Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Full length article

A fatal flaw: Positive leadership style research creates causal illusions

Thomas Fischer[a,*], Joerg Dietz[b], John Antonakis[b]

[a] University of Geneva Geneva School of Economics and Management (GSEM), Uni-Mail, Bd du Pont d'Arve, CH-1205 Geneva, Switzerland
[b] University of Lausanne, Faculty of Business and Economics (HEC), Department of Organizational Behavior, Internef CH-1015 Lausanne-Chamberonne, Switzerland

ABSTRACT

We argue and show empirically that constructs and measures of positive leadership styles, such as authentic, ethical, and servant leadership, are not veridical representations of leadership behaviors. Instead, these styles conflate behaviors with subjective evaluations of leaders. Labelling behaviors as, for example, "ethical" means evaluating leadership behaviors on positively valenced terms rather than describing these behaviors. Across four experiments, we show that positive leadership styles are outcomes that depend on non-behavioral, evaluative factors, such as information about a leader's previous success or value alignment between leaders and followers. More importantly, the measures of these leadership styles create causal illusions by spuriously predicting objective outcomes, even when leader behaviors and other leader-specific factors are kept constant. Furthermore, these measures have predictive properties similar to those of a purely evaluative measure of leadership. In conclusion, our studies cast serious doubts on previous research claiming that positive leadership styles cause positive outcomes. Moreover, positive leadership style research is not only wrong but also practically futile because its constructs and measures are amalgams that do not isolate concrete and learnable behaviors. We call for a radical reorientation of leadership style research and sketch out options for more solid future research.

Introduction

Valid constructs are the bedrock of social science. To serve as building blocks for explanatory theory, these constructs must be well defined, measurable, and causally linked to other constructs. However, many behavioral constructs do not meet these criteria. In a systematic review of leadership and organizational behavior research, Banks, Woznyj, and Mansfield (2021b) find that only 3 % of variables in these supposedly behavioral sciences capture the observable behaviors of individuals or groups. Instead, perceptions and evaluations, which refer to inner states, dominate. Ironically then, *behavioral scientists* rarely study *behaviors*, which is due to both improper conceptualization and incorrect measurement. As Ashford noted about leadership research: "[s]ome constructs out there […] are, to put it bluntly, a bit of a mess; and once they take root, we can't seem to get rid of them" (Ashford & Sitkin, 2019; p. 458).

This article demonstrates how a highly popular and allegedly behavioral stream of leadership research—positive leadership styles—has engendered such a "mess," conflating leadership behaviors and evaluations of these behaviors, and thereby leading to causal illusions. Through four studies, we offer empirical support for Fischer and Sitkin's (2023) claim that "the common finding that positive leadership styles lead to positive outcomes […] might be an artifact of conflation rather than a reflection of reality" (p. 1). In the present studies, we use parody to uncover an intractable but often-repeated error by reproducing the standard script of leadership style research (c.f. Hunter, Bedell-Avers, & Mumford, 2007) and then exposing its shortcomings. This script involves asking raters to evaluate a leader on a positive leadership style and then correlating the ratings with various outcomes, such as costly follower actions.

However, we add a critical twist to this script: we use an experimental design, which provides full control over information about the leader and variations in leader-level behavior. By doing so, we are able to demonstrate that positive leadership styles are conflated constructs that capture not only actual leader behaviors but also observers' subjective evaluations. We further demonstrate that the observer-level idiosyncrasies in these subjective evaluations produce causal illusions—that is, these idiosyncrasies predict objective leadership outcomes even when actual leader behavior does not vary or has been partialed out and when information about the leader is held constant. Stated differently, we observe leadership effects when there should be none. This fatal flaw has been and continues to be repeated with great regularity in empirical research and theory, and it ultimately leads to unwarranted advice for practice.

Setting the stage

In this section, after explaining our choice of authentic, ethical, and servant leadership styles as exemplars, we provide a preview of our work and reflect on its potential contributions for leadership research. We test

[*] Corresponding author.
E-mail address: thomas.fischer@unige.ch (T. Fischer).

https://doi.org/10.1016/j.leaqua.2023.101771
Received 24 December 2021; Received in revised form 6 December 2023; Accepted 27 December 2023

our propositions, including whether there are causal illusions, on authentic, ethical, and servant leadership, because these are particularly prominent positive leadership styles. The foundational articles about these styles, which are heavily cited, include Walumbwa, Avolio, Gardner, Wernsing, and Peterson (2008; on authentic leadership), Brown, Treviño, and Harrison (2005; on ethical leadership); and, on servant leadership, Liden et al. (2015) as well as Liden, Wayne, Zhao, and Henderson (2008). It is not surprising that Dinh et al. (2014) characterized these leadership styles as major streams of leadership research and that Lemoine, Hartnell, and Leroy (2019) classified them as the major representatives of moral forms of leadership.

In examining the roots of the causal illusions, we show that leadership style measures capture more than a leadership style and its associated behaviors. In addition to descriptions of behaviors, leadership styles contain subjective evaluations of these behaviors (e.g., about underlying intentions and the quality and effects of behaviors, Fischer & Sitkin, 2023). For instance, the servant-leadership dimension "behaving ethically" (Liden et al., 2008) is an attempt to describe leader behaviors but also requires an evaluation because classifying a behavior as ethical is a judgment call. Indeed, we observe that evaluation-relevant information such as a leader's previous success, previous normative achievements, and value alignment with the rater systematically affect positive leadership styles. In line with Fischer and Sitkin (2023), we regard description-evaluation conflation as a source of construct overlap that sits alongside previously identified behavioral overlaps (e.g., Banks, Gooty, Ross, Williams, & Harrington, 2018; Hoch, Bommer, Dulebohn, & Wu, 2018).

Furthermore, we elaborate on how description-evaluation conflation coincides with cause-effect conflation. Positive leadership styles also conflate causes (i.e., leader behaviors) with outcomes (i.e., observers' evaluations) of these behaviors. The cause-effect conflation is particularly evident when recognizing that behaviors are at the level of a leader whereas evaluations are at the level of the rater. Thus, leadership styles inadvertently merge leaders' exhibiting behaviors with raters' making judgement calls about these behaviors. Evaluators must judge whether, for example, a leader is believed to have an internalized moral perspective (Walumbwa et al., 2008), can be trusted (Brown et al., 2005), or is a person from whom one would seek personal help (Liden et al., 2015).

The tripartite conflation of description and evaluation, cause and effect, and leader-level and rater-level conceptualization and measurement implies that positive leadership styles cannot be unitary and meaningful constructs. Rather, this conflation makes them flawed constructs that obscure causality. The resulting causal illusions undermine the validity of past research in a way that goes beyond established concerns such as limited discriminant validity (Banks et al., 2018; Hoch et al., 2018). When the underlying causal mechanism is unclear, interpreting links between positive leadership styles and leadership outcomes as behavior-outcome relationships (c.f. Chiniara & Bentein, 2016; Hmieleski, Cole, & Baron, 2012; Lemoine et al., 2019; Ng & Feldman, 2015) is misplaced.

