Fostering Perceptions of Authenticity via Sensitive Self-Disclosure

Li Jiang
Leslie John
Reihane Boghrati
Maryam Kouchaki
Fostering Perceptions of Authenticity via Sensitive Self-Disclosure

Li Jiang
George Washington University

Leslie John
Harvard Business School

Reihane Boghrati
Arizona State University

Maryam Kouchaki
Northwestern University

Working Paper 20-070
Fostering Perceptions of Authenticity via Sensitive Self-disclosure

Li Jiang
Leslie John
Reihane Boghrati
Maryam Kouchaki

Li Jiang, Department of Marketing, School of Business, George Washington University; Leslie John, Department of Negotiation, Organizations & Markets, Harvard Business School; Reihane Boghrati, Department of Information Systems, Arizona State University; Maryam Kouchaki, Department of Management & Organizations, Kellogg School of Management, Northwestern University. The research is self-funded, and does not have additional funding sources.

Correspondence concerning this article should be addressed to Li Jiang, Department of Marketing, School of Business, George Washington University, Washington DC, 20052. E-mail: lijiang1@gwu.edu.
Fostering Perceptions of Authenticity via Sensitive Self-Disclosure

Abstract

Leaders’ perceived authenticity—the sense that they are acting in accordance with their “true self”—is associated with positive outcomes for both employees and organizations alike. How might a leader foster this impression? We show that sensitive self-disclosure, in the form of revealing weaknesses, makes leaders come across as authentic (Studies 1 & 2)—because observers infer that the discloser is not engaging in strategic self-presentation (Study 3). Further, we show that the authenticity gains of sensitive self-disclosure have positive downstream consequences, such as enhancing employees’ desire to work with the leader (Studies 4A and 4B). And, as our conceptual account predicts, these benefits emerge when the revealed weakness is made voluntarily (as opposed to by requirement) (Study 5), and by a relatively high-status person (Study 6). Finally, we conclude by presenting anecdotal field evidence (Study 7) consistent with the causal effects identified in Studies 1-6.

Keywords: authenticity, weaknesses, self-disclosure, leaders’ interpersonal perception
Public Significance Statement

When a leader self-discloses a weakness, s/he can be perceived as authentic, leading to positive downstream consequences, such as enhancing employees’ desire to work with the leader. This research suggests that leaders can consider sensitive self-disclosure as a tool to achieve positive outcomes—for themselves, their employees, and the firm alike.
**Fostering Perceptions of Authenticity via Sensitive Self-Disclosure**

Authenticity has become increasingly important (Sargent, 2016; Szalai, 2015; Talbot-Zorn & Marz, 2016; Zimmer, 2016; Zogby, 2016). Research in organizational behavior indicates that employees prefer leaders whom they perceive to be authentic (e.g., Clapp-Smith et al., 2009; D. S. Wang & Hsieh, 2013; H. Wang et al., 2014), with Generation Z being particularly likely to prioritize authenticity over other factors when choosing whom to work with (Cronin, 2019; Laudert, 2018). Consistent with these preferences, perceived authenticity—the perception that leaders are being genuine, acting in accordance with their true selves (Cha et al., 2019; George et al., 2007; Lehman et al., 2019)—is associated with positive outcomes for both employees and organizations. When followers perceive leaders to be authentic, they experience greater well-being (Rahimnia & Sharifirad, 2015; H. Wang et al., 2014), are more trusting of the organization (Avolio et al., 2004; Norman et al., 2010), perform better (Hannah et al., 2011; Leroy et al., 2012; Lyubovnikova et al., 2017; Rego et al., 2013, 2015), work harder (Hirst et al., 2016), and make more ethical decisions (Cianci et al., 2014; Zhu et al., 2011).

Despite these benefits, research also suggests that leaders struggle to come across as authentic (Hahl et al., 2017; Hahl & Zuckerman, 2014). Leaders are sometimes seen as manipulating their public images to seek power and status—regardless of whether they are actually engaging in such manipulation—which poses a barrier to being perceived as authentic (Fine, 2003; Hahl & Zuckerman, 2014; Zukin, 2008). Thus, the question arises: What can help leaders to come across as authentic? We propose that leaders can foster perceptions of authenticity by engaging in sensitive self-disclosure, which we operationalize in this context as revealing work-related weaknesses. In the following sections, we review prior literature that forms the basis of our predictions, and provide an overview of our empirical work.
Conceptual Development

“Sensitive self-disclosure” refers to the revelation of self-relevant information that makes a person vulnerable to being judged negatively by others (Derlega et al., 1993; Kelly & McKillop, 1996; Laurenceau et al., 1998; Moon, 2000). Such information tends to be unfavorable or socially undesirable, though this definition is naturally context-dependent. Within our context of interest—leaders’ disclosures within the workplace—we define “sensitive self-disclosure” as revealing work-related weaknesses. This is because disclosing job-related weaknesses, such as not being good at public speaking, not being good at time management, or lacking a vision, plausibly makes a leader vulnerable to being judged negatively by followers—it may threaten followers’ perceptions of that leader’s ability to lead effectively.

People often shy away from disclosing sensitive personal information (Bruk et al., 2018; Gromet & Pronin, 2009; John et al., 2016)—sometimes for good reason, given the risks. Disclosure of unsavory self-relevant facts can make a bad impression, especially given that people tend to overweigh negative information relative to positive information (Baumeister et al., 2001; Herr et al., 1991). Of particular relevance, revealing weaknesses can diminish others’ perceptions of the discloser’s status (Gibson et al., 2018).

Perhaps as a result of these negative outcomes, and motivated by the desire to be viewed favorably by others, people tend to disclose favorable, and withhold unfavorable, personal information (De Angelis et al., 2012; Leary & Allen, 2011; Paulhus & Reid, 1991; Turnley & Bolino, 2001). This is especially true in the workplace, where self-presentation concerns loom large. Positive perceptions indeed result in social and material rewards (Gilmore & Ferris, 1989; Leary, 1996; Schlenker, 1975), and thus individuals are strongly motivated to manipulate their image. For instance, in job interviews, people engage in extensive image creation, to the point of
making up fictional stories to showcase their strengths (Levashina & Campion, 2007). Relatedly, as organizational theorists have long noted, there is often a gap between the frontstage—i.e., a person’s public persona—and their backstage, in which frontstage is manipulated by the actor to gain extrinsic rewards (Trilling, 1972; Turner, 1976).

But we posit that this reticence to reveal unfavorable personal information comes with a cost: it contributes to perceptions of inauthenticity. Self-presentational acts are often subject to assessments of authenticity (Buss & Briggs, 1984; Leary, 1993; Schlenker, 1975; Tesser & Moore, 1986), as observers infer whether the actor appears to be presenting her true self or not (Deci & Ryan, 2000, 2002; Kernis & Goldman, 2006; Lenton et al., 2013; Sedikides et al., 2017). If we only reveal our desirable qualities, we are only showing a very narrow “sample” of our true selves. Observers who notice such selective presentation may infer that the actor must be motivated to impress others and thus presenting an insincerely positive image to others.

In contrast, when a person engages in sensitive self-disclosure, observers may infer that the actor has not filtered out information. This, we argue, creates the impression that the actor is revealing himself in a more complete, comprehensive, or unbiased way. As a result, we argue, observers perceive that actor as authentic. In making this proposition, we draw on seminal work in sociology on “staged authenticity”—the notion that access to “back regions” can enhance the intimacy, and perceived authenticity, of an experience (MacCannell, 1973)—as when, for example, a diner enters the kitchen area of a restaurant. Here, we posit that in interpersonal interactions, voluntarily allowing a person into one’s “backstage” by revealing something sensitive, can foster perceptions of authenticity.

Central to our account, we propose that the capacity for sensitive self-disclosure to foster perceptions of authenticity is driven by observers’ inferences about the discloser’s motives. Prior
work indicates that observers routinely make inferences about the motives that underlie others’ behavior (Campbell & Kirmani, 2000; Heider, 1958; Pizarro et al., 2003). For example, when a salesperson flatters a consumer prior to making a sale, consumers perceive that salesperson as insincere—because consumers infer that the salesperson has an ulterior motive (i.e., the salesperson is being complimentary only to make the sale; Campbell & Kirmani, 2000). Conversely, when a person uses politically incorrect (vs. correct) language, she comes across as authentic because she is perceived to lack strategic motives (Rosenblum et al., 2020).

What motivational inferences might observers make from a leader’s sensitive self-disclosures? We posit such disclosures to affect inferences of the leader’s self-presentation motives—or rather, the lack of such motives. Self-presentation refers to people’s fundamental motive to be seen positively by others (Baumeister, 1982; Leary & Kowalski, 1990; Tetlock, 2002); people tend to manipulate their images in an effort to be perceived in a desirable light (De Angelis et al., 2012; Leary & Allen, 2011; Paulhus & Reid, 1991; Turnley & Bolino, 2001). As such, self-presentation is often considered as strategic (Eastman, 1994). Drawing on the motivational inference literature, we propose that witnessing a leader self-disclose a weakness underscores an implicit assumption held by observers—that the person who discloses a weakness must not have filtered out information. Therefore, we propose that when a leader disclose a weakness, observers infer that that leader is not engaging in strategic self-presentation. As a result, the leader is perceived as relatively authentic.

