate awareness and social perceptions (Brion et al., 2015; Church, 1997; Long & Andrews, 1990; Ohtsubo et al., 2009), or has examined how the induction of self-awareness and/or perspective-taking affects one's own behaviors in socially relevant situations (Batson et al., 1997; Diener & Wallbom, 1976; Galinsky et al., 2005, 2008; Trötschel et al., 2011; Wang, Kenneth, et al., 2014; Wang, Ku, et al., 2014). By contrast, little research has isolated the causal impact of a target's external exhibition of self-awareness on others' judgments of that target, while controlling for the target's other behaviors that might be correlated with self-awareness. Our research attempts to fill this gap. In doing so, we suggest that social self-awareness sends not only these signals of potential prosocial behavior, but also an additional signal that can, in some cases, lead to *lower* perceptions of trustworthiness.

Social Self-Awareness as a Signal of Intentionality

Perhaps not surprisingly, inferring others' intentions behind their actions—rather than merely judging the actions themselves—is a key component of social judgment (Hackel et al., 2020; J. Landy & Uhlmann, 2018; Levine & Schweitzer, 2015; Malle & Knobe, 1997; Maselli & Altrocchi, 1969; Uhlmann et al., 2013, 2015; Uhlmann & Zhu, 2014). Intentionality is important because it allows people to make inferences about a target's general character beyond a single action, which in turn can help people predict how the target will behave in the future. Even in the absence of specific information about a target's intentionality, people are generally disposed to infer something about a target's intentions from their mere behavior (Gilbert & Malone, 1995).

The link between social self-awareness and perceived intentionality is suggested by past research on how people form inferences of others' intentions, which has examined people's lay definition of intentionality. This research has found that one component of lay definitions is a target person's *awareness* of their actions during the act itself (Malle & Knobe, 1997).

Accordingly, we expect observers to use social self-awareness as one cue from which they infer the degree of intentionality behind a target's actions, with higher social self-awareness leading observers to infer that the target's behavior is more intentional. That is, if a target person seems to be aware of how others will perceive his or her behavior, then we expect observers to infer that this behavior is more in line with the target's true intentions toward others, and thus a more revealing representation of the target's true character. Put another way, observers may infer that when a target seems high in social self-awareness, their behaviors are more *diagnostic* of their true intentions toward others, which in turn should lead observers to weight this information more heavily in their evaluations of the target (Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1987, 1989).

Given that we expect social self-awareness to signal intentions toward others, we expect social self-awareness to primarily affect benevolence-based trust, i.e., trust based on expectations of a target's intentions to behave kindly toward others (Mayer et al., 1995). By contrast, other theorized antecedents of trust—namely, perceived ability and integrity (Mayer et al., 1995)—should be less affected by a target's display of social self-awareness, given that they do not correspond as directly to the target's perceived positive intentions toward others.

How might greater perceived intentionality—inferred via the target's social self-awareness—affect observers' trust toward that target? We propose that the direction of the effect will depend on the target's specific behaviors, and in particular, how those behaviors *impact others*. If the target's actions create a positive impact on others, then greater perceived intentionality (signaled through high self-awareness) is likely to yield an *increase* in trust toward the target. Indeed, previous research suggests that when a target's behavior has a positive impact on others, people tend to evaluate the target more positively when the target's actions are more intentional (Swap, 1991; Tesser et al., 1968), presumably because this perceived intentionality serves as a more diagnostic indicator of positive future behavior. As a result, we expect that when a target exhibits high self-awareness while engaging in positive behaviors, observers will trust this target more (relative to targets who exhibit low self-awareness), as they will interpret this self-awareness as a stronger signal that the target will be reliably prosocial. We expect this increase in trust to be primarily driven by increases in perceived benevolence, more so than by increases in perceived ability or integrity.

On the other hand, if the target's actions create a negative impact on others, then greater perceived intentionality is likely to yield a *decrease* in trust toward the target. Research in the moral domain has found that the same moral transgression is perceived more negatively when it

is perceived as more intentional (Cushman, 2008; Schaich Borg et al., 2006; Young & Tsoi, 2013). Once again, this is because greater intentionality may signal more about the target's general character, and thus, the target's likelihood of behaving negatively in the future. We therefore expect that exhibiting high self-awareness while engaging in behaviors that negatively impact others will decrease observers' trust toward the target, as it serves as a more diagnostic signal that the target is likely to behave harmfully in the future. Again, we expect this decrease in trust to occur primarily because of a decrease in perceived benevolence.

Following from these predictions, we also expect that any factors that moderate perceptions of intentionality—such as situational attributions for a target's behaviors—should moderate the effect of social self-awareness on trust. For instance, we expect inferences of intentionality to be moderated by the perceived mutability of the target's behavior, i.e., how easy or difficult it would be for the target to change their behavior. People tend to be more likely to attribute others' behavior to internal characteristics, such as intentions, when there are fewer constraints on the target's behavior (E. E. Jones & Davis, 1965; Kelley, 1987)—meaning, when their behavior is high in mutability. In a context in which mutability is low—meaning the target had little opportunity to act differently even if they wanted to—the signal of intentionality is likely to be obscured, making it less influential in people's evaluations of the target's trustworthiness. As a result, the target's degree of social self-awareness may influence trust less in such cases. We test this possibility in two of our studies.

It is worth noting that we expect social self-awareness to be most closely tied to trust rather than other interpersonal judgments, such as liking. While liking may be informed by, and related to, expectations of a target's interpersonal behavior, it may also be influenced by other unrelated factors, such as attractiveness (Stroebe et al., 1971) or humor (Treger et al., 2013). We

focus primarily on trust because it most closely tracks the social judgment that we propose social self-awareness impacts.

Social Self-Awareness vs. Other Types of Self-Awareness

Why do our predictions focus on social self-awareness rather than internal or external self-awareness? We propose that social self-awareness should most strongly signal a target's intentions toward others, and trust is driven by expectations of an individual's intentions and future behavior toward others. For instance, a target could be aware that they feel frustrated (internal self-awareness) or that their face appears red (external self-awareness), but neither of these things in isolation suggests something about the target's intentions toward others. While internal and external self-awareness may still signal something about the target's cognitive abilities, such as emotional intelligence (Salovey et al., 2004; Salovey & Mayer, 1990), the effect on trust should be weaker given that they do not signal intentionality toward others. By contrast, if the target is aware that another person perceives them as being curt, abrupt, or rude—and is still behaving this way nonetheless—the target's social self-awareness is likely to signal a greater likelihood of behaving rudely toward others in the future, which should decrease trust.

Further, we expect social self-awareness to have the strongest effect on trust when the behavior actually has an *impact* on others. A target might be aware that others perceive her as incompetent, but this may affect others' outcomes more directly if they are work colleagues rather than friends. Drawing on interdependence theory (Thibaut & Kelley, 1959), we propose that social self-awareness will have the strongest effect on trust when observers' own outcomes are more dependent on the target's behavior. We test this possibility in our final study.

Overview of Current Research

Across seven studies, we test whether exhibiting social self-awareness, henceforth referred to as “self-awareness,” affects trust differently depending on whether the target’s behavior has a positive or negative impact on others, due to the fact that self-awareness signals greater intentionality. In Studies 1-4, we test whether targets who exhibit high self-awareness are trusted more when engaging in positive behaviors and trusted less when engaging in negative behaviors. We also test whether this effect is a function of perceived intentionality (Studies 1, 3, and 4). In Studies 5-7, we test theory-driven moderators, predicting that self-awareness will have less of an impact on trust when inferences of negative intentions toward others are reduced (i.e., when the target’s behavior is perceived to be low in mutability or does not have any impact on others). For all studies, we predetermined our sample size based on expected effect sizes. We report all data exclusions, all manipulations, and all measures in each study. All studies were preregistered, and our study materials, data, and preregistrations can be found on OSF: https://osf.io/f6cr9/?view_only=3f38d8736ab543b6acd5a3a729db5ef5.

Study 1: Social Self-Awareness is Good, but Only if You’re Nice to Me!

In Study 1, we tested whether perceived self-awareness affects trust differently depending on the valence of the target’s behavior, and whether this effect is driven by perceptions of intentionality. We introduced participants to a hypothetical target person and manipulated whether this target was friendly or unfriendly toward the participant. We also manipulated whether the target was high or low in self-awareness, and then measured trust toward the target. We predicted an interaction between self-awareness and valence of behavior, such that high self-awareness would yield higher trust when the target’s behavior was positive (friendly), but lower trust when the target’s behavior was negative (unfriendly). We predicted that these effects would

| Study | Sample size | Method | IV(s) | DV(s) | Main finding |
|-------|-------------|--|---|--------------------------------------|--|
| 1 | 399 | Scenario (Prolific) | Self-awareness (high vs. low); Behavior valence (positive vs. negative) | Trust; Intentionality (mediator) | Participants trusted a target more if they had high (vs. low) self-awareness while being friendly toward the participant, but the reverse was true if the target was being unfriendly toward the participant. |
| 2 | 1,455 | Scenario (Prolific) | Self-awareness (high vs. low); Behavior valence (positive vs. negative) | Desire to spend time with; Liking | Participants most wanted to spend time with, and most liked, a target who was aware (vs. unaware) of being rated positively on a given trait, but the reverse was true if the target was aware of being rated negatively on a given trait. |
| 3 | 307 | Recall (MBA students and downtown Chicago community) | Self-awareness (high vs. low) | Trust; Benevolence (mediator) | When recalling real workplace colleagues, participants trusted the colleague less if the colleague seemed aware of being rude toward them than if they seemed unaware. |
| 4 | 397 | Lab (Downtown Chicago community) | Self-awareness (high vs. low); Behavior valence (positive vs. negative) | Trust; Intentionality (mediator) | Participants trusted a confederate more if they had high (vs. low) self-awareness of being a good listener to the participant, but the reverse was true if the confederate was being a bad listener. |
| 5 | 1,203 | Scenario (Prolific) | Self-awareness (high vs. low); Mutability of behavior (high vs. low) | Trust; Benevolence (mediator) | Participants trusted a target less if they had high (versus low) self-awareness of being too talkative/unsympathetic/unappreciative, but this difference attenuated when the target was unable to change their behavior. |
| 6 | 1,603 | Scenario (Prolific) | Self-awareness (high vs. low); Mutability of behavior (high vs. low) | Trust; Benevolence (mediator) | Participants trusted a target less if they had high (versus low) self-awareness of blocking the participant's view or taking up their seat space, but this difference attenuated when the target was unable to change their behavior. |
| 7 | 800 | Scenario (Prolific) | Self-awareness (high vs. low); Impact on others (impact vs. no impact) | Trust; General intentions (mediator) | Participants trusted a target less if they had high (versus low) self-awareness of slacking off on a joint project, but this difference attenuated when the participant's outcomes did not depend on the target's performance. |

Table 2.1. Summary of studies and main findings in Chapter 2.

be driven by the target's perceived intentionality. This study's hypotheses and design were pre-registered at https://aspredicted.org/VGV_1T9.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$1.25. We preregistered that we would collect 400 participants after excluding those who failed the attention check or comprehension checks and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 419 participants who started the survey, we ended up with a final sample of 399 participants (58.15% female, 1.75% other gender, 18.05% non-White, $M_{\text{age}} = 36.01$, $SD_{\text{age}} = 13.10$) that fit these criteria.

Procedure. Participants read that they would learn about Taylor, one of their coworkers. We randomly assigned participants to one of four conditions in a 2 (Target self-awareness: high vs. low) x 2 (Target behavior valence: positive vs. negative) between-subjects design. We counterbalanced the gender of Taylor for stimulus sampling. In all conditions, participants read that whenever they interacted with Taylor, they tried to be friendly toward him/her. In the positive valence conditions, participants read that Taylor seemed to be very friendly back to them, listening carefully to the participant and responding enthusiastically and warmly. In the negative valence conditions, participants read that Taylor seemed to be very unfriendly toward them, not listening carefully to the participant and responding dismissively and coldly.

On the next page, all participants learned additional information about Taylor. In the high self-awareness conditions, participants read that Taylor had very high social self-awareness, meaning that (s)he is very good at knowing how (s)he is perceived by other people and how his/her actions toward others would affect them. They further read that everyone in the office had recently taken a "Social Self-Awareness Skills" assessment, which measured how good one is at knowing how one is perceived by other people, and that Taylor had scored a 90/100, indicating

that (s)he is very accurate in knowing how (s)he is perceived by others. In the low self-awareness conditions, we changed the text accordingly: Participants read that Taylor had very low social self-awareness and was thus bad at knowing how (s)he is perceived by other people, and had scored a 10/100 on the skills assessment.

On the next page, participants responded to two comprehension checks about the information they had read. If they did not answer both questions correctly after the second try, the study automatically ended. Otherwise, participants proceeded to respond to a series of pages with our manipulation checks and dependent measures. The two pages with the manipulation checks, mediator measure, and main dependent measures were presented first, with the order randomized between the two pages and with the order of questions on each page randomized. After that, participants responded to three pages of exploratory measures with the pages presented in random order, and the question order randomized on each page.

We asked the following two questions as manipulation checks: “How aware or unaware do you think Taylor is that you perceive [him/her] as [un]friendly?” (-3 = *extremely unaware*, 3 = *extremely aware*) and “To what extent do you believe Taylor’s behavior toward you has a negative or positive impact on you?” (-3 = *extremely negative*, 3 = *extremely positive*). On the same page, we also asked the following to measure our hypothesized mediator, the target’s perceived intentionality: “To what extent do you think Taylor is intentionally being [un]friendly toward you?” (1 = *not at all intentionally*, 7 = *extremely intentionally*).

On a separate page, participants responded to the following three questions, intended to measure overall trust (all -3 = *strongly disagree*, 3 = *strongly agree*): “In general, I would trust Taylor”; “I feel like I could rely on Taylor when it comes to matters that are important to me”; and “I consider Taylor to be an untrustworthy person” (reverse-coded).

We also included several additional measures, preregistered as exploratory, to capture additional perceptions of Taylor, which we report in the Supplemental Material for brevity (see Appendix B). Results on these measures were consistent with those on our other measures, and with the results of subsequent studies that included some of these measures.

Participants reported demographic information (gender, age, race) and were allowed to provide any optional feedback on the study in a free-response text box.

Results

Manipulation checks. As intended, participants perceived Taylor to be more aware that the participant perceived him/her as being (un)friendly in the high self-awareness condition ($M = 2.35$, $SD = 0.77$) compared to the low self-awareness condition ($M = -1.71$, $SD = 1.32$), $t(397) = 37.66$, $p < .001$. In addition, participants perceived Taylor's behavior as having a more positive impact on them in the positive valence condition ($M = 1.73$, $SD = 0.96$) compared to the negative valence condition ($M = -1.47$, $SD = 0.82$), $t(397) = -35.87$, $p < .001$.

We conducted two-way ANOVAs with self-awareness condition, valence condition, and their interaction as factors on our measure of intentionality and our composite of trust, respectively.

Intentionality. As expected, there was a main effect of self-awareness condition, $F(1, 395) = 201.35$, $p < .001$, $\eta_p^2 = .34$, such that participants perceived Taylor's behavior as more intentional in the high self-awareness condition than the low self-awareness condition. Unexpectedly, we also observed a main effect of valence condition, $F(1, 395) = 35.07$, $p < .001$, $\eta_p^2 = .08$, and a significant interaction, $F(1, 395) = 60.45$, $p < .001$, $\eta_p^2 = .13$, such that there was a greater difference in perceived intentionality between self-awareness conditions when the

behavior was negative ($M_{high} = 5.58, SD_{high} = 1.30; M_{low} = 2.66, SD_{low} = 1.11$) than when it was positive ($M_{high} = 5.33, SD_{high} = 1.21; M_{low} = 4.48, SD_{low} = 1.65$).

Overall trust. As preregistered, we combined our three trust measures into a composite given that they loaded together highly ($\alpha = 0.94$). Means for the trust composite in each condition are depicted in Figure 2.1. There was no main effect of self-awareness condition, $F(1, 395) = 0.73, p = .393, \eta_p^2 < .01$, but there was a main effect of valence condition, $F(1, 395) = 836.83, p < .001, \eta_p^2 = .68$. Most importantly, there was a significant Self-awareness x Valence interaction, $F(1, 395) = 19.50, p < .001, \eta_p^2 = .05$, as hypothesized. When Taylor was friendly, participants trusted Taylor more when Taylor had high ($M = 1.80, SD = 0.90$) compared to low ($M = 1.29, SD = 0.97$) self-awareness, $t(395) = 3.73, p < .001, 95\% CI = [0.24, 0.78], d = 0.53$, but when Taylor was unfriendly, participants trusted Taylor *less* when Taylor had high ($M = -1.44, SD = 1.02$) compared to low ($M = -1.09, SD = 0.98$) self-awareness $t(395) = -2.51, p = .012, 95\% CI = [-0.62, -0.08], d = -0.36$.

Moderated mediation. Next, we tested whether intentionality mediated the effect of self-awareness condition on trust—and whether it did so differently for positive versus negative behaviors—by conducting a moderated mediation model with self-awareness condition as the independent variable, intentionality as the mediator, valence condition as the moderator, and trust as the outcome, using 5,000 bootstrapped samples. Consistent with our hypotheses, we observed significant moderated mediation, $b = -0.31, 95\% CI = [-0.56, -0.049], p = .020$, such that perceived intentionality mediated the effect of self-awareness condition on trust in opposite directions for positive, $b = -0.03, 95\% CI = [-0.13, 0.05], p = .423$, versus negative, $b = 1.24, 95\% CI = [0.90, 1.58], p < .001$, target behaviors.

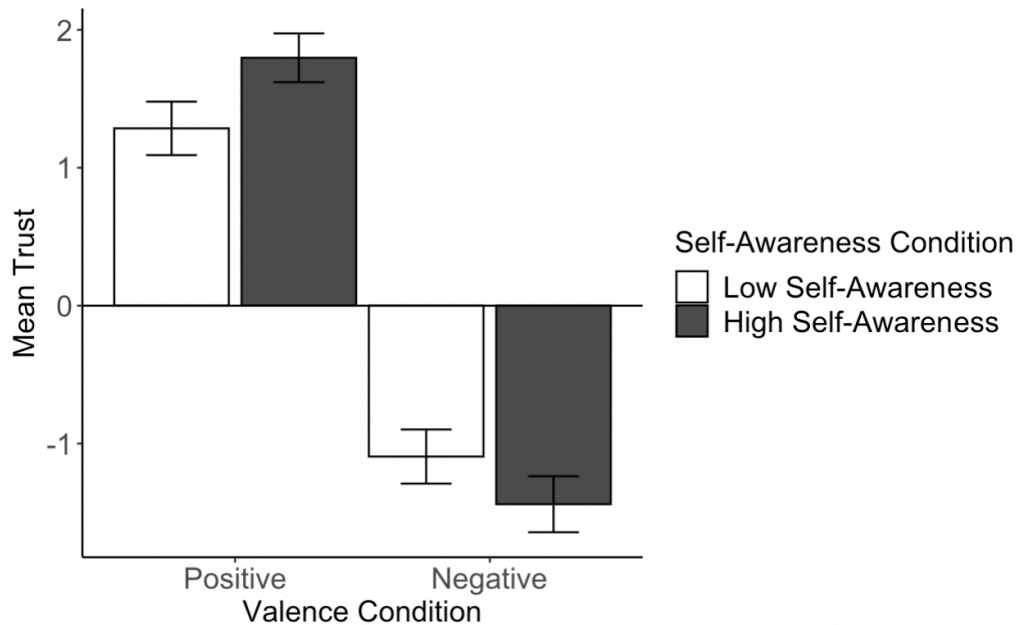


Figure 2.1. Mean trust in Study 1, Chapter 2. Error bars represent 95% confidence intervals.

Discussion

Study 1 provides initial evidence that self-awareness does not always increase trust; instead, it can decrease trust if the target’s behavior has a negative impact on others, as this behavior is seen as more intentional, and thus, as more of a diagnostic signal of the target’s character.

Study 2: The Price Rating is Right

In Study 2, we sought to replicate the findings from Study 1 with a manipulation of self-awareness of the target’s particular behavior, rather than general self-awareness. We varied one hypothetical person’s impression of a hypothetical target on a given trait, and then manipulated self-awareness by varying how accurate that target was at guessing what the other person thought of them. We used a more behaviorally oriented dependent measure—desire to spend time with the target—as an indicator of trust. We measured general liking as well, since liking may at least in part be driven by perceived trustworthiness. As before, we predicted that greater self-

awareness of positive traits would lead to greater desire to spend time with, and liking of, the target person compared to lower self-awareness, but that the reverse would be true for negative traits.¹ This study's hypotheses and design were preregistered at

https://aspredicted.org/RTH_6DH.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$0.75. We preregistered that we would collect 1,440 participants after excluding those who failed the attention check or comprehension checks, and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 1,848 participants who started the survey, we ended up with a final sample of 1,455 participants (48.73% female, 1.44% other gender, 26.53% non-White, $M_{\text{age}} = 35.20$, $SD_{\text{age}} = 12.36$) that fit these criteria.

Procedure. Participants read that they would be introduced to two individuals, Jordan and Taylor, who were acquaintances and had known each other for a little while. We randomized the gender of Jordan and Taylor (both male or both female) for stimulus sampling, and also randomized which of the two names went with which role. In the versions where Jordan was the target for participants to evaluate, participants read that Taylor had been asked to privately record their impression of Jordan as part of a research study; specifically, Taylor had rated Jordan on a specific trait (on a scale of 1 to 7). We randomly assigned participants to view one of eight traits (between-subjects) for stimulus sampling: friendly, kind, trustworthy, honest, competent, arrogant, lazy, and irritable. We chose these traits as those that could possibly have a positive or negative impact on the other person.

Next, participants read that Jordan had separately been asked to privately *predict* how they believed Taylor had rated them on the given trait (on the same 1-7 scale), and to try to guess

¹ This prediction was different from our preregistered prediction at the time, as we revised our theorizing based on the results.

as accurately as possible. Participants were told that they would view both Taylor's actual rating of Jordan on the given trait and Jordan's guess about this rating, and that we would then ask participants what they thought about Jordan (the guesser).

On the next page, participants viewed two images. One image showed Taylor's rating of Jordan on the trait on a Likert scale (1 = *not at all*, 7 = *extremely*). The second showed Jordan's *guess* of Taylor's rating on the exact same Likert scale. See Figure 2.2 for an example image of these stimuli.

We randomly assigned participants to one of six conditions in a 3 (Target self-awareness: high vs. moderate vs. low) x 2 (True rating valence: positive vs. negative) between-subjects design. In the positive true rating conditions, Taylor rated Jordan as a 7 ("extremely") if the trait was friendly, kind, trustworthy, honest, or competent (i.e., a desirable trait), or rated Jordan as a 1 ("not at all") if the trait was arrogant, lazy, or irritable (i.e., an undesirable trait). In the negative true rating conditions, Taylor rated Jordan as a 1 ("not at all") if the trait was friendly, kind, trustworthy, honest, or competent, or rated Jordan as a 7 ("extremely") if the trait was arrogant, lazy, or irritable. Each participant saw only one of these traits and ratings. We pretested these traits ($N = 75$) to determine that people viewed a 1 on all of these traits as approximately equivalently positive/negative as a 7 on the same trait (e.g., a 1 on friendliness is about as bad as a 7 is good, while a 1 on laziness is about as good as a 7 is bad). We also confirmed that people perceived each trait's valence as intended.

In the high target self-awareness conditions, Jordan's guess of Taylor's rating was exactly the same as Taylor's actual rating (e.g., a 7 if Taylor had rated them a 7). In the moderate self-awareness conditions, Jordan's guess was always a 4 (so it was either three scale points too high if the true rating was a 1, or three scale points too low if the true rating was a 7). In the low

self-awareness conditions, Jordan’s guess was at the opposite end of the scale from the true rating (a 1 if the true rating was a 7 and a 7 if the true rating was a 1). Thus, our manipulation of self-awareness manipulated how accurate the target was at knowing what another person thought of him/her.

First: Here is **Taylor’s ACTUAL rating** of how **friendly** he thought **Jordan** was (on a 1-7 scale):

| | | | | | | |
|----------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Not at all 1 | 2 | 3 | Moderately 4 | 5 | 6 | Extremely 7 |
| <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Second: Here is **Jordan’s GUESS** as to how **friendly Taylor had rated him** as being (on a 1-7 scale):

| | | | | | | |
|----------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| Not at all 1 | 2 | 3 | Moderately 4 | 5 | 6 | Extremely 7 |
| <input checked="" type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 2.2. Example rating stimulus shown in Study 2, Chapter 2.

On the next page, we included a reminder of these rating images, and then asked participants our two key dependent measures: “Based on all of the information we have given you, how much would you want to spend time with Jordan if you had the opportunity to do so in real life?” (1 = *not at all*, 7 = *a lot*) and “Based on all of the information we have given you, how much do you like Jordan?” (-3 = *dislike very much*, 3 = *like very much*). Both questions included a reminder that Jordan was the one who had guessed Taylor’s rating of them. We included one comprehension check on the same page, asking whether Jordan had provided the original rating or guessed the other person’s rating. Finally, participants provided demographic information

(gender, age, race). In the other versions, we reversed the names Taylor and Jordan across the entire study.

Results

Valence condition describes each “actual” (observer) rating in terms of whether it is positive or negative. “Positive” ratings include both high ratings on desirable traits (e.g., high on friendliness) and low ratings on undesirable traits (e.g., low on arrogance). Negative ratings include ratings of the reverse pattern (e.g., low on friendliness, high on arrogance). We collapsed across traits and conducted mixed models with valence condition, self-awareness condition, and their interaction as factors on each of our dependent variables (desire to spend time with and liking), with random effects for trait.²

We observed main effects of self-awareness condition and trait valence condition on both desire to spend time with (self-awareness: $F(2,1449) = 13.24, p < .001, \eta_p^2 = .02$; trait valence: $F(2,1449) = 759.86, p < .001, \eta_p^2 = .35$) and liking (self-awareness: $F(2,1449) = 15.60, p < .001, \eta_p^2 = .02$; trait valence: $F(2,1449) = 681.28, p < .001, \eta_p^2 = .32$).

Most importantly, in line with our hypotheses—and replicating the effects of Study 1—there was a significant interaction between self-awareness condition and trait valence on our measures of both desire to spend time with, $F(2,1449) = 74.06, p < .001, \eta_p^2 = .093$, and liking, $F(2,1449) = 47.01, p < .001, \eta_p^2 = .061$, suggesting that self-awareness of a trait affects judgments differently depending on whether the trait rating is positive or negative. For positive trait ratings, participants most wanted to spend time with the target who had high self-awareness ($M = 5.38, SD = 1.34$), significantly more than both moderate self-awareness ($M = 4.68, SD =$

² We report the results of ANOVAs on these mixed models here for ease of interpretation, though we did not specify using ANOVAs in our preregistration. Results remain unchanged whether we analyze the mixed models or ANOVAs on the mixed models.

1.26), $t(1444.93) = 5.78, p < .001, 95\% \text{ CI} = [0.43, 1.02], d = 0.54$, and low self-awareness ($M = 3.89, SD = 1.37$), $t(1444.09) = 11.85, p < .001, 95\% \text{ CI} = [1.19, 1.78], d = 1.11$. See Figure 2.3 for a depiction of the means of “desire to spend time with.” Similarly, liking was highest when the target had high self-awareness ($M = 1.51, SD = 1.27$), significantly higher than both moderate self-awareness ($M = 1.01, SD = 1.09$), $t(1444.59) = 4.63, p < .001, 95\% \text{ CI} = [0.25, 0.77], d = 0.43$, and low self-awareness ($M = 0.36, SD = 1.17$), $t(1443.83) = 10.40, p < .001, 95\% \text{ CI} = [0.89, 1.41], d = 0.97$.

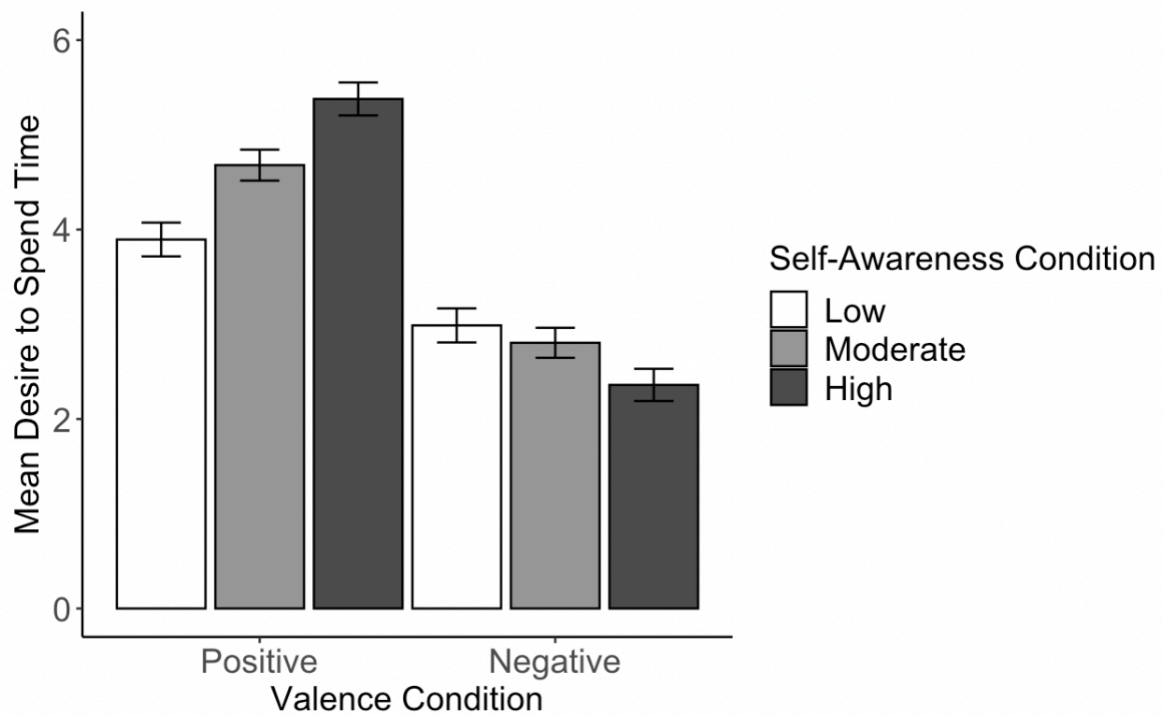


Figure 2.3. Mean desire to spend time with the target in Study 2, Chapter 2. Error bars represent 95% confidence intervals.

For negative trait ratings, however, we found the opposite pattern. Participants *least* wanted to spend time with the target who had high self-awareness ($M = 2.36, SD = 1.40$), significantly less than both moderate self-awareness ($M = 2.80, SD = 1.29$), $t(1443.41) = -3.65, p < .001, 95\% \text{ CI} = [-0.71, -0.15], d = -0.32$, and low self-awareness ($M = 2.99, SD = 1.44$),

$t(1444.96) = -5.14, p < .001, 95\% \text{ CI} = [-0.89, -0.33], d = -0.46$ (with no difference between moderate and low self-awareness, $p = .284$). Similarly, liking was *lowest* when the target had high self-awareness ($M = -0.87, SD = 1.28$), significantly lower than both moderate self-awareness ($M = -0.57, SD = 1.09$), $t(1443.23) = -2.74, p = .017, 95\% \text{ CI} = [-0.53, -0.04], d = -0.24$, and low self-awareness ($M = -0.53, SD = 1.25$), $t(1444.60) = -3.10, p = .006, 95\% \text{ CI} = [-0.57, -0.08], d = -0.27$, (with no difference between moderate and low self-awareness, $p = .926$).