Taken together, these arguments have profound implications by questioning seemingly well-established "truths" about positive leadership styles. We contradict past research and challenge conventional wisdom by showing that these styles are not clean behavioral constructs and that empirical evidence suggesting that these styles predict leadership outcomes is causally elusive. In other words, our research elucidates that much past research has theorized A, namely specific causal effects of positive leader behaviors (see, e.g., Lemoine et al., 2019 for a review), while in fact testing B, that is, associations between (i) ambiguous constructs that conflate leadership behaviors and evaluations and (ii) leadership outcomes (Kerr, 1975; Schriesheim, Castro, Zhou, & Yammarino, 2001). Consequently, although the data, which researchers have accumulated and alleged to be evidence for the convergent, discriminant, and predictive validity of these styles, are real, these data are misinterpreted. These data do *not* validate effects of leader behaviors. Rather, they "validate" a construct that conflates leader behaviors and evaluations of those behaviors, and such research neither properly informs leadership practice nor theory building.

These consequences are as far-reaching as those of the arguments, set out in past scathing critiques, that research on positive leadership styles implies a feel-good world in which good deeds lead to good outcomes, when the reality of leadership is also hard-nosed and power based (Alvesson, 2020; Alvesson & Einola, 2019; Pfeffer, 2015). This disconnection from reality is not our primary concern, however. Rather, we argue, in line with Fischer and Sitkin (2023), that current leadership style research can neither validate nor invalidate popular claims about leaders and their good deeds, even if positive leadership had a place in the real world.

Moreover, and at the heart of our contribution, we conducted four experiments, in which we put our arguments about the shortcomings of positive leadership style research to the test and thus go beyond purely conceptual critiques. We empirically and rigorously tested our claims that conflated constructs create causal illusions by isolating key effects, fully controlling the information environment, and ruling out alternative explanations. From this empirical contribution follows a practical contribution. Support for our arguments regarding construct conflation and causal illusion would cast stark doubts on the utility of positive leadership style research for informing or even guiding practice. Lastly, as we discuss in the concluding section, our research can be an eye opener for revitalizing leadership style research towards increased conceptual and methodological rigor.

Positive leadership styles as conflated constructs

Scholars have previously pointed to weaknesses in conceptualizing leadership styles. For instance, Yukl (1999) notes that both constructs and measures of charismatic and transactional-transformational leadership are highly ambiguous and that it is unclear which concrete behaviors belong to these styles of leadership. Van Knippenberg and Sitkin (2013), as with earlier critiques (Antonakis, Fenley, & Liechti, 2011), add that these leadership styles confound behaviors with their effects. Regarding ethical leadership, Banks et al. (2021a) lament conceptual imprecision. Furthermore, Alvesson and Einola (2019; see also Pfeffer, 2015) call authentic leadership an "excessively positive" hotchpotch that does not reflect organizational reality.

We build on and go beyond these previous critiques, just as Fischer and Sitkin (2023) do in their identification of conceptual conflation and causal indeterminacy in ten leadership styles. The starting point for our critique is the observation that leadership styles are defined as *patterns* (or characteristic modes or manners) of behaviors. For example, according to the Merriam-Webster Dictionary (2023a), a *style* is a "distinctive manner or custom of behaving or conducting oneself." Academic definitions are consistent with the dictionary use of the term *style*. For example, Bass and Bass's leadership handbook (2008) describes "leadership styles" as "alternative ways that leaders […] pattern their interactive behaviors with those they influence" (see also Appendix 1 for key definitions of authentic, ethical, and servant leadership).

However, a pattern of behaviors is not identical to behaviors per se, but rather a characterization of a set of behaviors as conveying a common theme. Inevitably, such a characterization is judgmental and carries an evaluative component. However, behaviors and evaluations are different concepts. *Behavior* means, for instance, "anything that an organism does involving action and response to stimulation" (Merriam-Webster, 2023b) or "[t]he internally coordinated responses of whole living organisms (individuals or groups) to internal or external stimuli, excluding responses more easily understood as developmental changes" (Levitis, Lidicker Jr, & Freund, 2009; p. 103). By contrast, *evaluation* means the "determination of the value, nature, character, or quality of something or someone" (Merriam-Webster, 2023c) or "the imputation of some degree of goodness or badness to an entity" (Eagly & Chaiken, 1993; p. 3). Behaviors are objective, whereas evaluations are subjective.

In the context of our studies, behaviors are leader-level constructs, and evaluations are observer-level constructs. Leadership style research, however, mostly overlooks the fact that the notion of a pattern of behavior conflates actual behaviors with evaluations of these behaviors. For instance, Hunter et al., (2007: 438) note that in the typical leadership study "it is assumed that such instruments [i.e., measures of leadership styles] capture the most critical and essential leadership behaviors.".

This behavioral notion is also evident in research on authentic, ethical, and servant leadership styles. Despite the potential problems we have noted, authors such as Lemoine et al., (2019: 177) claim that a "comparative review of the three dominant moral approaches [i.e., ethical, authentic, and servant leadership] clearly indicates that moral leadership *behaviors* positively impact a host of desirable organizationally relevant outcomes" (italics added). There are further examples of a behavioral view of authentic leadership (e.g., Banks, McCauley, Gardner, & Guler, 2016; Gardner, Avolio, Luthans, May, & Walumbwa, 2005; Gardner, Cogliser, Davis, & Dickens, 2011; Hmieleski et al., 2012), ethical leadership (e.g., Den Hartog, 2015; Kalshoven, Den Hartog, & De Hoogh, 2011; Mayer, Aquino, Greenbaum, & Kuenzi, 2012; Ng & Feldman, 2015; Walumbwa, Morrison, & Christensen, 2012), and servant leadership (e.g., Chiniara & Bentein, 2016, 2018; Ehrhart, 2004; Hu & Liden, 2011; Hunter et al., 2013; Lemoine & Blum, 2021). The behavioral view of leadership styles also dominates the literature on training for positive leadership styles (e.g., Avolio, Griffith, Wernsing, & Walumbwa, 2010; Cooper, Scandura, & Schriesheim, 2005). Apparently, leaders can learn authentic, ethical, and servant-like behaviors, and enacting these behaviors is supposed to have positive effects on outcomes. But if, as we say, the behavioral view of positive leadership styles is false due to the conflation of behavior and evaluation, then behavioral training for positive leadership styles no longer rests on a solid scientific basis.

The proclivity for the behavioral view in past research on positive leadership styles runs deep. And this penchant is also profoundly perturbing because styles represent evaluated patterns of behaviors, not concrete—i.e., objective—behaviors or an average of these behaviors. Whereas the descriptions of behaviors can be made according to an objective referent (e.g., whether a behavior was displayed or not), evaluations of leadership require a subjective referent (e.g., whether a behavior is "good" in the sense of being authentic, ethical, or servant-like). This evaluative requirement of leadership styles is a fundamental conceptual shortcoming that goes beyond perceptual biases such as stereotyping or halo effects (Fischer, 2023).

To provide an interim summary, the problem that we raise is not one of distorted perceptions of leadership styles, but of leadership styles as concepts that conflate behaviors and evaluations (Fischer, 2023; Fischer & Sitkin, 2023). Even undistorted and perfectly accurate perceptions of leadership styles would not solve this problem. Having made the point that leadership styles have an evaluative component, we now further elaborate on three sources of this evaluative component: first, the positive valence of positive leadership styles; second, nonbehavioral leader features that shape leadership style judgments above and beyond leader behaviors; and third, evaluator attributes that predispose these evaluators towards specific leadership style judgements.