The Stereotype Content Model (SCM), Fiske et al.’s (2002) influential model of person perception, is also potentially relevant here; it is plausible that a leader’s sensitive self-disclosure could affect how warm and competent she comes across—and these effects could potentially “crowd out” that of authenticity. For example, given prior work showing a link between self-
disclosure and liking (Collins & Miller, 1994; Cozby, 1972; Dalto et al., 1979; Jourard, 1959; Worthy et al., 1969), sensitive self-disclosure could increase perceived warmth. Moreover, the SCM would seem to treat the construct of “authenticity” as being a component of “warmth;” Fiske et al. (2002) include “sincere” in their multi-item measure of warmth. However, more recent work suggests that perceived authenticity is distinct from that of warmth. For example, Rosenblum, Schroeder, and Gino (2020) showed that speaking in politically incorrect tones makes people come across as authentic but *not* warm. Thus, we predict that sensitive self-disclosure will increase perceived authenticity even when controlling for the SCM’s two dimensions of person perception: warmth and competence.

Finally, past research has shown that negative information can have positive effects in the context of interpersonal attractiveness (Aronson et al., 1966; Collins & Miller, 1994), on the management of malicious envy from peers (Brooks et al., 2019), and on enhancing the effectiveness of persuasive appeals (e.g., two-sided messaging; see Crowley & Hoyer, 1994 as an example). We extend these findings and demonstrate that sensitive self-disclosure can enhance perceptions of authenticity. And, as we delineate in the next section, we further distinguish our account from related work by showing that it makes unique predictions about when sensitive self-disclosure will—vs. will not—foster perceptions of authenticity.

**Moderators**

*Voluntariness*

We suggest that for leaders’ disclosure of their weaknesses to boost perceived authenticity, they must be made *voluntarily*. This prediction stems from the fact that in making dispositional inferences about a person, observers take intentions into account. For example, actors are judged to be more moral and less blameworthy when they inadvertently, as opposed to
intentionally, cause something bad to happen (Greene et al., 2009; Pizarro et al., 2003). And, of particular relevance to the domain of self-disclosure, the negative signal that can arise from explicitly withholding information (e.g., refusing to answer a direct question) is restricted to situations in which a person volitionally withholds, as when, for example, they refuse to answer a question (as opposed to not answering simply because they did not see the question) (John et al., 2016). Analogously, we propose that for leaders to reap the authenticity benefits of revealing weaknesses, followers must perceive those leaders to be revealing on their own accord. Thus, it is not enough for followers to have awareness of their leaders’ weaknesses; the act of voluntary self-disclosure is crucial to boosting perceptions of authenticity.

**Status**

We posit the authenticity gains from self-disclosing weaknesses in organizations to be pronounced for high-status individuals—i.e., leaders within the organization. First, this prediction is rooted in work at the intersection of social identity and leadership. Specifically, as Giessner & van Knippenberg (2008) demonstrated, leaders are sometimes given a “license to fail” (Giessner & van Knippenberg, 2008): relative to low status individuals, high status individuals were treated more favorably after they failed to achieve a goal. Applied to the present context, this suggests that when leaders reveal weaknesses, they may be particularly poised to reap the benefits of doing so, and to avoid potential pitfalls.

Second, prior work indicates that leaders are particularly driven to present themselves in a favorable light (Bolino et al., 2008; Peck & Hogue, 2018)—for example, when managers have information that makes them look bad, they are particularly likely to keep this information private (Harrison & Harrell, 1993). Thus, given leaders’ particular reticence to reveal negative self-relevant information, we reason that when they do reveal such information, they are
particularly likely to come across as authentic. Indeed, the results of a pilot study point to leaders’ reticence to reveal weaknesses. We asked 110 full-time U.S. managers ($M_{age} = 37.2$ years, $SD = 10.5$; Male: 50.0%; White: 80.9%; Median income: $90,000 - $99,999) to write three pieces of self-relevant information: something favorable (i.e., something they are good at in the workplace), something neutral (i.e., hobbies), and something unfavorable (i.e., a workplace weakness). We then asked them which, if any, of these three facts about themselves they would include when introducing themselves to a new hire at work. Only 34.5% chose to disclose the weakness; by comparison, 96.3% chose to include the strength, and 64.5% chose to include the neutral fact (details in the supplement).

### Overview of Studies

Our empirical package consists of four sections. In Section 1, we demonstrate the effect of a leader’s disclosure of a weakness on perceptions of authenticity (Studies 1 and 2, Supplemental Studies 1A-1D) and the motivational inferences that we posit to underlie it (Study 3). We begin by presenting vignette studies (Study 1 and S1A-S1D), followed by a more naturalistic study in which participants watch a video of a Google executive who either discloses or does not disclose a weakness within a self-introduction (Study 2). In Studies 1 and 2, we also show that the effect holds when controlling for perceived warmth and competence.\(^1\) Next, we show that the capacity for a leader’s sensitive self-disclosure to foster authenticity is driven by the perception that he is not engaging in strategic self-presentation (Study 3).

The goals of Section 2 were twofold: to further increase realism by using live interaction paradigms; and to assess positive behavioral consequences of leaders’ self-disclosure of

---

\(^1\) We also measured competence in some of the studies in Sections 2 and 3. Consistent with Studies 1 and 2, the results hold when we control for this. Due to space constraints, however, we report these secondary results in the supplement.
weaknesses—consequences that are downstream from the effect on perceived authenticity identified in Section 1. We show that followers are more likely to put their own earnings at risk in the hands of (Study 4A), and to choose to work with (Studies 4A & 4B), leaders who disclose weaknesses. Tying these patterns back to the basic effect identified in Section 1, in Section 2, we also show that these positive behavioral consequences are mediated by perceived authenticity.

Section 3 tests our theory-derived moderators: voluntariness (Study 5) and status (Study 6). Specifically, these studies demonstrate that downstream positive consequences of revealing weaknesses in the workplace are limited to situations in which the disclosure is made voluntarily (as opposed to by requirement, Study 5) and by a person occupying a high-status role within the organization (Study 6). Again, we tie these patterns back to the basic effect identified in Section 1; here, by measuring perceived authenticity and documenting, moderated mediation.

In Section 4, we conclude by presenting anecdotal field evidence consistent with the causal effects identified in Sections 1-3. Specifically, using actual disclosures from a professional social networking app, we show that there is indeed a positive association between sensitive self-disclosure and subordinates’ reactions to those revelations (Study 7).

In total—across the main manuscript and supplement—we report the results from twelve studies (N = 38,354; consisting of n = 3,281 from eleven experiments and n = 35,073 from a field study). Online subjects were recruited online via Amazon Mechanical Turk or Prolific; Lab subjects were recruited from a northeastern university. As for data collection stopping rules, for our pre-registered, online studies (Studies 1, 2, 3, 4B, and Supplemental Study 1C), we adhered to our pre-registered sample size and exclusion criteria. Supplemental Studies 1A, 1B, 1D, and Study 5 were collected before the rise of the pre-registration procedure, and we aimed to collect 100 participants per condition. For Studies 4A and 6, the lab studies, we collected as much data
as we could within a preset number of sessions of data collection. We did not analyze data until the targets were met. For the field data (Study 7), we analyzed all of the data that our field partner gave us. We report all manipulations and measures; for brevity, some measures are only reported in the supplement. Any measures not reported in the manuscript are reported in the supplement. See Table 1 for a design overview of each study, and the supplement for more information about our supplemental studies, samples and designs.

Section 1: Basic Effect and Mechanism

Study 1. Vignettes

In a series of vignette studies, we experimentally test the idea that when leaders disclose their weaknesses, they come across as authentic. Here, we present one of these studies in full. As for the others, here, we provide details on their relative contribution as well as a meta-analysis of their effect sizes; we present them in full in the supplement.

Method

Participants. Two hundred and ninety-eight Prolific working professionals (full-time employed, 147 males; \( M_{age} = 32.8 \) years, \( SD = 9.8 \); White: 81.8%) participated. We prescreened participants for employment status, and only recruited full-time employed working professionals. We pre-registered our hypotheses and measures (https://aspredicted.org/F9N_FWS).

Materials and Procedure. Participants imagined that they were a new employee of a (fictitious) company called RockInvest and they met different managers who they would have the opportunity to work with if they wanted to. Participants were randomized to one of two conditions: a control condition, in which the manager did not disclose a weakness, or an experimental condition, in which he disclosed a work-related weakness.