Discussion

Study 2 provides further evidence that self-awareness can decrease trust-driven behavioral intentions toward the target when the target is aware of being perceived negatively by others. One limitation of this study is that we did not directly measure the perceived intentionality of the target. Thus, we cannot rule out that our effects are driven by other inferences that participants might make about the target's behavior. For instance, participants might have believed that targets who were rated negatively but displayed low social self-awareness (i.e., thought they were perceived positively) actually *did* behave slightly more positively than those who were rated negatively and knew it. In all of our other studies, we include measures of perceived intentionality to overcome this limitation.

Study 3: Recalling Non-Collegial Colleagues

In Study 3, we further tested our theory of self-awareness as a signal of intentionality using a recall study design. We focused only on negatively valenced behavior in this study. Further, we measured ability, benevolence, and integrity as antecedents to trust (Mayer et al., 1995), predicting that benevolence would be most predictive of trust because it represents the target's perceived positive intentions toward others. Participants were asked to think of someone they knew from a previous or current professional experience who was rude or unfriendly toward

them and seemed to have either high or low self-awareness of their behavior (between-subjects). We predicted that participants assigned to think of someone with high self-awareness of their negative behavior would trust the target person less than those assigned to think of someone with low self-awareness of their negative behavior. This study's hypotheses and design were pre-registered at https://aspredicted.org/1ZP_CJM.

Participants. Participants were recruited from two sources that allowed us to capture a sample of adults with diverse full-time work experiences: Some were recruited from MBA classes ($n = 68$) and the rest were recruited from a public behavioral science museum and laboratory run by a university in a large Midwestern city ($n = 239$). Participants in the MBA class completed the study in exchange for entry into a raffle to win a \$25 Amazon gift card, while participants from the public behavioral laboratory were compensated with points that are redeemable for prizes at the laboratory. We preregistered that we would collect 300 participants, after excluding those who were unable to recall a target person according to the instructions we gave them. Of the 353 participants who started the survey (across both samples), we ended up with a final sample of 307 participants (49.19% female, 0.01% other gender, 42.35% non-White, $M_{\text{age}} = 32.37$, $SD_{\text{age}} = 10.87$) who fit these criteria.

Procedure. Participants were asked to think of someone from a previous or current work experience (or another professional setting, like school) who had generally been unfriendly or rude toward the participant.³ Participants were randomly assigned to either the high or low self-awareness condition. In the high self-awareness condition, participants were asked to think of

³ Before we instructed participants to choose a target, we asked an exploratory question about how mutable (i.e., easily changeable) participants perceived unfriendliness or rudeness to be in general. Based on the theory, the impact of social self-awareness of a negative trait on trust should be moderated by the perceived mutability of that negative trait. However, we did not observe significant results on this measure in this study (reported in the Supplement). We suspect that this null result may have to do with the correlational nature of the data in this study. In Studies 6-7, in which we manipulated perceived mutability, we did find the expected moderation effect.

someone who they believed was *aware* of being generally unfriendly or rude toward the participant. In the low self-awareness condition, participants were asked to think of someone who they believed was *unaware* of being unfriendly or rude toward the participant. In both conditions, we gave participants some examples of unfriendly or rude behaviors, such as ignoring the participant, excluding them, or being critical. The examples were the same in both self-awareness conditions.

On the next page, we asked participants to write the initials of the person they had thought of in an open text box, or if they were unable to think of anyone who fit this description, they could check a box saying so. If they checked the box, the next page of the survey notified the participant that they would be unable to proceed but would be compensated the full amount for their participation. Thirty-five participants were unable to think of a target and are excluded from all analyses.

Participants who could think of a target proceeded to the next page and responded to two questions about the target they named: “[target initials] would best be described as your...” (multiple choice: *colleague, boss, subordinate, other [explain]*) and “How well do you know [target initials]?” (1 = *not well at all*, 7 = *extremely well*).

On the next several pages, participants responded to our dependent measures and manipulation checks, with both the pages and the questions within each page presented in random order. On one page, participants responded to our manipulation check measures: “How aware or unaware do you think [target initials] was that (s)he was being unfriendly or rude toward you?” (-3 = *extremely unaware*, 3 = *extremely aware*) and “To what extent did [target initials]’s unfriendliness or rudeness have a negative impact on you?” (1 = *not at all*, 7 = *a lot*).

This latter question was intended to ensure that participants across both conditions perceived the target's behavior as having a negative impact on them (i.e., above the midpoint of the scale).⁴

On a separate page, participants answered a series of questions measuring their perceptions of the target's benevolence, ability, and integrity (i.e., antecedents to trust). Specifically, we asked participants: "Please rate your (dis)agreement with the following statements: [target initials] is..."; then, participants responded to a list of trait words and were asked to indicate how much they agreed with each (-3 = *strongly disagree*, 3 = *strongly agree*). To measure benevolence, we used the following traits: "benevolent," "kind," "good-natured," and "well-intended." We treated the benevolence scale as another way of capturing perceived intentionality—in this case, the target's general positive intentions toward others.

To measure ability, we used the following: "competent," "resilient," "perseverant," and "high in self-control." To measure integrity, we used the following: "honest," "authentic," "sincere," and "truthful." We preregistered the ability and integrity measures as exploratory because we hypothesized that self-awareness would be most relevant to perceived benevolence as an indicator of intentions toward others, which in turn would have the largest effect on trust. On a separate page, participants responded to the exact same three questions measuring overall trust as in Study 1.

On a separate page, participants also responded to a set of exploratory measures of overall workplace trust, adapted from previous research (Mayer et al., 1995) (all -3 = *strongly disagree*, 3 = *strongly agree*): "I would be comfortable giving [target initials] a task or problem even if I could not monitor [his/her] actions"; "If I had my way, I wouldn't let [target initials]

⁴ We also included an exploratory measure of the perceived mutability of the target's specific behaviors, which we report in the Supplement, along with the exploratory measure of the perceived general mutability of those behaviors noted in the previous footnote.

have any influence over issues that are important to me”; “I would be willing to let [target initials] have complete control over my future career”; and “I really wish I had a good way to keep an eye on [target initials].”

Finally, participants in the MBA class sample provided demographic information (gender, age, race, number of years of work experience) and were optionally allowed to provide any feedback on the study in a free-response space. In the public behavioral laboratory sample, participants had already provided demographic information in a separate survey. Participants in both samples then read a debrief about the study purpose, and were able to enter their email address to be considered for the raffle (MBA sample) or were given their points for completing the study (public behavioral laboratory sample).

Results

Manipulation checks. As intended, participants perceived the target as more aware of their behavior in the high self-awareness condition ($M = 1.47$, $SD = 1.33$) than in the low self-awareness condition ($M = -0.76$, $SD = 1.54$), $t(305) = 13.66$, $p < .001$, and also perceived the target’s behavior as having a negative impact on them across both conditions (significantly above the midpoint of the scale; $M = 4.48$, $SD = 1.72$), $t(306) = 4.92$, $p < .001$.

As preregistered, we combined our benevolence measures ($\alpha = 0.82$) and our overall trust measures ($\alpha = 0.83$) into their respective composites given that they loaded highly together.

Benevolence. As expected, targets who seemed higher in self-awareness of their negative behavior were perceived as *less* benevolent ($M = -0.87$, $SD = 1.29$) than targets who seemed

lower in self-awareness ($M = -0.40$, $SD = 1.12$), $t(305) = -3.43$, $p < .001$, 95% CI = [-0.74, -0.20], $d = -0.39$.⁵

Overall trust. Further, targets who seemed higher in self-awareness were trusted *less* ($M = -1.20$, $SD = 1.47$) than targets who seemed lower in self-awareness ($M = -0.43$, $SD = 1.48$), $t(305) = -4.56$, $p < .001$, 95% CI = [-1.10, -0.44], $d = -0.52$.

Mediation. To determine whether perceptions of benevolence—an indicator of intentions—were driving the effect of self-awareness condition on trust, we ran a mediation model with self-awareness condition as the independent variable, the benevolence composite as the mediator, and the overall trust composite as the dependent variable. As hypothesized, we observed significant mediation, $b = 0.36$, 95% CI = [0.15, 0.56], $p < .001$.

Exploratory measure: Workplace trust composite. The alpha on our composite of workplace trust was relatively low ($\alpha = 0.51$); nevertheless, the results on this measure paralleled those of the overall trust measure, such that targets who seemed higher in self-awareness were trusted less ($M = -0.95$, $SD = 1.00$) than targets who seemed lower in self-awareness ($M = -0.60$, $SD = 0.99$), $t(305) = -3.07$, $p = .002$, 95% CI = [-0.57, -0.12], $d = -0.35$.

Exploratory measures: Ability and integrity. Interestingly, targets who seemed higher in self-awareness were also perceived as lower in ability, $t(305) = -2.68$, $p = .008$, 95% CI = [-0.66, -0.10], $d = -0.31$, and integrity, $t(305) = -3.23$, $p = .001$, 95% CI = [-0.87, -0.21], $d = -0.37$, than targets who seemed low in self-awareness.

Discussion

⁵ We preregistered that we would conduct a regression predicting benevolence and (separately) trust from self-awareness condition, but in hindsight, a simple t-test seemed more sensible. Results on the regressions confirm the results of the t-tests (benevolence: $b = 0.47$, $t(305) = 3.43$, $p < .001$; overall trust: $b = 0.77$, $t(305) = 4.56$, $p < .001$).

Study 3 extends our findings to real relationships, again showing that people perceive those with high self-awareness of negative behaviors as *less* trustworthy than those with low self-awareness of those behaviors. While these perceptions were driven at least in part by perceived benevolence—which represents beliefs about the target’s general positive intentions toward others—we also found effects on perceived ability and integrity, possibly due to a halo effect. In Studies 5-6, however, we find that our effects are driven by benevolence only, suggesting that the effect of self-awareness on benevolence is more consistent than that on ability and integrity.

Study 4: If You’re Not Going to Listen, At Least be Unaware

Study 4 sought to test our hypotheses within the context of live interactions. Participants were told they would take part in a pair study, during which they would tell a story to another participant, the “target.” In reality, the target was a confederate, and was instructed to display good or bad active listening skills while the participant told their story in order to manipulate the valence of the target’s behavior. To manipulate the target’s self-awareness, we varied whether the confederate was accurate or inaccurate in guessing what the participant thought of them. As in the previous studies, we predicted that higher self-awareness would increase trust when the target behaved positively (was a good active listener), but that the reverse would be true when the target behaved negatively (was a poor active listener). This study’s hypotheses and design were preregistered at https://aspredicted.org/KFZ_TMh.

Participants. Participants were recruited from the same public behavioral science museum and laboratory as in Study 3 and were compensated 400 points in base pay, plus 100 bonus points as described below (points are redeemable for prizes at the museum). We preregistered that we would collect 400 participants after excluding those for whom the study procedure was significantly interfered with. We excluded one participant who was reported by

the research assistant as being non-compliant with the survey procedures (e.g., clicking random buttons in the survey) and one participant who asked the confederate if she was a confederate during the conversation, to which the confederate responded in the affirmative. Sixteen participants were run in a non-randomly assigned valence condition, either because the confederate switched from the negative to positive condition to avoid undue distress if the participant started telling a truly heavy and emotional story (15 participants) or because the confederate mistakenly switched from the positive to negative condition (1 participant). We preregistered that we would run analyses with and without these non-randomly assigned participants, and since results remain unchanged, we include these participants in our sample. We ended up with a final sample of 397 participants (61.52% female, 3.80% other gender, 50.13% non-White, $M_{\text{age}} = 32.19$, $SD_{\text{age}} = 13.47$).

Procedure. See Figure 2.4 for an overview of the study procedure. Participants were told that they would be participating in a pair study. To begin, a research assistant led them to a room by themselves with a computer, in which they completed the first part of the survey on their own. The survey explained that they would participate in an exercise with the other participant, and that one of them would be assigned to the role of “storyteller” while the other would be assigned to the role of “listener.” In reality, all participants found out on the next page that they had been assigned to the role of “storyteller,” and that they would have three minutes to tell the “listener” about a time when they faced something challenging. We chose this topic so that participants would speak about something that had at least some personal significance to them. Participants were asked to jot down 1-2 sentences about what they planned to share with the other participant.

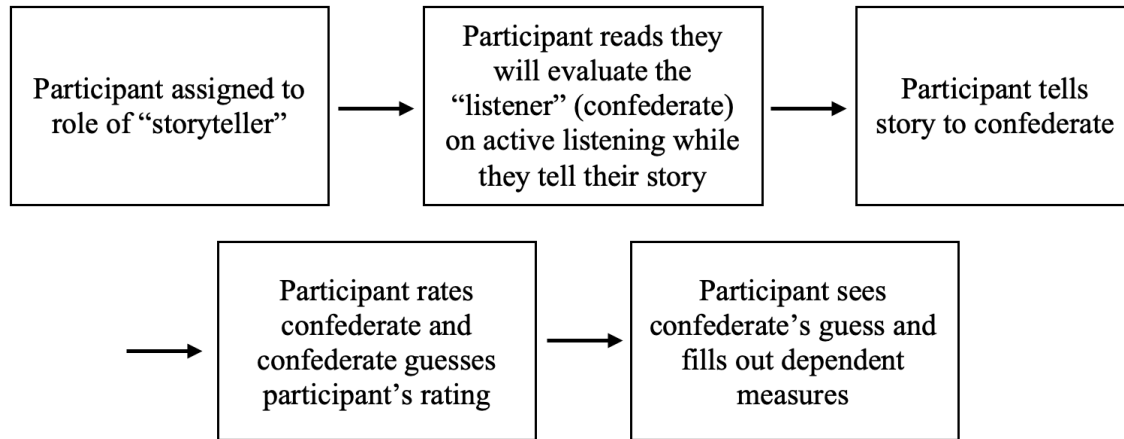


Figure 2.4. Overview of study procedures in Study 4, Chapter 2.

Active listening explanation. On the next page, participants read that they would be asked to evaluate the “listener” on their active listening skills while the participant shared their story. They then read specific information about what active listening is (instructions were taken directly from an online resource: <https://asana.com/resources/active-listening>). Specifically, participants read that the goal of active listening was to make them feel heard, valued, and understood while they shared something personal about themselves. They also read about specific behaviors that are involved in active listening, such as asking open-ended questions, being non-judgmental, offering positive nonverbal communication, and avoiding distractions. See OSF for the survey with the full explanation that participants saw about active listening. To ensure that participants had paid attention to this information, we asked them to summarize active listening and why it is important in 2-3 sentences. In addition, we told participants that the other participant would not be given any specific instructions on active listening prior to the interaction (in order to make it more believable that the target could truly be unaware of how good or bad at active listening they were).

On the next page, we told participants that active listening is generally easy for most people to implement when they try to, and then asked them the following: “Based on this, how easy or difficult do you think it will be for the other participant to change how well they are actively listening during your story (if they try to)?” (-3 = *extremely easy*, 3 = *extremely difficult*). We wanted to ensure that participants would perceive active listening as something that would be easy for the target to implement in order to make it more likely that they would infer intentionality behind the target’s behaviors, as described in the Introduction; we return to this notion in Studies 5-6.

Next, participants were given a reminder of what they planned to tell the other participant and that they should evaluate the other participant’s active listening abilities. They were instructed to notify the experimenter, who then brought the confederate into the room with the real participant. The experimenter gave a reminder of the instructions and asked if there were any questions before leaving the participant and the confederate alone in the room for three minutes, during which the participant told their story.

Valence manipulation: Confederate’s listening skills. Participants were randomly assigned to one of four conditions in a 2 (Self-awareness condition: high vs. low) x 2 (Target behavior valence: positive vs. negative) design. To manipulate the valence of the target’s behavior, we manipulated how good or bad at active listening the confederate was while the participant shared their story. In the positive conditions, the confederate demonstrated good active listening skills by using positive nonverbal cues (e.g., eye contact, empathetic facial expressions), focusing on the other person’s story and avoiding distractions, and asking follow-up questions to the participant’s story. In the negative conditions, the confederate demonstrated poor active listening skills by avoiding positive nonverbal cues, appearing disengaged and

distracted, and failing to ask follow-up questions. For snippets of video examples of each of these conditions, see our OSF page. We used three different confederates (two female, one male) for stimulus sampling across study sessions.

Self-awareness manipulation: Confederate's guess of participant's rating. Once the three minutes were up, the experimenter came back into the room and handed each participant a piece of paper. The experimenter explained that the participant was being asked to rate the other participant (i.e., confederate) on their active listening skills, and that the confederate was being asked to *guess* how the other participant would rate them on their active listening skills. The participant's piece of paper asked the following question: "During the 3 minutes you were given to share your experience, how good at active listening did you think the other participant was overall?" (1 = *not very good*, 7 = *extremely good*). This question served partly as a manipulation check. The confederate's piece of paper asked: "How good at active listening do you think the other participant will rate you as being overall, during the 3 minutes in which they were sharing their experience?" and used the same scale as the participant's question. Both pieces of paper included an explanation of what active listening is (for the real participant, this was presented as a reminder of the instructions they had already seen). We had the participant and the confederate fill out their sheets while in the same room so that it would not seem like the confederate had time to reflect on their behavior and perhaps become more self-aware later on than they had been in the moment.

After the participant and confederate had each filled out their sheet (on clipboards to allow for more privacy), the experimenter led the confederate out of the room. A few moments later, the experimenter returned with the confederate's piece of paper, which they handed to the participant.

To manipulate target self-awareness condition, we manipulated whether the confederate's guess was accurate or inaccurate relative to the participant's actual rating. The confederate's guess was always either a 2 (i.e., a poor rating) or a 6 (i.e., a good rating) on the 7-point scale. In the high self-awareness conditions, the confederate's rating was a 2 if in the negative condition or a 6 if in the positive condition, thus (most of the time) being either exactly matched to the real participant's rating or at least being on the same side of the scale as it. In the low self-awareness conditions, the confederate's rating was a 6 if in the negative condition or a 2 if in the positive condition, thus (most of the time) being on the opposite side of the scale from the participant's true rating.

Once the participant had viewed the confederate's guess, the experimenter instructed them to proceed with the computer survey, which asked them to enter both their own rating and the other participant's guess from their respective pieces of paper; this was done to ensure that both ratings would be top of mind for the participant as they filled out the rest of the survey. To further reinforce this, we also asked participants to "jot down 1-2 sentences about any thoughts or reactions you have after learning this participant's guess of your rating of them."

Measures. Next, we told participants that they would answer several questions about the other participant, which we told them we would not share with the other participant. On the next page, participants answered the following as manipulation checks: "To what extent do you think the other participant was aware that they were doing a [poor/moderately good/good, piped in based on the participant's actual rating] job at active listening?" (-3 = *extremely unaware*, 3 = *extremely aware*) and "Overall, would you say that the other participant's behavior while you shared your experience had a positive or negative impact on you?" (-3 = *extremely negative*, 3 = *extremely positive*). They answered the following as our mediator: "To what extent do you think

the other participant intended to come across as a [poor/moderately good/good] active listener?” (1 = *not at all intentionally*, 7 = *extremely intentionally*).

On the next page, participants responded to our primary measure of trust: “In general, how trustworthy do you think the other participant is?” (-3 = *extremely untrustworthy*, 3 = *extremely trustworthy*).

Exploratory behavioral measure of trust: Trust game. We also included an exploratory behavioral measure of trust. Participants read that they would play a game with the other participant that would determine their actual bonus payment for the study. The game was set up like a typical trust game: The participant read that they would be designated 50 bonus points, of which they could decide to send anywhere between 0 and 50 of those points to the other participant. Whatever amount they sent would triple in value, and then the other participant would decide how many points (out of the tripled number) to send back to them, thus determining their ultimate bonus payment. Participants responded to three comprehension check questions about the trust game, which they were given two tries to answer correctly. If they did not answer correctly on the second try, we told them the correct answers and allowed them to proceed. On the next page, we asked: “How many (if any) of the 50 points do you choose to send to the other participant?” (slider scale from 0 to 50).

On the following page, we asked participants to predict how many points the other participant would send back to them (out of the tripled number of points that the participant just sent to them), on a slider scale from 0% to 100%. This measure was intended to be an additional and more direct way of capturing the target’s perceived trustworthiness (and particularly benevolence-based trust), as some participants might be inclined to send points to the target out of politeness even if they did not trust the other participant to send them points back. If

participants had initially chosen to send zero points to the other participant, this question was not displayed.

On the next page, we asked two free-response questions to better understand participants' thought processes during the study: "What's the main thing you considered when you chose to send the other participant [number] points? Briefly explain below (in up to one sentence)" and "Overall, what did you think about the other participant in this study? Please jot down any thoughts in the space below."

Finally, participants were taken to a debrief page that explained the deception in the study and told them they would earn a 100-point bonus since they were not playing the trust game with a real participant. Once the participant notified the experimenter that they were finished, the experimenter came in and repeated the key parts of the debrief in case the participant had any questions.

Results

We preregistered that we would run analyses with and without participants who reported a rating of the confederate's behavior that did not correspond to their assigned condition. Participants whose ratings did not correspond to their assigned condition included those who gave a positive rating when the confederate's behavior was supposed to be negative, and vice versa, as well as those who gave a completely neutral rating (at the midpoint of the scale). If the results remained substantively unchanged when excluding these participants, we preregistered we would report the analyses with all participants included; otherwise, we would report both. Results do change somewhat when excluding these participants, so we report both sets of results on all measures below. We refer to our full sample as the "intent-to-treat" sample ($N = 397$) and

the sample with exclusions based on ratings of the confederate as our “successfully treated” sample ($n = 280$).

Manipulation checks. Results on our manipulation checks followed the expected patterns in both samples—see Table 2.2.

| | Intent-to-treat ($N = 397$) | | | Successfully treated ($n = 280$) | | |
|-----------------------|---|--|------------------------------------|---|--|------------------------------------|
| | <u>Condition</u> 1 | <u>Condition</u> 2 | <u>Significance</u> | <u>Condition</u> 1 | <u>Condition</u> 2 | <u>Significance</u> |
| Self-Awareness | Low SA: $M = -0.59,$ $SD = 1.89$ | High SA: $M = 0.86,$ $SD = 1.70$ | $p < .001$ | Low SA: $M = -1.04,$ $SD = 1.78$ | High SA: $M = 1.50,$ $SD = 1.27$ | $p < .001$ |
| Impact | Positive: $M = 1.64,$ $SD = 1.17$ | Negative: $M = -0.17,$ $SD = 1.49$ | $p < .001$ | Positive: $M = 1.89,$ $SD = 0.97$ | Negative: $M = -0.89,$ $SD = 1.17$ | $p < .001$ |
| Mutability | Overall: $M = -0.66, SD = 1.29$ | | $p < .001$ (vs. scale midpoint) | Overall: $M = -0.75, SD = 1.30$ | | $p < .001$ (vs. scale midpoint) |

Scales ranged from -3 to 3 for each measure.

Table 2.2. Results on manipulation check measures in intent-to-treat and successfully treated samples in Study 4, Chapter 2.

Intentionality. In our intent-to-treat sample, a 2 (Self-awareness condition: high vs. low) x 2 (Valence condition: positive vs. negative) ANOVA revealed a non-significant effect of self-awareness condition, $F(1,393) = 2.61, p = .107, \eta_p^2 < .01$. However, there was a significant interaction with valence condition, $F(1,393) = 4.74, p = .030, \eta_p^2 = .01$, such that when the target was a good active listener, participants perceived the target as more intentional when they had high self-awareness ($M = 5.18, SD = 1.71$) than low self-awareness ($M = 4.42, SD = 2.01$), $t(393) = -2.82, p = .005, d = -0.38$, but when the target was a poor active listener, there was no difference in perceived intentionality ($M_{high} = 3.51, SD_{high} = 2.00; M_{low} = 3.62, SD_{low} = 2.26$),

$t(393) = 0.38, p = .705, d = 0.06$. There was also a significant main effect of valence condition, $F(1,393) = 38.16, p < .001, \eta_p^2 = .09$, such that active listeners were seen as more intentional than inactive listeners. Thus, in our intent-to-treat sample, results on this measure only partially supported our hypothesis. This result may be in part because this sample included participants who did not rate the valence of the confederate's behavior as intended, and thus, those participants may have answered a different question than the others (e.g., they would have answered how intentionally the target came across as a "good" active listener if they rated the target positively but were in the negative condition).

Our successfully treated sample, however, fully followed the predicted pattern: There was a significant main effect of self-awareness condition, $F(1,276) = 7.69, p = .006, \eta_p^2 = .03$, with no interaction, $F(1,276) = 0.00, p = .956, \eta_p^2 < .01$: Both when the target was a good active listener and when the target was a poor active listener, the target was perceived as more intentional in the high ($M_{positive} = 5.48, SD_{positive} = 1.42; M_{negative} = 3.41, SD_{negative} = 2.16$) compared to low self-awareness condition ($M_{positive} = 4.82, SD_{positive} = 1.88; M_{negative} = 2.77, SD_{negative} = 2.28$), $t_{positive}(276) = -2.38, p = .018, d = -0.35; t_{negative}(276) = -1.69, p = .092, d = -0.34$. There was also a main effect of valence condition, $F(1,276) = 76.46, p < .001, \eta_p^2 = .22$, again indicating that the active listener was seen as more intentional overall.

Overall trust. Means for overall trust are displayed in Figure 2.5. In our intent-to-treat sample, there was no main effect of self-awareness condition, $F(1,393) = 0.01, p = .909, \eta_p^2 < .01$, but there was a main effect of valence condition, $F(1,393) = 61.26, p < .001, \eta_p^2 = .13$, and—critically—a Self-Awareness x Valence interaction, $F(1,393) = 4.15, p = .042, \eta_p^2 = .01$. These results are consistent with the predicted pattern, as are the results in our successfully treated sample: There was no main effect of self-awareness condition, $F(1,276) = 0.19, p = .666,$

$\eta_p^2 < .01$, but there was a main effect of valence condition, $F(1,276) = 140.24, p < .001, \eta_p^2 = .34$, and the key Self-Awareness x Valence interaction, $F(1,276) = 5.14, p = .024, \eta_p^2 = .02$. In both samples, the pairwise contrasts were non-significant, though the means followed the predicted pattern: When the target was a good active listener, participants perceived them as (directionally) more trustworthy when they had high ($M_{intent} = 1.54, SD_{intent} = 1.15; M_{successful} = 1.69, SD_{successful} = 1.04$) compared to low ($M_{intent} = 1.30, SD_{intent} = 1.03; M_{successful} = 1.43, SD_{successful} = 0.94$) self-awareness (intent-to-treat: $t(393) = -1.43, p = .154, d = -0.19$; successfully treated: $t(276) = -1.55, p = .123, d = -0.23$), but when the target was a bad active listener, participants perceived them as (directionally) less trustworthy when they had high ($M_{intent} = 0.32, SD_{intent} = 1.44; M_{successful} = -0.28, SD_{successful} = 1.41$) compared to low ($M_{intent} = 0.58, SD_{intent} = 1.29; M_{successful} = 0.09, SD_{successful} = 1.21$) self-awareness (intent-to-treat: $t(393) = 1.46, p = .146, d = 0.22$, successfully treated: $t(276) = 1.68, p = .095, d = 0.34$).

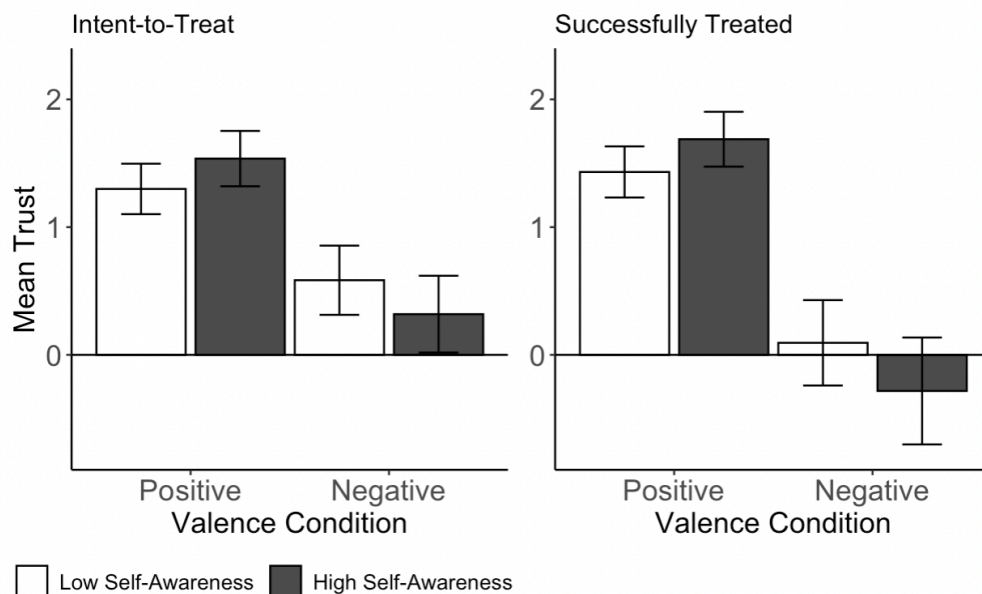


Figure 2.5. Mean trust in intent-to-treat and successfully treated samples in Study 4, Chapter 2. Error bars represent 95% confidence intervals.

Moderated mediation. We ran a model with self-awareness condition as the independent variable, perceived intentionality as the mediator, valence condition as the moderator, and perceived trustworthiness as the outcome. In our intent-to-treat sample, we observed marginally significant moderated mediation, $b = -0.09$, 95% CI = [-0.19, 0.00], $p = .059$, such that perceived intentionality mediated when the target was a good active listener, $b = -0.07$, 95% CI = [-0.15, -0.01], $p = .031$, but—contrary to predictions—not when the target was a bad active listener, $b = 0.01$, 95% CI = [-0.06, 0.09], $p = 0.74$. In our successfully treated sample, contrary to predictions, we did not observe significant moderated mediation, $b = 0.00$, 95% CI = [-0.09, 0.08], $p = .979$. These insignificant results may be the result of the small overall effect sizes.