1. Positive valence: Saying that a person leads authentically, ethically, or in a servant-like way carries an extremely positive valence (Alvesson, 2020; Antonakis, 2017; Fischer & Sitkin, 2023). It is only fitting that scholars have categorized authentic, ethical, and servant leadership as moral leadership styles (e.g., Lemoine et al., 2019). An assessment of these forms of leadership requires a *judgment* about their positivity—authentic, ethical, or servant-like—which is by definition evaluative (Fischer, 2023). These evaluations inevitably sully seemingly behavioral measures.

2. Nonbehavioral leader features: Furthermore, these evaluations rest not only on the observed behaviors but also on other observable leader-specific factors (Banks et al., 2021b). For instance, evidence shows that a leader's facial appearance affects observers' evaluations of a leader (e.g., Todorov, Mandisodza, Goren, & Hall, 2005; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015), which has important consequences for leader outcomes (Antonakis & Eubanks, 2017). Similarly, knowledge about a leader's previous performance strongly shapes leadership style evaluations (Gioia & Sims, 1985; Lord, Binning, Rush, & Thomas, 1978). Ultimately, evaluations of leader behaviors are not merely judgments about the observed behaviors. Instead, these evaluations are rather holistic assessments of the leader as a person. Moreover, these judgments are embedded in the observer's larger knowledge base and entail inferences congruent with the general impression about the leader (Hansbrough, Lord, & Schyns, 2015). Because such variables (e.g., leader characteristics and their past performance) affect the evaluation of the style and, potentially, the outcome predicted, they become omitted variables that create an intractable endogeneity problem in estimation (Shaver, 2020).

3. *Evaluator attributes*: This point presents an even bigger problem. Evaluator-specific factors influence judgments about leaders too; affect (Martinko et al., 2018) as well as gender and personality (Wang, Van Iddekinge, Zhang, & Bishoff, 2019) are well-established exemplars. Moreover, evaluators differ in their understanding of authenticity (Lehman, O'Connor, Kovacs, & Newman, 2018) or ethics (Banks et al., 2021a). Therefore, even if two evaluators observe the same leader behaviors and characteristics, they might arrive at different judgments about whether a leader's style is authentic or ethical.

The impact of both nonbehavioral leader features and evaluator attributes on leadership style judgments is not only troublesome because it renders these judgements evaluative and introduces unobserved variability in these judgments (e.g., due to leaders' facial attractiveness, knowledge of performance outcomes, or followers' personality). Equally importantly, this unobserved variability may, in turn, explain variance in outcomes, whether these outcomes are objectively or subjectively measured, creating serious endogeneity bias. Even in light of only this latter argument, there is an obvious conceptual problem with leadership styles that translates into endogenous estimates due to omitted variables in statistical models. However, the problem is evidently much larger than endogeneity bias. The crux of our argument, which we translate into concrete propositions in the next section, is that the conceptualization of leadership styles encompasses a positive evaluation that will correlate positively with other positive outcomes *without causing them.*

Deriving propositions

On the basis of our preceding line of reasoning, we derive three general propositions that serve as a basis for testable arguments in our empirical studies. First, leadership styles, as behavior-evaluation conflations, are caused only partially by behaviors and, more importantly, are also caused by numerous other factors. Second, leadership styles predict outcomes even when leader behaviors do not vary (i.e., causal illusions). Third, leadership styles function similarly to an entirely evaluative measure of leadership.

Leadership style ratings as outcomes of leadership

Leadership style ratings reproduce the conceptual conflation of leadership style constructs. The evaluative component of these styles has many antecedents beyond the leader's behavior. These previously mentioned antecedents include nonbehavioral leader properties and follower characteristics. Importantly, any antecedent of evaluations—whether located in the evaluator, in the leader's nonbehavioral properties, or in the context—can produce variation in ratings of leadership styles.

Two examples are alignment in values between the follower and the leader, and observers' knowledge of previous outcomes attributed to the leader. We expect that an alignment in values between the follower and the leader explains whether followers view leaders as having a more authentic, ethical, or servant-like leadership style. The values that followers hold shape their evaluations of a leader's behaviors, with shared values leading to more favorable evaluations than do diverging values. Regarding knowledge of previous leadership outcomes, as Yukl (2008) indicates, leaders are supposed to deliver results, and on that basis, followers are likely to formulate more positive evaluations of leaders who have delivered results in the past (Lord et al., 1978). On the basis of the above arguments, we formulate the following proposition:

Proposition 1. *Measures of positive leadership styles capture evaluations of leader behaviors. That is, non-leader-behavioral antecedents such as follower properties or contextual factors predict scores on these measures.*

Illusory causation by positive leadership styles

Proposition 1 implies that leadership style ratings are at best ambiguous indicators of leader behaviors. Hence, typical measures such as the Authentic Leadership Questinnaire (ALQ; Walumbwa et al., 2008), the Ethical Leadership Scale (ELS; Brown et al., 2005), and the Servant Leadership Questionnaire (SL-7; Liden et al., 2015), cannot be clean indicators of leader behaviors and capture multiple confounds, rendering these measures opaque and uninterpretable in terms of variation on leader behaviors.

As proximal outcomes of leader behaviors and other factors, positive leadership styles should be related with other outcomes of leadership. What is more, the relationship of leadership styles to other outcomes should exist even if there is no variation in behaviors due to the nonbehavioral antecedents of these styles. Evaluating a leader's style as authentic, ethical, or servant-like carries a positive connotation and should relate positively to other positively valenced leadership outcomes, such as leadership effectiveness. Just as a person judged to be friendly for whatever reason (e.g., has a physical appearance that makes them seem friendlier, Todorov, 2017) would be assumed to have more friends, a leader judged to have an ethical style (e.g., because of positive performance signals or physical appearance) would be assumed to be more trustworthy (compare with Brown et al., 2005). In summary, we argue that relationships between positive leadership styles and outcomes can exist independently of variation in leader behaviors.

Proposition 2. *Measures of positive leadership styles illusorily cause outcomes of interest. That is, these measures predict leadership outcomes, even if there is no variation in leader behaviors.*

Leadership style ratings function as purely evaluative measures of leadership

Measures of positive leadership styles—among them the ALQ, ELS, and SL-7—contain an amalgam of items; some are relatively descriptive of behaviors, whereas others are more evaluative (Fischer, 2023). An example of a relatively descriptive item from the ELS is "When making decisions, asks 'what is the right thing to do?'" An example of an evaluative item from the ELS is "can be trusted" (Brown et al., 2005; p. 125).