Specifically, in the control condition, participants were told:
"I began my career as a mortgage trader at RockInvest. The company, launched in 1988, initially focused on bonds. But thanks to shrewd acquisitions, the firm is now the world’s largest asset manager, with $870 billion, offering a slew of equity funds and multi-asset funds. I take care of my staff, offering health benefits even to part-timers. I like to climb mountains in Colorado and collect American folk art."

In the experimental condition, we appended this sentence, in which the manager disclosed a weakness: “Even if I am a manager of a multi-billion company, I am not good at public speaking. When I make a speech, my mouth gets dry and I sometimes start to panic.”

**Measures.** We assessed perceived authenticity by asking participants to rate the CEO on six items (α = .95: authentic, real, sincere, genuine, inauthentic (reverse-coded), phony (reverse-coded), which, when combined, form a measure of perceived authenticity. Unless noted otherwise, all items used a 7-point scale ranging from 1 (not at all) to 7 (very much). The items were adapted from established perceived authenticity scales (see Cheshin et al., 2018; Gershon & Smith, 2020; Grandey et al., 2005; Hahl & Zuckerman, 2014). We selected competence (α = .92: competent, efficient, intelligent) and warmth measures that prior scholars have used (α = .93, warm, kind, easygoing) (Aaker et al., 2010; Fiske et al., 2007; Goodwin et al., 2014). The order of the authenticity, competence, and warmth measures was randomized.

**Results**

**Perceived authenticity.** The manager was perceived as more authentic when he disclosed a weakness relative to when he did not (M<sub>experimental</sub> = 5.65, SD = 0.92; M<sub>control</sub> = 4.97, SD = 1.11), t(296) = 5.69, p < .0001, Cohen’s d = 0.56.

**Perceived competence and warmth.** The manager was perceived to be just as competent when he disclosed a weakness relative to when he did not (M<sub>experimental</sub> = 5.41, SD =
0.85; \( M_{\text{control}} = 5.33, SD = 0.90 \), \( t(296) = .77, p = .44 \), Cohen’s \( d = 0.09 \); he was also perceived to be just as warm when he disclosed a weakness relative to when he did not (\( M_{\text{experimental}} = 5.02, SD = 1.02 \) vs. \( M_{\text{control}} = 4.80, SD = 1.08 \), \( t(296) = 1.81, p = .07 \), Cohen’s \( d = 0.25 \). Moreover, the effect of condition on perceived authenticity held when controlling for both perceived competence and warmth (\( t(294) = 6.31, p < .0001 \)), suggesting the effect of disclosure on perceptions of authenticity is independent of perceptions of warmth and competence.

Finally, given that some of the prior work in person perception has treated authenticity as a sub-dimension of warmth, we also conducted a factor analysis of our authenticity and warmth measures. As further evidence of the distinctiveness of the authenticity construct, the factor analysis revealed two factors, with all of the authenticity items loading on one factor, and all of the warmth factors loading on the other factor (see the supplement for details).

**Conceptual replications**

**Different weaknesses.** Supplemental Studies 1A-1D are conceptual replications of the basic effect, showing that it emerges across a variety of weaknesses (see Table 1 for the weakness used in each study).

**Sequence.** In Supplemental Study 1A (S1A), we manipulated whether the weakness was presented at the beginning versus end of the leader’s statement; the effect emerged in both cases. This robustness is noteworthy, as it distinguishes our effect from a related phenomenon, namely, how incorporating a small dose of negative information in product descriptions can lead to positive evaluations (Ein-Gar et al., 2011). Ein-Gar et al. (2011) show that this effect arises because negative information, when placed after positive information, makes the positive information more salient. Their effect does not emerge when the negative information is
presented first. By contrast, our effect holds regardless of whether the weakness is disclosed upfront versus prefaced with the disclosure of neutral or desirable information.

**Disclosure length.** In Study 1, the leader’s statement was longer in the experimental condition relative to the control condition. Therefore, we ran Supplemental Study 1C (S1C), in which we replicated the basic effect, this time keeping the length of the disclosure constant.

**Gender.** In Supplemental Study 1D (S1D), we demonstrate that the effect is robust to the leader’s gender—i.e., disclosing a weakness increases perceived authenticity both when the leader is male, as well as when the leader is female.

**Meta-analysis.** Finally, we performed a meta-analysis of the effect of self-disclosure of a weakness on perceived authenticity using the data from all of the above-mentioned studies. We used the R package meta (Schwarzer, 2007, v. 4.19-1) and used a random effects model by using the inverse variance method. The test of heterogeneity ($Q(6) = 2.85, p = .827$) was non-significant, suggesting that the studies consistently documented a significant condition effect of self-disclosure on perceived authenticity. The average effect size is 0.55 (95% CI = [0.45, 0.65]). The results are shown in Figure 1.

**Study 2. Increasing Realism**

To increase realism, in Study 2 we invited a Google executive to record a video in which we instructed him to introduce himself, and to include a weakness. We did not give the executive guidance on what weakness to disclose, because we wanted the stimuli to be as naturalistic as possible. We then edited the video to create two clips; in the experimental condition we included the self-disclosed weakness, and in the control condition we simply edited this part of the introduction out. We recruited working professionals, asked them to imagine that they had recently joined the company, and randomized them to view one of the two versions of the video.
Method

Participants. As outlined in our pre-registration (https://aspredicted.org/QJY_DNV), we recruited 400 working professionals from Prolific (U.S., full-time employed; 203 females; $M_{age}$ = 32.5 years, $SD = 9.6$; White 76.4%).

Materials and Procedure. The executive was instructed to think about how he may introduce himself to new employees at his company, and to include anything he would like to in this self-introduction. In addition, we asked him to disclose a weakness of himself; he disclosed that he had joined the company “after applying to nearly 36 other roles and consequently receiving 35 other rejections.” See the supplement for the full video transcript. Between-subjects, we manipulated whether this sentence was present. Next, we recruited working professionals and instructed them to imagine that they had just joined a company and were meeting different managers who they could choose to work with (or not). Participants were randomly assigned to watch one of the two videos of the executive—the only difference being that in the experimental condition, the manager disclosed a weakness; whereas in the control condition, he did not.

Measures. We used the same measures of authenticity ($\alpha = .94$), competence ($\alpha = .91$), and warmth ($\alpha = .91$) as in Study 1, question order randomized.

Results

Perceived authenticity. The manager was perceived as more authentic when he disclosed a weakness relative to when he did not ($M_{experimental} = 5.51, SD = 1.10$; $M_{control} = 5.24, SD = 1.16$), $t(398) = 2.41, p = .016$, Cohen’s $d = 0.24$.

Perceived competence and warmth. The manager was perceived as just as competent when he disclosed a weakness relative to when he did not ($M_{experimental} = 6.06, SD = 0.83$; $M_{control} = 6.01, SD = 0.81$), $t(398) = .67, p = .50$, Cohen’s $d = 0.06$. He was also perceived as just as
warm when he disclosed a weakness relative to when he did not \((M_{\text{Experimental}} = 4.85, SD = 1.16 \text{ vs. } M_{\text{Control}} = 4.68, SD = 1.22), t(398) = 1.43, p = .15, \text{Cohen's } d = 0.14\). Moreover, the effect of condition on perceived authenticity held when controlling for warmth and competence \((t(396) = 2.50, p = .013)\).

**Study 3. Perceived Strategic Self-Presentation as Mechanism**

We propose that when a leader discloses a weakness, it makes observers less likely to perceive that leader to be acting strategically, in turn fostering perceptions of authenticity. Thus, in Study 3, we test whether the effect of disclosing a weakness on perceived authenticity is mediated by inferences of strategic self-presentation. We use procedures similar to Study 1.

**Method**

**Participants.** Three hundred working professionals from Prolific (U.S., full-time employed, 146 males; \(M_{\text{age}} = 31.9 \text{ years, } SD = 8.6; 78.5\% \text{ White}\) participated, as pre-registered (https://aspredicted.org/blind.php?x=/TBG_TG8).

**Materials and Procedure.** Participants imagined that they were a new employee of a (fictitious) company called RockInvest and were meeting different managers with whom they could choose to work. Participants were randomized to one of two conditions: a control condition, in which the manager did not disclose a weakness, or the experimental condition, in which he disclosed a weakness. Specifically, in the control condition, participants were told the same description of the manager as in Study 1. In the experimental conditions, this sentence was appended: “Even though I have managed the company for many years, I struggle with adapting to new technologies, and as a manager I am not on top of technological changes.” Participants then completed the dependent measures and provided demographic information.
**Measures.** We used the same measure of authenticity ($\alpha = .94$) as in Study 1, and added a measure of perceived strategic self-presentation, adapted from Rosenblum et al. (2020). Specifically, participants were asked: “to what extent does the manager’s self-introduction seem to be strategic?” on a 7-point scale from 1 (*not at all*) to 7 (*very much*).

**Results**

**Perceived authenticity.** The manager was perceived as more authentic when he disclosed a weakness relative to when he did not ($M_{\text{experimental}} = 5.59, SD = 0.92; M_{\text{control}} = 4.79, SD = 1.25$), $t(298) = 6.37, p < .0001$, Cohen’s $d = 0.73$).