Exploratory trust game measures. On our measure of how many bonus points the participant wanted to send to the target as a behavioral measure of trust, our intent-to-treat sample yielded no main effect of self-awareness condition, $F(1,393) = 0.12$, $p = .728$, $\eta_p^2 < .01$ and no significant interaction, $F(1,393) = 1.11$, $p = .292$, $\eta_p^2 < .01$. There was a significant main effect of valence condition, $F(1,393) = 8.10$, $p = .005$, $\eta_p^2 = .02$, such that participants sent more points when the target was a good active listener than a bad active listener. However, in our successfully treated sample, there was a marginally significant interaction, $F(1,276) = 3.03$, $p = .083$, $\eta_p^2 = .01$, with no main effect of self-awareness condition, $F(1,276) = 0.00$, $p = .946$, $\eta_p^2 < .01$, and a main effect of valence condition, $F(1,276) = 11.32$, $p < .001$, $\eta_p^2 = .04$. Again, the pairwise contrasts were non-significant, but the means followed the predicted pattern: When the target had been a good active listener, participants were willing to trust the target with bonus points (directionally) more when the target had high self-awareness ($M = 44.53$, $SD = 9.50$) compared to low self-awareness ($M = 41.85$, $SD = 11.58$), $t(276) = -1.53$, $p = .129$, $d = -0.23$, but when the target had been a bad active listener, participants trusted them with (directionally)

fewer bonus points when the target had high ($M = 36.98$, $SD = 14.77$) compared to low ($M = 39.45$, $SD = 12.91$) self-awareness, $t(276) = 1.04$, $p = .299$, $d = 0.21$.

Results on our measure of what percentage of points participants expected the target to send back to them followed the predicted pattern even more strongly.⁶ In our intent-to-treat sample, there was no main effect of self-awareness condition, $F(1,392) = 0.03$, $p = .873$, $\eta_p^2 < .01$, but there was a main effect of valence condition, $F(1,392) = 7.68$, $p = .006$, $\eta_p^2 = .02$, and the key Self-Awareness x Valence interaction, $F(1,392) = 7.23$, $p = .007$, $\eta_p^2 = .02$. Similarly, in our successfully treated sample, there was no main effect of self-awareness condition, $F(1,275) = 0.14$, $p = .714$, $\eta_p^2 < .01$, but there was a main effect of valence condition, $F(1,275) = 9.31$, $p = .003$, $\eta_p^2 = .03$, and the key Self-Awareness x Valence interaction, $F(1,275) = 4.58$, $p = .033$, $\eta_p^2 = .02$. Once again, the pairwise contrasts did not all reach significance, but the pattern of means was consistent with hypotheses: In both cases, when the target had been a good active listener, participants thought the target would send a greater percentage of points back to them when the target had high self-awareness ($M_{intent} = 51.61$, $SD_{intent} = 19.89$; $M_{successful} = 51.98$, $SD_{successful} = 19.13$) compared to low self-awareness ($M_{intent} = 46.11$, $SD_{intent} = 16.50$; $M_{successful} = 47.64$, $SD_{successful} = 16.13$), (intent-to-treat: $t(392) = -2.12$, $p = .035$, $d = -0.19$; successfully treated: $t(275) = -1.50$, $p = .136$, $d = -0.22$), but when the target had been a bad active listener, participants thought the target would send a *smaller* percentage of points back when the target had high ($M_{intent} = 41.08$, $SD_{intent} = 17.44$; $M_{successful} = 39.26$, $SD_{successful} = 21.46$) compared to low ($M_{intent} = 45.95$, $SD_{intent} = 22.40$; $M_{successful} = 45.40$, $SD_{successful} = 23.28$) self-awareness (intent-to-treat: $t(392) = 1.71$, $p = .088$, $d = 0.22$; successfully treated: $t(275) = 1.56$, $p = .121$, $d = 0.31$).

⁶ One participant sent zero points to the target, meaning we did not show them the question about what percentage of points they thought the target would send back. We excluded this participant from the analyses on this measure.

Discussion

Study 4 demonstrated that even within live interactions, people use others' self-awareness to infer something about their trustworthiness and form different inferences depending on whether the target has a positive or negative impact on them. In our study setup, participants had a multitude of additional cues that they could use to evaluate the target besides our manipulations (e.g., appearance, tone of voice, etc.), potentially creating much of the noise that exists in real-world interactions. Nevertheless, even with these additional cues, we still observed an effect of self-awareness on trust, especially among those for whom our valence manipulation was successful. Moreover, although our results on the behavioral measure of trust were weaker, they are suggestive of potential downstream behavioral consequences based on these judgments.

Study 5: It's Better if I Know You Can't Change

In our last three studies, we turned to testing theory-driven moderators of the effect of perceived self-awareness on trust. In particular, any factor that reduces the extent to which observers can infer the target's intentions toward others from their behavior should in turn attenuate the effect of self-awareness on trust. In Study 5, we tested one such moderator: the perceived mutability of the target's behavior. We conducted a scenario study in which we manipulated whether a target seemed to have high or low self-awareness about a behavior that was having a negative impact on the participant. We also manipulated whether this behavior seemed easy or hard for the target to change. We predicted that targets who seemed more self-aware of their negative behavior would be trusted less than targets who seemed less self-aware (as before), but that the difference between high and low self-awareness would be larger when the behavior was perceived as high in mutability (easy to change) than low in mutability. This study's hypotheses and design were preregistered at https://aspredicted.org/QSL_NV7.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$1.75. We preregistered that we would collect 1,200 participants after excluding those who failed the attention check or comprehension checks, and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 1,520 participants who started the survey, we ended up with a final sample of 1,203 participants (75.15% female, 2.33% other gender, 25.44% non-White, $M_{\text{age}} = 26.68$, $SD_{\text{age}} = 8.25$) who fit these criteria.

Procedure. Participants were assigned to one of four conditions in a 2 (Self-awareness condition: high vs. low) x 2 (Mutability condition: high vs. low) between-subjects design, and read one of three scenarios (randomly assigned for stimulus sampling). All three scenarios described Taylor, a work colleague (whose gender was randomly assigned—male or female—for stimulus sampling). In one scenario, participants read that they were chatting with Taylor about an upcoming discussion at work about how to improve their organization. The participant mentioned that they had a lot of ideas they wanted to bring up during the discussion. During the discussion the next day, however, Taylor was so talkative that (s)he dominated the entire conversation with his/her own ideas and prevented the participant from participating in the conversation, which left the participant feeling frustrated. In another scenario, participants read that they were talking with Taylor and told Taylor that they would be really upset if they lost their job. A few weeks later, the participant lost their job and told Taylor, but Taylor acted extremely unsympathetic in response, leaving the participant feeling even sadder. In the third scenario, participants read that they told Taylor that they had been putting a lot of effort into choosing a gift for Taylor. A few days later, the participant gave the gift to Taylor, but Taylor acted very unappreciative, leaving the participant feeling sad.

In all three scenarios, Taylor was told at the end of the scenario by another work colleague that (s)he was acting really talkative, unsympathetic, or unappreciative (depending on the scenario), which the participant overheard. We manipulated Taylor's response according to the condition the participant was assigned. Participants in the high self-awareness condition read that Taylor responded: "Yeah, I know I was." Participants in the low self-awareness condition read that Taylor responded: "Oh, I was? I didn't realize that."

On the next page, participants read some additional information about Taylor to manipulate the perceived mutability of Taylor's behavior. To make our manipulation as convincing as possible, we presented participants with a strong reason for why it was easy or difficult for Taylor to change their behavior. In all conditions, they read that Taylor recently had a medical procedure that affected some parts of his/her brain, which had a particular effect on his/her personality. In the high mutability conditions, participants read that it was now very easy for Taylor to change how talkative/unsympathetic/unappreciative (s)he was. In the low mutability conditions, participants read that it was now very hard for Taylor to change how talkative/unsympathetic/unappreciative (s)he was.

Before proceeding, participants had to correctly answer three comprehension checks about the scenario. They were given two tries to answer correctly. If they failed at least one question on the second try, they were unable to proceed with the rest of the study.

Once participants passed the comprehension checks, they proceeded to answer several pages of our dependent measures and manipulation checks, with both the pages and the questions within each page presented in random order. On one page, participants responded to our manipulation check measures: "Do you agree or disagree with the following statement? It seems like during the [discussion/conversation/exchange], Taylor was fully aware of how

[talkative/unsympathetic/unappreciative] (s)he was being” (-3 = *strongly disagree*, 3 = *strongly agree*); “Based on what you know from the scenario, how easy or difficult do you think it would have been for Taylor to change how [talkative/unsympathetic/unappreciative] (s)he was during the [discussion/conversation/exchange]?” (-3 = *extremely easy*, 3 = *extremely difficult*); “To what extent do you believe that Taylor’s [talkativeness/lack of sympathy/lack of appreciation] had a negative impact on you?” (1 = *not at all*, 7 = *a lot*). As in the previous studies, this last question was asked in order to confirm that participants perceived Taylor’s behavior as having a negative impact on them (i.e., above the midpoint of the scale).

On a separate page, participants answered a series of questions measuring their perceptions of Taylor’s benevolence, ability, and integrity, using the exact same measures as in Study 3. Once again, we preregistered the ability and integrity measures as exploratory given that our primary hypotheses concern perceptions of benevolence, as an indicator of perceived intentions toward others. On a separate page, participants also responded to the same set of workplace trust measures as in Study 3.

After completing all of the above measures, participants completed an exploratory trust game measure, similar to that in Study 4. They were told that they would hypothetically be given \$5, could send any amount of this money to Taylor, and that whatever money they sent would triple in value. Taylor would then decide how much to send back to them. Once participants passed three comprehension checks regarding the trust game instructions, we asked participants: “How much of the \$5 would you choose to send to Taylor?” Again, this measure was intended to serve as an additional and more behaviorally oriented measure of trust.

At the end, participants reported demographic information (gender, age, race), were asked to briefly write what they thought we were studying with the survey, and were given an optional space to provide any feedback on the study.

Results

As intended, participants perceived the target as more aware of their behavior in the high self-awareness conditions ($M = 2.45$, $SD = 0.90$) than in the low self-awareness conditions ($M = 1.40$, $SD = 1.74$), $t(1201) = -48.29$, $p < .001$, perceived the target's behavior as more easily changeable in the high mutability conditions ($M = -2.12$, $SD = 1.27$) than in the low mutability conditions ($M = 2.55$, $SD = 0.89$), $t(1201) = -73.76$, $p < .001$, and perceived the target's behavior as having a negative impact on them across all conditions (significantly above the midpoint of the scale; $M = 4.84$, $SD = 1.54$), $t(1202) = 18.90$, $p < .001$.

We preregistered to combine our benevolence, ability, and integrity scales, as well as our overall trust scale, into their respective composites if they yielded α 's > 0.70 . Only the benevolence ($\alpha = 0.88$) and integrity ($\alpha = 0.86$) scales met this criterion. However, because each individual item on our trust scale ($\alpha = 0.48$) exhibited the same pattern of means between conditions, and because our competence measures ($\alpha = 0.66$) were only exploratory, we nevertheless report results on the composites here for ease of communication. Results for the individual measures are also available in the Supplement (see Appendix B).

We collapsed across scenarios and computed OLS regressions with self-awareness condition (high vs. low), mutability condition (high vs. low), and their interaction as between-subjects predictors, fixed effects for scenario (talkative vs. unsympathetic vs. unappreciative), and the benevolence composite or workplace trust composite as the outcome measure.

Benevolence. On our measure of benevolence, we found a main effect of self-awareness condition: As before, participants perceived the target as *less* benevolent when the target was high in self-awareness of their negative behavior than when the target was low in self-awareness, $b = -1.35$, $t(1197) = -16.00$, $p < .001$. There was also a main effect of mutability condition, $b = 0.47$, $t(1197) = 5.57$, $p < .001$. However, critically, we also found a significant interaction between self-awareness condition and mutability condition, $b = 0.68$, $t(1197) = 5.67$, $p < .001$. When the behavior was perceived as high in mutability, the difference between self-awareness conditions was larger ($M_{high_self-awareness} = -0.91$, $SD_{high_self-awareness} = 1.07$; $M_{low_self-awareness} = 0.51$, $SD_{low_self-awareness} = 1.14$; $d = 1.31$ using post-hoc t-tests comparing the means) than when the behavior was perceived as low in mutability ($M_{high_self-awareness} = 0.23$, $SD_{high_self-awareness} = 1.07$; $M_{low_self-awareness} = 0.94$, $SD_{low_self-awareness} = 0.97$; $d = 0.66$). In other words, participants were less sensitive to the target's self-awareness when the behavior would have been hard for the target to change.

Workplace trust. The same pattern emerged for our measure of workplace trust (means are visualized in Figure 2.6). There was a main effect of self-awareness condition, such that the target was trusted less when they were high in self-awareness than when they were low in self-awareness, $b = -0.57$, $t(1197) = -7.88$, $p < .001$, as well as a main effect of mutability condition, $b = 0.26$, $t(1197) = 3.51$, $p < .001$. Yet there was also an interaction between self-awareness condition and mutability condition, $b = 0.33$, $t(1197) = 3.25$, $p = .001$, such that the gap in trust between self-awareness conditions was larger when the behavior was perceived as high in mutability ($M_{high_self-awareness} = -0.99$, $SD_{high_self-awareness} = 0.85$; $M_{low_self-awareness} = -0.41$, $SD_{low_self-awareness} = 0.86$; $d = 0.56$ using post-hoc t-tests comparing the means) compared to low in

mutability ($M_{high_self-awareness} = -0.41$, $SD_{high_self-awareness} = 0.94$; $M_{low_self-awareness} = -0.16$, $SD_{low_self-awareness} = 0.96$; $d = 0.23$).

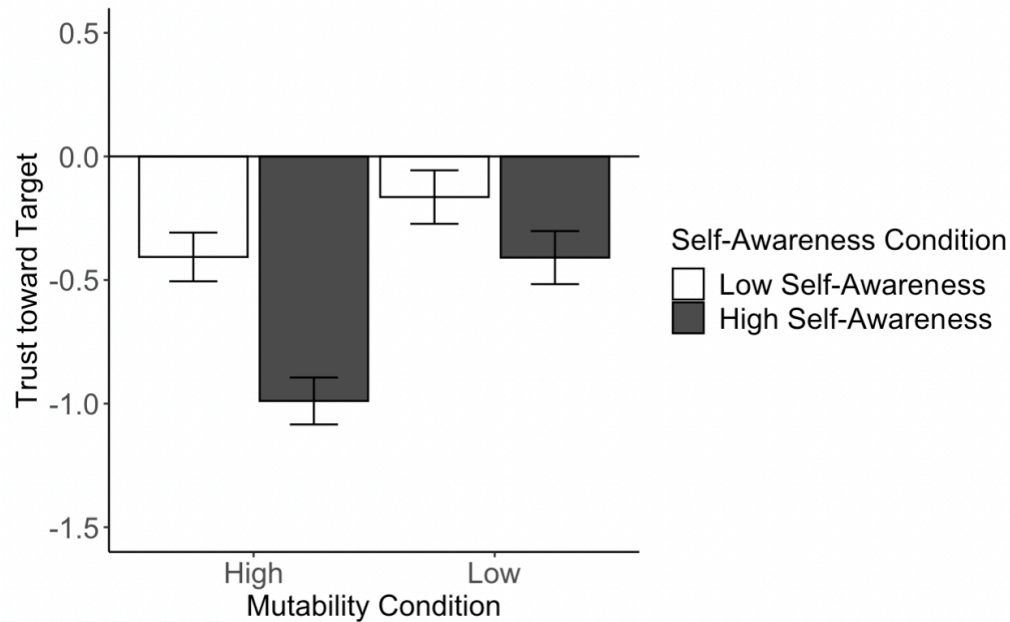


Figure 2.6. Mean trust toward target in Study 5, Chapter 2. Error bars represent 95% confidence intervals.

Moderated mediation. Finally, we found that perceived benevolence significantly mediated the effect of self-awareness condition on trust, but this was moderated by mutability condition, $b = 0.21$, 95% CI = [0.11, 0.32], $p < .001$, such that benevolence was a stronger mediator when the behavior was perceived as high in mutability, $b = 0.46$, 95% CI = [0.36, 0.56], $p < .001$, compared to low in mutability, $b = 0.20$, 95% CI = [0.13, 0.27], $p < .001$.

Exploratory measures: Ability and integrity. We observed only marginal effects of self-awareness condition, $b = 0.18$, $t(1197) = 1.75$, $p = .081$, and mutability condition, $b = 0.19$, $t(1197) = 1.83$, $p = .067$, with no interaction, $b = 0.04$, $t(1197) = 0.28$, $p = .783$, on our composite of ability. We observed a main effect of mutability condition, $b = 0.66$, $t(1197) = 7.28$, $p < .001$, but no main effect of self-awareness condition, $b = 0.12$, $t(1197) = 1.31$, $p = .189$ and no

interaction, $b = -0.03$, $t(1197) = -0.23$, $p = .816$, on our composite of integrity. Finally, we did not observe significant moderated mediation for either ability ($p = .711$) or integrity ($p = .866$).

Exploratory measure: Trust game decision. In line with our results on the trust scale measure, we found main effects of both self-awareness condition, $b = -0.75$, $t(1197) = -6.02$, $p < .001$, and mutability condition, $b = 0.25$, $t(1197) = 2.01$, $p = .044$, as well as an interaction, $b = 0.61$, $t(1197) = 3.44$, $p < .001$, such that the gap between self-awareness conditions was larger when the behavior was perceived as high in mutability ($M_{high_self-awareness} = 2.02$, $SD_{high_self-awareness} = 1.59$; $M_{low_self-awareness} = 2.82$, $SD_{low_self-awareness} = 1.54$; $d = 0.73$) compared to low in mutability ($M_{high_self-awareness} = 2.86$, $SD_{high_self-awareness} = 1.62$; $M_{low_self-awareness} = 3.03$, $SD_{low_self-awareness} = 1.43$; $d = 0.14$).

Discussion

The results of Study 5 lend further support to our theory that self-awareness affects trust via inferences of intentionality: The effect of self-awareness on trust is moderated when we induce ambiguity about the target's intentions—in this case, due to the perceived immutability of the behavior. That is, if the target has little control over their behavior, then it is less clear to observers whether the target intended to behave this way, even if they appear to have high self-awareness. As a result, the target's self-awareness does not make their behavior seem more diagnostic—and thus, has less of an influence on whether observers trust that target—in contrast to cases where the target has greater perceived control over their behavior.

Study 6: It's Better if I Know You Can't Change, Take Two

In Study 6, we used a very similar design to Study 5, except that we used different scenarios in order to allow for greater generalizability and to allow for a different manipulation of perceived mutability. As in Study 5, we predicted that targets who seemed more (versus less)

self-aware of their negative behavior would be trusted less, but that this effect would be weaker when the behavior was perceived as low (versus high) in mutability. This study's hypotheses and design were preregistered at https://aspredicted.org/1T5_8K5.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$1.75. We preregistered that we would collect 1,600 participants after excluding those who failed the attention check or comprehension checks, and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 1,727 participants who started the survey, we ended up with a final sample of 1,603 participants (62.69% female, 2.06% other gender, 20.15% non-White, $M_{\text{age}} = 34.01$, $SD_{\text{age}} = 13.62$) who fit these criteria.

Procedure. Similar to Study 5, participants were assigned to one of four conditions in a 2 (Self-awareness condition: high vs. low) x 2 (Mutability condition: high vs. low) between-subjects design, and read one of two scenarios, randomly assigned for stimulus sampling. Both scenarios randomly manipulated the gender of the target to be male or female for stimulus sampling. In one scenario, participants read that they were on an airplane in a middle seat and that one of the people sitting next to them was taking up some of their seat space for the entire flight, making it hard for the participant to sit comfortably. In the high mutability conditions, participants read that this person was leaning into their seat even though this person had plenty of extra space in their own seat area that they could be using instead. In the low mutability conditions, participants read that this person was leaning into their seat because they were a very broad-framed person, and therefore could not stay out of the participant's seat area even if they tried.

In the other scenario, participants read that they were sitting in a park watching a play being performed on an outdoor stage, and that there was a person right in front of them blocking

most of their view of the show, preventing them from enjoying the play. In the high mutability conditions, participants read that this person was blocking their view because the person kept choosing to stand up, even though (s)he had a seat. In the low mutability conditions, participants read that the person was blocking their view because they were very tall and would block the participant's view no matter how they were sitting.

We manipulated self-awareness in the same way as in Study 5. The target person was told at the end by a family member that they appeared to be taking up the participant's seat or blocking the participant's view, depending on the scenario, and the target responded according to condition. As before, in the high self-awareness conditions, the target responded: "Yeah, I know I was." In the low self-awareness conditions, the target responded: "Oh, I was? I didn't realize that."

Participants then responded to two comprehension checks about the scenario, which they were given two tries to pass (if they failed after the second try, they were unable to proceed with the rest of the study). Once participants passed these checks, they responded to several pages of our dependent measures and manipulation checks, with both the pages and the questions within each page presented in random order.

Our manipulation checks and dependent measures were nearly the same as in Study 5, except for a few changes. First, we adapted the phrasing to refer to "this person" instead of "Taylor," and modified wording slightly as needed to fit the specific scenarios. Second, we slightly changed the wording of the mutability manipulation check with a parenthetical clarification at the end: "Based on what you know from the scenario, how easy or difficult do you think it would have been for this person to [take up less of your seat space on the plane/avoid blocking your view during the show] (if he/she tried to)?" Third, we used the scale

measuring general overall trust from Studies 1 and 3, and did not include the workplace trust scale from Studies 3 and 5, both because of the relatively low alpha in Studies 3 and 5 and because the workplace-oriented trust measures made less sense in this context. Otherwise, the rest of our measures, including the trust antecedent measures, remained the same as those of Study 5.

After completing all of these measures, participants completed the same exploratory trust game measure as in Study 5, and then provided demographic information (gender, age, race) at the end, along with optional feedback.

Results

As intended, participants perceived the target as more aware of their behavior in the high self-awareness conditions ($M = 2.61, SD = 0.67$) than in the low self-awareness conditions ($M = 1.21, SD = 1.67$), $t(1601) = -60.14, p < .001$, perceived the target's behavior as more easily changeable in the high mutability conditions ($M = -2.37, SD = 1.03$) than in the low mutability conditions ($M = 1.34, SD = 1.43$), $t(1601) = -59.79, p < .001$, and perceived the target's behavior as having a negative impact on them across all conditions (significantly above the midpoint of the scale; $M = 5.11, SD = 1.47$), $t(1602) = 30.35, p < .001$.

As in Study 5, we combined our benevolence ($\alpha = 0.91$), ability ($\alpha = 0.77$), integrity ($\alpha = 0.88$), and overall trust ($\alpha = 0.92$) scales into their respective composites. We again collapsed across scenarios and computed OLS regressions with self-awareness condition (high vs. low), mutability condition (high vs. low), and their interaction as between-subjects predictors, fixed effects for scenario (airplane vs. play), and the benevolence composite or overall trust composite as the outcome measure.

Benevolence. On our measure of benevolence, we found a main effect of self-awareness condition: As before, participants perceived the target as *less* benevolent when the target was high in self-awareness of their negative behavior than when the target was low in self-awareness, $b = -1.84$, $t(1598) = -25.79$, $p < .001$. There was also a main effect of mutability condition, $b = 0.81$, $t(1598) = 11.30$, $p < .001$. As expected, there was also a significant interaction between self-awareness condition and mutability condition, $b = 0.51$, $t(1598) = 5.09$, $p < .001$. When the behavior was perceived as high in mutability, the difference between self-awareness conditions was larger ($M_{high_self-awareness} = -2.11$, $SD_{high_self-awareness} = 0.78$; $M_{low_self-awareness} = -0.26$, $SD_{low_self-awareness} = 1.12$; $d = 1.82$ using post-hoc t-tests comparing the means) than when the behavior was perceived as low in mutability ($M_{high_self-awareness} = -0.78$, $SD_{high_self-awareness} = 1.15$; $M_{low_self-awareness} = 0.55$, $SD_{low_self-awareness} = 0.93$; $d = 1.31$).

Overall trust. Once again, the same pattern emerged for our measure of trust: There was a main effect of self-awareness condition, such that the target was trusted less when they were high in self-awareness than when they were low in self-awareness, $b = -1.45$, $t(1598) = -17.84$, $p < .001$, as well as a main effect of mutability condition, $b = 0.89$, $t(1598) = 10.92$, $p < .001$. Yet there was also the hypothesized interaction between self-awareness condition and mutability condition, $b = 0.52$, $t(1598) = 4.54$, $p < .001$, such that the gap in trust between self-awareness conditions was larger when the behavior was perceived as high in mutability ($M_{high_self-awareness} = -1.84$, $SD_{high_self-awareness} = 0.98$; $M_{low_self-awareness} = -0.39$, $SD_{low_self-awareness} = 1.16$; $d = 1.26$), compared to low in mutability ($M_{high_self-awareness} = -0.43$, $SD_{high_self-awareness} = 1.36$; $M_{low_self-awareness} = 0.50$, $SD_{low_self-awareness} = 1.07$; $d = 0.81$).

Moderated mediation. Finally, we found that perceived benevolence significantly mediated the effect of self-awareness condition on trust, but that this was moderated by

mutability condition, $b = 0.42$, 95% CI = [0.24, 0.60], $p < .001$: Benevolence was a stronger mediator when the behavior was perceived as high in mutability, $b = 1.39$, 95% CI = [1.25, 1.54], $p < .001$, compared to low in mutability, $b = 1.15$, 95% CI = [1.01, 1.29], $p < .001$.

Exploratory measures (ability, integrity, and trust game decision). Results on all of our exploratory measures followed the same pattern as those in Study 5, so we report them in the Supplement (see Appendix B).

Discussion

The results of Study 6 replicate those of Study 5, providing further support to our theory that the perceived mutability of a target's behavior moderates the extent to which self-awareness signals a target's intentionality, and thus, moderates the effect of self-awareness on trust.

Study 7: Do you really want to hurt me?

Our final study sought to isolate the importance of perceived intentions *toward others* as key for the effect of social self-awareness on trust. We manipulated whether the target was aware or unaware of a negative behavior (poor work ethic on a work project), and whether they were working jointly with the participant (thereby affecting the participant's own evaluations and bonus at work) or separately from the participant (thereby having no effect on the participant). We expected that self-awareness of a behavior that does *not* affect others should send a weaker signal about the target's intended future behavior toward others—even if the behavior itself is still perceived as more intentional—thereby attenuating the effect of self-awareness on trust. This study's hypotheses and design were preregistered at https://aspredicted.org/W9X_YHR.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$1.00. We preregistered that we would collect 800 participants after excluding those who failed the attention check, comprehension checks, and/or provided gibberish or bot-

like responses to a free-response question at the beginning of the study. Of the 843 who started the survey, we ended up with a final sample of 800 participants (43.00% female, 1.88% other gender, 25.87% non-White, $M_{age} = 38.13$, $SD_{age} = 13.09$) that fit these criteria.

Procedure. Participants read that they were at work chatting with Taylor, one of their coworkers. Again, we counterbalanced the gender of Taylor for stimulus sampling. We randomly assigned participants to one of four conditions in a 2 (Target self-awareness: high vs. low) x 2 (Impact condition: impact vs. no impact) design. In the impact conditions, participants read that they were working jointly on a work project with Taylor, such that the participant and Taylor would be evaluated jointly and the participant's annual bonus would depend on the quality of both their own work and Taylor's work. In the no impact conditions, participants read that they were working on a separate project from Taylor, such that the participant and Taylor would each be evaluated separately for their work and the participant's annual bonus would only depend on the quality of their own work and not Taylor's. We asked participants a comprehension check to ensure that they understood this information; if they failed the comprehension check after two tries, the study automatically ended.

On the next page, participants read that Taylor had been slacking off on the project lately, and that this seemed to be due to laziness rather than anything personal going on with Taylor. During the conversation, the participant mentioned that Taylor had seemed negligent on the project lately. Like in Studies 5-6, we manipulated Taylor's response according to self-awareness condition. In the high self-awareness conditions, Taylor seemed unsurprised and responded: "Yeah, I know I have been." In the low self-awareness conditions, Taylor seemed surprised and responded: "Oh, really? I didn't realize I seemed that way." We again asked participants to

complete a comprehension check about this information; if they failed the comprehension check after two tries, the study automatically ended.

We randomized the order of the page with our manipulation check and mediator questions versus the page with our main dependent measures. The manipulation check questions were similar to those of the previous studies, with slight modifications to fit the particular context: “How aware or unaware do you think Taylor was that you perceived [him/her] as negligent on the project (before you said so)?” (-3 = *extremely unaware*, 3 = *extremely aware*); “To what extent do you believe Taylor’s negligence on the project has a positive or negative impact on you?” (-3 = *extremely negative*, 3 = *extremely positive*).

We included the following measure of Taylor’s perceived intentions toward the participant as our mediator measure: “To what extent do you believe Taylor’s general intentions toward you are positive or negative?” (-3 = *extremely negative*, 3 = *extremely positive*). We also included the following exploratory measure of the intentionality behind Taylor’s specific actions (slacking off on the project): “To what extent do you think Taylor is intentionally slacking off on the project?” (1 = *not at all intentionally*, 7 = *extremely intentionally*). We predicted that self-awareness would signal greater intentionality behind Taylor’s specific actions regardless of whether that behavior impacted the participant or not, but that it would only influence perceptions of Taylor’s general (positive or negative) intentions *toward the participant* when the behavior actually impacted the participant.

Our main dependent measures of trust consisted of the same three-item trust scale used in Studies 1, 3, and 6. On the last page, participants responded to the same (exploratory) measures of benevolence, ability, and integrity as in the previous studies. Finally, participants provided

demographic information (gender, age, race) at the very end, along with an optional space to provide feedback.

Results

As intended, participants perceived Taylor as more aware of their behavior in the high self-awareness condition ($M = 1.94$, $SD = 0.93$) than in the low self-awareness condition ($M = -1.36$, $SD = 1.52$), $t(659.23) = -36.96$, $p < .001$, and perceived Taylor as having more of a negative impact on the participant in the impact condition ($M = -2.11$, $SD = 0.84$) compared to the no impact condition ($M = -0.22$, $SD = 0.84$), $t(797.85) = -31.92$, $p < .001$.

As in the previous studies, we combined our overall trust scale ($\alpha = 0.87$) into a composite. We conducted a two-way ANOVA with self-awareness condition (high vs. low), impact condition (impact vs. no impact), and their interaction as between-subjects predictors, with perceived intentions or overall trust as the outcome measure.

General intentions. There was a main effect of self-awareness condition, $F(1, 796) = 48.68$, $p < .001$, $\eta_p^2 = .06$, and a main effect of impact condition, $F(1, 796) = 54.56$, $p < .001$, $\eta_p^2 = .06$, qualified by a significant interaction, $F(1, 796) = 17.74$, $p < .001$, $\eta_p^2 = .02$: As expected, there was a greater difference in perceived intentions toward the participant (between self-awareness conditions) in the impact condition ($M_{high} = -0.78$, $SD_{high} = 1.15$; $M_{low} = 0.06$, $SD_{low} = 1.02$), $t(796) = 7.92$, $p < .001$, 95% CI = [0.63, 1.05], $d = 0.79$, than in the no impact condition ($M_{high} = 0.09$, $SD_{high} = 1.07$; $M_{low} = 0.30$, $SD_{low} = 1.01$), $t(796) = 1.95$, $p = .051$, 95% CI = [0.00, 0.42], $d = 0.20$.