If measures of positive leadership styles are not only behavioral but also evaluative (Fischer & Sitkin, 2023), they should share empirical properties with purely evaluative measures of leadership. We constructed such a measure, which we call the Evaluative Questionnaire (EvalQ). The exercise of creating the EvalQ and correlating it with outcomes is intentionally parodic and might appear ludicrous at first glance, given its five items: the leader "is an interesting human being," "is a special individual," "is distinctive," "has a unique character," and "is a real leader." These items have zero behavioral content, and there is no clear mapping between leader behaviors and follower evaluations (e.g., whether a leader is an interesting human being). Thus, even if leaders' behaviors affect responses on the EvalQ, the EvalQ is not a behavioral measure, but an evaluative one. Thus, the EvalQ allowed us to compare empirical properties of a purely evaluative measures with those of the measures of authentic, ethical, and servant leadership that are seemingly behavioral. If the EvalQ and the three leadership style measures function in alike fashion, then there is solid evidence for a sizeable evaluative component in the three leadership style measures. Such evidence would be damning to past claims that leadership style measures are representations of leader behaviors.

Proposition 3. *Measures of positive leadership styles have empirical properties that are similar to those of a positively evaluative, nonbehavioral measure of leadership.*

In Fig. 1, we illustrate nonbehavioral links between leadership styles, antecedents, and outcomes (Proposition 1), which are associated with illusory causation and misleading predictions (Proposition 2). Fig. 1 also points to the link between leadership style ratings and a purely evaluative measure of leadership (Proposition 3).

Empirical strategy

Across four experiments, we reproduce under laboratory conditions the typical structure of research on authentic, ethical, and servant leadership styles (Hunter et al., 2007; Lemoine et al., 2019). This structure is as follows:

1. Observers are given information, via various modes, on a target leader;
2. Observers are asked to rate the target leader on leadership style measures; and
3. Variability in these leadership style ratings is used to predict an objective and costly outcome.

In this section, we first provide a brief overview of our four studies. Then, we explain our methodological and statistical approaches to testing the three propositions. Next, we elaborate four additional features that characterize our experiments: (a) differential degrees of controlling information about the leader, (b) objective outcome measures, (c) ruling out alternative explanations, and (d) replication across contexts.

Overview of the four studies

In Study 1 participants watched one of two versions of a video that had quasi-identical information content (but varied on rhetorical tactics); in Study 2 they just saw one version of the video (i.e., the behavior is constant). All the videos used in Studies 1 and 2 show a leader motivating workers to work hard for a charity. Participants subsequently assessed the leadership style of this leader. Then, they decided whether to donate money to the charity advocated by the leader or to keep the money for themselves. In Study 3, participants read an inaugural address given by a U.S. president (either Bill Clinton, George W. Bush, or an unnamed previous president) to citizens. Participants rated the president's leadership and decided whether to donate to the president's charity (Study 3) or to keep the money for themselves. Finally, in Study 4, participants watched one of two versions of a leader's speech motivating them to undertake a real-effort task, the more of which they completed, the greater the payoff they received. The two versions were identical content-wise but differed in their use of signals about completing the task ethically. Participants then worked on this task. At the end, they self-declared how much of the task they had completed, meaning participants had the opportunity to cheat to increase their financial compensation. The data and replication material can be found here: https://osf.io/hjbqt/?view_only=3892c09a9a224fd2967be10f79be067f.

Design for proposition testing

Design for Proposition 1: Construct-irrelevant prediction of positive leadership style ratings. We designed our studies to shed light on the conflation of behaviors and evaluations in positive leadership style measures, allowing us to test Proposition 1. Study 1 contained two experimental manipulations: information about a leader's previous success (high or low) and the degree of use of rhetorical tactics (high or low) indicative of charisma. If measures of authentic, ethical, and servant leadership styles only capture corresponding leadership behaviors, neither manipulation should affect the measures. An effect would indicate that these measures pick up variance that could not stem from leadership styles as behavioral constructs.

More specifically, the previous-success manipulation is nonbehavioral but relevant for evaluating leaders. Thus, this manipulation allowed us to test whether evaluative and nonbehavioral factors systematically influence presumably behavioral leadership style ratings. The rhetorical-tactics manipulation is behavioral but conceptually unrelated to the measured leadership styles. Thus, this manipulation allowed us to test whether construct-unrelated behavioral factors systematically influence leadership style ratings.

Because the use of charismatic tactics did not affect positive leadership style ratings, we discontinued using a behavioral yet construct-unrelated manipulation in the subsequent studies. In Study 2, we experimentally manipulated information about a leader's ethical achievements (high or low). Like the previous-success manipulation in Study 1, the ethical-achievement manipulation enabled us to test whether nonbehavioral yet evaluation-relevant variation affects ratings of leadership styles.

Study 3 was a constructive replication of Studies 1 and 2 in a political setting. In Study 3, we manipulated the identity of a U.S. president and measured raters' political preferences to study the impact of political value alignment on leadership style ratings.[1] Akin to the previous-success and ethical-achievement manipulations in the two previous studies, political value alignment is a nonbehavioral yet evaluation-relevant factor. We expected value alignment between a rater and a president to lead to higher leadership style ratings, which would further undermine the traditional view of positive leadership ratings as behavioral measures.

Lastly, in Study 4, we manipulated the extent to which a leader used behaviors that signal ethical leadership. Hence, in contrast to the preceding studies, this study manipulates construct-relevant behaviors. It is obvious that behaviors signaling ethical leadership are construct-relevant behaviors for ethical leadership. Moreover, to the extent that authentic and servant leadership overlap with ethical leadership (Banks et al., 2018; Hoch et al., 2018), this manipulation should influence construct-relevant behaviors for these styles too. We statistically isolated variance caused by this manipulation from variance caused by raters' evaluative idiosyncrasies. Predicting leadership style ratings from these evaluative idiosyncrasies would additionally weaken the claim that positive leadership styles are behavioral constructs.

Design for Proposition 2: Demonstrating causal illusions through experimental control of leadership and its associated behaviors. Our experimental design also allowed us to test our proposition that measures of leadership styles predict objective outcomes even if there is no variation in leadership behaviors. If leadership style ratings predict outcomes in the absence of behavioral variation and any other leader-level variation, these ratings must be systematically driven by nonleadership-related, rater-level evaluative factors.

Design for Proposition 3: Comparing positive leadership style measures to a purely evaluative measure of leadership. We used our purely evaluative measure of leadership (i.e., the EvalQ) in all four studies in addition to the positive leadership style measures. Subsequently, we compared the results obtained from testing Propositions 1 and 2 via the EvalQ with those obtained via the positive leadership style measures to check for similar empirical properties. Finding that the results from testing Propositions 1 and 2 via the EvalQ reproduced the pattern of significance and directionality of the results obtained via the positive leadership style measures would indicate that the EvalQ and the leadership style measures share empirical properties.

Statistical Proposition Testing

Testing Proposition 1. We needed to assess the effect of construct-irrelevant manipulations, such as information about past success, on leadership style ratings. We did so through OLS regression, in which manipulations were modeled as predictors of the criterion of positive leadership style. In doing so, in Studies 1 to 3, we conducted separate analyses for each leadership style. Study 4 did not contain a construct-irrelevant manipulation.