**Perceived strategic self-presentation.** The manager was perceived as less likely to be engaging in strategic self-presentation when he disclosed a weakness relative to when he did not ($M_{\text{experimental}} = 5.08, SD = 1.18; M_{\text{control}} = 5.73, SD = 1.23$), $t(298) = -4.67, p < .0001$, Cohen’s $d = 0.54$.

**Mediation.** Bootstrapping analyses (with 10,000 resamples) showed that perceived strategic self-presentation mediated the relationship between weakness disclosure and perceived authenticity: the index of indirect effect excluded zero ($b = .100, SE = .046$, 95% CI = [.027, .211]), suggesting a significant indirect effect (Hayes, 2017). Specifically, self-disclosure of a weakness decreased perceptions of strategic self-presentation ($b = -.650, SE = .139$), $t(298) = -4.67, p < .0001$, which in turn heightened perceived authenticity ($b = -.153, SE = .052$), $t(298) = -2.96, p = .003$. Perceptions of strategic self-presentation explain 14.5% of the variance in perceived authenticity.

In sum, and consistent with our theorizing, Study 3 suggests that when a leader reveals a weakness, it dampens the perceptions that he is engaging in strategic self-presentation, which, in turn, increases perceived authenticity.
Section 2: Live Interactions and Behavioral Outcomes

So far, via vignette and video paradigms, we have documented that leaders can come across as authentic when they reveal weaknesses, and that this effect is driven by dampened perceptions of strategic self-presentation. And, pointing to the distinctiveness of the effect, it holds when controlling for perceptions of warmth and competence. The goals of Section 2 were twofold: to further increase realism by using live interaction paradigms; and to assesses positive behavioral consequences of leaders’ self-disclosure of weaknesses—consequences downstream from the effect on perceived authenticity identified in Section 1. Specifically, we show that leaders’ self-disclosure produces credible positive outcomes: employees are willing to risk their own money at the hands of the leader (Study 4A), and to choose to work with that leader for a subsequent task (Study 4B). Moreover, tying these patterns back to the basic effect identified in Section 1, in Section 2, we also show that the positive behavioral consequences of a leader’s sensitive self-disclosure are mediated by perceived authenticity.

Study 4A: In-person Interaction Study

In Study 4A, participants engaged in a face-to-face, simulated employment task. In this interaction, the participant randomized to the role of manager was privately instructed to disclose—or to not disclose—a weakness to the other participant, who was randomized to the role of prospective employee. In an incentive compatible task, we then assessed whether employees were willing to entrust the manager with their money. Secondarily, we also assessed whether managers instructed to disclose a weakness would accurately predict that doing so would lead to positive behavioral outcomes (relative to managers randomized to not disclose a weakness).

Method
Participants. We recruited students and community members \( N = 218; 99 \) males; \( M_{\text{age}} = 22.3 \text{ years}, SD = 4.3; \) White: 40.8%; Part-time employed: 42%; Full-time employed: 4%) to come to a lab at a northeastern university. Participants received a $15 base payment plus study earnings, as described below.

Materials and Procedure. In a simulated hiring task, we randomized half of participants to the role of manager and the other half to the role of prospective employee, and randomly grouped participants into manager-employee dyads. Participants began the session in individual cubicles, where they were informed of their assigned role and given information on the task to follow.

At the start of the study, prospective employees were told that they would be participating in a simulated employment task and that in a moment, they would meet their potential manager, who would have a task for them to complete. Therefore, they were told, the manager would be evaluating the prospective employee’s performance on the task.

Managers were informed that they would be meeting their potential employee for their team and would assign the employee a ten-item “word correction” task. In reality, we used the word-correcting task as a cover story. Managers were further informed that in a few minutes, they would meet their employee, at which point they should introduce themselves using a script provided for this purpose. Critically, the script manipulated whether managers would disclose a weakness. Specifically, in the control condition, managers were instructed to introduce themselves by saying:

Hi, I am [name], the manager. I am going to direct the task and the standards by which the work is to be evaluated. In addition, I will also evaluate you at the end of the session in a private questionnaire. Let me introduce myself a little bit: I am the president of the
graduate student association at the university. I get to travel often to cities across the
country to give presentations. I enjoy what I do.

In the experimental condition, the script was the same, except that the following sentences were appended to the end: “I’m quite shy. I am nervous about public speaking, and I have a habit of cracking my knuckles.”

Next, participants were randomly assigned to manager-employee dyads; each dyad was ushered into their own private room to complete the task. Managers were given a few minutes to practice the script so that they could deliver it from memory, without a written script, when they introduced themselves to the employees. Next, the manager assigned the employee the task and, using a stopwatch, gave the employee one minute to complete it.

**Survey measures.** After the task, participants returned to their individual cubicles. Each prospective employee assessed their manager’s authenticity ($\alpha = .88$) as in Study 1, and their desire to work with the manager: “Would you want to be paired with this manager again for a subsequent task?” measured on a 7-point scale from 1 (want to work with a different manager) to 7 (want to work with this manager).

Managers indicated how they thought the prospective employee viewed them; specifically, we asked managers: “Do you think the prospective employee would want to be paired with you as a manager again for a subsequent task?” on a scale from 1 (want to work with a different manager) to 7 (want to work with this manager).

Finally, both managers and prospective employees indicated whether they knew each other before the experiment; three pairs did, and therefore were excluded from the data analysis, leaving 212 participants. The results are substantively equivalent when these three dyads are included.
Incentive-compatible behavioral outcome measure. Participants engaged in a trust game (Berg et al., 1995), which served as an incentive compatible measure of cooperative behavior (Cesarini et al., 2008; Epley et al., 2006). Essentially, we were interested in whether employees’ positive assessments of a leader who self-discloses a weakness might manifest in a willingness to entrust the manager with their money; the trust game measures the extent to which people are willing to put their own money at risk by entrusting their counterpart (in this case, the manager) with it.

We explained the game to all participants, telling them that the employee would receive an initial endowment of $3 (in quarters), and would have to decide how much, if any, of this money to transfer to the manager. Any amount transferred would be tripled. Next, the manager would decide how much, if any, of this tripled amount s/he would like to send back to the employee.

Participants were encouraged to ask questions or re-read instructions if they did not understand how the game worked. Upon checking a box labelled “I understand how the game works,” participants proceeded to the game, with each employee indicating how much, if any, money to transfer to their manager, and with the manager then indicating how much, if any, of this money to return to their employee. Participants were given real money to play the trust game, and could earn as much as $9 (plus the $15 base payment). At the end of the experiment, participants provided demographic information and were debriefed.

Results

Employees.

Perceived authenticity. Prospective employees perceived their manager as more authentic when their manager disclosed a weakness relative to when s/he did not (Mexperimental =
5.43, $SD = 1.10; M_{\text{control}} = 4.92, SD = 1.09), t(104) = 2.42, p = .017, \text{Cohen’s } d = 0.49.

**Willingness to work with the manager.** Prospective employees were more interested in continuing to work with their manager when that manager disclosed a weakness relative to when s/he did not ($M_{\text{experimental}} = 5.41, SD = 1.34; M_{\text{control}} = 4.78, SD = 1.38), $t(104) = 2.37, p = .019, \text{Cohen’s } d = 0.46.$

**Incentive compatible behavioral outcome.** Prospective employees transferred more money to the manager when their manager disclosed a weakness relative to when s/he did not ($M_{\text{experimental}} = $2.39, SD = 0.84; M_{\text{control}} = $2.00, SD = 1.07), $t(104) = 2.04, p = .043, \text{Cohen’s } d = 0.41.$

**Mediation.** We conducted a mediation analysis with willingness to work as the dependent variable to test mediation by authenticity. A 10,000-sample bootstrap analysis (Hayes, 2017, Model 4) showed that the index of mediation excluded zero ($b = .330, SE = .146, 95\% \text{CI} = [.078, .652]), indicating a significant indirect effect. Perceived authenticity explained 30.2\% of the variance in willingness to work. We observed similar mediation analyses results with incentive-compatible behavioral outcome (i.e., money allocation, $b = .130, SE = .069, 95\% \text{CI} = [.028, .315]) as dependent variable, again indicating a significant indirect effect. The direction of the mediations indicates that revealing a weakness increased perceived authenticity, in turn increasing both willingness to work with the manager, and money transferred. Perceived authenticity explained 11.8\% of the variance in money allocation.

**Managers.** There were no differences between conditions in managers’ predictions of whether the prospective employee would want to be paired with them again for a subsequent task ($M_{\text{experimental}} = 4.81, SD = 1.03; M_{\text{control}} = 4.56, SD = 1.19), $t(104) = 1.16, p = .25, \text{Cohen’s } d = 0.22.$ This result suggests that would-be disclosers may be unaware of the benefits of sensitive
self-disclosure: managers induced to disclose a weakness did not appear to anticipate that doing so would cause their employees to want to work with them. This result is consistent with the pilot study reported in the introduction, in which the majority of managers chose to not disclose a weakness in a self-introduction to a prospective employee.