Overall trust. Means for the trust composite are shown in Figure 2.7. There was a main effect of self-awareness condition, $F(1, 796) = 46.68$, $p < .001$, $\eta_p^2 < .06$, and a main effect of impact condition, $F(1, 796) = 32.36$, $p < .001$, $\eta_p^2 = .04$, qualified by a significant interaction,

$F(1, 796) = 6.01, p = .014, \eta_p^2 < .01$. As hypothesized, when Taylor’s work on the project impacted the participant, participants trusted Taylor less when Taylor had high ($M = -1.55, SD = 1.12$) compared to low self-awareness ($M = -0.77, SD = 1.25$), $t(796) = 6.57, p < .001, 95\% CI = [0.55, 1.01], d = 0.66$, but this gap became smaller when Taylor’s work had no impact on the participant ($M_{high} = -0.87, SD_{high} = 1.20; M_{low} = -0.50, SD_{low} = 1.19$), $t(796) = 3.09, p = .002, 95\% CI = [0.13, 0.60], d = 0.31$.

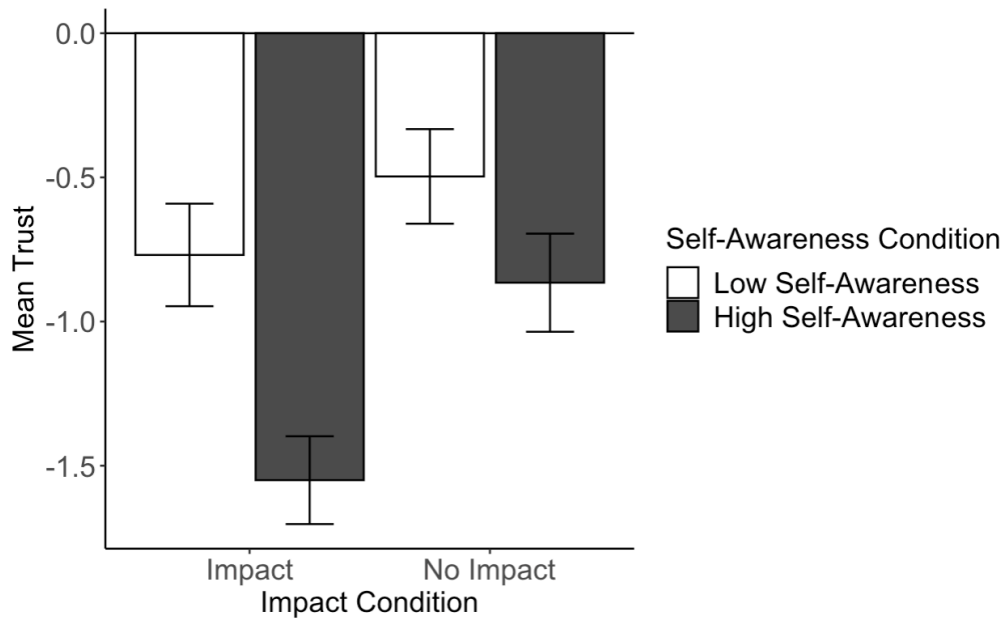


Figure 2.7. Mean trust in Study 7, Chapter 2. Error bars represent 95% confidence intervals.

Moderated mediation. In line with hypotheses, we found that perceived general intentions toward the participant significantly mediated the effect of self-awareness condition on trust, but that this was moderated by impact condition, $b = -0.33, 95\% CI = [-0.50, -0.17], p < .001$: General intentions mediated the effect of self-awareness on trust more strongly in the impact condition, $b = 0.45, 95\% CI = [0.32, 0.60], p < .001$, than in the no impact condition, $b = 0.11, 95\% CI = [0.00, 0.21], p = .048$.

Exploratory measure of intentionality of specific behavior. As expected, there was a main effect of self-awareness condition, $F(1,796) = 455.58, p < .001, \eta_p^2 = .36$, but no main effect of impact condition, $F(1,796) = 2.22, p = .136, \eta_p^2 < .01$, and no interaction, $F(1,796) = 1.23, p = .268, \eta_p^2 < .01$, suggesting that self-awareness still increases the perceived intentionality behind Taylor's particular behavior, regardless of its effect on others. However, in line with hypotheses, our results on the primary dependent measures suggest that this increase in perceived intentionality only translates to perceptions of overall negative intentions toward others *if* the behavior impacts others negatively.

Exploratory benevolence, ability, and integrity measures. We report results on these measures in the Supplement.

Discussion

Study 7 provides further evidence that the effect of self-awareness on trust is driven by perceptions of the target's intentions toward others; thus, when a target exhibits self-awareness about a behavior that has no impact on others, observers are less sensitive to the target's self-awareness in forming judgments of trustworthiness.

General Discussion

Self-awareness is frequently thought of as a desirable quality in others, yet across seven studies, we find that the effect of self-awareness on trust is nuanced. While exhibiting self-awareness can increase others' trust toward the target when the target engages in positive behaviors, it can also decrease trust when the target engages in negative behaviors (Studies 1-4). We find this effect both for self-awareness as a general trait (Study 1) and for self-awareness that is specific to a given behavior (Study 2-4). Importantly, we document that this phenomenon occurs even in real, face-to-face interactions (Study 4). These effects are driven by perceptions of

greater intentionality that accompany highly self-aware targets. When the signal of intentionality is weakened, however—such as by increasing the perception that the target had little ability to change their behavior (Studies 5-6)—or when negative intentions do not seem to be targeted toward others (Study 7), self-awareness of negative behaviors does not decrease trust as much.

Our findings make several theoretical contributions to prior literature. First, we document important *interpersonal* consequences of self-awareness, i.e., consequences beyond the influence of self-awareness on a target's own subsequent experiences and behaviors. While self-awareness, and perspective-taking more generally, have been shown to affect how people view themselves and regulate their own behavior (Diener & Wallbom, 1976; Galinsky et al., 2005; Wang, Kenneth, et al., 2014), our research examines how a target's external display of self-awareness affects *others'* inferences about the target. Moreover, we provide causal, rather than correlational, evidence for the effect of self-awareness on these inferences by holding constant a target's exact behaviors and simply manipulating their apparent self-awareness. This causal finding contrasts with much of the prior work on self-awareness and/or meta-perception, which has provided correlational evidence of the relationship between knowing what others think of oneself and subsequent interpersonal outcomes (Brion et al., 2015; Ohtsubo et al., 2009).

We also contribute to literature on person perception more broadly. Prior literature has documented the importance of a target's perceived intentions when forming judgments based on their actions, rather than the mere actions themselves (Hackel et al., 2020; J. Landy & Uhlmann, 2018; Levine & Schweitzer, 2015; Malle & Knobe, 1997; Maselli & Altrocchi, 1969; Uhlmann et al., 2013, 2015; Uhlmann & Zhu, 2014). Our findings provide evidence that a target's apparent self-awareness serves as one cue from which observers form inferences about the target's intentionality. In turn, these inferences of intentionality lead observers to perceive the target's

behavior as more diagnostic of their character, and thus weight this behavior more heavily in their perceptions of the target (Mende-Siedlecki et al., 2013; Skowronski & Carlston, 1987, 1989). Accordingly, we draw an important link between self-awareness and interpersonal judgment.

Finally, our findings contribute to literature on trust formation. While past research has examined a variety of target characteristics that may affect others' perceptions of how trustworthy they are (Deutsch, 1958; Gabarro, 1978; McKnight et al., 1998; Rempel et al., 1985), no research, to our knowledge, has examined a target's apparent self-awareness as an input to trust. Given the importance of trusting decisions to virtually all human interpersonal contexts—ranging from workplace or professional settings (Dirks & Ferrin, 2001; McKnight et al., 1998; Whitener et al., 1998) to romantic relationships (Kim et al., 2015; Rempel et al., 1985) to interactions with strangers (Foddy et al., 2009; Ho & Weigelt, 2005; Macy & Skvoretz, 1998)—our research illuminates an additional factor behind these ubiquitous decisions. Moreover, we highlight the psychological process by which social self-awareness affects trust: It influences perceptions of the target's positive or negative intentions toward others. As such, social self-awareness influences perceptions of benevolence-based trust most directly and has a less consistent effect on ability- or integrity-based trust.

Practically, our findings suggest that exhibiting self-awareness may not universally increase others' trust, and thus, that people may be well-served to adjust their displays of self-awareness according to the situation to maximize others' perceptions of their trustworthiness. In general, people may trust one more when one exhibits high self-awareness to others, e.g., by explicitly communicating that one knows how others perceive or might perceive oneself, by regulating one's behaviors according to others' perceptions, and so on. However, if one commits

a negative act without knowing how it will be perceived, making it clear to others that one did not know how others would perceive oneself may make one seem more trustworthy, as it may reduce others' perceptions of intentionality from that act.

Our findings have several limitations that open up directions for future research. For instance, we expect that additional factors may moderate our observed effects, such as those that further affect inferences of a target's intentionality. One possibility is that high self-awareness about a negative behavior does not decrease trust (or does not decrease it as much) when there is some other positive or justifiable reason for the action, even if it does not directly benefit the observer. An extreme example might be stealing food in order to save one's own life, while still showing self-awareness regarding the negative impact this may have on others. In such cases, people might infer less about a target's negative intentions toward other people when they realize that other—less harmful—motives could have been behind the act, and that the negative impact could have been an undesired consequence of the pursuit of some other benefit. We report an initial test of this hypothesis in a supplemental study (see Appendix B). Another possibility is that high self-awareness of positive behaviors could decrease trust if it appears strategic or manipulative. For instance, if observers know that a target is seeking some benefit for themselves, such as a promotion, and perceive the target as high in self-awareness while behaving kindly to their boss, they might trust the target *less*. By contrast, a target in the same situation who appears low in self-awareness may seem more authentically kind. Finally, future research should examine what specific cues people are most likely to use to form inferences about others' degree of self-awareness in the first place (see Chapter 3 for some initial findings related to this question).

Overall, our research suggests that social self-awareness serves as a signal with which others evaluate a target's trustworthiness. While many popular perceptions portray self-awareness in a positive light, our research cautions that the full story is more nuanced.

Chapter 3:

When and Why People Discern Others' Degree of Social Self-Awareness

Abstract

People frequently evaluate a variety of characteristics in others, such as competence, morality, and sociability. In the current research, we examine whether, and when, people evaluate how *self-aware* a target person is (i.e., whether the target seems to be aware of what others think of them or not). In contrast to evaluating many other traits, evaluating self-awareness requires modeling the target's mind (in order to discern what the target thinks others think of them). Thus, we propose that observers will be most likely to spontaneously evaluate a target's self-awareness when the observer is prompted to undergo a more careful attribution process. Based on past research, we expect this to occur when observers are surprised by the target's behavior, and/or when observers evaluate the target negatively on other attributes or behaviors. In such cases, observers will seek to make sense of the target's behavior, and in doing so, will sometimes bring to mind the target's degree of self-awareness as one possible explanation (e.g., inferring the target must lack self-awareness if their behavior defies expectations). We test these hypotheses across several studies, and find initial support for this framework.

Introduction

The first two chapters of this dissertation sought to answer the following question: When people have explicit indicators about another person's degree of (social) self-awareness, how do they use this information in forming impressions of that person? A separate question, however, is: In the absence of explicit indicators of another person's degree of self-awareness, when and why do people form an inference about whether a person is socially self-aware or not? Is self-awareness, in fact, a quality that people spontaneously evaluate in others at all? These questions are the focus of Chapter 3.

We propose that people are more likely to spontaneously evaluate a target person's degree of self-awareness when they are prompted to undergo a more thoughtful attribution process about that person's behavior. Evaluating self-awareness stands in contrast to evaluating many other traits, such as warmth and competence, in that it requires the observer to model the target person's mind (i.e., to infer something about what the target *thinks others think of them*). In order to engage in such cognitive processing, observers likely have to undergo a more elaborated attribution process when forming an impression of the target.

When might observers engage in such a process? Existing research suggests at least two contexts in which this should occur: when observers experience surprise at the target's behavior, and/or when observers evaluate the target negatively on other attributes or behaviors (Wong & Weiner, 1981). First, surprise or uncertainty can lead people to undergo a more thoughtful attribution process in order to understand why the unexpected outcome occurred (Wong & Weiner, 1981). Similar research has shown that violations of social expectations tend to induce greater arousal and a more elaborated appraisal process (Burgoon, 2015), and that—more generally—people often find uncertainty unpleasant (Buhr & Dugas, 2002; Grupe & Nitschke,

2011), which can lead them to attempt to reduce the uncertainty by making sense of the situation or finding meaning in it (C. G. Davis & Nolen-Hoeksema, 2001; Kay et al., 2008; Smyth et al., 2001). This, again, can occur within the context of social judgment specifically (FeldmanHall & Shenhav, 2019). Relatedly, when observers are suspicious of the motives behind a target's behavior, they tend to engage in more careful attribution processes, and thus are less likely to automatically assume that the target's underlying intentions match their behavior (Hilton et al., 1993).

Second, observing negative outcomes can also lead people to undergo a more careful attribution process. While Chapter 2 of this dissertation showed that creating a negative impression *alongside explicit cues about high or low self-awareness* affects perceived trustworthiness, it could be that in the absence of these explicit cues to self-awareness, negative behavior alone prompts spontaneous thoughts about a target's self-awareness as a way to explain that behavior and/or to rectify it with one's previous impression of the target. In general, negative information is more salient and more influential in people's judgments (Rozin & Royzman, 2001; Vaish et al., 2008). Within interpersonal judgment, negative traits are especially influential in people's impressions, especially in the moral domain (Skowronski & Carlston, 1987, 1989). Negative behaviors may also be more surprising in and of themselves, since norms exist around politeness and kindness in social interactions (Fraser, 1990)—thus further drawing people's attention to them and prompting an attribution process. It is perhaps not unexpected, then, that existing research has indeed documented that people are more likely to engage in a thoughtful attribution process when they observe negative rather than positive outcomes (Wong & Weiner, 1981).

Following from these findings, we propose that when observers undergo a more thoughtful attribution process because of surprise and/or negativity, one explanation that they may bring to mind to explain the target's behavior is whether or not the target is socially self-aware. To illustrate with an example, imagine that you know someone who behaves really kindly and pleasantly all of the times you have interacted with them, but who, in one conversation, says something you find a bit offensive. Because this statement stands in contrast to the rest of the target's kind behavior (and because it is a negative behavior in particular), you may be especially surprised by this statement and seek to explain why they would say something offensive. In searching for an explanation, one possibility you might think of is that the target was *unaware* that their statement could be perceived as offensive (i.e., that the target was lacking in self-awareness)—thus making this statement seem less contradictory to their apparent overall character (because it no longer implies negative intentions as strongly).

As an alternative example, imagine that you know someone who speaks to you in a curt and gruff manner each time you see them, but who one day acts very nice and friendly. Again, you might seek to explain the apparent contradiction in their behavior; one explanation you might come up with is that the target is now trying to manipulate your impression of them (perhaps for some self-interested motive)—and thus, is very *aware* of how you might be perceiving them. In both cases, we suggest that the observer will be more likely to consciously consider the target's degree of self-awareness than if the target's behavior did not seem surprising or contradictory, and/or if it did not result in negative outcomes.

This hypothesis yields an additional prediction: People may be more likely, on average, to spontaneously evaluate a target's self-awareness when that person seems to have *low* self-awareness compared to high self-awareness. To help illustrate, imagine another scenario: You

are meeting a new person for the first time, and they behave in an unfriendly manner toward you. You may be surprised by their behavior (even though this is a first-time interaction), again because positive (or neutral) actions are more normative and more common, and you may also perceive them negatively. In your attempt to reconcile this perception with your general expectations that most people have positive (or neutral) intentions toward others, you might attribute their behavior to *low* self-awareness in particular (e.g., this person might not have wanted to seem unfriendly, but simply did not realize they were coming across as unfriendly). Thus, negativity in particular (and the surprise that may often accompany it) might lead people to be more likely, on average, to think about others' self-awareness when they perceive it as *low* than high. Nevertheless, our theory predicts that people should still think of others' self-awareness at other times they are surprised, which in some cases might lead them to form an inference that others are *high* in self-awareness.

We tested our key prediction—that people spontaneously evaluate others' degree of self-awareness when they are surprised by the person's behavior and/or perceive their behavior negatively—in one pilot study and three main studies. First, we conducted a pilot study to explore the types of situations in which people report that a target's degree of self-awareness is most salient to them. This pilot study helped us generate the key hypothesis that we tested in the subsequent studies. In Study 1, we provide correlational evidence for our hypothesis using realistic stimuli (actual dating profiles from a publicly available dataset): We show that spontaneous thoughts about a target's self-awareness are most likely when the observer finds the profile surprising and/or forms a negative impression of the target. In Study 2, we test our hypothesis using a recall study design in which we manipulate positive or negative evaluations of the target on other traits (likability and competence) as a direct way of manipulating negativity

(and perhaps also, implicitly, surprise). In Study 3, we experimentally test our hypothesis in a scenario study in which we orthogonally manipulate the coherence of the target's behavior (to create surprise or lack thereof) as well as the positivity or negativity of the target's behavior.

Pilot Study: Your (Lack of) Self-Awareness is Showing

We ran a pilot study in order to get an initial sense of when people are most likely to think about others' degree of social self-awareness. We gave participants our definition of social self-awareness, and then asked them to think about a situation in which another person's degree of social self-awareness seemed very salient to them. We asked them to describe this situation and to answer a few questions about it.

Participants. Participants were recruited from a University-run behavioral science museum and laboratory in downtown Chicago and were compensated 100 points for their participation (points are redeemable for prizes at the museum). We preregistered that we would collect 50 participants and ended up with a final sample of 51 participants (64.71% female, 5.88% other gender, 62.75% non-White, $M_{\text{age}} = 29.80$, $SD_{\text{age}} = 10.66$).

Procedure. First, we gave participants a thorough definition of social self-awareness. We defined it as “accurately knowing what other people think of you,” and gave examples of high and low social self-awareness, both through bullet points and through graphic images (see Appendix C for the full set of instructions). We also specified that social self-awareness does not include other types of awareness, such as awareness of one's own internal experiences.

On the next page, participants answered a comprehension check question about the definition of social self-awareness. 46 participants (90.20%) answered correctly. Those who failed the check were told the correct answer and were allowed to proceed.

Next, participants were asked to think about a situation in which another person's degree of social self-awareness was very salient or noteworthy to the participant. We explained that this could include someone who was both high or low in self-awareness, or someone whose degree of self-awareness was salient for a different reason. We also told participants that if they could not think of a real past situation like this, they could simply imagine a situation in which another person's degree of social self-awareness would be salient or noteworthy.

On the following page, we asked participants to describe the situation they were thinking of (free-response). We then asked whether the situation they were thinking of was a real past experience or an imagined situation (multiple choice), as well as the following: "In the situation you wrote about, did the other person seem to have high, moderate, or low social self-awareness?" (*high, moderate, low, other [please explain]*).

Lastly, participants read the study debrief and were compensated with points. Participants had already provided demographic information (gender, age, race, etc.) in an earlier survey that we linked to their responses.

Results

Only 7 participants (13.73%) imagined a situation rather than recalling a real past experience. We included all participants in our analyses as we expected people's imagined situations to still be informative about when people find self-awareness most salient.

Level of self-awareness. One notable finding from this pilot is that low self-awareness seemed to be more salient than high self-awareness: A majority of participants (64.71%) reported that the target they were thinking of had low self-awareness, relative to 19.61% who said they were thinking of someone with high self-awareness, 13.73% who said moderate self-awareness,

| <i>Theme</i> | <i>Example</i> | <i>Level of Self-Awareness</i> |
|---|---|--------------------------------|
| Self-Awareness of Positive Qualities | “Zak [m]et his girlfriend's parents for the first time and made a good impression on them over lunch. The parents thought of him with high regards. Zak was aware of this.” | High |
| | “At work while working [o]n a project a colleague was extremely [sic] and contributed immensely to the best possible outcome, she did not feel she was helpful at all and did not think much of it.” | Low |
| | “I often run into coworkers that are well liked but think others do not like them for a variety of reasons.” | Moderate |
| Self-Awareness of Negative Qualities | “A coworker (Bob) consistently believes themselves to be a funny individual and brags about their ability to make others laugh, but other coworkers avoid them and only laugh at their jokes to placate them and move on.” | Low |
| | “A guy was talking about something that was upsetting me and he didn't realize that I was really annoyed. He kept talking about it and thinks we're closer friends than we are.” | Low |
| | “A fellow teacher in my school constantly sings out loud at any given social opportunity. For example, during our professional development on the last day of school, this Mr. S decided to sing out loud for over a minute for the entire staff, when everyone was trying to get the meeting over with. I don't know anyone on our staff who truly thinks he is a talented singer, and no one likes it when he bursts into song (think singing Wind Beneath My Wings during a large meeting in the auditorium), and yet he does it anyways.” | Low |
| | “...she was wearing a shirt and I asked her what it was and she immediately thought ‘well now my friends are gonna think I'm lame.’ The shirt was kind of not trendy so it made sense that she made that comment.” | High |
| Misunderstanding a Relationship | “One of my friends had hung out with a girl he liked 2 or 3 times. At the end of the last time they hung out, he said ‘I really enjoyed going on these dates with you.’ The girl responded by saying that she had no idea that they were dates. Then, it was incredibly awkward because my friend became flustered. He described to me thinking the entire time that the girl had liked him and that she was flirting with him. And she had had no idea that he was interested in her.” | Low |
| | “Someone was unaware of their relationship with a friend and overstepped their boundaries.” | Low |

Table 3.1. Selected responses from the Pilot Study, Chapter 3.

and 1.96% who said “other.” Although very preliminary, these findings might provide some initial suggestive evidence in support of our hypothesis that people spontaneously think about others’ degree of self-awareness more often when it is low than when it is high (because this inference of low self-awareness often accompanies negative behaviors, more so than positive or neutral ones).

Examples of responses. Some common themes, as well as examples of those themes, are reported in Table 1. All examples in the table are real (rather than imagined) experiences.

Discussion

Our pilot study showed that people observe others’ degree of social self-awareness in a wide variety of situations, and provided some initial suggestive evidence that it may be more common for people to notice others’ self-awareness when the target has low self-awareness than moderate or high self-awareness. In the rest of our studies, we sought to test our hypothesis that people spontaneously evaluate a target’s self-awareness when they are surprised by the target’s behavior and/or perceive it negatively.

Study 1: Self-Awareness—Online Dating Edition

In our first study, we tested whether thinking about others’ self-awareness correlates with being surprised by their behavior and/or perceiving their behavior negatively. We showed participants the text from real online dating profiles—a naturalistic context in which people frequently form consequential impressions of others—and asked them to list the first five thoughts that came to mind about the writer of the dating profile. We then asked them to categorize whether each thought was related to any of several traits we gave them, including self-awareness. We also measured perceptions of surprise by directly asking how surprised participants were by the content of the profile, and negativity by asking how negatively and how

positively participants perceived the person who wrote the profile. We predicted that listing more thoughts related to self-awareness would correlate with being surprised by the information in the profile, and with evaluating the profile negatively overall. This study's design and hypotheses were preregistered at https://aspredicted.org/T4N_JFV.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$2.00. We preregistered that we would collect 500 participants after excluding those who failed the attention check and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Because we were yoking each participant to one of 500 dating profiles, we preregistered to collect more respondents if we were left with some profiles that did not get evaluated by anyone, and to include only the first evaluation of each profile in our final dataset. Of the 596 who started the survey, we ended up with a final sample of 500 participants (34.40% female, 3.40% other gender, 29.80% non-White, $M_{\text{age}} = 30.83$, $SD_{\text{age}} = 8.71$) that fit these criteria.

Procedure. *Stimuli curation.* We obtained the text from real online dating profiles from the dating site OkCupid using a public dataset available at <https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles> (Albert & Escobedo-Land, 2015). The dataset was originally posted in 2015, so all profiles are from prior to 2015. From the approximately 60,000 dating profiles available, we randomly selected 500 profiles. Each profile had up to 10 columns of free-response text, each of which responded to a different prompt by the site (e.g., “About Me / Self summary,” “Current Goals / Aspirations,” “My Golden Rule / My Traits,” etc.). Although the dataset creators explicitly chose not to disclose which column corresponded to which prompt, we surmised that the first column seemed to be the most general prompt (e.g., “About Me / Self summary”), and chose to select all of our profile texts from this

column—i.e., we ended up with 500 short essays, each of which came from a different dater, and we disregarded any additional short essays that that same dater had written. We also limited our selection to essays at least 30 characters long so that our participants would have enough substance to evaluate. Occasionally, the profile text would include a link that the writer of the profile included (e.g., to their SoundCloud or YouTube page); we replaced each link with “[link redacted]” to protect anonymity. Otherwise, we did no other cleaning or alteration to the profile text. In our final sample, the median profile length was 461 characters. We did not include any other information along with the profile text (e.g., no demographic information about the writer).

Profile evaluation. After selecting these profiles, we conducted an online study. In order to ensure our study population would be well-positioned to evaluate dating profiles, we used Prolific’s prescreen functionality to limit our study population to those who had self-identified as single and who had answered yes to the following question: “Have you used any smartphone enabled dating apps?”

Once they started the study, we told participants that we would show them the text from a real person’s dating profile, and that we would ask for their impressions of this person based on the profile text. On the next page, we showed the actual text from one of the 500 profiles, and asked the participant to list five thoughts that came to mind about this person after reading their profile. We told participants that there were no right or wrong answers, and that we simply wanted them to write down anything that came to mind about this person. They were given five spaces to list each thought, and instructed to write the thoughts as complete sentences (with a minimum of 10 characters).

Self-coding of thoughts. On the next page, we gave participants instructions for how to self-code (i.e., categorize) their thoughts. We told them that they would be asked to label

whether their thoughts related to certain characteristics of the person. The characteristics we gave them were: trustworthiness, competence, likability, self-awareness, and kindness. For each characteristic, we gave them two examples of the types of thoughts that would fall under that category. For self-awareness, we gave examples specifically referring to social self-awareness (i.e., an awareness of what others think of oneself), in order to nudge participants toward thinking of self-awareness in the social sense; however, we wanted to avoid drawing too much attention to the self-awareness category and therefore did not try to emphasize a specific definition beyond participants' lay definitions. We explained that for each thought they wrote, participants could select one or more characteristic, and could also select an "other" option that would allow them to write a free-response characteristic.

Before asking participants to categorize their own thoughts, we had them practice categorizing three example thoughts in order to ensure that they understood the instructions. We gave them feedback after each practice question based on whether they categorized it as we intended or not. To see these practice questions and feedback, as well as the full categorization instructions, see Appendix C.

Once participants completed all three practice questions (regardless of whether they got them "correct" or not), participants proceeded to categorize their own thoughts. We showed them each of the five thoughts they had listed in sequence (on separate pages), and asked them to check the box next to any and all characteristics that their thought was about (*trustworthiness, competence, likability, self-awareness, kindness, other [please specify]*). On each page, we provided a reminder of the example thoughts we gave for each characteristic, for ease of reference.

Impressions measures. Next, we asked participants the following questions about their overall impressions of the profile: “Was there information in the profile that you were surprised this person included in their dating profile?”; “Based on this profile, how negatively do you perceive this person overall?”; and “Based on this profile, how positively do you perceive this person overall?” (all 1 = *none* or *not at all*, 7 = *a lot* or *extremely*). We also asked the following two questions, preregistered as exploratory: “How aware or unaware do you think the person who wrote this profile is of how their profile will be perceived by others?” (this was intended as a measure of overall degree of self-awareness) and “If you were actively dating and encountered this profile on a dating app or website—assuming you found this person attractive-looking—how likely would you be to try to date this person (based on just this profile text)? (*Note: For this question, please try to disregard gender preferences and answer solely based on your impression of the type of person this is.*)” (both -3 = *extremely unaware* or *extremely unlikely*, 3 = *extremely aware* or *extremely likely*). This last question was intended as a measure of behavioral intentions toward the profile writer. We included the parenthetical note because even though we did not provide explicit gender information with the profile, some profiles contained information that made it very clear what gender the writer was or what gender they preferred to date.

Finally, participants provided demographic information (gender, age, race) at the very end, along with an optional space to provide feedback.

Research assistant coders. The limitation of analyzing participants’ self-coded thoughts is that we cannot guarantee they were thinking of self-awareness in a social sense when categorizing thoughts as “self-awareness.” Even though we tried to suggest this by providing example thoughts of self-awareness in a social sense, it is possible that participants categorized some thoughts as “self-awareness” that referred to other types of self-awareness (e.g., awareness

of one's own internal emotional states), or otherwise did not understand what we meant by the term "self-awareness." To help address this shortcoming, we preregistered a secondary analysis in which we recruited two research assistants, blind to hypotheses, to conduct an additional round of coding of the participants' thoughts. These third-party codings are limited in that the coders have far less insight into what was in participants' minds when they wrote down the thoughts compared to the participants themselves, and thus may not realize what trait associations lay beneath the participants' explicit thoughts. Nevertheless, we used these codings to conduct a much more conservative test of our hypotheses. We expect the ground truth to lie somewhere between participants' self-coding and the third-party coding.

Results

For all analyses, we totaled the number of thoughts categorized as "self-awareness" for each participant (regardless of whether those thoughts were also categorized as another trait or not) and conducted our analyses on this total count of self-awareness thoughts for each participant; thus, each participant's value ranged from 0 to 5.

First, we examined a basic question: Do people tend to spontaneously evaluate others' self-awareness *at all*? To answer this, we simply took the overall average of each participant's number of self-awareness thoughts. We found that participants recorded an average of 1.83 self-awareness-related thoughts ($SD = 1.36$), which was significantly greater than 0, $t(499) = 29.95, p < .001$. Thus, participants do seem to spontaneously bring to mind self-awareness-related thoughts in the context of evaluating online dating profiles. The full distribution of number of thoughts is shown in Figure 3.1.

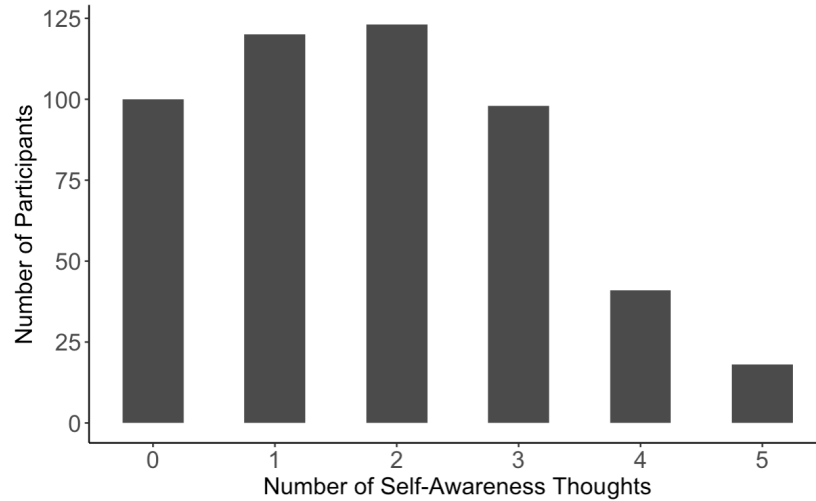


Figure 3.1. Histogram of the number of participants who listed each of the possible total numbers of self-awareness-related thoughts in Study 1, Chapter 3.