Testing Proposition 2. We needed to examine whether leadership style ratings predicted an outcome (making donations in Studies 1 to 3, and lying in Study 4), even when there was no behavioral variation at all (Study 2) or when variance stemming from the manipulations was partialed out (Studies 1, 3, and 4). Across studies, these manipulations were evaluation-relevant yet construct-irrelevant (e.g., information about previous success), except for the manipulation of ethical signaling in Study 4. Our logic for isolating this leader-related variance was that the remaining systematic variance would have to originate from rater-level idiosyncrasies. To isolate manipulation-related variance, we conducted within-condition analyses for evaluation-relevant and nonbehavioral manipulations (previous success in Study 1, normative achievements in Study 2, and value alignment in Study 3), and we used residualization for behavioral manipulations (i.e., construct-unrelated rhetorical tactics in Study 1, marginally different speeches in Study 3, and ethical leadership signaling in Study 4). Within-condition analysis ensured that information about the leader was kept constant, whereas residualization served the same purpose for the leader's behaviors. We next explain within-condition analysis and then residualization in more detail.

Within-condition analysis. We study the relationship between leadership styles and leadership outcomes within experimental conditions that contained variation in information about the leader, and we report results for the experimental conditions both separately and jointly. Within the experimental conditions, participants were exposed to perfectly identical leader characteristics and information about the leader. Under such conditions, if leadership styles predict leadership outcomes, the effects have to be causal illusions in terms of information about the leader, because leadership-relevant factors are kept identical. We test the effect of the manipulations by testing whether the dependent variable is significantly different across experimental conditions (see Online Appendix 1 for further details).

Residualization. To isolate the variance caused by behavioral manipulations (Study 4, and to minor degrees Studies 1 and 3), we used residualization as an estimation technique, which allowed us to separate variance explained by leader behaviors from variance that is orthogonal to the behavioral manipulations in the rated leader behavior. The use of such an estimation technique was not necessary in Study 2 because the methodological design's use of only one video held leader behaviors and other leader properties perfectly constant. For Study 4, by contrast, residualization is necessary because we experimentally manipulated conceptually relevant behaviors. For Studies 1 and 3, residualization and OLS estimation give similar results because the small degrees of behavioral variation were deliberately conceptually irrelevant and proved to be empirically irrelevant too. For consistency, we report the results of Studies 1 and 3 with residualization (results with OLS estimation can be found in Online Appendices 8 and 9).

When using the residualization technique, the variance orthogonal to the behavioral manipulations, which is exogenous, is captured by the residuals. In the analyses, we only use these residuals to predict

[1] In Study 3, we also used two versions of inaugural speeches to rule out the possibility that results would hold only for a particular speech. The two versions which were excerpts from the first actual inaugural speeches of Bill Clinton and George W. Bush were highly similar to each other in terms of rhetorical content.

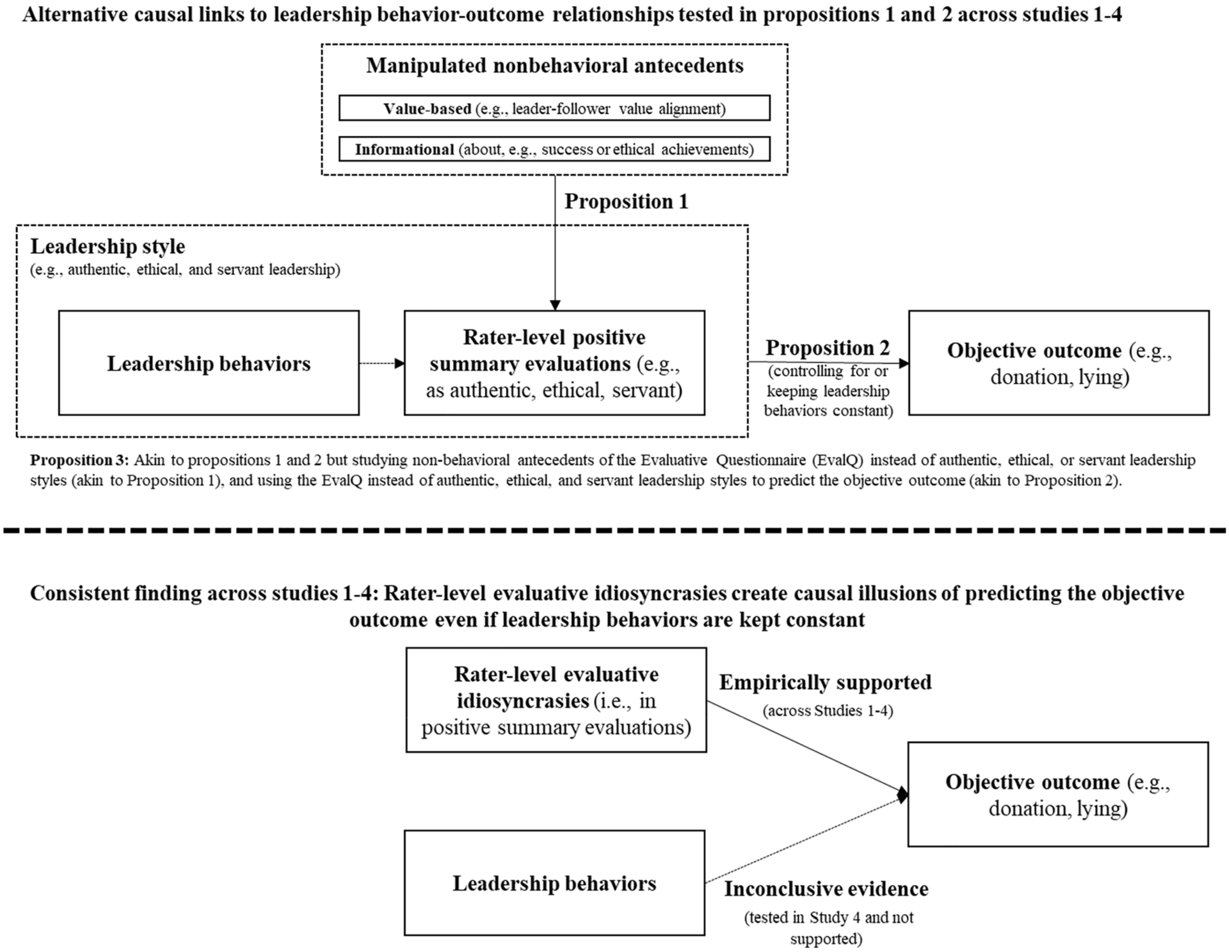
Alternative causal links to leadership behavior-outcome relationships tested in propositions 1 and 2 across studies 1-4
Manipulated nonbehavioral antecedents
Value-based (e.g., leader-follower value alignment)
Informational (about, e.g., success or ethical achievements)
Proposition 1
Leadership style
(e.g., authentic, ethical, and servant leadership)
Leadership behaviors
Rater-level positive summary evaluations (e.g., as authentic, ethical, servant)
Proposition 2
(controlling for or keeping leadership behaviors constant)
Objective outcome (e.g., donation, lying)
Proposition 3: Akin to propositions 1 and 2 but studying non-behavioral antecedents of the Evaluative Questionnaire (EvalQ) instead of authentic, ethical, or servant leadership styles (akin to Proposition 1), and using the EvalQ instead of authentic, ethical, and servant leadership styles to predict the objective outcome (akin to Proposition 2).
Consistent finding across studies 1-4: Rater-level evaluative idiosyncrasies create causal illusions of predicting the objective outcome even if leadership behaviors are kept constant
Rater-level evaluative idiosyncrasies (i.e., in positive summary evaluations)
Empirically supported
(across Studies 1-4)
Objective outcome (e.g., donation, lying)
Leadership behaviors
Inconclusive evidence
(tested in Study 4 and not supported)

Fig. 1. Our Empirical Model and the Origins of Illusory Causation.

outcomes, which owing to the study design cover all rater-level evaluative idiosyncrasies and no leader-level and behavioral variation. If the residuals of these leadership style ratings predict the outcome of interest, this relationship can evidently only be a causal illusion of an effect caused by leader behaviors. Stated differently, the relationship could falsely be taken as resulting from a leader's behavioral style, when in fact it stems from nonbehavioral sources. The relationship cannot stem from systematic informational distortions either, because within-condition analyses keep such information constant. Thus, systematic causal illusions have to be due to rater-level evaluative idiosyncrasies.