**Study 4B: Online Interaction Study with Working Professionals**

Study 4B replicates and extends Study 4A in several ways. In Study 4B, we further enhanced realism in two ways. First, instead of giving participants a script, in Study 4B participants’ weakness disclosures were self-generated. Second, we recruited working professionals (as opposed to students and community members, as we had in Study 4A); participants in Study 4B worked in various industries, in a wide range of professions. In addition, Study 4B examined the effect in the context of a manager-employee relationship rather than in an interview context.

**Method**

**Participants.** We recruited working professionals from Prolific ($N = 400; 187$ males; $M_{\text{age}} = 30.5$ years, $SD = 10.7$; White: 72.1%), as indicated in our pre-registration (https://aspredicted.org/Z8N_GND).

**Materials and Procedure.** We randomized half of participants to the role of manager and the other half to the role of employee, randomly grouping them into manager-employee dyads in real-time. We used Qualtrics and SMARTRIQS software to perform this matching (c.f. Molnar, 2019). The matching procedure operates by having participants enter the survey at the same time—participants wait for up to two minutes until another participant joins; if no participant joins within two minutes, then the participant is thanked and paid for their participation. This procedure resulted in 80.3% of participants being matched (as indicated in our
pre-registration, our target sample size was 400 matched participants). Importantly, this matching procedure occurred prior to randomization.

Each dyad was given four minutes to chat, ostensibly to get to know each other before engaging in a task together. Critically, it was during this “get-acquainted” chat session that we induced the managers to either disclose a weakness (experimental condition) or not (control condition). Thus, prior to the chat session, participants were given the following information.

**Employees.** At the start of the study, employees were told that they would be participating in a task together with the manager. Before the task, they would chat with their manager using a chat window.

**Managers.** Managers were instructed to start the chat with the employee by introducing themselves, and were randomly assigned to one of the two conditions. In both conditions, before entering the chat, managers were told: “First, we would like you to chat with the employee, to get to know each other a bit. As the manager, you will start the conversation by telling your employee a bit about yourself. Please chat in a natural way and make sure you respond to your employee.” Further, managers were given specific information on what to include in their self-introduction, as follows:

“What to include in your introduction:

1. Your profession (but not where you work).

2. What you are good at.

For managers assigned to the experimental condition, there was a third bullet point, which read: “3. A work-related weakness.” To help managers come up with a weakness, we further told these participants:
“For the work-related weakness, sometimes it’s hard for people to come up with this. Here are some prompts that might help you come up with a weakness to reveal.

- Do you sometimes procrastinate? If so, you could say something like “I sometimes procrastinate and do things last minutes.”
- Do you sometimes let your personal life interfere with your performance? If so, you could say something like “I have to admit that sometimes my personal life interferes with my job.”
- Do you sometimes arrive late? If so, you could say something like “I am only human... occasionally I start work a little late.”

Participants were told that their payment was contingent on complying with these instructions.

Next, each pair was given four minutes to chat freely in a chat box in real time. After the chat, employees completed several measures, described next. Managers provided demographic information, were debriefed that there was no additional task, and paid.

**Measures.** Employees rated their managers’ perceived authenticity, as in Study 4A. Employees were also asked to choose whether they would like to work with the manager to complete the task: “For a subsequent task you are going to complete, you have the option to choose whether you want to work with this manager or to be paired with another manager. How much are you willing to work with the manager in the subsequent task?” on a scale of 1 (not at all) to 7 (very much). The order of authenticity and willingness to work questions were counterbalanced. Next, all participants were asked: “Did you feel the chat to be natural?” (1 = yes, 2 = no), provided demographic information, and were debriefed.

**Results**
Managers. All matched managers chatted with their employees. 92.3% of participants indicated that they thought the chat to be natural. Compliance was high: in the control condition, no managers disclosed a weakness, whereas in the experimental condition, 94.0% did so. A research assistant coded the managers’ disclosures; they disclosed a wide range of work-related weaknesses: 22.3% disclosed weaknesses in time management (e.g., procrastinating, being late for work), 19.1% in stress management, 18.1% in public speaking, 10.6% in social struggle, 7.4% in project management skills, 6.4% in the ability to focus, 5.3% in being patient, and 1.1% in being overconfident. We adopt an intent-to-treat approach to data analysis, whereby all employees were included in analysis, regardless of whether their manager actually complied. The results remained significant if we only included complied participants.

Employees.

Perceived authenticity. Employees perceived their manager as more authentic when their manager was instructed to disclose a weakness relative to when s/he was not ($M_{\text{experimental}} = 4.96, SD = 1.73; M_{\text{control}} = 4.43, SD = 1.35$), $t(198) = 2.33, p = .021$, Cohen’s $d = 0.37$.

Willingness to work with the manager for a subsequent task. Employees were more interested in working with their manager when their manager was instructed to disclose a weakness relative to when s/he was not ($M_{\text{experimental}} = 5.41, SD = 1.43; M_{\text{control}} = 4.90, SD = 1.91$), $t(198) = 2.14, p = .034$, Cohen’s $d = 0.30$.

Mediation. We conducted a mediation analysis with willingness to work as the dependent variable to test mediation by authenticity. A 10,000-sample bootstrap analysis (Hayes, 2017, Model 4) showed that the index of mediation excluded zero ($b = .359, SE = .157, 95\% CI = [.083, .703]$), suggesting a significant indirect effect. The direction of the mediation indicates that revealing a weakness increased perceived authenticity, which in turn increased willingness to
work with the manager. Perceived authenticity explained 39.1% of the variance in willingness to work.

In sum, Studies 4A and 4B provide converging evidence that sensitive self-disclosure can make leaders come across as authentic, resulting in positive downstream consequences, such as a heightened interest in working for that leader.

Section 3: Moderators

In Section 3, we test theory-derived moderators: voluntariness (Study 5) and status (Study 6). We predicted that downstream positive consequences of revealing weaknesses in the workplace are limited to situations in which the disclosure is made voluntarily (as opposed to by requirement, Study 5) and by a person occupying a high-status role within the organization (Study 6).

Study 5: Voluntariness

In Study 5, participants read a manager’s disclosure and indicated their willingness to work with that manager, as well as their perceptions of that manager’s authenticity. The study was a 2x2 between-subjects design in which we manipulated the content of the disclosure (weakness disclosed vs. no weakness disclosed), as well as the voluntariness of the disclosure (voluntary vs. required). Participants were randomized to one of the four conditions.

We predicted an interaction such that interest in working for the manager would be heightened only in the condition in which the weakness was disclosed voluntarily, as opposed to by requirement. We also predicted that this effect would arise via gains in perceived authenticity—i.e., we predicted moderated mediation.

Method
Participants. Three hundred and ninety-two U.S. MTurk workers (174 males; $M_{age} = 29.8$ years, $SD = 12.3$; White: 80%; Median income: $50,000-$99,999) participated.

Materials and Procedure. Participants read how a previous participant had ostensibly introduced themselves in a prior experiment:

I am a manager of a technological company. I began my career as an engineer at this company. Thanks to shrewd acquisitions, the firm is now one of the big companies in the field. As a manager, I take care of my staff, offering health benefits even to part-timers. I like to climb mountains in Colorado and collect American folk art.

We manipulated whether the manager disclosed a weakness by, for half of participants, appending the following self-disclosure: “Even though I am a manager of the company, I am nervous about public speaking and I have a habit of cracking my knuckles.”

We manipulated whether the manager’s disclosures were made voluntarily by informing half of participants that the disclosure had been required. (In the voluntary condition, we simply omitted this note, on the assumption that, unless stated otherwise, participants would assume that the disclosure had been made voluntarily.) This notice was necessarily different depending on whether the manager had disclosed a weakness. Specifically, when the manager had disclosed a weakness, participants randomized to the required disclosure condition were further told that: “In the previous study, the individual was required to include some negative self-relevant information in the introduction.” Similarly, when the manager had not disclosed a weakness, participants randomized to the required disclosure condition were told that “In the previous study, the individual was required to include only positive self-relevant information in the introduction.”
**Measures.** We measured participants’ perceptions of the manager’s authenticity ($\alpha = .92$) as in Study 1. We measured participants’ willingness to work with the manager. Specifically, participants were asked “If you were looking for a job and were offered a job from this manager, how likely would you accept the job and work for the manager?” on a 7-point scale from 1 (*not at all*) to 7 (*very much*). The order of the two sets of questions was randomized.