We also examined how the average number of thoughts related to self-awareness compared to the average number of thoughts related to other traits—see Appendix C for this analysis.

Next, we turned to analyzing our key hypothesis: whether listing thoughts related to self-awareness was associated with being surprised by the information in the profile and/or evaluating the profile negatively. As hypothesized, there was a significant positive correlation between number of self-awareness-related thoughts and our measure of surprise, $r(498) = .20, p < .001$. Thus, participants were more likely to spontaneously think about the target’s self-awareness when they were surprised by the profile.

There was also a significant positive correlation between number of self-awareness-related thoughts and our measure of how negatively participants evaluated the profile, $r(498) = .18, p < .001$, as hypothesized. Similarly, there was a significant negative correlation between number of self-awareness-related thoughts and our measure of how positively participants evaluated the profile, $r(498) = -.19, p < .001$. Thus, participants were more likely to

spontaneously think about the target's self-awareness when they perceived the target negatively overall, and were less likely to think about self-awareness when they perceived the target positively overall.

As one might expect, participant ratings of surprise were also positively correlated with negative evaluations, $r(498) = .30, p < .001$, and were negatively correlated with positive evaluations, $r(498) = -.20, p < .001$. To see if surprise and valence of evaluations remained significant predictors even when controlling for the other, we conducted a non-preregistered analysis: We ran two regressions with number of (participant-coded) self-awareness thoughts as the outcome variable, one with negative ratings and surprise as simultaneous predictors, and the other with positive ratings and surprise as simultaneous predictors. In both regressions, both predictors remained significant (with negative ratings: $b_{\text{surprise}} = 0.12, p < .001, b_{\text{negative_rating}} = 0.10, p = .005$; with positive ratings: $b_{\text{surprise}} = 0.12, p < .001, b_{\text{positive_rating}} = -0.12, p < .001$), suggesting independent effects.

On our exploratory measure of the profile writer's degree of self-awareness, we observed a significant negative correlation with number of self-awareness-related thoughts listed, $r(498) = -.22, p < .001$. In other words, participants' thoughts about self-awareness tended to be associated with low, rather than high, self-awareness in this context, similar to what we observed in the pilot study, and further supporting our hypotheses.

On our exploratory measure of likelihood of dating the profile writer, we observed a significant negative correlation with number of self-awareness-related thoughts listed, $r(498) = -.15, p < .001$. In other words, participants were less likely to want to date the target the more they spontaneously thought about that target's degree of self-awareness, which is consistent with the

fact that participants were also more likely to evaluate the target negatively overall—and as lower in self-awareness—when they thought more about that target’s self-awareness.

Research assistant coding. For the research assistant coding, we averaged the number of thoughts coded as each trait between the two research assistant coders, $r(498)$'s > 0.29 , p 's $< .001$. Using this metric, participants still listed significantly more than 0 thoughts about self-awareness, $t(499) = 6.70$, $p < .001$, though not surprisingly, the mean is much smaller in this case ($M = 0.19$, $SD = 0.62$). When analyzing the correlation with surprise ratings, we did not observe a significant correlation, $r(498) = .06$, $p = .154$, perhaps due to the lower overall mean. Similarly, there were no significant correlations with negativity ratings, $r(498) = -.01$, $p = .892$, or positivity ratings, $r(498) = -.01$, $p = .902$. Our exploratory measures of degree of self-awareness and dating likelihood also yielded non-significant correlations (p 's $> .251$).

Discussion

Study 1 provided correlational evidence in support of our main hypothesis: that people are more likely to think about a target’s self-awareness when they are surprised by that target’s behavior and/or perceive the target negatively overall, at least when analyzing their self-coding. In one regression analysis, we found that negative/positive valence remained a significant predictor of self-awareness thoughts even after controlling for surprise, and vice versa. This provides evidence for two independent effects that spark thoughts about self-awareness. In our next two studies, we sought to further test our hypotheses using recall and experimental paradigms.

Study 2: Negative Qualities Beget Thoughts on Self-Awareness

In Study 2, we tested whether perceiving someone negatively would be associated with spontaneously thinking more about that person’s self-awareness (compared to perceiving them

positively) using a recall paradigm. We asked participants to think of someone they knew whom they evaluated either negatively or positively, and to list five thoughts that came to mind when they thought of that person. For stimulus sampling, we varied the domain by asking for someone either high/low in likability or high/low in competence (between-subjects). We predicted that those who were asked to think of someone low in likability and/or competence would list more thoughts related to self-awareness than those asked to think of someone high in likability and/or competence. This study's design and hypotheses were preregistered at https://aspredicted.org/YBQ_Z1D.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$1.60. We preregistered that we would collect 400 participants after excluding those who could not think of their assigned target person, failed the attention check, failed the comprehension checks, and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 471 who started the survey, we ended up with a final sample of 400 participants (46.50% female, 1.75% other gender, 17.25% non-White, $M_{\text{age}} = 38.28$, $SD_{\text{age}} = 12.07$) that fit these criteria.

Procedure. In all conditions, we asked participants to bring to mind someone they know (e.g., a relative/family member, friend, coworker, acquaintance, or anyone else they had interacted with). We randomly assigned participants to one of four conditions in a 2 (Trait: likability vs. competence) x 2 (Level: high vs. low) between-subjects design. In the high likability condition, participants were asked to think of someone they knew who was very likable. In the low likability condition, participants were asked to think of someone they knew who was not very likable. Similarly, participants were asked to think of someone very competent or not very competent in the high and low competence conditions, respectively. We asked

participants to record the initials of the person they brought to mind. If participants were unable to think of anyone who met the description we gave, they were given the option to check a box saying they could not think of anyone. If they checked this box, they were taken directly to the final page of the survey where they reported demographics and then were still compensated with the full study payment.

Otherwise, participants proceeded to the next page of the study, where they were asked to write down the first five thoughts that came to mind when they thought about their target person. Our procedure was exactly the same as in Study 1: We told them that there were no right or wrong answers, and that we simply wanted them to write down anything that came to mind about this person. They were given five spaces to list each thought, and instructed to write the thoughts as complete sentences (with a minimum of 10 characters).

We then gave participants the exact same instructions as in Study 1 for how to self-code (i.e., categorize) their thoughts. Similarly, we had participants practice the coding using three example thoughts before actually coding their own thoughts. As in Study 1, we then showed them each of the five thoughts they had listed (on separate pages), and asked them to check the box next to any and all characteristics that their thought was about (*trustworthiness, competence, likability, self-awareness, kindness, other [please specify]*).

Finally, participants provided demographic information (gender, age, race) at the very end, along with an optional space to provide feedback.

Research assistant coders. Just like in Study 1, we preregistered a secondary analysis in which we recruited two research assistants, blind to hypotheses, to conduct an additional round of coding of the participants' thoughts. Once again, we expect the ground truth to lie somewhere between participants' self-coding and the (much more conservative) third-party coding.

Results

As in Study 1, we conducted our analyses by totaling the number of thoughts categorized as “self-awareness” for each participant (regardless of whether those thoughts were also categorized as another trait or not) and conducted our analyses on this total count of self-awareness thoughts (0 to 5) for each participant.

First, as in Study 1, we analyzed how often participants listed thoughts related to self-awareness overall. We did this by taking the overall average of each participant’s number of self-awareness thoughts, collapsing across all conditions. We found that participants recorded an average of 1.51 self-awareness-related thoughts ($SD = 1.42$), which was significantly greater than 0, $t(399) = 21.29, p < .001$. Thus, participants do seem to spontaneously bring to mind self-awareness-related thoughts when thinking about the targets we prompted them to think of. The full distribution of number of thoughts is shown in Figure 3.2.

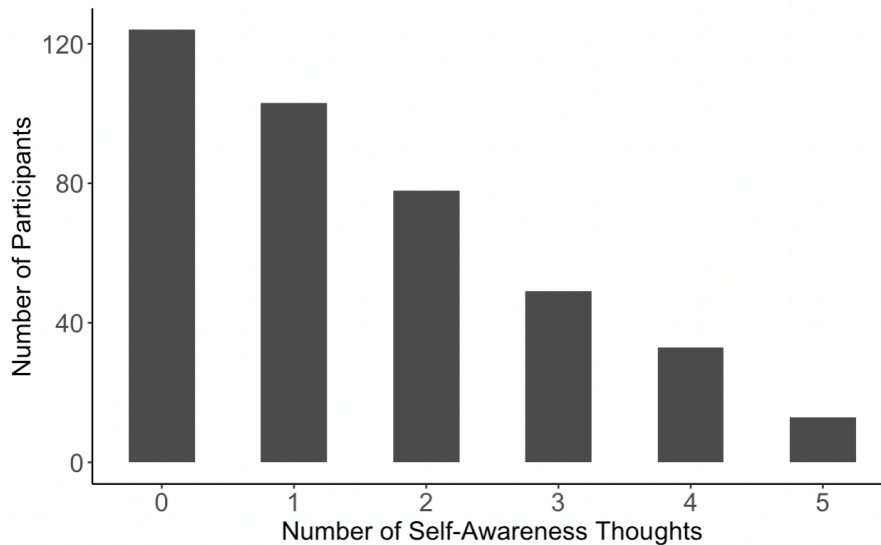


Figure 3.2. Histogram of the number of participants who listed each of the possible total numbers of self-awareness-related thoughts in Study 2, Chapter 3.

As in Study 1, we also examined how the average number of thoughts related to self-awareness compared to the average number of thoughts related to other traits—see Appendix C for this analysis.

Next, we tested our key hypothesis: whether the number of self-awareness-related thoughts varied by condition. We conducted an ANOVA with target trait condition (likability vs. competence), level condition (high vs. low), and their interaction as factors on our dependent measure, total number of self-awareness-related thoughts. There was no main effect of trait, $F(1,396) = 0.11, p = .740, \eta_p^2 < .01$, but a main effect of level condition, $F(1,396) = 46.86, p < .001, \eta_p^2 = .11$, with no interaction, $F(1,396) = 0.93, p = .337, \eta_p^2 < .01$.¹ As shown in Figure 3.3, participants spontaneously thought about the target’s self-awareness more often when they were asked to think of someone *low* in either likability or competence compared to someone *high* in either trait (likability: $t(396) = -5.58, p < .001, 95\% \text{ CI} = [-1.42, -0.68], d = -0.78$; competence: $t(396) = -4.12, p < .001, 95\% \text{ CI} = [-1.17, -0.41], d = -0.59$), with no difference between the two traits (p ’s $> .354$).

Research assistant coding. For the research assistant coding, we averaged the number of thoughts coded as each trait between the two research assistant coders, $r(398)$ ’s $> 0.40, p$ ’s $< .001$. Using this metric, participants still listed significantly more than 0 thoughts about self-awareness, $t(399) = 10.34, p < .001$, though once again, not surprisingly, the mean is much smaller in this case ($M = 0.21, SD = 0.41$).

¹ The lack of interaction was contrary to our preregistered hypothesis, as we expected participants to report the most self-awareness-related thoughts when they were asked to think of a target who was low in likability, followed by low in competence, followed by the two “high” conditions. As described in the text, we did observe this interaction when analyzing the research assistant codings instead.

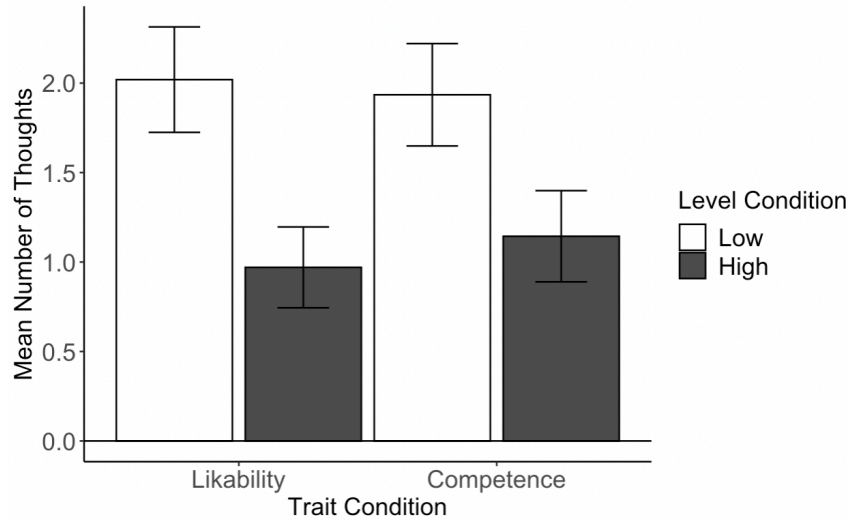


Figure 3.3. Mean number of (self-coded) self-awareness-related thoughts by trait and level condition in Study 2, Chapter 3. Error bars represent 95% confidence intervals.

When analyzing the number of self-awareness thoughts by condition, we observed a main effect of trait condition, $F(1,396) = 3.91, p = .049, \eta_p^2 < .01$, a main effect of level condition, $F(1,396) = 65.14, p < .001, \eta_p^2 = .14$, and a significant interaction, $F(1,396) = 8.09, p = .005, \eta_p^2 = .02$: For both traits, participants listed more thoughts having to do with self-awareness when the target was low in that trait ($M_{competence} = 0.27, SD_{competence} = 0.39; M_{likability} = 0.45, SD_{likability} = 0.58$) compared to high in that trait ($M_{competence} = 0.08, SD_{competence} = 0.23; M_{likability} = 0.04, SD_{likability} = 0.14$), but this difference was even larger for likability, $t(396) = -7.80, p < .001, 95\% CI = [-0.51, -0.30], d = -1.09$, than for competence, $t(396) = -3.66, p < .001, 95\% CI = [-0.30, -0.09], d = -0.52$.

Discussion

Study 2 provided additional support for our hypotheses, showing that when observers evaluate a target negatively, they are more likely to think about that target's self-awareness. It is possible that—as in Study 1—evaluating the target negatively is also associated with some degree of surprise, though our analyses in Study 1 suggest that each still exerts an independent

effect on self-awareness thoughts. In Study 3, we tested the role of surprise and negativity more directly by experimentally manipulating whether the target person's behavior seemed contradictory or not.

Study 3: Why are you behaving that way?

Study 3 orthogonally manipulated surprise and negativity in order to examine the causal role of each in prompting thoughts about self-awareness. Specifically, we gave participants information about a hypothetical target person's behavior, and varied whether the behavior was consistent with, versus contradictory to, the participant's expectations, in order to create surprise. Separately, we also manipulated whether the target's behavior was positive or negative. We predicted that both manipulations would exert independent effects on thoughts about the target's self-awareness. This study's design and hypotheses were preregistered at https://aspredicted.org/GV9_D93.

Participants. Participants were recruited online via Prolific and completed the study in exchange for \$0.80. We preregistered that we would collect 800 participants after excluding those who failed the attention check, comprehension checks, and/or provided gibberish or bot-like responses to a free-response question at the beginning of the study. Of the 872 who started the survey, we ended up with a final sample of 801 participants (50.31% female, 1.37% other gender, 21.10% non-White, $M_{age} = 39.40$, $SD_{age} = 13.59$) that fit these criteria.

Procedure. We told participants to imagine that they had a work colleague named Jamie. To enrich participants' experience of "meeting" Jamie, we first had them watch a short (~5 second) video clip of Jamie talking to them (with no sound). We randomly counterbalanced Jamie's gender and thus had one male and one female version of the clips. We obtained the clips

from a free stock footage website called Pexels (pexels.com). See Appendix C for links to the exact videos we used.

On the next page, participants read more information about Jamie. We randomly assigned participants to one of four conditions in a 2 (Focal behavior: positive vs. negative) x 2 (Contradiction: not contradictory vs. contradictory) between-subjects design. In all conditions, participants first read about their general overall impression of Jamie: They either read that they perceived Jamie as someone likable and kind-hearted, or as someone not likable and not kind-hearted. They then read about Jamie's behavior whenever the participant engaged in casual conversations with Jamie: Jamie was either described as asking the participant questions about themselves and trying to involve the participant as much as possible (i.e., desirable conversational behaviors), or as monologuing about themselves without involving the participant much in the conversation (i.e., undesirable conversational behaviors). Thus, we varied whether Jamie's behavior in conversations (the focal behavior) matched, or was contradictory to, the participant's overall impression of Jamie, and we varied whether it was positive or negative. For instance, in the negative focal behavior and contradictory condition, participants read that they liked Jamie and that Jamie seemed like a genuinely kind-hearted person who had done and said nice things to the participant, but that when they got into casual conversations with Jamie, Jamie would often launch into monologues without involving the participant in the conversation much, leaving the participant puzzled given that Jamie seemed genuinely kind and well-intentioned. In the other conditions, we varied the participant's general impression of Jamie and Jamie's behavior during the conversation accordingly (i.e., we fully crossed them to create all four possible combinations).

On the next page, we asked participants a comprehension check about the information we gave them about Jamie. Participants were given two tries to answer this correctly, and if they failed after the second try, the study automatically ended.

On the next page, participants responded to our key dependent measure. We asked participants the following about Jamie's behavior during conversations with the participant: "In your opinion, why do you think Jamie [frequently asks you questions and involves you in the conversation so nicely / often monologues so much without involving you more in the conversation]?" We asked participants to write 1-3 sentences in response.

On the next page, we asked an additional exploratory measure about Jamie's self-awareness during the conversations: "How aware or unaware do you think Jamie is of you how perceive [him/her] during your conversations with [him/her]?" (-3 = *extremely unaware*, 3 = *extremely aware*).

Finally, participants provided demographic information (gender, age, race) at the very end, along with an optional space to provide feedback.

We had two research assistants code participants' open-ended explanations for Jamie's behavior. Our main dependent measure was whether the explanation included a reference to Jamie's self-awareness (1) or not (0). We instructed the research assistants to categorize something as self-awareness if it included any reference to apparent awareness *or* lack of awareness of what others thought of Jamie. For instance, if the participant wrote that Jamie was involving them in the conversation nicely because Jamie was trying to make the participant like them or manipulate the participant's impression of them, this would count as self-awareness. On the other hand, if the participant wrote that Jamie was probably unaware they were monologuing so much, this would also count as self-awareness.

Results

We averaged the coding of the two research assistants, such that each participant's explanation was assigned either a 0 (if neither research assistant coded it as a 1), a 1 (if both coded it as a 1), or a 0.5 (if one research assistant coded it as a 1 and the other a 0, suggesting partial indication of a self-awareness reference). See Barneron, Choshen-Hillel, and Yaniv (2021) for a similar approach to averaging binary coding.

We conducted an ANOVA with focal behavior condition (positive vs. negative), contradiction condition (not contradictory vs. contradictory), and their interaction as factors on self-awareness references (treated as continuous). As predicted, we observed main effects of both focal behavior condition, $F(1, 797) = 42.67, p < .001, \eta_p^2 = .05$, and contradiction condition, $F(1, 797) = 35.48, p < .001, \eta_p^2 = .04$: Participants were more likely to think of Jamie's degree of self-awareness as an explanation for Jamie's behavior when the behavior seemed contradictory to their prior impression of Jamie than when it did not, and when the behavior was negative than positive. There was no interaction, $F(1, 797) = 2.23, p = .136, \eta_p^2 < .01$. See Figure 3.4.

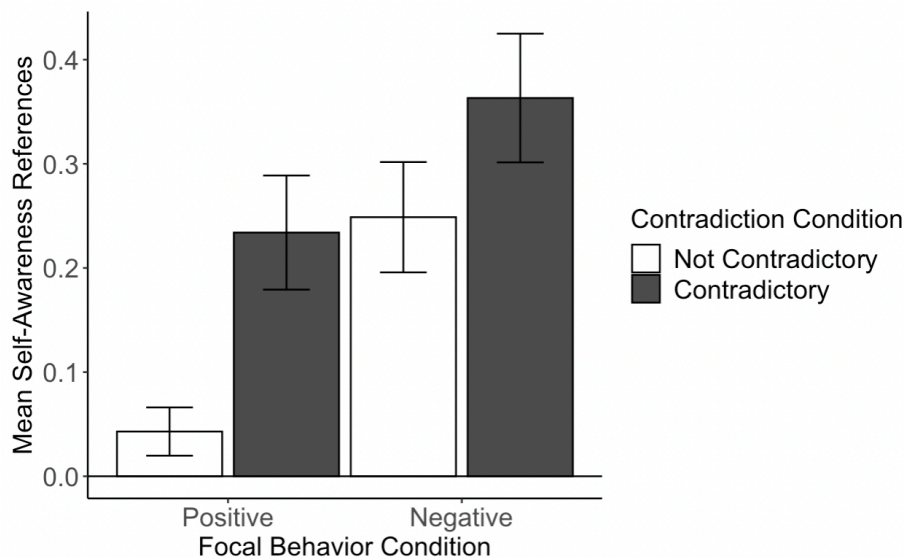


Figure 3.4. Mean number of references to the target's self-awareness in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

On our exploratory measure of the target's rated self-awareness during the conversation (the question we explicitly asked after the free-response question), there were main effects of both focal behavior condition, $F(1, 797) = 508.03, p < .001, \eta_p^2 = .39$, and contradiction condition, $F(1, 797) = 72.59, p < .001, \eta_p^2 = .08$, as well as a significant interaction, $F(1, 797) = 41.61, p < .001, \eta_p^2 = .05$. When the focal behavior was positive, the target was perceived as more self-aware when their behavior was not contradictory ($M = 1.43, SD = 1.04$) compared to contradictory ($M = -0.01, SD = 1.67$), $t(797) = 10.59, p < .001, d = 1.06$, but when the focal behavior was negative, there was little difference in the perceived self-awareness of the target whether their behavior was not contradictory ($M = -1.37, SD = 1.45$) or contradictory ($M = -1.57, SD = 1.23$), $t(797) = 1.46, p = .144, d = 0.15$. There was also a significant negative correlation between ratings of self-awareness and number of self-awareness references, $r(799) = -.22, p < .001$, again suggesting that people are more likely to think about a target's self-awareness when the target seems to lack self-awareness than to have high self-awareness.

We also asked the research assistants to code for several other types of explanations of Jamie's behavior, besides self-awareness. We report these results in Appendix C.

Discussion

Study 3 isolated the causal role of both surprise and negativity in sparking spontaneous thoughts about the target's self-awareness. As in Study 1, we showed that both surprise and negativity exert independent effects.

General Discussion

Across one pilot study and three full studies, we examined when people are most likely to spontaneously think about others' (social) self-awareness. We found that people are most likely to do so when they are surprised by the target person's behavior (Studies 1 and 3) and/or when

they perceive that target's behavior negatively (Studies 1-3). We propose that in such cases, the observer undergoes a more thoughtful attribution process in order to make sense of the target's behavior, and considers the target's degree of self-awareness as one possible explanation.

Our findings contribute theoretically by underscoring the role of perceived self-awareness in person perception, and by illuminating the underlying attribution process behind perceiving another person's degree of self-awareness. We proposed a theory in which self-awareness may spontaneously come to mind when observers are prompted to pay more attention to the target and to spend more time evaluating their behavior. While much past work has uncovered how we evaluate a variety of traits in others, such as competence, sociability, and morality (Brambilla et al., 2011, 2019; Fiske et al., 2007; Goodwin et al., 2014; Koch et al., 2016), no past work, to our knowledge, has examined how (and whether) people evaluate self-awareness in others. As described, evaluating a target's self-awareness—in contrast to evaluating these other traits—requires the observer to model the target's mind (i.e., to infer something about how that target *thinks others' see them*). Thus, the process of perceiving others' self-awareness may be unique relative to other traits in that it requires circumstances that prompt a more thoughtful attribution process (e.g., surprise or negativity).

When taken together with other research on the interpersonal consequences of self-awareness presented in Chapters 1-2 of this dissertation, these findings have practical implications for optimal impression management. If a target creates circumstances in which observers are more likely to consciously evaluate that target's degree of self-awareness (e.g., by behaving unpredictably), then that target's subsequent behavior may be judged differently. In particular, perceiving a target as self-aware tends to amplify observers' existing perceptions of the target's trustworthiness, while perceiving a target as *lacking* self-awareness leads observers

to perceive the target's behavior as less diagnostic of their overall character (see Chapter 2). As a result, people may want to adjust their behavior—or to choose how and when to provide explicit explanations for it—if it may draw attention to their self-awareness (or lack thereof), which in turn could affect impressions of subsequent (unrelated) behaviors.

These findings open up a number of interesting questions for future research to explore. We explored a possible trigger of self-awareness-related thoughts in observers—surprise and negativity—but surely there must be others; future research could uncover some of these additional triggers, including those that prompt similar attribution processes. Relatedly, future research could examine moderation by certain types of interactions, relationships, or contexts that might make others' self-awareness more or less salient and discernable.

Our data may also provide some initial suggestive evidence that surprise about positive behaviors prompts thoughts not only about self-awareness, but also about ulterior or self-interested motives. In Study 3, we found that people who observed positive behavior that contradicted a negative prior impression not only thought about the target's self-awareness more, but also thought about the target's potential ulterior motives substantially more than in all of the other conditions (see Appendix C). This pattern may suggest that when people experience surprise about a positive behavior, they are more likely to think that the target is intentionally considering how others' view them in order to gain something strategic or self-interested, rather than because the target genuinely wants to be kind to others. However, more research is needed to systematically examine when people do or do not infer motives, including self-interested ones, alongside inferences about self-awareness. Overall, our findings provide initial evidence that observers seek to make sense of surprising and/or negative target behaviors by bringing to mind that target's degree of self-awareness.

Appendix A: Supplemental Material, Additional Tables, and Stimuli for Chapter 1

Supplemental material for Chapter 1 can be found on OSF:

https://osf.io/y7bj8/?view_only=1f71e792fafa452687f5e1439cb5e5aa

Main Effects Tables for Studies 2-4 in Chapter 1

| | | <u>Main Effect of Disclaimer Condition</u> | <u>Main Effect of Credibility Condition</u> |
|------------------|----|---|--|
| Self-Awareness | 2a | $F(2,594) = 0.70, p = .402, \eta_p^2 < .01$ | $F(2,594) = 41.39, p < .001, \eta_p^2 = .12$ |
| | 2b | $F(2,593) = 2.68, p = .102, \eta_p^2 < .01$ | $F(2,593) = 54.84, p < .001, \eta_p^2 = .16$ |
| Warmth Composite | 2a | $F(2,594) = 0.00, p = .995, \eta_p^2 < .01$ | $F(2,594) = 89.59, p < .001, \eta_p^2 = .23$ |
| | 2b | $F(2,593) = 0.40, p = .525, \eta_p^2 < .01$ | $F(2,593) = 101.26, p < .001, \eta_p^2 = .25$ |
| Competence | 2a | $F(2,594) = 1.92, p = .166, \eta_p^2 < .01$ | $F(2,594) = 243.76, p < .001, \eta_p^2 = .45$ |
| | 2b | $F(2,593) = 3.28, p = .070, \eta_p^2 < .01$ | $F(2,593) = 201.19, p < .001, \eta_p^2 = .40$ |

Table A.1.1. Main effects in Studies 2a-b, Chapter 1. Significant effects are in bold.

| | | <u>Main Effect of Disclaimer Condition</u> | <u>Main Effect of Credibility Condition</u> |
|------------------|------------------|---|--|
| Self-Awareness | Sociability (4a) | $F(1,400) = 0.02, p = .889, \eta_p^2 < .01$ | $F(1,400) = 0.11, p = .735, \eta_p^2 < .01$ |
| | Generosity (4b) | $F(1,397) = 1.05, p = .306, \eta_p^2 < .01$ | $F(1,397) = 59.26, p < .001, \eta_p^2 = .13$ |
| | Music (4c) | $F(1,396) = 0.65, p = .420, \eta_p^2 < .01$ | $F(1,396) = 210.77, p < .001, \eta_p^2 = .35$ |
| | Athletic (4d) | $F(1,392) = 14.79, p < .001, \eta_p^2 = .04$ | $F(1,392) = 119.92, p < .001, \eta_p^2 = .23$ |
| Believability | Sociability (4a) | $F(1,400) = 6.26, p = .013, \eta_p^2 = .02$ | $F(1,400) = 93.64, p < .001, \eta_p^2 = .19$ |
| | Generosity (4b) | $F(1,397) = 0.70, p = .403, \eta_p^2 < .01$ | $F(1,397) = 158.89, p < .001, \eta_p^2 = .29$ |
| | Music (4c) | $F(1,396) = 0.22, p = .640, \eta_p^2 < .01$ | $F(1,396) = 549.95, p < .001, \eta_p^2 = .58$ |
| | Athletic (4d) | $F(1,392) = 29.54, p < .001, \eta_p^2 = .07$ | $F(1,392) = 506.57, p < .001, \eta_p^2 = .56$ |
| Warmth Composite | Sociability (4a) | $F(1,400) = 1.50, p = .221, \eta_p^2 < .01$ | $F(1,400) = 22.73, p < .001, \eta_p^2 = .05$ |
| | Generosity (4b) | $F(1,397) = 1.23, p = .268, \eta_p^2 < .01$ | $F(1,397) = 342.99, p < .001, \eta_p^2 = .46$ |
| | Music (4c) | $F(1,396) = 0.00, p = .948, \eta_p^2 < .01$ | $F(1,396) = 218.11, p < .001, \eta_p^2 = .36$ |
| | Athletic (4d) | $F(1,392) = 15.87, p < .001, \eta_p^2 = .04$ | $F(1,392) = 175.26, p < .001, \eta_p^2 = .31$ |
| Competence | Sociability (4a) | $F(1,400) = 2.31, p = .129, \eta_p^2 < .01$ | $F(1,400) = 5.82, p = .016, \eta_p^2 = .01$ |
| | Generosity (4b) | $F(1,397) = 3.28, p = .071, \eta_p^2 < .01$ | $F(1,397) = 98.82, p < .001, \eta_p^2 = .20$ |
| | Music (4c) | $F(1,396) = 0.27, p = .601, \eta_p^2 < .01$ | $F(1,396) = 315.93, p < .001, \eta_p^2 = .44$ |
| | Athletic (4d) | $F(1,392) = 14.03, p < .001, \eta_p^2 = .03$ | $F(1,392) = 241.25, p < .001, \eta_p^2 = .38$ |

Table A.1.2. Main effects in Studies 3a-d, Chapter 1. Significant effects are in bold.

| | <u>Main Effect of Disclaimer Condition</u> | <u>Main Effect of Credibility Condition</u> |
|--|---|--|
| Awareness of How Others See Target | $F(1,393) = 0.69, p = .407, \eta_p^2 < .01$ | $F(1,393) = 8.07, p = .005, \eta_p^2 = .02$ |
| Sincerity | $F(1,393) = 0.50, p = .481, \eta_p^2 < .01$ | $F(1,393) = 14.35, p < .001, \eta_p^2 = .04$ |
| Warmth Composite | $F(1,393) = 0.50, p = .481, \eta_p^2 < .01$ | $F(1,393) = 24.38, p < .001, \eta_p^2 = .06$ |
| Competence | $F(1,393) = 1.44, p = .231, \eta_p^2 < .01$ | $F(1,393) = 95.19, p < .001, \eta_p^2 = .19$ |
| Exploratory Task Choice (Continuous) | $F(1,393) = 0.09, p = .759, \eta_p^2 < .01$ | $F(1,393) = 47.95, p < .001, \eta_p^2 = .11$ |
| Exploratory Task Choice (Binary) | $b = -0.02, p = .933$ | $b = -0.97, p < .001$ |
| Exploratory Social Interaction Measure | $F(1,393) = 0.65, p = .419, \eta_p^2 < .01$ | $F(1,393) = 19.61, p < .001, \eta_p^2 = .05$ |

Table A.1.3. Main effects in Study 4, Chapter 1. Significant effects are in bold.