To this end, we used insights from the Frisch-Waugh-Lovell theorem (see Davidson & MacKinnon, 1993; pp. 19-24). According to this theorem, the residual x_{resid} of x (from regressing x on z, and assuming an exogenous, randomized, z) contains variation in x that is orthogonal to z and includes all causes in x not due to z. Thus, in our case, the residual of the OLS-estimation with a single leadership style rating ($lead_{resid}$) as the dependent variable (x), and a leader-behavioral manipulation (LBM) as independent variable (z), is exactly the part of the leadership style measure that is orthogonal to and hence cannot be explained by the manipulation. In our case, this insight means that we can partition behavioral variation (i.e., manipulation) and idiosyncratic evaluative variation (i.e., $lead_{resid}$) by undertaking the following steps. First, we individually model the four leadership styles, k (i.e., authentic, ethical, servant, and in Study 4, also transformational leadership), as a function of the manipulation:

$$lead_k = \delta_{k0} + \delta_{k1} LBM_k + u_k \tag{1}$$

whereby u_k refers to variance that is not explained by the manipulation. Then we obtain the residuals ($lead_{k.resid}$) by subtracting the predicted value for each style ($\widehat{lead_k}$) from the observed value for each style ($lead_{k.observed}$) for each observation:

$$lead_{k.resid} = lead_{k.observed} - \widehat{lead_k} \tag{2}$$

Then we estimate (using Study 4 as an example, where lying was the objective outcome)

$$lying_k = \beta_{k0} + \beta_{k1} lead_{k.resid} + \beta_{k2} LBM_k + Controls_k + e_k \tag{3}$$

whereby e_k refers to the error term. By specifying the regression model in Eq. (3), we test Proposition 2 ("causal illusion"). This model is correctly specified given that the residual of the leadership rating is orthogonal to the manipulation (LBM). If we find β_{k1} to be a significant predictor of the outcome variable (donating in Studies 1 to 3 and lying in Study 4), we can deduce that this variation is wholly rater-idiosyncratic and as such evaluative. Furthermore, in the case of a small or even insignificant effect of LBM on $lead_k$, there is particularly strong evidence for our proposition, because then $lead_k$(the leadership style rating) has to be driven by idiosyncratic evaluative variation, which correlates with

y (lying). We further explain this procedure in Online Appendix 10 and show analytically and with simulated data why this method is ideal for testing Proposition 2.[2]

Testing Proposition 3. The statistical tests of Proposition 3 were analogous to those reported above for Propositions 1 and 2, except for our replacing the positive leadership style measures with the evaluative leadership questionnaire (EvalQ). The EvalQ could be classified as similar to the other leadership style measures only if the regression coefficient has the same directionality and if the *t*-test is equally (in)significant.

Different degrees of controlling information about the leader

In Studies 1, 2 and 4, we had perfect control over the information that participants were given about the leader, with participants obtaining this information only through the study materials. The advantage of this approach is that we could rule out that variation in terms of information about the leader (e.g., due to prior interactions or preexisting information) might explain results. However, a disadvantage is that participants have less information about the leader than they would in natural settings. Stated more generally, this low-information context in Studies 1, 2, and 4 enhances internal validity but potentially reduces external validity.

Therefore, we used former presidents as leaders in Study 3 so that participants would have more information—and more heterogeneous information—about the leader. Both former President Clinton and former President George W. Bush should be reasonably well-known to U. S. participants. Such a high-information environment creates additional idiosyncratic noise in statistical terms but strengthens realism, thereby enhancing external validity. Jointly, the varying degrees of controlling information about the leader across Studies 1 to 4 complement each other.

Objective outcome measures

Across the four studies, participants had to make costly choices, which we used as objective outcome measures: extra remuneration for themselves versus donating to a charity in Studies 1, 2, and 3, and the chance to make extra money by cheating in Study 4.[3] By using objectively measurable decisions as outcomes, we ruled out common method bias as a source of predictions. These decisions were not cheap talk or socially desirable responses, but rather real choices with monetary impact.

Ruling out alternative explanations

The controlled laboratory context enabled us to directly test the mechanism behind the relationships among leadership styles, their antecedents, and leadership outcomes. That is, we can show that the associations found between leadership styles and outcomes are causal illusions, inexplicable by alternative mechanisms such as the effects of leader behaviors or of information about the leader. The tested mechanism hinges on how evaluators form a judgment about leader behaviors rather than on the phenomenon of leadership per se. Hence, we examined potential shortcomings in the conceptualization of these styles that could lead to illusory causation, and we did not test the real-world tenability of the do-good logic of leadership.

Replication across contexts

We drew participants from diverse populations to replicate our findings across contexts. The U.S.-based participants of the samples of Studies 1, 3 and 4 came from a broad variety of backgrounds, whereas the Study 2 sample was drawn from the experimental participant pool of a business faculty at a Swiss public university. We thus tested the propositions in two different national cultures and on various age groups and socio-economic cohorts (for an overview, see Table 1). In addition, Study 3 is set in a political context, whereas Studies 1, 2, and 4 are set in organizational contexts. Lastly, participants were observers of leadership in Studies 1 and 2, prospective voters in Study 3, and followers in Study 4.

Study 1

Study 1 served to test all three propositions. To test Proposition 1 on leadership styles as partially nonbehavioral outcomes, we examined whether information about a leader's success as a nonbehavioral variable predicted measures of authentic, ethical, and servant leadership.

Hypothesis 1. *Information about a leader's previous success positively predicts ratings of authentic, ethical, and servant leadership.*

We also explored whether leaders' use of charismatic tactics, which ought not to affect the authenticity, ethicality, or servant-like character of their leadership, might predict the three positive leadership style measures. Our testing of Proposition 2 about illusory causation involved predicting donations to a leader's charity when the leader's behavior and information about the leader were held constant (i.e., using residualization and within-condition analysis):

Hypothesis 2. *Ratings of authentic, ethical, and servant leadership positively predict donations, even when leader behavior is held constant.*

Proposition 3, on positive leadership styles sharing empirical properties with the EvalQ, translated into the following hypothesis:

Hypothesis 3. *For the relationships specified in Hypothesis 1 and Hypothesis 2, ratings on the EvalQ behave in the same way as ratings of authentic, ethical, and servant leadership do.*

Moreover, we measured evaluators' trait affect and personality traits. We did so to control for these factors but also to explore whether they further explained spurious predictions by leadership styles and mismeasurement of these styles.