**Results**

**Perceived authenticity.** A 2x2 ANOVA revealed a main effect of disclosing a weakness ($F(1, 388) = 17.43, p < .0001$): the manager was viewed as more authentic when he disclosed a weakness relative to when he did not ($M_{weakness} = 5.57, SD = 1.04; M_{no\_weakness} = 5.11, SD = 1.06$), $t(388) = 4.17, p < .0001$, Cohen’s $d = 0.44$. However, this main effect was qualified by an interaction ($F(1, 388) = 4.90, p = .027$) which suggested that the authenticity benefit of revealing a weakness was driven by managers who had done so voluntarily, as opposed to by requirement. Specifically, when managers voluntarily revealed a weakness, they were perceived as more authentic relative to when they only voluntarily revealed their strengths ($M_{weakness\_volunteered} = 5.75, SD = 0.96; M_{no\_weakness\_volunteered} = 5.06, SD = 1.06$), $t(193) = 4.52, p < .0001$, Cohen’s $d = 0.68$. However, when managers were required to reveal a weakness, this difference disappeared ($M_{weakness\_required} = 5.39, SD = 1.17; M_{no\_weakness\_required} = 5.18, SD = 1.02$), $t(195) = 1.39, p = .17$, Cohen’s $d = 0.18$.

**Willingness to work with the leader.** Mirroring the authenticity results, a 2x2 ANOVA revealed that disclosing a weakness only increased willingness to work with the manager when he did so voluntarily ($M_{weakness\_volunteered} = 5.85, SD = 1.04; M_{no\_weakness\_volunteered} = 5.51, SD = 1.11$), $t(193) = 2.19, p < .03$, Cohen’s $d = 0.32$, as opposed to by requirement ($M_{weakness\_required} = 5.59, SD = 1.00; M_{no\_weakness\_required} = 5.66, SD = 1.15$), $t(195) = -0.43, p = .66$, Cohen’s $d = 0.07$. 
Moderated mediation. A moderated mediation analysis with willingness to work with the leader (10,000 sample bootstrap analysis, Hayes, 2017, Model 7) indicated that the index of moderated mediation excluded zero ($b = .301, SE = .132, 95\% CI = [.051, .570]$), suggesting a significant indirect effect. Authenticity mediated the relationship between disclosure of weaknesses and willingness to work when disclosure was done voluntarily ($b = .434, SE = .098, 95\% CI = [.256, .641]$) but not when disclosure was by requirement ($b = .133, SE = .101, 95\% CI = [-.059, .338]$). Perceived authenticity explained 39.0% of the variance in willingness to work.

In sum, Study 5 suggests that for leaders to reap benefits from self-disclosing weaknesses, they must do so voluntarily.

Study 6: Status

In Study 6, we test the hypothesis that positive outcomes arising from self-disclosing weaknesses are restricted to high-status disclosers. The study was a 2x2 between-subjects design in which we manipulated the discloser’s status (high versus low) as well as the disclosure content (weakness disclosed vs. no weakness disclosed). Participants were randomized to one of the four conditions. We predicted an interaction, such that there would be heightened interest in working for a high-status, but not low-status, colleague who disclosed a weakness. And, as in Study 5, we also predicted that this effect would arise via gains in perceived authenticity—i.e., we predicted moderated mediation.

Method

Participants. Undergraduate students and community members ($N = 183; 97$ males; $M_{age} = 31.5$ years, $SD = 12.0$) came to a behavioral lab at a northeastern university and participated in this study for pay. This study was part of an hour-long series of studies for which participants received $20$ as compensation.
Materials and Procedure. Participants supposed that they were prospective employees engaged in a job interview with a company employee named Elis. We manipulated discloser status by randomizing participants to imagine that Elis was either the CEO (high-status condition) or a junior analyst (low-status condition). All participants supposd that Elis said:

The company, launched in 1988, initially focused on bonds. But thanks to shrewd acquisitions, the firm is now the world's largest asset manager, with $870 billion, offering a slew of equity funds and multi-asset funds. I like to climb mountains and collect American folk art in my spare time.

Participants were further randomized to one of two conditions: a condition in which Elis disclosed a weakness (weakness), or a condition in which he did not disclose a weakness (no weakness). In the weakness condition, the following sentence was appended to the above script: “Even though I need to fly several times a month, I am actually afraid of flying and I avoid flying.” In the no weakness condition, this sentence simply read: “I need to fly several times a month, and I enjoy flying.”

Pretest. We conducted a pretest to ascertain whether the disclosed weakness was perceived as similarly sensitive across the status manipulation. We randomly assigned participants (N = 391 U.S. MTurk workers; 172 males; M\(_{age}\) = 37.0 years, SD = 11.8) to one of the above four conditions in a 2 (weakness vs. no-weakness) x 2 (high vs. low status) between-subjects design. We told these pretest participants: “We are interested in your judgment of the sensitivity of information that Elis disclosed to you above. By ‘sensitive’ we mean information that is risky for Elis to disclose, in the sense of making him vulnerable to negative consequences arising from that disclosure.” We then asked: “How vulnerable, if at all, is Elis making himself in disclosing this information?” on a scale from 1 to 5 (1 = not at all; 5 = extremely vulnerable).
Results of a 2x2 ANOVA revealed only a main effect of disclosing a weakness ($F(1, 387) = 18.31, p < .0001$): disclosing a fear of flying made Elis more vulnerable relative to when he did not do so ($M_{Weakness} = 2.15, SD = 0.85; M_{No-Weakness} = 1.79, SD = 0.81$), $t(387) = 4.38, p < .0001$, Cohen’s $d = 0.44$. Importantly, there was no other main effect or interaction, suggesting that the weakness disclosure was seen as equally sensitive across status, hence, the disclosure manipulation was equally strong as a function of status.

**Measures.** Participants provided their perceptions of the discloser’s authenticity ($\alpha = .93$) as in Study 1, and willingness to work with the manager as in Study 5. The order of the two sets of questions was randomized.

**Results**

**Perceived authenticity.** A 2x2 ANOVA revealed only an interaction ($F(1, 179) = 4.59, p = .033$). Specifically, disclosing a weakness boosted perceived authenticity of the high-status discloser ($M_{Weakness_{HighStatus}} = 5.01, SD = 1.16; M_{NoWeakness_{HighStatus}} = 4.28, SD = 1.50$), $t(76) = 2.34, p = .02$, Cohen’s $d = 0.54$, but not of the low-status discloser ($M_{Weakness_{LowStatus}} = 4.48, SD = 1.46; M_{NoWeakness_{LowStatus}} = 4.63, SD = 1.34$), $t(103) = 1.34, p = .18$, Cohen’s $d = 0.11$.

**Willingness to work with the discloser.** A 2x2 ANOVA revealed an interaction ($F(1, 179) = 4.01, p = .047$): when the discloser was high status, disclosing a weakness enhanced participants’ interest in working with him ($M_{Weakness_{HighStatus}} = 5.35, SD = 1.57$; $M_{NoWeakness_{HighStatus}} = 4.61, SD = 2.02$), $t(76) = 1.99, p = .05$, Cohen’s $d = 0.41$ – this effect was not observed when the discloser was low-status ($M_{Weakness_{LowStatus}} = 4.87, SD = 1.36$; $M_{NoWeakness_{LowStatus}} = 5.12, SD = 1.66$), $t(103) = 0.77, p = .44$, Cohen’s $d = 0.16$.

**Moderated mediation.** We conducted a moderated-mediation analysis with willingness to work with the discloser as the dependent variable. A 10,000-sample bootstrap analysis (Model
showed that the index of moderated mediation excluded zero \((b = -.535, SE = .258, 95\% CI = [-1.099, -0.075])\), suggesting a significant indirect effect (Hayes, 2017). Authenticity mediated the relationship between disclosure of weaknesses and willingness to work for high status disclosers \((b = .443, SE = .195, 95\% CI = [.101, .871])\) but not for low status disclosers \((b = -.092, SE = .169, 95\% CI = [-.437, .228])\). Perceived authenticity explained 25.7% of the variance in willingness to work.

In sum, Study 6 suggests that positive outcomes arising from self-disclosing weaknesses are restricted to high-status disclosers.

**Section 4: Field Evidence**

In Section 4, we conclude by presenting anecdotal field evidence consistent with the causal effects identified in Sections 1-3. Specifically, using actual disclosures from a professional social networking app, we show that there is indeed a positive association between sensitive self-disclosure and subordinates’ reactions to those revelations. In addition, we used a broader definition of sensitive self-disclosure: whereas the studies so far have operationalized sensitive self-disclosure in terms of revealing a weakness, here, we code any disclosure judged as making the discloser vulnerable to negative judgment, as sensitive (c.f. Derlega et al., 1993). As such, this study speaks to the broader generalizability of our effects.

**Study 7. Leader’s Sensitive Self-Disclosure on a Social Networking Platform**

We obtained a large dataset from a social-networking platform that allows professionals to connect with other relevant professionals both within their company and across their industry. We assessed the positivity of reactions to senior leaders’ posts as a function of whether the post contained a sensitive self-disclosure. This platform presents an ideal context for our research, because in addition to housing common, non-revelatory, and even self-promoting posts as on
LinkedIn, self-disclosive posts are also prevalent.