Study Stimuli and Scenarios in Studies 2-3, Chapter 1

Studies 2a-b:

Imagine that you work at a mid-size corporation. You are at work, having a conversation with one of your colleagues in the office break room.

[High/low-credibility conditions]: You know that this colleague has a high-ranking [low-ranking] position at the company, is [not] very well-respected, and is [not] very competent at their job. During the conversation, this colleague says to you:

[Control condition]: You do not know this colleague's position at the company, how well-respected they are, or how competent they are at their job. During the conversation, this colleague says to you:

“I won the yearly national manager award at my old job.”

[or]

“I graduated with honors when I was in college.”

Disclaimers:

[Study 2a]

“I know this may seem hard to believe, but...”

“This may sound strange to you, but...”

[Study 2b]

“I'm not that smart, but...”

“I'm no genius, but...”

Studies 3a-d:

Imagine that you work at a mid-size corporation. You are at work, having a conversation with one of your colleagues in the office break room.

[Sociability version]: You know that this colleague is outgoing and charming [shy and socially awkward]. They tend [not] to make friends easily and always seem to be going to [don't seem to go to many] social gatherings.

[Generosity version]: You know that this colleague is generous [selfish] and always [never] seems to consider other people.

[Musical talent version]: You know that this colleague is [not] very talented at playing the guitar. They played [tried playing] their guitar at an office gathering one time and it blew everyone away [it was so bad, people felt embarrassed for them].

[Athletic skill version]: You know that this colleague is [not] very athletically skilled and seems to be really [un]coordinated. Whenever your office holds informal social gatherings, this colleague seems to be the best [worst] at whatever activity is available, whether it's frisbee-throwing, whiffle ball, or running relay races.

During the conversation, this colleague says to you:

[Sociability version]

"I was voted most popular in high school."

[or]

"I've been invited to 7 different parties next weekend."

[Generosity version]

"I've volunteered for at least 10 different charities this year."

[or]

"I've donated more money to charity than I've spent on myself in the last month."

[Musical talent version]

"I once won an award for my guitar playing."

[or]

"I was chosen to be the lead guitar player in a band that was signed with one of the major record labels."

[Athletic skill version]

"I have the fastest mile running time in my running group."

[or]

"I was recruited by multiple colleges who wanted me to play on their Division 1 soccer teams."

Disclaimers:

[All study versions]

“I know this may seem hard to believe, but...”

“This may sound strange to you, but...”

[Sociability version]

“I’m no social butterfly, but...”

“I’m not that outgoing, but...”

[Generosity version]

“I’m no saint, but...”

“I’m not that generous, but...”

[Musical talent version]

“I’m no virtuoso, but...”

“I’m not that musically talented, but...”

[Athletic skill version]

“I’m no star athlete, but...”

“I’m not that athletically skilled, but...”

Stimuli in Study 4, Chapter 1

Essay prompt (displayed to participants):

Q1. INSTRUCTIONS: *In the space below, please write an essay that describes a movie that you would recommend, and explains WHY this movie is a good movie. Please write it as though you are writing a movie review for an audience that hasn't seen the movie.*

Low-credibility condition essay:

My favorite movie is called like sunday, like rain. I think its a really good movie, although there are lots of other movies I like to. There are a lots of things I like about the movie, for example the movie itself, which is really fun to watch. Also the story reminds me of something that happened when I was a kid. But I would want to watch it again if I could. I think another thing I liked was the fact that the movie was a better length than some of the other movies i've watched, even though I don't know if that was for sure or not, I would've' wanted to know more about the characters but I don't think the director wanted to get in to that too much, if I had to guess, I would say it went through a lot of different revisions. Another thing I like, though, is that it doesn't drag on too long, which I think is not very common these days but is important because a lot of people have busy schedules and need something they can watch in a shorter amount of time so that they enjoy something and then get on with their days, which people don't always think about when their making these movies. If you asked me to recommend another movie, though, i would have to think about it some more, I'm not really sure what my second favorite movie would be, but i do think this one is probably my faverite movie.

High-credibility condition essay:

For those looking to see an emotional and thought-provoking movie, I recommend a movie called Like Sunday, Like Rain. Set in modern-day New York City, the story follows a young woman named Eleanor who becomes the nanny for a boy from a wealthy family. At the time of their meeting, both protagonists are facing their own individual struggles, and despite coming from vastly different backgrounds, they each become unlikely sources of support and understanding for one another. Three aspects in particular stand out as especially compelling about this movie. First, the characters themselves are well-crafted. Their actions and motivations are believable, and the ways they experience their struggles - from financial difficulties to tense family dynamics - are likely to be relatable to any viewer, regardless of the viewer's own background. Second, and related to the first point, the actors' performances are especially effective in bringing the characters to life. This is particularly impressive given that one of the main characters is played by a child actor - yet he carries his role convincingly and naturally. Finally, the soundtrack of the movie serves to elevate the movie's storyline. Indeed, music is an important motif throughout the movie's plot, and thus, part of the soundtrack is interwoven with the music that exists within the movie's storyline. Moreover, the music evokes a strong emotional response, thereby heightening the movie's impact. Overall, then, I recommend this movie for its strong writing, acting, and music.

Low-credibility condition survey responses (shown to participants):

Q325. Have you ever had any formal instruction in writing?

no I was never formally taught how to write

Q326. Have you ever submitted any writing to a publication (e.g. newspaper, magazine, or journal)? If so, was it accepted?

yes, I sent something to my local newspaper once, they didn't publish it

Q327. Have you ever had a full-time or part-time job that involved a lot of writing? If so, please describe what kinds of writing you did.

no

Q328. Do you give permission for us to share these responses with the other participant in this study?

Yes

No

Figure A.1.1. Low-credibility survey responses shown in Study 4, Chapter 1.

High-credibility condition survey responses (shown to participants):

Q325. Have you ever had any formal instruction in writing?

Yes, I'm studying English in college and have participated in several writing workshops.

Q326. Have you ever submitted any writing to a publication (e.g. newspaper, magazine, or journal)? If so, was it accepted?

Yes, I've submitted to a few literary magazines and one local newspaper. All were accepted.

Q327. Have you ever had a full-time or part-time job that involved a lot of writing? If so, please describe what kinds of writing you did.

Yes, I work part-time as a contributor to a content website geared toward college students. I write essays and stories about my experiences as a college student and young adult.

Q328. Do you give permission for us to share these responses with the other participant in this study?

Yes

No

Figure A.1.2. High-credibility survey responses shown in Study 4, Chapter 1.

Procedure in Study 5b, Chapter 1

Below are snapshots of our job post on Upwork:




| | |
|--|---|
| <p>Recruiting & Talent Sourcing Invite-Only Renewed 3 months ago</p> <p>📍 Only freelancers located in the U.S. may apply.</p> | <p>🔄 Reuse posting 👁️ View posting ✅ View hires</p> |
| <p>Needs to hire 99 Freelancers</p> <p>We are looking for HR professionals, managers, researchers, and/or those with relevant hiring-related or research-related expertise to help us evaluate candidates for an entry-level research assistant position.</p> <p>This task will be very brief: You will simply review 1 candidate's profile and tell us whether you would recommend we hire the candidate. We expect that this evaluation should take no more than 10 minutes in total, as the profile is very brief.</p> <p>You will be one of a number of professionals we hire, each of whom will review one profile selected from our large pool of applicants. Your recommendation will help us determine which candidates to move to the second round of review, at which point we will collect more information from them that we will use to make our final evaluation (this will not be part of your job - you will only be asked to assist with this first round of review, and will thus only evaluate 1 candidate's short profile).</p> | <p>About the client </p> <p>✅ Payment method verified ★★★★★ 4.95 of 204 reviews</p> <p>United States Chicago 9:26 am</p> <p>1 job posted 100% hire rate, 1 open job</p> <p>\$3K total spent 316 hires, 0 active</p> <p>Member since Apr 18, 2023</p> |
| <p>Featured Job  \$10.00 Fixed-price</p> <p>Entry level  I am looking for freelancers with the lowest rates</p> | <p>Job link</p> <p>https://www.upwork.com/jobs</p> <p>Copy link</p> |
| <p>Project Type: One-time project</p> | |

Figure A.1.3. First half of job posting for Study 5b, Chapter 1.

Skills and Expertise

Recruiting & Talent Sourcing Deliverables

- Human Resource Management
- Candidate Evaluation
- LinkedIn Recruiting
- Candidate Recommendation

Industry

- HR & Business Services

Recruiting & Talent Sourcing Skills

- Candidate Interviewing
- Candidate Sourcing
- Recruiting
- Staff Recruitment & Management
- Candidate Management

Other

- Communications
- Resume Screening
- Administrative Support
- Candidate Mangement
- Research Methods
- Research Papers
- Research Proposals
- Training
- Human Resources Consulting
- Resume Screening
- Academic Research
- Academia

Preferred qualifications

English level:  Conversational

Activity on this job

Proposals:  5 to 10

Last viewed by client:  1 month ago

Interviewing: 6

Invites sent: 1,337

Unanswered invites: 735

Figure A.1.4. Second half of job posting for Study 5b, Chapter 1.

Below were the steps of our procedure for hiring each worker on Upwork:

1. We made the job posting public and sent out invitations to people (“participants”) to review the job posting.

Hello!

I'd like to invite you to take a look at the job I've posted about evaluating a candidate's profile for a research assistant position (\$10 for a 10-minute task). Please submit a proposal if you're available and interested!

<https://www.upwork.com/jobs/~01c1826882daac9602>

[RA Name]

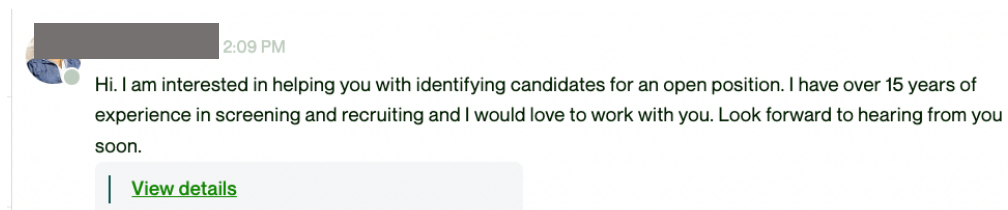
To do this, we used Upwork’s search function, typing in keywords such as “recruiter,” “talent acquisition,” “HR,” “PhD researcher,” etc. We also specified the talent type to be “freelancer,” English level as “conversational,” and location as “US only.”



We sent out invitations when a participant’s Upwork job title (see below for an example), resume, or employment history (not the Upwork project working history) showed that they had relevant experiences.



2. Participants who accepted our invitation submitted their proposals with a proposed bid and reasons for being interested in the job. Some uninvited participants who saw the job posting on their own (e.g., through searching jobs on the site) also submitted a proposal.



3. We reviewed the proposals and sent out the offer.

RA Name sent an offer 3:32 PM

We are looking for HR professionals, managers, researchers, and/or those with relevant hiring-related or research-related expertise to help us evaluate candidates for an entry-level research assistant position.

This task will be very brief: You will simply review 1 candidate's profile and tell us whether you would recommend we hire the candidate. We expect that this evaluation should take no more than 10 minutes in total, as the profile is very brief.

You will be one of a number of professionals we hire, each of whom will review one profile selected from our large pool of applicants. Your recommendation will help us determine which candidates to move to the second round of review, at which point we will collect more information from them that we will use to make our final evaluation (this will not be part of your job - you will only be asked to assist with this first round of review, and will thus only evaluate 1 candidate's short profile).

Est. Budget: \$10.00

Milestone 1: Help evaluate job candidates

Due: Friday, Jul 28, 2023

Amount in escrow: \$10.00

[View details](#)

We rejected a proposal when the proposal, the participant's resume, employment history, and job title all showed that they had no relevant experience. We also rejected a proposal if the bid price was too high, and the participants refused to negotiate over the payment. (Usually, we reached out to participants when they proposed over \$10. We ultimately contracted with one participant at \$12 and another for \$11, thus slightly over our set price of \$10. There were also two participants who proposed \$5, and we just contracted with them at \$5. All other participants were paid exactly \$10.)

4. Participants reviewed the offer and accepted it. When they submitted the proposal but didn't accept the offer (i.e., pending offer), we sent multiple reminders to them. If, despite the reminders, they did not take action, we would reject their proposals.

Hi! Thanks for submitting the proposal. If you're still interested in the task, please accept the offer as soon as possible. Thanks!

5. We sent out a document where participants could find the task instructions and candidate information through Upwork's Message function and recorded the date/time in the Study Log.

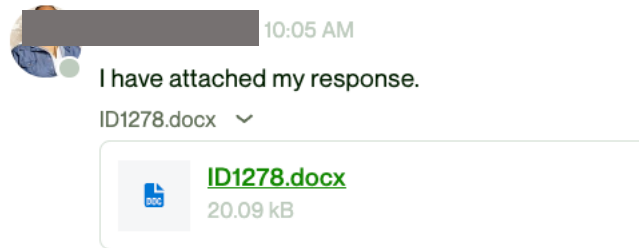
Thanks for helping with the task!

For this job, we would like you to review 1 candidate's application profile and tell us whether you would recommend we hire the candidate.

I will send you a document in a minute. Please refer to it for more information about the task.

(Document attached)

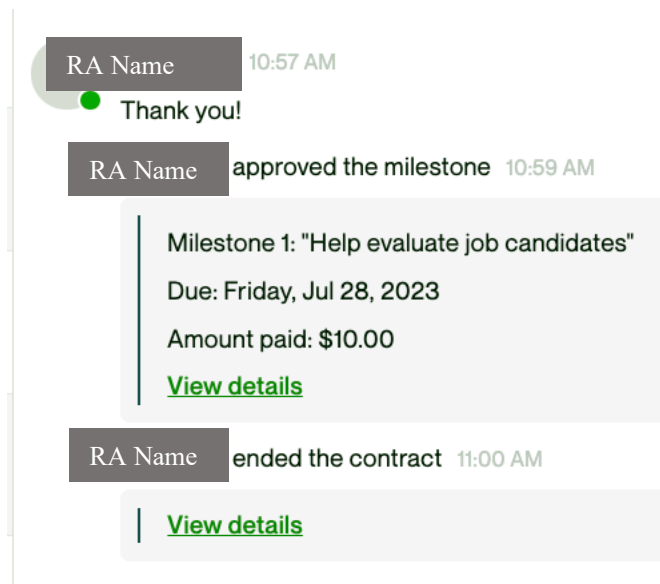
- Participants returned the document through Upwork’s Message function or directly typed their responses into the chat box (and we would help them fill in the document). We recorded the date/time and their responses in the Study Log.



When the participants did not return the document 2-3 days after receiving it, we sent them a reminder. If they didn’t complete the task after we sent multiple reminders, we would withdraw the contract.

Hello! Thank you once again for agreeing to assist with the task. I wanted to provide a friendly reminder that the deadline for the task has passed. Please complete it as soon as possible. If you have any questions, please don't hesitate to reach out. Thanks!

- After participants successfully completed the task, we sent out the payment, ended the contract, and completed the evaluation form required by Upwork.



Appendix B: Supplemental Material for Chapter 2

Supplemental Results from Studies in Chapter 2

Study 1

Additional measures. We asked the following exploratory measures about general liking: “I think Taylor is a likable person” and “I would choose to spend more time with Taylor in the future, if I had the opportunity” (both -3 = *strongly disagree*, 3 = *strongly agree*). We also asked the following exploratory measures to capture perceived social skill: “I believe that Taylor would be skilled at navigating future social interactions”; “I believe that Taylor is generally skilled at anticipating how others will respond to [him/her]”; and “I believe that Taylor is generally skilled at knowing how [his/her] behavior is affecting others” (all -3 = *strongly disagree*, 3 = *strongly agree*). Finally, we also asked the exact same measures of benevolence, ability, and integrity as in Studies 3 and 5-7 in Chapter 2.

We combined our measures of social skill ($\alpha = 0.96$), benevolence ($\alpha = 0.95$), ability ($\alpha = 0.85$), and integrity ($\alpha = 0.92$) into their respective composites. Results on the two liking measures and each of these composites are visualized in Figures A.2.1-A.2.2.

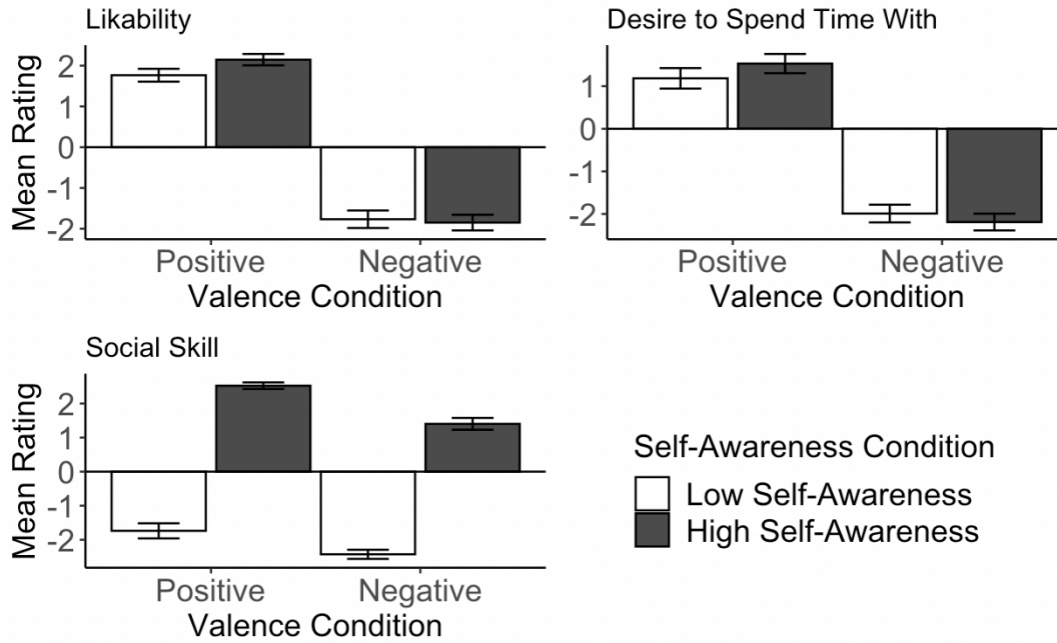


Figure A.2.1. Results on likability, desire to spend time with, and perceived social skill in Study 1, Chapter 2.

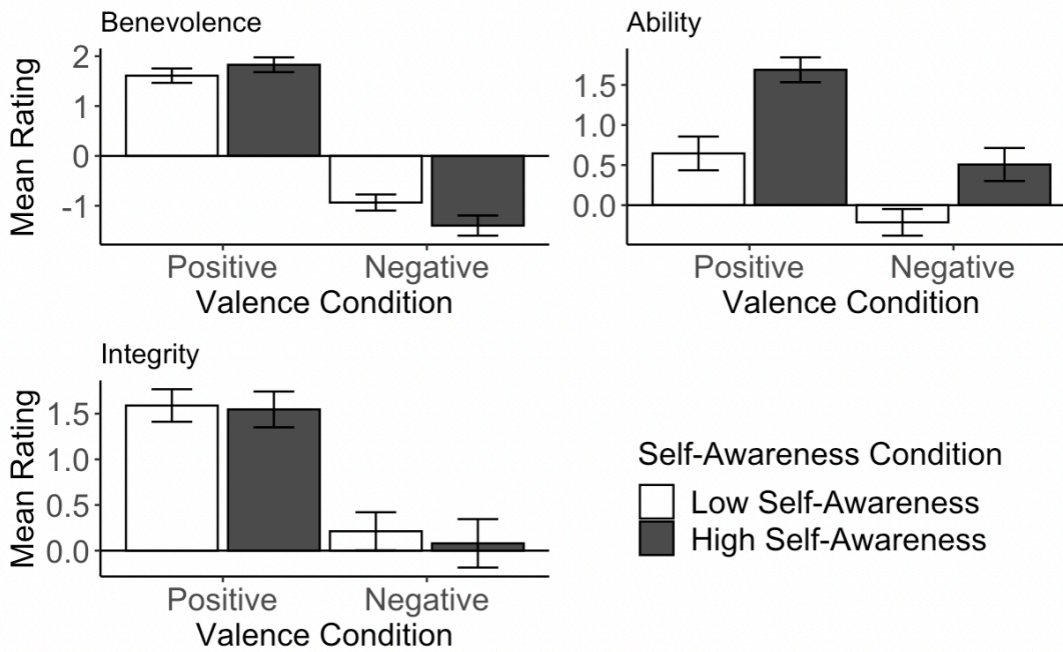


Figure A.2.2. Results on benevolence, ability, and integrity composites in Study 1, Chapter 2.

Study 3

Mutability measures. We asked the following question to measure perceptions of how mutable (i.e., easily changeable) unfriendliness/rudeness is in general: “When thinking about unfriendliness/rudeness, how easy or difficult do you think it would be for someone to change how unfriendly/rude they are toward others in general?” (-3 = *extremely easy*, 3 = *extremely difficult*). We asked the following question to measure perceptions of the mutability of the specific target’s unfriendly/rude behavior: “How easy or difficult do you think it would have been for [initials] to change how unfriendly/rude (s)he was being toward you (if he/she tried to change it)?” (-3 = *extremely easy*, 3 = *extremely difficult*).

We did not observe a significant interaction between self-awareness condition and either measure in predicting overall trust toward the target (general mutability: $b = -0.00$, $t(303) = -0.03$, $p = .979$; specific mutability: $b = -0.11$, $t(303) = -1.24$, $p = .217$).

Study 5

Results on individual measures in trust and ability composites. The results on the individual items in the trust and ability scales are visualized in the Figures A.2.3-A.2.4.

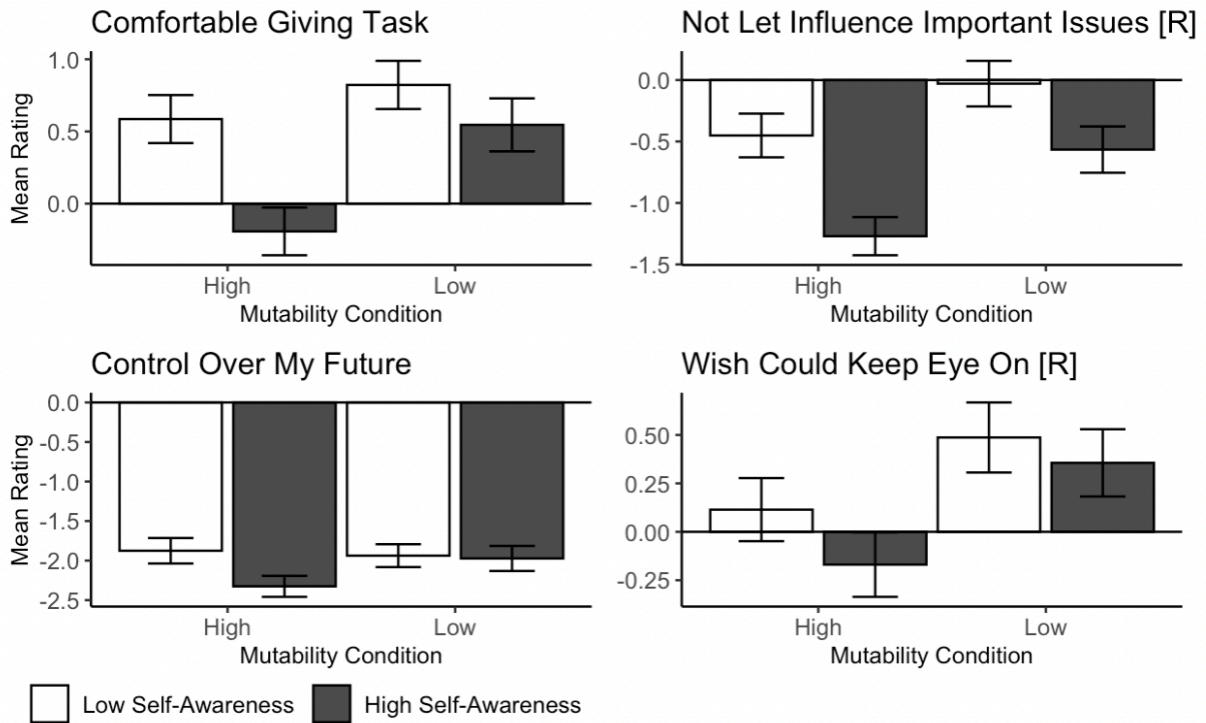


Figure A.2.3. Results on individual items in trust composite in Study 5, Chapter 2. Reverse-coded measures are labeled with [R].

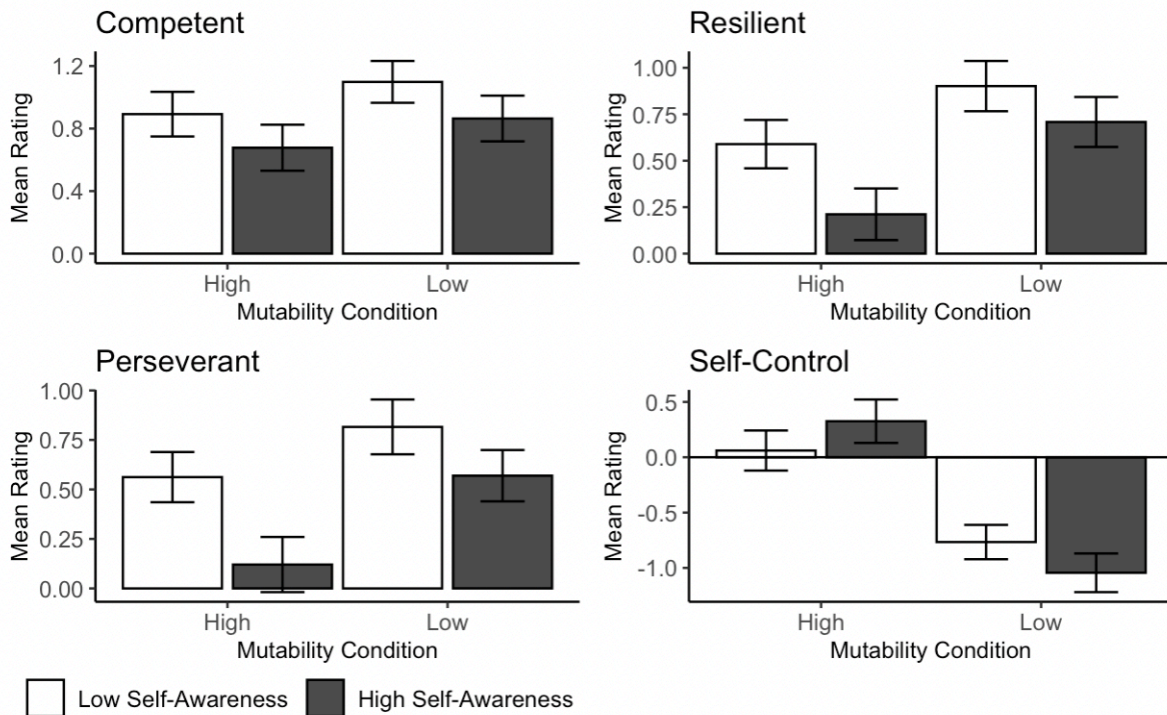


Figure A.2.4. Results on individual items in ability composite in Study 5, Chapter 2.

Study 6

Results on exploratory measures. On our composite measure of ability, there was a main effect of self-awareness condition, such that the target was perceived as less competent when they were high in self-awareness than when they were low in self-awareness, $b = -0.23$, $t(1598) = -3.59$, $p < .001$, and a main effect of mutability condition, $b = 0.47$, $t(1598) = 7.40$, $p < .001$, with no interaction, $b = 0.06$, $t(1598) = 0.65$, $p = .514$.

On our composite measure of integrity, there was a main effect of self-awareness condition, such that the target was perceived as lower in integrity when they were high in self-awareness than when they were low in self-awareness $b = -0.28$, $t(1598) = -3.22$, $p = .001$, and a main effect of mutability condition, $b = 0.81$, $t(1598) = 9.45$, $p < .001$, with no interaction, $b = 0.07$, $t(1598) = 0.57$, $p = .566$.

On our trust game decision measure, there was a main effect of self-awareness condition, such that targets wanted to send less money to the target when the target had high self-awareness compared to low self-awareness, $b = -1.00$, $t(1598) = -8.22$, $p < .001$, and a main effect of mutability condition, $b = 0.75$, $t(1598) = 6.14$, $p < .001$, with a significant interaction, $b = 0.36$, $t(1598) = 2.08$, $p = .038$, such that the gap between self-awareness conditions was larger when the behavior was high in mutability than low in mutability.

Study 7

Results on exploratory measures. We combined the benevolence ($\alpha = 0.88$), ability ($\alpha = 0.83$), and integrity ($\alpha = 0.91$) scales into their respective composites.

On our composite measure of benevolence, there were main effects of both self-awareness condition, $F(1, 796) = 85.63$, $p < .001$, $\eta_p^2 = .10$, and impact condition, $F(1, 796) = 51.43$, $p < .001$, $\eta_p^2 = .06$, with a significant interaction, $F(1, 796) = 15.82$, $p < .001$, $\eta_p^2 = .02$.

On our composite measure of ability, there were main effects of both self-awareness condition, $F(1, 796) = 28.40$, $p < .001$, $\eta_p^2 = .03$, and impact condition, $F(1, 796) = 11.25$, $p < .001$, $\eta_p^2 = .01$, with a significant interaction, $F(1, 796) = 11.38$, $p < .001$, $\eta_p^2 = .01$.

On our composite measure of integrity, there were main effects of both self-awareness condition, $F(1, 796) = 15.25$, $p < .001$, $\eta_p^2 = .02$, and impact condition, $F(1, 796) = 26.68$, $p < .001$, $\eta_p^2 = .03$, with no interaction, $F(1, 796) = 2.49$, $p = .115$, $\eta_p^2 < .01$.

Supplemental Study: Will you take this (annoying) survey?

In our Supplemental study, we tested whether showing self-awareness of a negative behavior would not decrease, and might even increase, trust when the negative act is perceived as justifiable. We had research assistants walk up to people at a train station and ask them if they would be willing to provide their email address so that we could email them a survey. We varied the exact language that the research assistant used while asking this, such that the research assistant either expressed self-awareness about the fact that their request might seem burdensome or not. Because we expected people to perceive a request to take a survey as justifiable—even if still burdensome—we predicted that exhibiting self-awareness of the burden of the request would no longer decrease, and might even increase, trust toward the requestor. This study’s preregistration is available at: https://aspredicted.org/8RC_CFW.