Method

Participants. We recruited 420 U.S.-based participants via Amazon's Mechanical Turk. Participants received a remuneration of USD 2.00 at the beginning of the experiment. After the experiment, participants received an additional USD 0.50, which they could either keep for

[2] It may be tempting to suppose that an instrumental-variable (IV) estimation technique should be used to "take out" endogeneity in $lead_k$. Although IV estimation can remove the endogeneity bias in $lead_k$ by regressing the outcome on $\widehat{lead_k}$, such an estimator shows how LBM might affect the outcome via $lead_k$. This estimator thus effectively shows the effect of LBM on the outcome by isolating the variation of $lead_k$ that overlaps with LBM and the outcome. This estimate is a legitimate one to obtain a consistent estimator. However, such an estimator removes the predictive properties of idiosyncratic evaluations we are attempting to showcase. Our residualization procedure allows for the modeling of LBM simultaneously with $lead_{k.resid}$ to predict the outcome; the coefficient of LBM is, of course, causally interpretable, because LBM is randomized. However, the coefficient of $lead_{k.resid}$ captures exactly the role of evaluations we wish to showcase (e.g., idiosyncratic rater effects not due to LBM). To the extent that the coefficient of $lead_{k.resid}$ is significantly related to the outcome allows us to demonstrate illusory causation.

[3] In Studies 1, 2, and 3, we also measured subjective outcomes, namely generalized leadership impression and trust in Study 1, and generalized leadership impression and evaluation of communication effectiveness in Studies 2 and 3. We report the results in Online Appendices 5a, 5b, 6a, 6b, 7a, and 7b. The results for the subjective outcomes are even stronger than those for objective outcomes and thus offer more support our propositions. However, common method bias might have inflated these results. For this reason, in the manuscript we only report results for objective outcomes as more conservative and thus stronger tests of the three propositions.

Table 1
Overview of Studies and Findings.

	Manipulated nonbehavioral antecedents of leadership styles	Other variables (exploratory, control)	Proposition 1 ("leadership styles are nonbehavioral evaluative outcomes")	Proposition 2 ("misleading predictions by leadership styles")	Proposition 3 ("leadership styles are empirically similar to positive evaluations")
Study 1 (MTurk; 408 participants)	Leader's previous success (high vs low)	Charismatic tactics, positive / negative trait affect, Big-5, age, gender, work experience, education, industry	✓	✓	✓
Study 2 (laboratory subject pool; 367 participants)	Leader's previous ethical achievements (high vs low)	Positive / negative trait affect, Big-5, age, gender, education, faculty, language	✓	(✓)	✓
Study 3 (Prolific Academic; 689 participants)	Presumed identity of speaker (Bush, Clinton, neutral; interacted with the rater's conservatism)	Social dominance orientation, actual speech (Bush, Clinton), age, gender, education, race	✓	✓	✓
Study 4 (Prolific Academic; 555 participants)	— *(the study does not replicate tests of* Proposition 1 *but whether Propositions 2 & 3 hold also when there is construct-relevant behavioral variation)*	Ethical leadership signaling, age, gender, education, race, U.S. nationality, mental ability	not applicable	✓*	no support

Note. The ticks indicate whether a proposition is fully supported across leadership styles. A tick in a bracket indicates partial support for the respective proposition. Please note that across the four studies, Proposition 1 has been tested with Hypothesis 1, Proposition 2 with Hypothesis 2, and Proposition 3 with Hypothesis 3.
* Please note that the effect holds for authentic, ethical, and servant leadership but not for transformational leadership.

themselves or donate to the leader's charity. The study materials included four questions to test whether participants responded seriously. We restricted the final sample to those participants who responded properly these four questions and dropped one more candidate who indicated a wrong age (160 years), resulting in a final sample of 408 participants (97.1%). In the final sample, 213 participants were female (52.1%), the average age was 40.16 years (SD = 10.77), and the average number of years of work experience was 17.98 years (SD = 10.67). We also measured participants' highest educational degree (from high school to PhD) and the industry of their current occupation.

Procedure. The online study consisted of five steps. First, participants read the instructions. Second, they responded to questions about their demographic background. Third, they watched a video of a leader who sought to motivate real workers to prepare as many letters as possible for a fundraising campaign to benefit a charity. Fourth, they rated the leader and themselves on multiple questionnaires. Fifth, at the end of the experiments, they were offered an unexpected bonus sum of USD 0.50; they had to decide whether to keep it or donate it to a charity (see Online Appendix 2 for more details on instructions and materials).

Manipulation and measures

Manipulation of previous success. Before watching the video of the leader, participants read a text telling them that the average performance among workers was 200 letters. In the high-success condition, we told participants to assume that the workers in the leader's team had completed 300 letters on average. In the low-success condition, the workers had completed 100 letters on average.

Donation. We coded donations as a binary variable (0 = no donation, 1 = donation).

Leadership. We measured authentic leadership with Walumbwa et al.'s (2008) scale, ethical leadership with Brown et al.'s (2005) scale, and servant leadership with Liden et al.'s (2015) scale, keeping their original scaling. For our self-developed evaluative questionnaire (EvalQ), we used a five-point Likert-type scale (going from "strongly disagree" to "strongly agree").

Additional variables. We explored the effects of a set of different variables. First, building on Martinko et al.'s (2018) finding that state affect influences leadership ratings, we measured participants' positive and negative trait affect. However, so as to circumvent the endogenous nature of state affect when doing so, we used Watson, Clark, and Tellegen's (1988) scale, which assesses general feelings on 10 positive and 10 negative affective adjectives on scales ranging from 1 ("very slightly or not at all") to 5 ("very much").

Second, to explore the effect described by Felfe and Schyns (2006)—that is, personality variables influence ratings of leadership styles—while seeking to keep the number of items limited, we gauged participants' personality (i.e., the big five) using the short 10-item measure formulated by Rammstedt and John (2007). Each trait was measured using two items on five-point scales, ranging from 1 ("strongly disagree") to 5 ("strongly agree"). Third, we collected information on the demographic variables participant age, being male (0 = female, 1 = male), work experience, education, and industry background (using the Standard Industry Classification scheme), using these as classical control variables.

Fourth, we manipulated the leader's rhetorical tactics (also known as charismatic leadership tactics, or CLTs) using previously established operationalizations (Antonakis, d'Adda, Weber, & Zehnder, 2022; Meslec, Curseu, Fodor, & Kenda, 2020). The rationale for including this manipulation was to check whether rhetorical techniques, which are conceptually unrelated to authentic, ethical, and servant leadership styles, would nonetheless affect ratings of these styles. We also expected that the manipulation of previous success would have the same causal effects across levels of charisma. In both experimental conditions the leader was the same person and the information given in the speech was extremely similar. However, in one experimental condition the leader used more rhetorical tactics (e.g., more analogies, contrasts, and

Table 2
Means, Standard Deviations, and Intercorrelations of Study 1.