We obtained a large anonymized dataset of posts and comments from senior leaders on this platform and the reactions (i.e., “likes” and comments) that these posts garnered. Using machine learning, we trained a classifier on a set of human-coded data and predicted labels for the remaining set of the data. In other words, we first had human coders code a small subset of posts and comments for the presence versus absence of self-disclosive content. Next, we trained a machine learning algorithm to categorize the remaining sample as either sensitive or non-sensitive disclosure. We then tested whether reactions to, and comments on, these posts and comments differed as a function of whether the content was self-disclosive. We hypothesized that self-disclosive content would garner more positive reactions and comments relative to less disclosive content.

**Method**

**Data.** Our dataset consists of posts and comments from senior leaders on a professional social-networking platform, as well as reactions to, and comments on, those posts. A post is the initiation of a new topic or thought. As on Facebook, fellow users can then respond to that post in two ways: by reacting and/or commenting. A reaction entails pressing a button to choose one of five possible reactions, all of which are positive: “like” (the default), “helpful,” “funny,” “uplifting,” or “smart.” A comment is a written response to the post (and, in turn, people can react to, and comment on, comments). Unlike Facebook, this social network is exclusive; at the time of data collection, it only granted access to those employed at certain firms within consulting and advertising. The platform only requires that users reveal their rank and/or employer. We requested the posts and comments from all senior leaders (operationalized as Director and above) within consulting firms, and the reactions and comments accompanying
these posts and comments. We focused on consulting firms because of its strict hierarchy (advertising firms tend to have flatter organizational structures).

Our dataset consisted of 1,484 posts, which, collectively, garnered 159,221 reactions and 33,589 comments. The vast majority (93%) of reactions were “likes”; the remaining 7% were distributed as follows: 6.1% “funny,” 0.98% “smart,” 0.73% “helpful,” and 0.73% “uplifting.” The number of reactions garnered by any given post or comment ranged from zero to 1,121; however, most (78%) garnered between zero and five reactions. Posts generally received more reactions than comments: the modal number of reactions was one for posts and zero for comments; the average number of reactions was nine for posts and four for comments.

Procedure. First, one of the authors worked iteratively with three research assistants to develop a binary coding scheme to code all (1,484) posts and a random selection of 2,000 comments. We conducted a sensitivity analysis using GPower (Faul et al., 2009) to find the low bound of effect size. The analysis, assuming two-tailed $\alpha = 0.05$ and 80% power, revealed a minimum effect size of $d = 0.05$ (this sensitivity analysis is a rough estimate only, as our analysis was non-parametric). Hereafter we refer to these posts and comments as “observations.” This quantity of observations was large enough to train a machine learning algorithm, yet reasonable for human coders to code manually. Observations containing sensitive self-disclosure were coded as 1; those not displaying such disclosure were coded as 0.

Per prior work, we defined sensitive self-disclosure as information that made the discloser vulnerable to being judged negatively by others (c.f. Derlega et al., 1993; Kelly & McKillop, 1996; Laurenceau et al., 1998; Moon, 2000). The human coding process went as follows: The team of four coders independently coded approximately ten randomly-selected observations, resolved disagreements by discussion, and used that discussion to build a refined
understanding of what, within this context, qualifies as sensitive self-disclosure. The team repeated this process twice for a total of approximately 30 randomly-selected observations. Next, two of the research assistants independently coded approximately 50 additional randomly-selected observations. Their agreement rate was 79.4%. Disagreements were resolved via discussion. One of the research assistants then coded the remaining ~3,400 observations.

Next, we used the 3,484 human-coded observations to train a classifier capable of predicting labels for the remaining 31,589 observations that had not been human-coded. To do so, we used BERT (Bidirectional Encoder Representations for Transformers, Devlin et al., 2018), the state-of-the-art deep-learning model in natural language processing which has recently received attention and been applied in different setting by researchers (Hartmann et al., 2021; Puranam et al., 2021). BERT learns contextual relations between words in text data. When used as a classifier, BERT adds a neural layer on top of the base model and predicts a label for a given input text.

We tested the predictive validity of the classifier by training it on a randomly-selected sample of 80% of the 3,484 human-coded observations and applying it to the 20% holdout sample. The classifier achieved 96% accuracy in this holdout sample—i.e., for 96% of observations, the classifier’s categorization agreed with that of the human-coder classification.

**Results**

Posts and comments categorized as self-disclosive garnered more (positive) reactions relative to non-self-disclosive ones \( (M_{	ext{disclosive}} = 6.98, SD = 16.15; M_{\text{non-disclosive}} = 4.44, SD = 14.04), t(1529.1) = -5.89, p < .001, \) Cohen’s \( d = 0.17; \) specifically, they garnered more likes \( (M_{	ext{disclosive}} = 6.41, SD = 14.38; M_{\text{non-disclosive}} = 4.06, SD = 12.06), t(1517.1) = -5.94, p < .001, \) Cohen’s \( d = 0.18. \) Other reactions, though used relatively infrequently, were also more prevalent
for self-disclosive observations; specifically, such observations were more likely to be deemed helpful ($M_{\text{disclosive}} = 0.09, SD = 0.51; M_{\text{non-disclosive}} = 0.04, SD = 0.30$), $t(1477.9) = -3.41, p < .001$, Cohen’s $d = 0.12$, uplifting ($M_{\text{disclosive}} = 0.10, SD = 0.68; M_{\text{non-disclosive}} = 0.03, SD = 0.29$), $t(1457.8) = -3.61, p < .001$, Cohen’s $d = 0.13$, and smart ($M_{\text{disclosive}} = 0.06, SD = 0.30; M_{\text{non-disclosive}} = 0.03, SD = 0.23$), $t(1494) = -3.4, p < .001$, Cohen’s $d = 0.11$. The prevalence of funny reactions was not different as a function of observation type ($M_{\text{disclosive}} = 0.32, SD = 1.90; M_{\text{non-disclosive}} = 0.27, SD = 2.87$), $t(1727.3) = -0.94, p = .35$, Cohen’s $d = 0.02$. These results hold when analyzing the two types of observations—posts versus comments—separately (see the supplement).

The above results are based on a binary coding scheme—each observation was coded as either containing, or not containing, a sensitive self-disclosure. In a supplementary analysis (see full details in the supplement), we asked two research assistants to code all 1,484 posts on a scale of 1 (not sensitive at all) to 5 (very much sensitive); their ratings were highly correlated ($r = .84$) and disagreements were not more than one score apart. We then used the average of the two coders’ scores for each post to train a BERT model to predict the sensitivity scores for the remaining data. There was a positive correlation between sensitivity score and positive reactions ($r = .08, p < .001$). These results held when analyzing the two types of observations—posts versus comments—separately (see the supplement).

In sum, these results are consistent with our basic hypothesis, revealing an association between leaders’ propensity to engage in sensitive self-disclosure, and the positive reactions that this activity appears to garner. That said, we acknowledge that positive reactions on social media may not always be indicative of positive reactions in face-to-face interactions.

**General Discussion**
Although authenticity in organizations has benefits, leaders, in particular, face barriers to being perceived as such (Hahl & Zuckerman, 2014). We show that leaders can increase perceptions of their authenticity by engaging in sensitive self-disclosure—and that this effect is mediated by dampened perceptions of strategic self-presentation. Moreover, the increased perceptions of authenticity arising from leaders’ sensitive self-disclosures translate into broader desirable outcomes (e.g., willingness to work for the leader). Also consistent with our conceptual account, we documented that the self-disclosure needs to be made voluntarily, and from a high-status individual (e.g., leaders and managers).

**Contribution to Theory and Practice**

We contribute to the self-disclosure literature. Whereas past research has emphasized the relationship between self-disclosure and liking (Collins & Miller, 1994), we focus on the role of self-disclosure in work relationships, specifically in the context of leader-follower relationships, and we demonstrate the effect of self-disclosure on perceptions of authenticity and subsequent outcomes while controlling for warmth and competence perceptions. Similar to Gibson et al. (2018), our work broadens the scope of self-disclosure from dyad relationships to organizationally-relevant settings. Complementing Gibson et al. (2018), our work suggests that perceived authenticity is a unique input to work-relevant interpersonal outcomes in addition to perceived status.

Our research contributes to the leadership literature by providing guidance on how leaders can enhance perceived authenticity. Given that leaders are prone to being perceived as inauthentic (Hahl & Zuckerman, 2014) and given the organizational benefits of having authentic leaders (Avolio et al., 2004; Norman et al., 2010), uncovering how, when, and whose self-
disclosure can boost perceived authenticity is important. We suggest that for leaders to realize the benefits of sensitive self-disclosure, the disclosure has to be voluntary in nature.