Participants. Participants were individuals in the public waiting areas at a busy train station in downtown Chicago. We preregistered to obtain 400 participants after exclusions and ended up with a final sample of 403 participants. We provided no compensation for this study.

Procedure. Two trained research assistants ran the study at a time. Each research assistant carried a tablet and approached people in separate areas of the train station. For each person that the research assistant approached, we randomly varied the script they used such that they either expressed self-awareness or not. In the low self-awareness condition, the research assistant simply said the following: “Excuse me, would you be willing to take a short survey for a school research project? I would collect your email address now and then we’d send you the survey later.” In the high self-awareness condition, we changed the first sentence as follows (keeping the second sentence the same): “Excuse me, I know you probably hate getting approached and asked to take surveys, but would you be willing to take a short survey for a

school research project?” Thus, in the high self-awareness condition, the research assistant acknowledged that the participant might find the request annoying or burdensome.

Our main dependent measure was whether the participant agreed to provide their email address (coded as 1) or not (coded as 0). If the participant said no, the research assistant recorded their response in the data sheet after the interaction ended. If the participant said yes, the research assistant handed them the tablet, on which they responded to a brief survey that had them enter their email address. Once the interaction had ended, the research assistant entered the participant’s ID number into the survey in order to link it to their email address, and recorded their response (1 for yes) in the data sheet.

We gave the research assistants specific instructions for whom to approach and how to answer any of the participants’ questions about the survey, in order to standardize their responses. For the full instructions, please refer to the study script on our OSF page (https://osf.io/f6cr9/?view_only=3f38d8736ab543b6acd5a3a729db5ef5).

After all data collection had finished, we actually sent out a survey to all participants who had provided their email address. The survey contained two exploratory measures unrelated to the current study, as well as one exploratory measure related to this study: “Why did you agree to provide your email address in order to take this survey?” (multiple choice: *I was curious about the research, I enjoy taking surveys, I wanted to help the person who asked me, I knew I’d have some spare time, other [please describe]*). We included this question in order to see if more people (out of those who opted to take the survey) agreed to take the survey because they wanted to help the requestor when the requestor expressed high (versus low) self-awareness.

Results

We computed the proportions of people who agreed to provide their email address out of the total who were approached in each condition. A chi-squared test revealed a non-significant difference between the high (44.3%) and low (38.0%) self-awareness conditions, $\chi^2 (1, N = 403) = 1.42, p = .234$ (though the pattern of results was directionally consistent with the prediction that more people would provide their email in the high compared to the low self-awareness condition). At the very least, our results do not demonstrate a penalty for high self-awareness of a negative behavior in this case, as hypothesized—though it is also possible that this result is due to a lack of sufficient statistical power.

Response rates on our actual survey were too low to allow for meaningful analyses (35 responses in total). Within the high self-awareness condition, 10 out of 18 (55.56%) respondents indicated that they provided their email address in order to help the requestor, and within the low self-awareness condition, 7 of 11 (63.64%) respondents indicated such.

Appendix C: Study Materials and Additional Results in Chapter 3

In the pilot study, social self-awareness was defined to participants as follows:

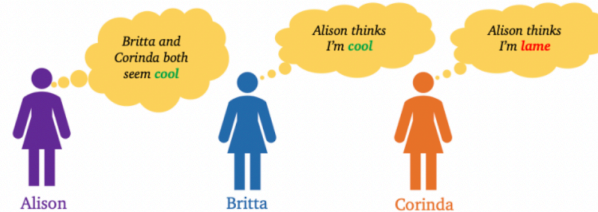
In this study, we are going to ask you to think about **social self-awareness**.

First, here is the definition of social self-awareness that we will be using throughout this study; take a few minutes to read it carefully:

- Social self-awareness can be thought of as **accurately knowing what other people think of you**.
- For instance, someone who has **high** social self-awareness might be **aware** that **other people** view them as friendly/rude/intelligent/arrogant/etc.
- On the other hand, someone who has **low** social self-awareness might be **unaware** that **other people** view them as friendly/rude/intelligent/arrogant/etc.
- Social self-awareness can apply to **any trait** (not just those listed in the previous examples).
- Social self-awareness **does not include other types of self-awareness**, such as awareness of one's own emotions or internal experiences that are independent of what others think of oneself.

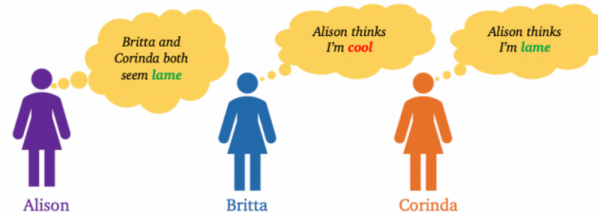
Example 1:

*In this case, Britta has **HIGH** social self-awareness, but Corinda has **LOW** social self-awareness.*



Example 2:

*In this case, Britta has **LOW** social self-awareness, but Corinda has **HIGH** social self-awareness.*



Make sure you understand this definition before proceeding.

Figure A.3.1. Definition of social self-awareness provided in Pilot Study, Chapter 3.

Stimuli and Additional Analyses in Study 1, Chapter 3

Self-coding instructions and examples. Instructions, practice questions, and feedback for participants' self-coding are shown below.

Next, we would like to ask you some questions about the thoughts you just listed. First, we will explain the questions we would like you to answer. **Please take your time and read all of this carefully.**

For each thought you listed, we would like you to tell us **the extent to which the thought is about certain characteristics of the person.** Each characteristic is listed below, along with examples of thoughts that would be about that characteristic.

Trustworthiness

Example thoughts about trustworthiness:

"This person seems like they are dishonest"

"I don't think I could rely on this person"

Competence

Example thoughts about competence:

"This person doesn't seem all that intelligent"

"This person sounds like a talented artist"

Likability

Example thoughts about likability:

"I would probably find this person annoying"

"This person seems fun to hang out with"

Self-Awareness

Example thoughts about self-awareness:

"This person must be oblivious to how they come across to others"

"This person seems to know how to make a good impression on others"

Kindness

Example thoughts about kindness:

"This person is probably very generous toward others"

"This person seems selfish"

As you can see, thoughts that fall under each category include both those about someone who **possesses that characteristic** (e.g., is very trustworthy) and about someone who does **NOT possess that characteristic** (e.g., is untrustworthy).

Figure A.3.2. Instructions for self-coding in Study 1, Chapter 3.

Some thoughts may also **fall under more than one category**. You will get a chance to select **multiple categories** when you categorize your thoughts.

Other thoughts may not fall under any category. You will also be given an **"Other"** option to select if you feel the thought is about a category not listed.

Please review all of this information very carefully. On the next few pages, we will ask you to complete a few trial questions to help make sure this task is clear. After that, we will ask you to categorize your own thoughts.

Figure A.3.2 (cont'd). Instructions for self-coding in Study 1, Chapter 3.

We then provided participants with the following example thoughts and feedback based on their responses:

- Example thought 1: “This person seems really smart”
 - Participants were given feedback indicating that the best answer was “competence,” and not the other characteristics (selecting “other” was acceptable as well).
- Example thought 2: “This person seems to think they’re being funny but they actually sound mean.”
 - Participants were given feedback indicating that the best answer was “likability,” “self-awareness,” and “kindness,” and not the other characteristics (selecting “other” was acceptable as well).
- Example thought 3: “This person seems like they are lying.”
 - Participants were given feedback indicating that the best answer was “trustworthiness,” and not the other characteristics (selecting “other” was acceptable as well).

Additional analyses. We conducted a one-way ANOVA to compare how the average number of thoughts related to self-awareness compared to the average number of thoughts related to the other traits. We observed significant variation in the number of thoughts categorized as each trait, $F(5, 494) = 249.18, p < .001, \eta_p^2 = .33$. As shown in Figure A.3.3, self-awareness thoughts were less common than likability thoughts, but more common than all the other traits.

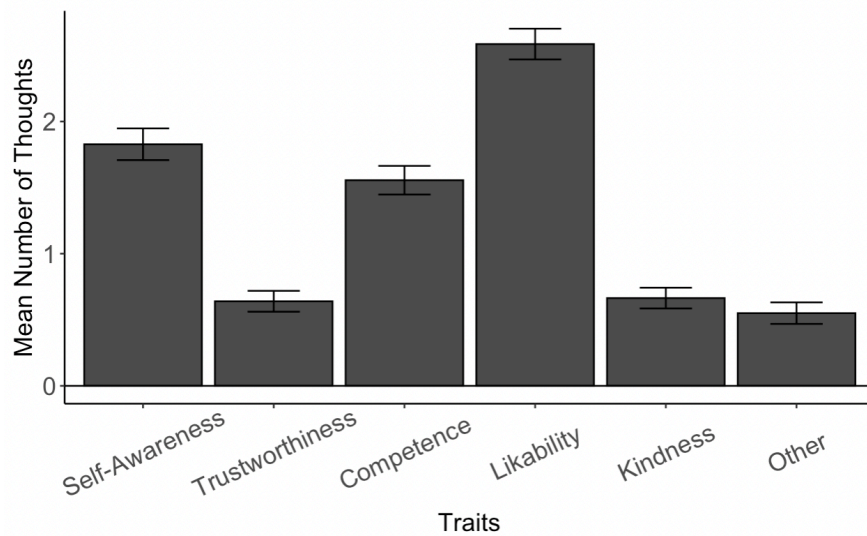


Figure A.3.3. Mean number of thoughts reported for self-awareness (self-coded) compared to each of the other traits in Study 1, Chapter 3. Error bars represent 95% confidence intervals.

When analyzing the research assistant coding, we observed significant variation in the number of thoughts categorized as each trait, $F(5,494) = 2452.66$, $p < .001$, $\eta_p^2 = .83$, such that thoughts about self-awareness were significantly less common than most of the other traits (see Figure A.3.4).

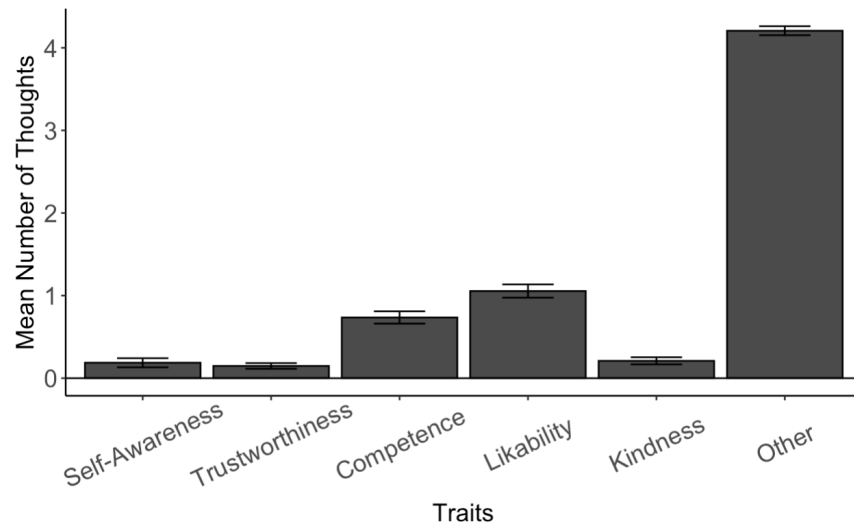


Figure A.3.4. Mean number of thoughts reported for self-awareness (research assistant coded) compared to each of the other traits in Study 1. Error bars represent 95% confidence intervals.

Additional Analyses in Study 2, Chapter 3

We conducted a one-way ANOVA to compare how the average number of thoughts related to self-awareness compared to the average number of thoughts related to the other traits, collapsing across all conditions. We observed significant variation in the number of thoughts categorized as each trait, $F(5, 394) = 128.23, p < .001, \eta_p^2 = .24$. As is shown in Figure A.3.5, self-awareness-related thoughts were less common than competence- and likability-related thoughts (perhaps in part because our prompts were about competence and likability), about as common as kindness-related thoughts, and more common than trustworthiness-related thoughts or thoughts categorized as “other.”

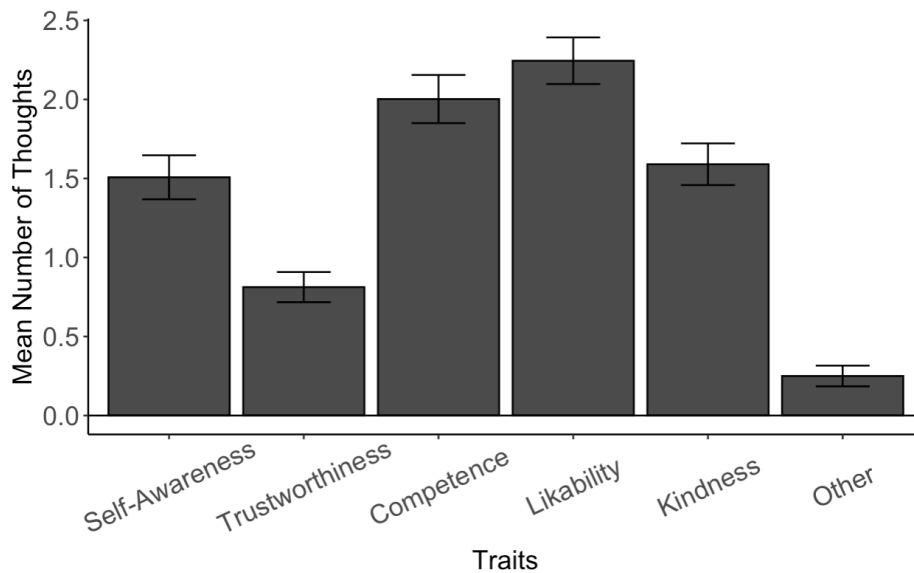


Figure A.3.5. Mean number of thoughts reported for self-awareness (self-coded) compared to each of the other traits in Study 2, Chapter 3. Error bars represent 95% confidence intervals.

When analyzing the research assistant coding, we observed significant variation in the number of thoughts categorized as each trait, $F(5,394) = 144.66, p < .001, \eta_p^2 = .27$, such that thoughts about self-awareness were significantly less common than each of the other traits (see Figure A.3.6).

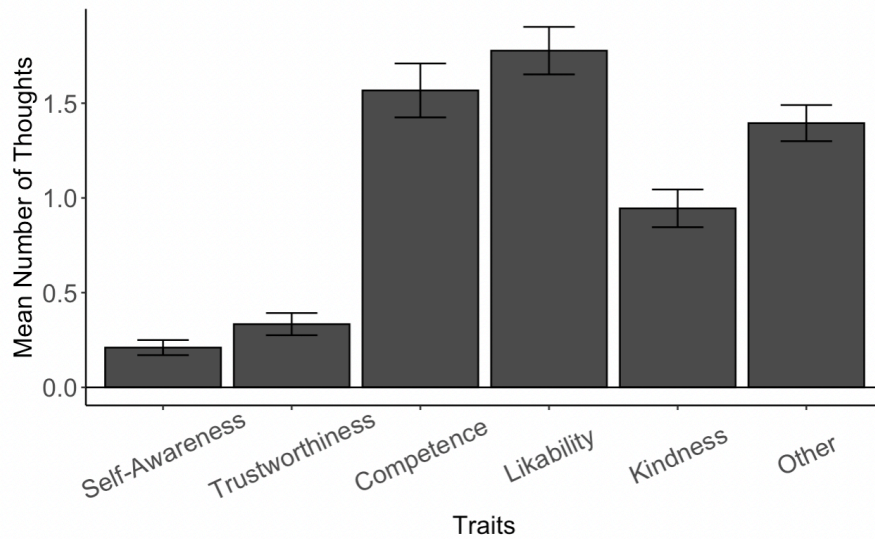


Figure A.3.6. Mean number of thoughts reported for self-awareness (research assistant coded) compared to each of the other traits in Study 2, Chapter 3. Error bars represent 95% confidence intervals.

Video Links and Results on Additional Explanations in Study 3, Chapter 3

Female version of video: <https://youtu.be/YFiThZLRkWA>

Male version of video: <https://youtu.be/k3PvikEgvmg>

We asked the research assistants in Study 3 to code for several other common explanations of the target's behavior, besides self-awareness. Each of these explanations, and the mean references by condition, are presented below.

Jamie has social anxiety/difficulty or is socially awkward:

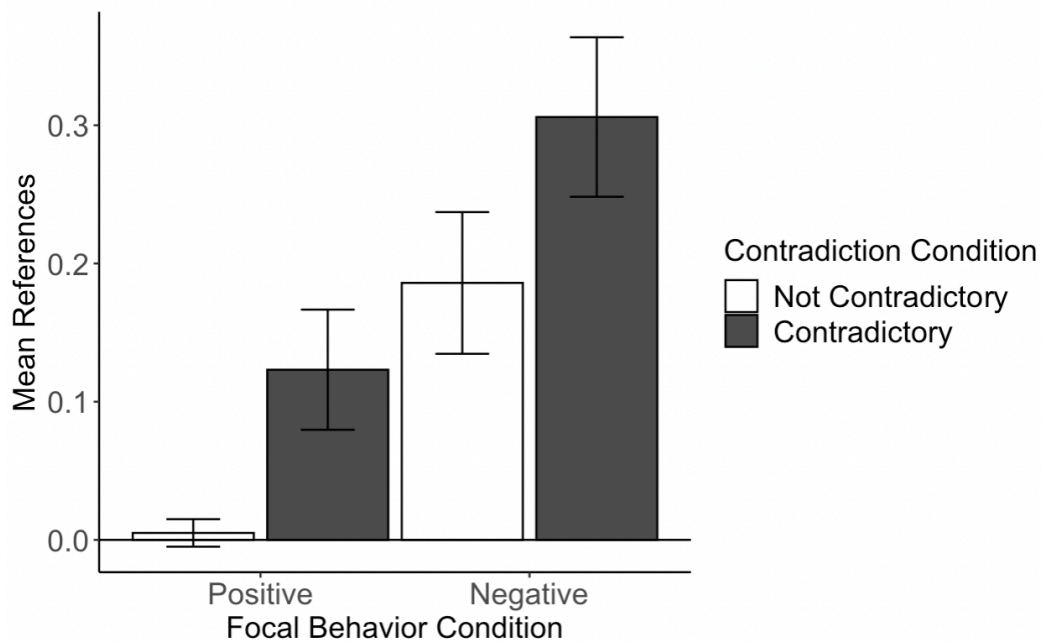


Figure A.3.7. Mean references to social anxiety/difficulty/awkwardness in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

Jamie's behavior matches their intentions/character:

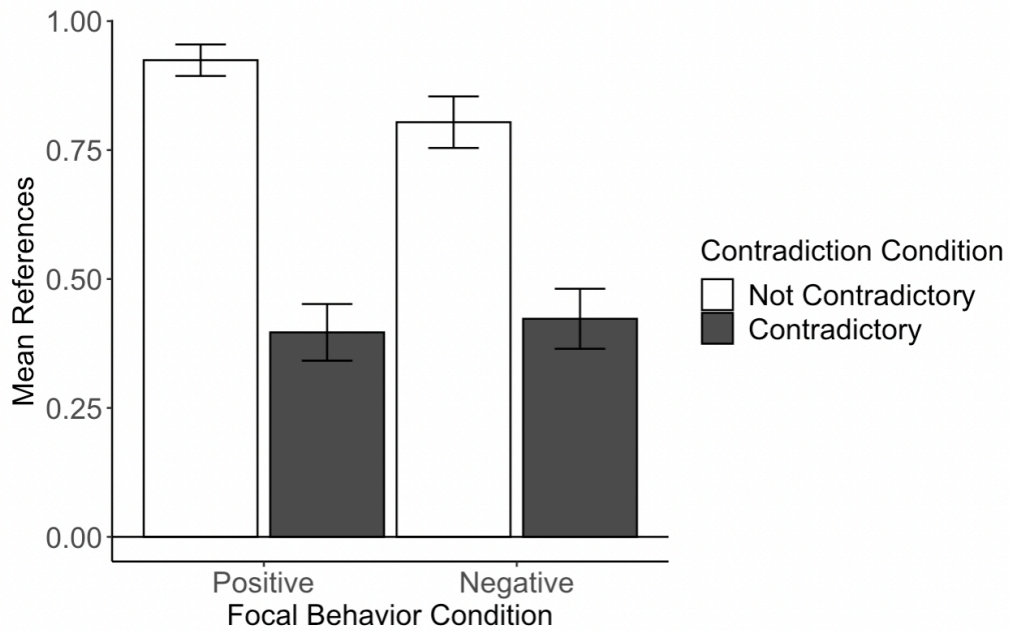


Figure A.3.8. Mean references to behavior matching intentions/character in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

Jamie has some other good (external) reason for their behavior:

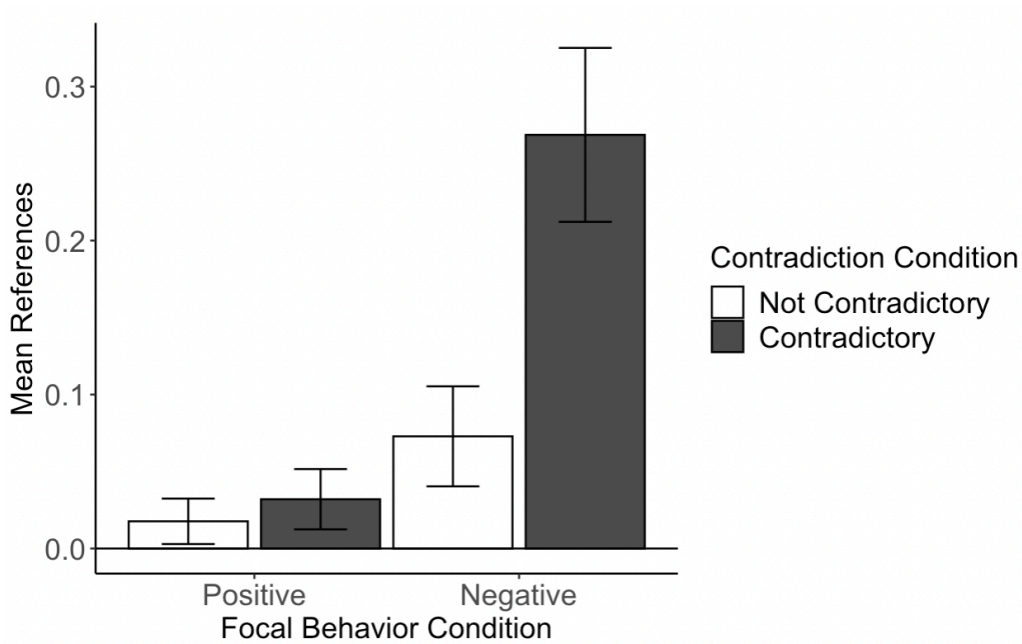


Figure A.3.9. Mean references to good (external) reasons for behavior in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

Jamie has an ulterior motive:

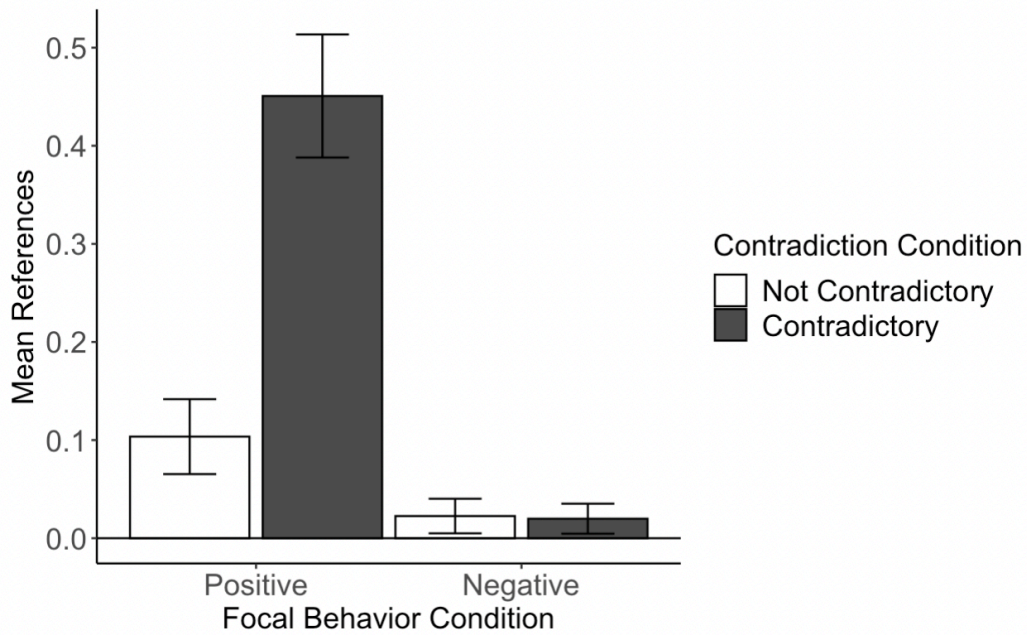


Figure A.3.10. Mean references to ulterior motives in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

The participant's prior impression of Jamie must have been incorrect:

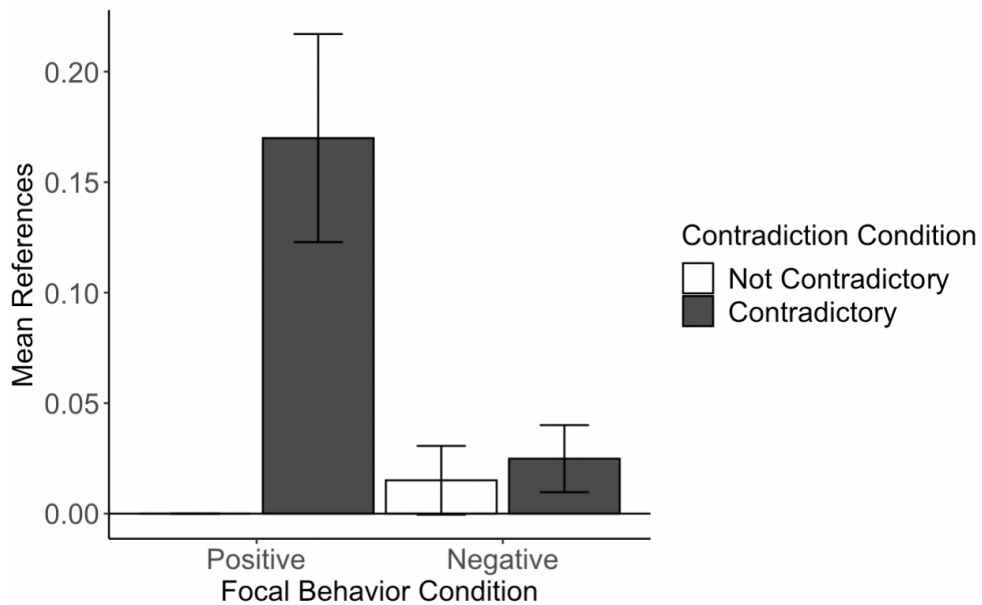


Figure A.3.11. Mean references to incorrect prior impressions in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

Other explanations (anything not in the prior categories):

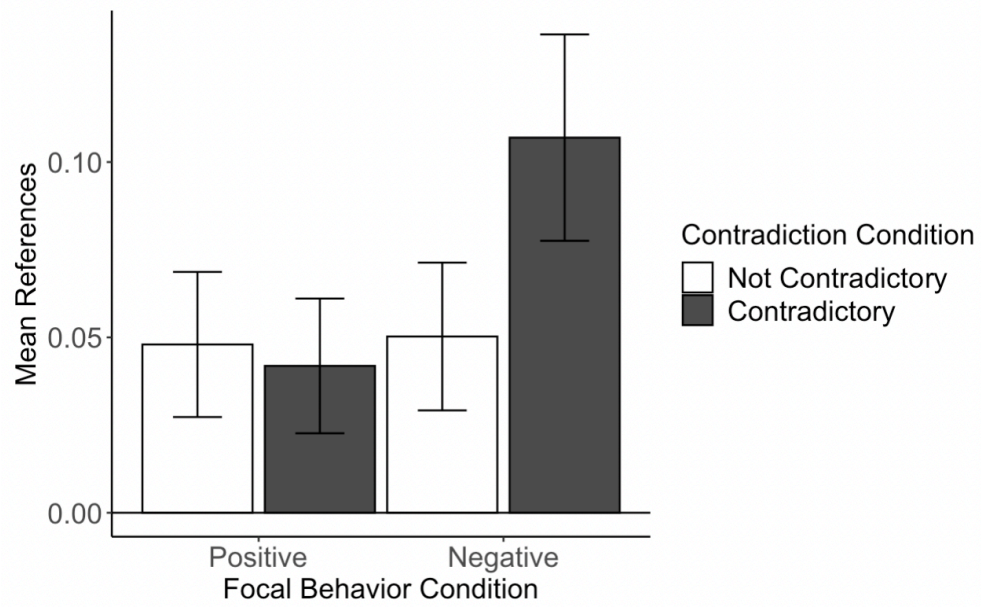


Figure A.3.12. Mean references to other explanations in Study 3, Chapter 3. Error bars represent 95% confidence intervals.