Finally, our research contributes to the self-presentation literature by uncovering one way to soften the "braggart" image that is associated with self-promotion, and focusing on the work-related relationships in organizations. Motivated by self-presentation concerns, actors seek to maximize their perceived competence by self-promoting; however, self-promotion can decrease liking without boosting perceived competence (Scopelliti et al., 2015). We suggest that by disclosing weaknesses, leaders may be able to come across as more authentic and generate more favorable outcomes without diminishing perceptions of their competence.

Limitations and Future Research Directions

One important future research direction would be to identify boundary conditions. Disclosure of serious weaknesses may not forge positive impressions, for the authenticity gains of doing so may be offset by negative impressions. Similarly, disclosure of morality-related weaknesses (e.g., I have committed a crime) might not forge positive impressions. Our findings imply that leaders can come across as authentic if they voluntarily self-disclose weaknesses. However, future research could test whether our effect holds when our research is used prescriptively—as in, when leaders are explicitly encouraged to reveal weaknesses for the express purpose of increasing perceived authenticity. In this case, it could be hard for leaders to "hide" that they are engaging in sensitive self-disclosure for the very purpose of strategic self-presentation. Relatedly, further research may further explore potential boundaries of the benefits of sensitive self-disclosure; for example, by manipulating the severity or nature of weaknesses, by manipulating perceived strategic self-presentation motives, or by measuring additional outcomes, including negative ones.
Future research could also assess whether self-disclosure of weaknesses works equally for different people. We found the effect to hold for both male and female disclosers (see Supplemental Study 1D) but future research may explore other dimensions, such as the age and cultural background of the discloser. We showed the effect to be restricted to high status disclosers. Although status (i.e., respect and admiration) and power (i.e., resource control) are distinct constructs, they often co-occur (e.g., Magee & Galinsky, 2008); that is, organizational leaders are often high in both status and power. Future work could therefore explore the “off-diagonals” (e.g., Blader et al., 2016; Fast et al., 2012)—i.e., what happens when a high-status but low-power individual makes sensitive self-disclosures relative to someone who is low in status but high in power.

Another future direction would be to examine how audience size and audience structure influence the observed effect. Does self-disclosing weaknesses to a large audience (e.g., broadcasting) lead to higher or lower perceptions of authenticity than self-disclosing to a small audience (e.g., one person, or narrowcasting)? Past research suggests that actors’ motives differ across audience size: actors tend to focus on the audience when talking to one or two people (narrowcasting) but on the self when talking to a large group (broadcasting, Barasch & Berger, 2014). However, from the recipients’ perspective, whether recipients perceive an actor who discloses to a small group (vs. large group) of people as more authentic is an empirical question for future research. Similarly, the audience structure might impact the observed effect. In Study 7 we focused on consulting firms because of their strict hierarchy, but the effects might weaken in flatter organization structures.

Finally, conceptually, authenticity may have some overlap with the construct of warmth. We demonstrated that our effect holds when controlling for perceived warmth and competence;
however, future research could aim to further distinguish warmth from authenticity. Such an investigation could further our understanding of the distinction between these two constructs and their differential effects on perception and behavior.

Conclusions

By making sensitive self-disclosures, leaders can enhance how authentic their followers perceive them to be, leading to positive interpersonal outcomes, and potentially organizational ones as well. In sum, we conclude that despite their apprehension to reveal weaknesses, leaders may reap surprising benefits from doing so.

Open Practices

Details of the stimuli and procedure used in these studies and data are available at https://www.dropbox.com/sh/v6ehroyynb52ke3/AAC6wjGml4NdRbzTL1x8ppEga?dl=0
References


https://doi.org/https://doi.org/10.1111/j.1467-6494.1982.tb00743.x


Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and


Gromet, D. M., & Pronin, E. (2009). What were you worried about? Actors’ concerns about
revealing fears and insecurities relative to observers’ reactions. Self and Identity, 8(4), 342–364.


https://doi.org/10.1177/00222437211037258


Heider, F. (1958). The psychology of Interpersonal Relations. Wiley.


https://doi.org/https://doi.org/10.1016/S0065-2601(06)38006-9


https://doi.org/10.1177/0963721417713296


http://www.wsj.com/articles/authenticity-in-the-2016-campaign-1442591796


Table 1
Overview of Study Designs, Materials, and Key Measures

<table>
<thead>
<tr>
<th>Study</th>
<th>Materials</th>
<th>Key dependent measures</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section 1: Effect of disclosure of a weakness on authenticity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 1 (pre-registered)</td>
<td>Not good at public speaking: “I am not good at public speaking. When I make a speech, my mouth gets dry and I sometimes start to panic.”</td>
<td>Authenticity</td>
<td>Prolific full-time employed</td>
</tr>
<tr>
<td>S1A</td>
<td>Speaking weakness: “I am nervous about public speaking and I have a habit of cracking my knuckles.”</td>
<td>Authenticity</td>
<td>Mturk</td>
</tr>
<tr>
<td>S1B (three conditions)</td>
<td>Speaking weakness: same as S1A Technological weaknesses: “I feel that as the company keeps growing, I feel a little under the water. The skills the company needs to succeed now are skills I do not seem to have. I am not able to keep track of the technological changes.”</td>
<td>Authenticity</td>
<td>Mturk</td>
</tr>
<tr>
<td>S1C (pre-registered)</td>
<td>Technological weakness: “as a manager I struggle with keeping track of technological changes” Control: “as a manager I keep track of technological changes”</td>
<td>Authenticity</td>
<td>Prolific full-time employed</td>
</tr>
<tr>
<td>S1D</td>
<td>Speaking weakness: same as S1A Gender of the discloser: male leader vs. female leader</td>
<td>Authenticity</td>
<td>Mturk</td>
</tr>
<tr>
<td>Study 2 (pre-registered)</td>
<td>Real leader’s video</td>
<td>Authenticity</td>
<td>Prolific full-time employed</td>
</tr>
<tr>
<td>Study 3 (pre-registered)</td>
<td>Technological weakness: same as S1B</td>
<td>Authenticity</td>
<td>Prolific full-time employed</td>
</tr>
<tr>
<td><strong>Section 2: The outcomes of disclosure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 4A</td>
<td>Dyad lab interaction study (with script): Speaking weakness as S1A</td>
<td>Authenticity</td>
<td>Students and community members from a university lab</td>
</tr>
<tr>
<td>Study 4B (pre-registered)</td>
<td>Dyad online chat study (naturalistic setting): Self-generated weaknesses</td>
<td>Authenticity</td>
<td>Prolific full-time and part-time employed</td>
</tr>
<tr>
<td><strong>Section 3: Moderators</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 5</td>
<td>Voluntary vs. requested disclosure: Speaking weakness as S1A</td>
<td>Authenticity</td>
<td>Mturk</td>
</tr>
<tr>
<td>Study 6</td>
<td>High vs. low status: Experimental: “Even though I need to fly several times a month, I am actually afraid of flying and I avoid flying.” Control: “I need to fly several times a month, and I enjoy flying.”</td>
<td>Authenticity</td>
<td>Students and community members from a university lab</td>
</tr>
<tr>
<td><strong>Section 4: Field Evidence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 7</td>
<td>Field study</td>
<td>Likes and engagement of the posts</td>
<td>NA</td>
</tr>
</tbody>
</table>
**Figure 1**

*Meta-analysis Results for Study 1, S1A-S1D*

<table>
<thead>
<tr>
<th>Study</th>
<th>Authenticity DV: All studies</th>
<th>SMD</th>
<th>95%-CI</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study I (not good at public speaking)</td>
<td>0.66 [0.42; 0.89]</td>
<td>0.66</td>
<td>[0.42; 0.89]</td>
<td>19.72%</td>
</tr>
<tr>
<td>S1A (Speaking weakness)</td>
<td>0.49 [0.20; 0.78]</td>
<td>0.49</td>
<td>[0.20; 0.78]</td>
<td>12.94%</td>
</tr>
<tr>
<td>S1B (Speaking weakness)</td>
<td>0.55 [0.26; 0.83]</td>
<td>0.55</td>
<td>[0.26; 0.83]</td>
<td>13.40%</td>
</tr>
<tr>
<td>S1C (Technological weakness)</td>
<td>0.65 [0.37; 0.94]</td>
<td>0.65</td>
<td>[0.37; 0.94]</td>
<td>13.32%</td>
</tr>
<tr>
<td>S1D (Speaking weakness; female)</td>
<td>0.54 [0.26; 0.82]</td>
<td>0.54</td>
<td>[0.26; 0.82]</td>
<td>13.47%</td>
</tr>
<tr>
<td>S1D (Speaking weakness; male)</td>
<td>0.51 [0.23; 0.80]</td>
<td>0.51</td>
<td>[0.23; 0.80]</td>
<td>13.37%</td>
</tr>
<tr>
<td>Overall</td>
<td>0.39 [0.11; 0.67]</td>
<td>0.39</td>
<td>[0.11; 0.67]</td>
<td>13.77%</td>
</tr>
</tbody>
</table>

**Overall** 0.55 [0.45; 0.65] 100.00%