References

- Albert, K., & Escobedo-Land, A. (2015). OkCupid Data for Introductory Statistics and Data Science Courses. *Journal of Statistics Education*, 23. <https://doi.org/10.1080/10691898.2015.11889737>
- Aloise-Young, P. A. (1993). The Development of Self-Presentation: Self-Promotion in 6- to 10-Year-Old Children. *Social Cognition*, 11(2), 201–222. <https://doi.org/10.1521/soco.1993.11.2.201>
- Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, 103(4), 718–735. <https://doi.org/10.1037/a0029395>
- Aronson, E., Willerman, B., & Floyd, J. (1966). The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science*, 4(6), 227–228. <https://doi.org/10.3758/BF03342263>
- Atwater, L. E., & Yammarino, F. J. (1992). Does Self-Other Agreement on Leadership Perceptions Moderate the Validity of Leadership and Performance Predictions? *Personnel Psychology*, 45(1), 141. <https://www.proquest.com/docview/220136438/abstract/9E95882301E84BF1PQ/1>
- Barasch, R., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or Selfless? On the Signal Value of Emotion in Altruistic Behavior. *Journal of Personality and Social Psychology*, 107(3), 393–413.
- Barneron, M., Choshen-Hillel, S., & Yaniv, I. (2021). Reaping a benefit at the expense of multiple others: How are the losses of others counted? *Organizational Behavior and Human Decision Processes*, 164, 136–146. <https://doi.org/10.1016/j.obhdp.2021.02.004>
- Baron, R. A. (1986). Self-Presentation in Job Interviews: When There Can Be “Too Much of a Good Thing.” *Journal of Applied Social Psychology*, 16(1), 16–28. <https://doi.org/10.1111/j.1559-1816.1986.tb02275.x>
- Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective Taking: Imagining How Another Feels Versus Imaging How You Would Feel. *Personality and Social Psychology Bulletin*, 23(7), 751–758. <https://doi.org/10.1177/0146167297237008>
- Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, 91(1), 3–26. <https://doi.org/10.1037/0033-2909.91.1.3>
- Baumeister, R. F., & Ilko, S. A. (1995). Shallow Gratitude: Public and Private Acknowledgement of External Help in Accounts of Success. *Basic and Applied Social Psychology*, 16(1–2), 191–209. <https://doi.org/10.1080/01973533.1995.9646109>

- Baumeister, R. F., & Jones, E. E. (1978). When self-presentation is constrained by the target's knowledge: Consistency and compensation. *Journal of Personality and Social Psychology*, 36(6), 608–618. <https://doi.org/10.1037/0022-3514.36.6.608>
- Bell, R. A., Zahn, C. J., & Hopper, R. (1984). Disclaiming: A test of two competing views. *Communication Quarterly*, 32(1), 28–36. <https://doi.org/10.1080/01463378409369528>
- Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart's Dilemma: On the Social Rewards and Penalties of Advertising Prosocial Behavior. *Journal of Marketing Research*, 52(1), 90–104. <https://doi.org/10.1509/jmr.14.0002>
- Bitterly, T. B., & Schweitzer, M. E. (2019). The impression management benefits of humorous self-disclosures: How humor influences perceptions of veracity. *Organizational Behavior and Human Decision Processes*, 151, 73–89. <https://doi.org/10.1016/j.obhdp.2019.01.005>
- Boothby, E. J., Cooney, G., Sandstrom, G. M., & Clark, M. S. (2018). The Liking Gap in Conversations: Do People Like Us More Than We Think? *Psychological Science*, 29(11), 1742–1756. <https://doi.org/10.1177/0956797618783714>
- Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology*, 80(1), 168–177. <https://doi.org/10.1037/0021-9010.80.1.168>
- Bradley, P. H. (1981). The folk-linguistics of women's speech: An empirical examination. *Communication Monographs*, 48(1), 73–90. <https://doi.org/10.1080/03637758109376048>
- Brambilla, M., Carraro, L., Castelli, L., & Sacchi, S. (2019). Changing impressions: Moral character dominates impression updating. *Journal of Experimental Social Psychology*, 82, 64–73. <https://doi.org/10.1016/j.jesp.2019.01.003>
- Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology*, 41(2), 135–143. <https://doi.org/10.1002/ejsp.744>
- Brion, S., Lount, R. B., & Doyle, S. P. (2015). Knowing If You Are Trusted: Does Meta-Accuracy Promote Trust Development? *Social Psychological and Personality Science*, 6(7), 823–830. <https://doi.org/10.1177/1948550615590200>
- Brooks, A. W., Huang, K., Abi-Esber, N., Buell, R. W., Huang, L., & Hall, B. (2019). Mitigating malicious envy: Why successful individuals should reveal their failures. *Journal of Experimental Psychology: General*, 148(4), 667–687. <https://doi.org/10.1037/xge0000538>
- Bruk, A., Scholl, S. G., & Bless, H. (2018). Beautiful mess effect: Self–other differences in evaluation of showing vulnerability. *Journal of Personality and Social Psychology*, 115(2), 192–205. <https://doi.org/10.1037/pspa0000120>

- Buhr, K., & Dugas, M. J. (2002). The intolerance of uncertainty scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, *40*(8), 931–945. [https://doi.org/10.1016/S0005-7967\(01\)00092-4](https://doi.org/10.1016/S0005-7967(01)00092-4)
- Burgoon, J. K. (2015). Expectancy Violations Theory. In *The International Encyclopedia of Interpersonal Communication* (pp. 1–9). John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781118540190.wbeic102>
- Carlson, E. N., & Furr, R. M. (2009). Evidence of Differential Meta-Accuracy: People Understand the Different Impressions They Make. *Psychological Science*, *20*(8), 1033–1039. <https://doi.org/10.1111/j.1467-9280.2009.02409.x>
- Carlson, E. N., Furr, R. M., & Vazire, S. (2010). Do We Know the First Impressions We Make? Evidence for Idiographic Meta-Accuracy and Calibration of First Impressions. *Social Psychological and Personality Science*, *1*(1), 94–98. <https://doi.org/10.1177/1948550609356028>
- Carlson, E. N., & Kenny, D. A. (2012). Meta-accuracy: Do we know how others see us? In *Handbook of self-knowledge* (pp. 242–257). The Guilford Press.
- Chaudhry, S. J., & Loewenstein, G. (2019). Thanking, apologizing, bragging, and blaming: Responsibility exchange theory and the currency of communication. *Psychological Review*, *126*(3), 313–344. <https://doi.org/10.1037/rev0000139>
- Chaudhry, S. J., & Wald, K. (2022). Overcoming listener skepticism: Costly signaling in communication increases perceived honesty. *Current Opinion in Psychology*, 101442. <https://doi.org/10.1016/j.copsyc.2022.101442>
- Chon, D., & Sitkin, S. B. (2021). Disentangling the Process and Content of Self-Awareness: A Review, Critical Assessment, and Synthesis. *Academy of Management Annals*, *15*(2), 607–651. <https://doi.org/10.5465/annals.2018.0079>
- Church, A. H. (1997). Managerial self-awareness in high-performing individuals in organizations. *Journal of Applied Psychology*, *82*, 281–292. <https://doi.org/10.1037/0021-9010.82.2.281>
- Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling Theory: A Review and Assessment. *Journal of Management*, *37*(1), 39–67. <https://doi.org/10.1177/0149206310388419>
- Cramer, R. J., Brodsky, S. L., & Decoster, J. (2009). Expert Witness Confidence and Juror Personality: Their Impact on Credibility and Persuasion in the Courtroom. *Journal of the American Academy of Psychiatry and the Law*, *37*, 63–74.

- Crant, J. M. (1996). Doing More Harm Than Good: When Is Impression Management Likely to Evoke a Negative Response? *Journal of Applied Social Psychology*, 26(16), 1454–1471. <https://doi.org/10.1111/j.1559-1816.1996.tb00080.x>
- Crowley, A. E., & Hoyer, W. D. (1994). An Integrative Framework for Understanding Two-sided Persuasion. *Journal of Consumer Research*, 20(4), 561–574. <https://doi.org/10.1086/209370>
- Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2), 353–380. <https://doi.org/10.1016/j.cognition.2008.03.006>
- Davis, C. G., & Nolen-Hoeksema, S. (2001). Loss and meaning: How do people make sense of loss? *The American Behavioral Scientist*, 44(5), 726–741. <https://doi.org/10.1177/00027640121956467>
- Davis, M. H., & Franzoi, S. L. (1999). Self-awareness and self-consciousness. In *Personality: Contemporary theory and research*, 2nd ed (pp. 307–338). Nelson-Hall Publishers.
- Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, 2(4), 265–279. <https://doi.org/10.1177/002200275800200401>
- Diener, E., & Wallbom, M. (1976). Effects of self-awareness on antinormative behavior. *Journal of Research in Personality*, 10(1), 107–111. [https://doi.org/10.1016/0092-6566\(76\)90088-X](https://doi.org/10.1016/0092-6566(76)90088-X)
- Dirks, K. T., & Ferrin, D. L. (2001). The Role of Trust in Organizational Settings. *Organization Science*, 12(4), 450–467. <https://doi.org/10.1287/orsc.12.4.450.10640>
- Donnelly, K., Moon, A., & Critcher, C. R. (2022). Do people know how others view them? Two approaches for identifying the accuracy of metaperceptions. *Current Opinion in Psychology*, 43, 119–124. <https://doi.org/10.1016/j.copsyc.2021.06.018>
- Duval, S., & Wicklund, R. A. (1972). *A theory of objective self awareness*. Academic Press.
- Eisenkraft, N., Elfenbein, H. A., & Kopelman, S. (2017). We Know Who Likes Us, but Not Who Competes Against Us: Dyadic Meta-Accuracy Among Work Colleagues. *Psychological Science*, 28(2), 233–241. <https://doi.org/10.1177/0956797616679440>
- El-Alayli, A., Myers, C. J., Petersen, T. L., & Lystad, A. L. (2008). “I Don’t Mean to Sound Arrogant, but. . .” The Effects of Using Disclaimers on Person Perception. *Personality and Social Psychology Bulletin*, 34(1), 130–143. <https://doi.org/10.1177/0146167207309200>

- Ellis, A., West, B., Ryan, A., & DeShon, R. (2002). The Use of Impression Management Tactics in Structured Interviews: A Function of Question Type? *The Journal of Applied Psychology, 87*, 1200–1208. <https://doi.org/10.1037/0021-9010.87.6.1200>
- Elsaadawy, N. (2018). *The Good Judge of Meta-perception* [M.A., University of Toronto (Canada)]. <https://www.proquest.com/docview/2139705324/abstract/1C27C7AF355343F0PQ/1>
- Elsaadawy, N., Carlson, E. N., & Human, L. J. (2021). Who influences meta-accuracy? It takes two to know the impressions we make. *Journal of Personality and Social Psychology, 121*(1), 201–214. <https://doi.org/10.1037/pspp0000376>
- Etgar, M., & Goodwin, S. A. (1982). One-Sided Versus Two-Sided Comparative Message Appeals for New Brand Introductions. *Journal of Consumer Research, 8*(4), 460–465. <https://www.jstor.org/stable/2489035>
- Exline, J. J., & Geyer, A. L. (2004). Perceptions of Humility: A Preliminary Study. *Self and Identity, 3*(2), 95–114. <https://doi.org/10.1080/13576500342000077>
- FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour, 3*(5), 426–435. <https://doi.org/10.1038/s41562-019-0590-x>
- Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology, 43*(4), 522–527. <https://doi.org/10.1037/h0076760>
- Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. <https://doi.org/10.1016/j.tics.2006.11.005>
- Foddy, M., Platow, M. J., & Yamagishi, T. (2009). Group-Based Trust in Strangers: The Role of Stereotypes and Expectations. *Psychological Science, 20*(4), 419–422. <https://doi.org/10.1111/j.1467-9280.2009.02312.x>
- Forsyth, D. R., Berger, R. E., & Mitchell, T. (1981). The Effects of Self-Serving vs. Other-Serving Claims of Responsibility on Attraction and Attribution in Groups. *Social Psychology Quarterly, 44*(1), 59–64. <https://doi.org/10.2307/3033865>
- Franzoi, S. L., & Brewer, L. C. (1984). The experience of self-awareness and its relation to level of self-consciousness: An experiential sampling study. *Journal of Research in Personality, 18*(4), 522–540. [https://doi.org/10.1016/0092-6566\(84\)90010-2](https://doi.org/10.1016/0092-6566(84)90010-2)
- Fraser, B. (1990). Perspectives on politeness. *Journal of Pragmatics, 14*(2), 219–236. [https://doi.org/10.1016/0378-2166\(90\)90081-N](https://doi.org/10.1016/0378-2166(90)90081-N)

- Gabarro, J. J. (1978). The development of trust, influence and expectations. *Interpersonal Behavior : Communication and Understanding in Relationships*, 290–303. <https://ci.nii.ac.jp/naid/20001538304/>
- Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-Taking and Self-Other Overlap: Fostering Social Bonds and Facilitating Social Coordination. *Group Processes & Intergroup Relations*, 8(2), 109–124. <https://doi.org/10.1177/1368430205051060>
- Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why It Pays to Get Inside the Head of Your Opponent: The Differential Effects of Perspective Taking and Empathy in Negotiations. *Psychological Science*, 19(4), 378–384. <https://doi.org/10.1111/j.1467-9280.2008.02096.x>
- Gangestad, S., & Thornhill, R. (2011). The Evolution of Social Inference Processes. In J. P. Forgas, M. G. Haselton, & W. von Hippel (Eds.), *Evolution and the Social Mind: Evolutionary Psychology and Social Cognition*. Psychology Press.
- Giacalone, R. A., & Rosenfeld, P. (1986). Self-Presentation and Self-Promotion in an Organizational Setting. *The Journal of Social Psychology*, 126(3), 321–326. <https://doi.org/10.1080/00224545.1986.9713592>
- Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, 117(1), 21-38. <https://doi.org/10.1037/0033-2909.117.1.21>
- Gilovich, T., Medvec, V. H., & 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(2), 211-222. <https://doi.org/10.1037/0022-3514.78.2.211>
- Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, 75(2), 332-346. <https://doi.org/10.1037/0022-3514.75.2.332>
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148–168. <https://doi.org/10.1037/a0034726>
- Gordon, R. A. (1996). Impact of ingratiation on judgments and evaluations: A meta-analytic investigation. *Journal of Personality and Social Psychology*, 71(1), 54–70. <https://doi.org/10.1037/0022-3514.71.1.54>
- Grupe, D. W., & Nitschke, J. B. (2011). Uncertainty Is Associated with Biased Expectancies and Heightened Responses to Aversion. *Emotion (Washington, D.C.)*, 11(2), 413–424. <https://doi.org/10.1037/a0022583>

- Hackel, L. M., Mende-Siedlecki, P., & Amodio, D. M. (2020). Reinforcement learning in social interaction: The distinguishing role of trait inference. *Journal of Experimental Social Psychology*, 88, 103948. <https://doi.org/10.1016/j.jesp.2019.103948>
- Hamilton, R., Vohs, K. D., & McGill, A. L. (2014). We'll Be Honest, This Won't Be the Best Article You'll Ever Read: The Use of Dispreferred Markers in Word-of-Mouth Communication. *Journal of Consumer Research*, 41(1), 197–212. <https://doi.org/10.1086/675926>
- Hass, R. G. (1984). Perspective taking and self-awareness: Drawing an E on your forehead. *Journal of Personality and Social Psychology*, 46(4), 788–798. <https://doi.org/10.1037/0022-3514.46.4.788>
- Heatherton, T. F., & Baumeister, R. F. (1991). Binge eating as escape from self-awareness. *Psychological Bulletin*, 110, 86–108. <https://doi.org/10.1037/0033-2909.110.1.86>
- Heck, P. R., & Krueger, J. I. (2016). Social Perception of Self-Enhancement Bias and Error. *Social Psychology*, 47(6), 327–339. <https://doi.org/10.1027/1864-9335/a000287>
- Hewitt, J. P., & Stokes, R. (1975). Disclaimers. *American Sociological Review*, 40(1), 1–11. <https://doi.org/10.2307/2094442>
- Hilton, J. L., Fein, S., & Miller, D. T. (1993). Suspicion and Dispositional Inference. *Personality and Social Psychology Bulletin*, 19(5), 501–512. <https://doi.org/10.1177/0146167293195003>
- Ho, T.-H., & Weigelt, K. (2005). Trust Building Among Strangers. *Management Science*, 51(4), 519–530. <https://doi.org/10.1287/mnsc.1040.0350>
- Hoffman-Graff, M. A. (1977). Interviewer use of positive and negative self-disclosure and interviewer-subject sex pairing. *Journal of Counseling Psychology*, 24(3), 184–190. <https://doi.org/10.1037/0022-0167.24.3.184>
- Jiang, L., John, L. K., Boghrati, R., & Kouchaki, M. (2022). Fostering perceptions of authenticity via sensitive self-disclosure. *Journal of Experimental Psychology: Applied*, 28(4), 898-915.
- John, L. K., Barasz, K., & Norton, M. I. (2016). Hiding personal information reveals the worst. *Proceedings of the National Academy of Sciences*, 113(4), 954–959. <https://doi.org/10.1073/pnas.1516868113>
- Jones, E. E. (1964). *Ingratiation*. Appleton-Century-Crofts.
- Jones, E. E., & Davis, K. E. (1965). From Acts To Dispositions The Attribution Process In Person Perception. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219–266). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60107-0](https://doi.org/10.1016/S0065-2601(08)60107-0)

- Jones, E. E., Gergen, K. J., Gumpert, P., & Thibaut, J. W. (1965). Some conditions affecting the use of ingratiation to influence performance evaluation. *Journal of Personality and Social Psychology, 1*(6), 613–625. <https://doi.org/10.1037/h0022076>
- Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self, 1*(1), 231–262.
- Jones, E., & Pittman, T. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self, 1*(1), 231-262.
- Kay, A., Gaucher, D., Napier, J., Callan, M., & Laurin, K. (2008). God and the Government: Testing a Compensatory Control Mechanism for the Support of External Systems. *Journal of Personality and Social Psychology, 95*, 18–35. <https://doi.org/10.1037/0022-3514.95.1.18>
- Kelley, H. H. (1987). Attribution in social interaction. In *Attribution: Perceiving the causes of behavior* (pp. 1–26). Lawrence Erlbaum Associates, Inc.
- Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin, 114*(1), 145–161. <https://doi.org/10.1037/0033-2909.114.1.145>
- Kim, J. S., Weisberg, Y. J., Simpson, J. A., Oriña, M. M., Farrell, A. K., & Johnson, W. F. (2015). Ruining it for Both of Us: The Disruptive Role of Low-Trust Partners on Conflict Resolution in Romantic Relationships. *Social Cognition, 33*(5), 520–542. <https://doi.org/10.1521/soco.2015.33.5.520>
- Knowles, E., & Linn, J. (2004). Alpha and Omega Strategies for Change. In E. S. Knowles & J. A. Linn (Eds.), *Resistance and Persuasion* (pp. 117–148). Lawrence Erlbaum Associates, Inc.
- Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675–709. <https://doi.org/10.1037/pspa0000046.supp>
- Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569–598.
- Kramer, R. M., & Lewicki, R. J. (2010). Repairing and Enhancing Trust: Approaches to Reducing Organizational Trust Deficits. *Academy of Management Annals, 4*(1), 245–277. <https://doi.org/10.5465/19416520.2010.487403>

- Ku, G., Wang, C. S., & Galinsky, A. D. (2015). The promise and perversity of perspective-taking in organizations. *Research in Organizational Behavior*, 35, 79–102. <https://doi.org/10.1016/j.riob.2015.07.003>
- Kumar, A., & Epley, N. (2018). Undervaluing Gratitude: Expressers Misunderstand the Consequences of Showing Appreciation. *Psychological Science*, 29(9), 1423–1435. <https://doi.org/10.1177/0956797618772506>
- Laing, R. D., Phillipson, H., & Lee, A. R. (1966). *Interpersonal perception: A theory and a method of research* (pp. vii, 179). Springer.
- Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology*, 29(3), 299–304. <https://doi.org/10.1037/h0036018>
- Landy, J., & Uhlmann, E. (2018). Morality is personal. In K. Gray & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 121–132). Guilford Publications.
- Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin*, 107(1), 34–47. <https://doi.org/10.1037/0033-2909.107.1.34>
- Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126, 88–106. <https://doi.org/10.1016/j.obhdp.2014.10.007>
- List of Weaknesses: Examples of What To Say in an Interview | Indeed.com.* (2022, August 31). Indeed Career Guide. <https://www.indeed.com/career-advice/interviewing/list-of-example-weaknesses-for-interviewing>
- Long, E. C., & Andrews, D. W. (1990). Perspective taking as a predictor of marital adjustment. *Journal of Personality and Social Psychology*, 59(1), 126–131. <https://doi.org/10.1037/0022-3514.59.1.126>
- Macy, M. W., & Skvoretz, J. (1998). The Evolution of Trust and Cooperation between Strangers: A Computational Model. *American Sociological Review*, 63(5), 638–660. <https://doi.org/10.2307/2657332>
- Malle, B. F., & Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, 33(2), 101–121. <https://doi.org/10.1006/jesp.1996.1314>
- Maselli, M. D., & Altrocchi, J. (1969). Attribution of intent. *Psychological Bulletin*, 71(6), 445–454. <https://doi.org/10.1037/h0027348>

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
- McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. *Academy of Management Review*, 23(3), 473–490. <https://doi.org/10.5465/amr.1998.926622>
- Mende-Siedlecki, P., Baron, S. G., & Todorov, A. (2013). Diagnostic Value Underlies Asymmetric Updating of Impressions in the Morality and Ability Domains. *Journal of Neuroscience*, 33(50), 19406–19415. <https://doi.org/10.1523/JNEUROSCI.2334-13.2013>
- Moon, A., Gan, M., & Critcher, C. R. (2020). The overblown implications effect. *Journal of Personality and Social Psychology*, 118(4), 720–742. <https://doi.org/10.1037/pspi0000204>
- Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*, 117(26), 14864–14872. <https://doi.org/10.1073/pnas.2001263117>
- Ohtsubo, Y., Takezawa, M., & Fukuno, M. (2009). Mutual liking and meta-perception accuracy. *European Journal of Social Psychology*, 39(5), 707–718. <https://doi.org/10.1002/ejsp.568>
- Ohtsubo, Y., & Watanabe, E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(2), 114–123. <https://doi.org/10.1016/j.evolhumbehav.2008.09.004>
- Pfeffer, J., Fong, C. T., Cialdini, R. B., & Portnoy, R. R. (2006). Overcoming the Self-Promotion Dilemma: Interpersonal Attraction and Extra Help as a Consequence of Who Sings One's Praises. *Personality and Social Psychology Bulletin*, 32(10), 1362–1374. <https://doi.org/10.1177/0146167206290337>
- Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1), 39–57. <https://doi.org/10.1002/bdm.460>
- Proost, K., Schreurs, B., De Witte, K., & Derous, E. (2010). Ingratiation and Self-Promotion in the Selection Interview: The Effects of Using Single Tactics or a Combination of Tactics on Interviewer Judgments. *Journal of Applied Social Psychology*, 40(9), 2155–2169. <https://doi.org/10.1111/j.1559-1816.2010.00654.x>
- Rempel, J., Holmes, J., & Zanna, M. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49, 95–112. <https://doi.org/10.1037/0022-3514.49.1.95>

- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Academy of Management Review*, 23(3), 393–404. <https://doi.org/10.5465/amr.1998.926617>
- Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, 5(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2
- Rucker, D. D., Petty, R. E., & Briñol, P. (2008). What’s in a frame anyway?: A meta-cognitive analysis of the impact of one versus two sided message framing on attitude certainty. *Journal of Consumer Psychology*, 18(2), 137–149. <https://doi.org/10.1016/j.jcps.2008.01.008>
- Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, 74(3), 629–645. <https://doi.org/10.1037/0022-3514.74.3.629>
- Salovey, P., Brackett, M. A., & Mayer, J. D. (2004). *Emotional Intelligence: Key Readings on the Mayer and Salovey Model*. National Professional Resources Inc./Dude Publishing.
- Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, 9(3), 185–211. <https://doi.org/10.2190/DUGG-P24E-52WK-6CDG>
- Savitsky, K., Epley, N., & Gilovich, T. (2001). Do others judge us as harshly as we think? Overestimating the impact of our failures, shortcomings, and mishaps. *Journal of Personality and Social Psychology*, 81(1), 44–56. <https://doi.org/10.1037/0022-3514.81.1.44>
- Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, 18(5), 803–817. <https://doi.org/10.1162/jocn.2006.18.5.803>
- Schlenker, B. R. (1980). *Impression management*. Monterey, CA: Brooks/Cole.
- Schlenker, B. R., & Leary, M. R. (1982). Audiences’ reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of Experimental Social Psychology*, 18(1), 89–104. [https://doi.org/10.1016/0022-1031\(82\)90083-X](https://doi.org/10.1016/0022-1031(82)90083-X)
- Scopelliti, I., Loewenstein, G., & Vosgerau, J. (2015). You Call It “Self-Exuberance”; I Call It “Bragging”: Miscalibrated Predictions of Emotional Responses to Self-Promotion. *Psychological Science*, 26(6), 903–914. <https://doi.org/10.1177/0956797615573516>
- Sezer, O. (2017). *Misguided self-presentation: The ironic consequences of humblebragging, backhanded compliments and namedropping*. Harvard University.

- Sezer, O., Gino, F., & Norton, M. I. (2018). Humblebragging: A distinct—and ineffective—self-presentation strategy. *Journal of Personality and Social Psychology, 114*(1), 52–74. <https://doi.org/10.1037/pspi0000108>
- Sezer, O., Prinsloo, E., Brooks, A., & Norton, M. I. (2019). *Backhanded Compliments: How Negative Comparisons Undermine Flattery* [SSRN Scholarly Paper]. <https://doi.org/10.2139/ssrn.3439774>
- Shapiro, D. L., & Bies, R. J. (1994). Threats, Bluffs, and Disclaimers in Negotiations. *Organizational Behavior and Human Decision Processes, 60*(1), 14–35. <https://doi.org/10.1006/obhd.1994.1073>
- Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*, 689–699. <https://doi.org/10.1037/0022-3514.52.4.689>
- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*, 131–142. <https://doi.org/10.1037/0033-2909.105.1.131>
- Smyth, J., True, N., & Souto, J. (2001). Effects of Writing About Traumatic Experiences: The Necessity for Narrative Structuring. *Journal of Social and Clinical Psychology, 20*(2), 161–172. <https://doi.org/10.1521/jscp.20.2.161.22266>
- Snyder, M. (1975). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*(4), 526–537. <https://doi.org/10.1037/h0037039>
- Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass, 11*(6), e12321. <https://doi.org/10.1111/spc3.12321>
- Stellar, J. E., & Willer, R. (2018). Unethical and inept? The influence of moral information on perceptions of competence. *Journal of Personality and Social Psychology, 114*(2), 195–210. <https://doi.org/10.1037/pspa0000097>
- Stroebe, W., Insko, C. A., Thompson, V. D., & Layton, B. D. (1971). Effects of physical attractiveness, attitude similarity, and sex on various aspects of interpersonal attraction. *Journal of Personality and Social Psychology, 18*, 79–91. <https://doi.org/10.1037/h0030710>
- Swap, W. C. (1991). When prosocial behavior becomes altruistic: An attributional analysis. *Current Psychology, 10*(1–2), 49–64. <https://doi.org/10.1007/BF02686780>
- Tal-Or, N. (2010a). Bragging in the right context: Impressions formed of self-promoters who create a context for their boasts. *Social Influence, 5*(1), 23–39. <https://doi.org/10.1080/15534510903160480>

- Tal-Or, N. (2010b). Direct and indirect self-promotion in the eyes of the perceivers. *Social Influence*, 5(2), 87–100. <https://doi.org/10.1080/15534510903306489>
- Taylor, S. N., Wang, M., & Zhan, Y. (2012). Going beyond self–other rating comparison to measure leader self-awareness. *Journal of Leadership Studies*, 6(2), 6–31. <https://doi.org/10.1002/jls.21235>
- Tedeschi, J. T. (2013). *Impression Management Theory and Social Psychological Research*. Academic Press.
- Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration Trumps Confidence as a Basis for Witness Credibility. *Psychological Science*, 18(1), 46–50. <https://doi.org/10.1111/j.1467-9280.2007.01847.x>
- Tesser, A., Gatewood, R., & Driver, M. (1968). Some determinants of gratitude. *Journal of Personality and Social Psychology*, 9(3), 233-236. <https://doi.org/10.1037/h0025905>
- Thibaut, J. W., & Kelley, H. H. (1959). *The Social Psychology of Groups*. Routledge. <https://doi.org/10.4324/9781315135007>
- Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology*, 69(6), 1120–1138. <https://psycnet.apa.org/doiLanding?doi=10.1037%2F0022-3514.69.6.1120>
- Treger, S., Sprecher, S., & Erber, R. (2013). Laughing and liking: Exploring the interpersonal effects of humor use in initial social interactions. *European Journal of Social Psychology*, 43(6), 532–543. <https://doi.org/10.1002/ejsp.1962>
- Trötschel, R., Hüffmeier, J., Loschelder, D. D., Schwartz, K., & Gollwitzer, P. M. (2011). Perspective taking as a means to overcome motivational barriers in negotiations: When putting oneself into the opponent’s shoes helps to walk toward agreements. *Journal of Personality and Social Psychology*, 101(4), 771–790. <https://doi.org/10.1037/a0023801>
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A Person-Centered Approach to Moral Judgment. *Perspectives on Psychological Science*, 10(1), 72–81. <https://doi.org/10.1177/1745691614556679>
- Uhlmann, E. L., & Zhu, L. [Lei]. (2014). Acts, Persons, and Intuitions: Person-Centered Cues and Gut Reactions to Harmless Transgressions. *Social Psychological and Personality Science*, 5(3), 279–285. <https://doi.org/10.1177/1948550613497238>
- Uhlmann, E. L., Zhu, L. (Lei), & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, 126(2), 326–334. <https://doi.org/10.1016/j.cognition.2012.10.005>

- Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. <https://doi.org/10.1037/0033-2909.134.3.383>
- Van Velsor, E., Taylor, S., & Leslie, J. B. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender, and leader effectiveness. *Human Resource Management*, *32*(2–3), 249–263. <https://doi.org/10.1002/hrm.3930320205>
- Vonk, R. (1999). Impression Formation and Impression Management: Motives, Traits, and Likeability Inferred from Self-Promoting and Self-Deprecating Behavior. *Social Cognition*, *17*(4), 390–412. <https://doi.org/10.1521/soco.1999.17.4.390>
- Wang, C. S., Kenneth, T., Ku, G., & Galinsky, A. D. (2014). Perspective-Taking Increases Willingness to Engage in Intergroup Contact. *PLOS ONE*, *9*(1), e85681. <https://doi.org/10.1371/journal.pone.0085681>
- Wang, C. S., Ku, G., Tai, K., & Galinsky, A. D. (2014). Stupid Doctors and Smart Construction Workers: Perspective-Taking Reduces Stereotyping of Both Negative and Positive Targets. *Social Psychological and Personality Science*, *5*(4), 430–436. <https://doi.org/10.1177/1948550613504968>
- Ward, A., & Brenner, L. (2006). Accentuate the Negative: The Positive Effects of Negative Acknowledgment. *Psychological Science*, *17*(11), 959–962. <https://doi.org/10.1111/j.1467-9280.2006.01812.x>
- Wegner, D. M., & Schaefer, D. (1978). The concentration of responsibility: An objective self-awareness analysis of group size effects in helping situations. *Journal of Personality and Social Psychology*, *36*, 147–155. <https://doi.org/10.1037/0022-3514.36.2.147>
- Whitener, E. M., Brodt, S. E., Korsgaard, M. A., & Werner, J. M. (1998). Managers as Initiators of Trust: An Exchange Relationship Framework for Understanding Managerial Trustworthy Behavior. *Academy of Management Review*, *23*(3), 513–530. <https://doi.org/10.5465/amr.1998.926624>
- Wicklund, R. A. (1975). Objective Self-Awareness. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 8, pp. 233–275). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60252-X](https://doi.org/10.1016/S0065-2601(08)60252-X)
- Wong, P. T., & Weiner, B. (1981). When people ask “why” questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology*, *40*(4), 650–663. <https://doi.org/10.1037/0022-3514.40.4.650>
- Wosinska, W., Dabul, A. J., Whetstone-Dion, R., & Cialdini, R. B. (1996). Self-Presentational Responses to Success in the Organization: The Costs and Benefits of Modesty. *Basic and Applied Social Psychology*, *18*(2), 229–242. https://doi.org/10.1207/s15324834basps1802_8

Young, L., & Tsoi, L. (2013). When Mental States Matter, When They Don't, and What That Means for Morality. *Social and Personality Psychology Compass*, 7(8), 585–604. <https://doi.org/10.1111/spc3.12044>

Zahavi, A., & Zahavi, A. (1999). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. Oxford University Press.

Zhao, X., & Epley, N. (2021). Insufficiently complimentary?: Underestimating the positive impact of compliments creates a barrier to expressing them. *Journal of Personality and Social Psychology*, 121(2), 239-256. <https://doi.org/10.1037/pspa0000277>