	M	SD	alpha	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Donation	0.38	.49		1.00																
2. Authentic l. (ALQ)	3.60	.79	.95	.13	1.00															
3. Ethical l. (ELS)	5.36	1.02	.91	.15	.83	1.00														
4. Servant l. (SL-7)	5.06	1.07	.87	.15	.83	.84	1.00													
5. Eval. quest. (EvalQ)	3.70	.99	.93	.19	.68	.66	.68	1.00												
6. Previous success	.46	.50		.00	.27	.20	.22	.29	1.00											
7. Charismatic tactics	.48	.50		.04	.03	.00	.01	.15	-.03	1.00										
8. Positive trait affect	3.14	.89	.93	.06	.31	.31	.30	.30	-.06	-.01	1.00									
9. Negative trait affect	1.43	.62	.93	-.04	-.09	-.12	-.08	-.09	.03	.03	-.25	1.00								
10. Extraversion	2.75	1.20	.79	.01	.19	.17	.16	.18	-.07	.04	.47	-.21	1.00							
11. Agreeableness	3.56	1.07	.62	.07	.18	.19	.20	.24	-.04	-.05	.43	-.40	.27	1.00						
12. Conscientiousness	4.19	.85	.66	.07	.14	.15	.11	.16	-.05	-.08	.50	-.43	.31	.29	1.00					
13. Neuroticism	2.39	1.21	.85	.00	-.15	-.16	-.13	-.09	.07	.03	-.45	.56	-.40	-.45	-.46	1.00				
14. Openness	3.85	.96	.54	-.03	.19	.15	.15	.10	-.02	.01	.25	-.06	.17	.11	.21	-.15	1.00			
15. Age	40.16	1.77		.19	-.01	-.03	-.04	-.04	.01	-.06	.05	-.10	.03	.06	.11	-.13	.00	1.00		
16. Male	.52	.50		-.08	.00	-.04	.01	-.04	.04	.09	-.06	.05	-.03	-.11	-.12	-.17	-.06	-.17	1.00	
17. Work experience	17.98	1.67		.13	-.02	-.04	-.03	-.05	-.01	-.03	.05	-.09	.04	.05	.09	-.12	.04	.88	-.10	1.00

Note. For ease of illustration, we do not include the categorical control variables of raters' educational and industry background. Male = 1 (else = 0). Correlations greater than |.10| are significant at $p < .05$; correlations greater than |.13| are significant at $p < .01$; correlations greater than |.17| are significant at $p < .001$.

collective sentiments) than did the leader in the other experimental condition.

Results

Descriptive statistics. Table 2 shows the means, standard deviations, reliabilities, and correlations among the variables.

Recall check. We tested whether participants remembered the manipulation of the leader's previous success on a three-item scale ("This leader was successful in obtaining high performance from his team"; "This leader got good results"; "This leader was effective in motivating workers") administered at the end of the experiment. The items used a scale ranging from 1 to 5 (from "strongly agree" to "strongly disagree"). Participants successfully recalled the manipulation (β = 2.48; p <.01).

Hypothesis testing. To examine Hypothesis 1, on whether previous success influenced ratings of authentic, ethical, and servant leadership, we specified three regression models. In each of these models, the dependent variable was one of the three leadership styles. Previous success was the focal independent variable, and the additional variables served as control variables. In support of Hypothesis 1, previous success of the leader was highly predictive of authentic, ethical, and servant leadership measures ($\beta = .48,\ p < .01,\ \beta = .47,\ p < .01,\ \beta = .54,\ p < .01$, respectively) (see Table 3).

In addition, we examined Hypothesis 2 about causal illusions by testing whether authentic, ethical, and servant leadership predicted donations. To do so, we specified six OLS regression models. Donations were the dependent variable in all six regressions, and the above-specified exploratory variables were the controls. In addition, in each regression one of the three measures of leadership styles (authentic, ethical, and servant leadership) was the focal independent variable; more specifically, we used the residuals of these leadership styles that are orthogonal to charismatic tactics (see our section "Statistical Proposition Testing"). For each of these three focal independent variables, we specified two regression models, one for each level of the experimental manipulation (i.e., low and high previous success).[4] Although donations, the dependent variable for testing Proposition 2, was binary, we used OLS instead of probit estimation for two reasons. First, OLS produces causally consistent results (Angrist & Pischke, 2008), even if a condition is fully determined (Caudill, 1988). Second, in contrast to probit, regression coefficients are easily interpretable; the observed coefficient is the marginal effect (Huang, 2019). In support of Hypothesis 2, authentic, ethical, and servant leadership measures predicted donations across conditions ($\beta = .08,\ p < .05, \beta = .07,\ p < .01, \beta = .07,\ p < .01$, respectively), but also within each of the two conditions (see Table 4).

To test Hypothesis 3, regarding whether the EvalQ and the three measures of positive leadership styles have similar empirical properties, we redid the tests for Hypotheses 1 and 2 using the evaluative questionnaire (EvalQ) instead of the leadership style measures (see Column 4 in Tables 3 and 4). To establish similar empirical properties, the regression coefficient of the EvalQ should have the same directionality as the regression coefficient of the three leadership styles. Moreover, as an even stronger test, the coefficients of both the EvalQ and the three leadership styles should be statistically indistinguishable from each other in predicting donations. In line with Hypothesis 3, previous success predicted the evaluative questionnaire ($\beta = .62,\ p < .01$; see Table 3), and the evaluative questionnaire predicted donations across and within conditions ($\beta = .08,\ p < .01$; see Table 4). We tested whether

[4] We specified separate statistical models per experimental condition. Such a within-condition analysis is statistically equivalent to interacting the experimental manipulation with all independent variables. In absence of specific interaction hypotheses, such a comprehensive model is ideal for ensuring consistent estimates (see Online Appendix 1).

Table 3
Nonbehavioral Prediction of Leadership Styles in Study 1 (i.e., Propositions 1 & 3).

	(1)	(2)	(3)	(4)
	Authentic leadership (ALQ)	Ethical leadership (ELS)	Servant leadership (SL-7)	Evaluative questionnaire (EvalQ)
Previous success	.48***	.47***	.54***	.62***
	(6.40)	(4.70)	(5.14)	(6.88)
Charismatic tactics	.05	.02	.03	.33***
	(.72)	(.23)	(.30)	(3.65)
Positive trait affect	.25***	.32***	.33***	.26***
	(4.42)	(4.32)	(4.28)	(3.94)
Negative trait affect	-.01	-.05	.01	.02
	(.09)	(.45)	(.07)	(.21)
Extraversion	.04	.03	.03	.06
	(1.04)	(.73)	(.69)	(1.38)
Agreeableness	.05	.06	.11*	.18***
	(1.17)	(1.04)	(1.90)	(3.52)
Conscientiousness	-.03	-.04	-.06	.05
	(.49)	(.49)	(.73)	(.76)
Neuroticism	.01	.00	.03	.09
	(.24)	(.08)	(1.43)	(1.63)
Openness	.10**	.09*	.10*	.03
	(2.43)	(1.77)	(1.85)	(.65)
Age	.00	.00	-.01	.00
	(.13)	(.39)	(.66)	(.24)
Male	.00	-.08	.01	-.05
	(.05)	(.72)	(.05)	(.51)
Work experience	.00	.00	.00	-.01
	(.15)	(.28)	(.20)	(1.05)
Education	Included	Included	Included	Included
Industry	Included	Included	Included	Included
Constant	2.65***	4.92***	4.00***	1.26*
	(4.53)	(6.35)	(4.90)	(1.78)
F-test for all controls	F(33,373) = 2.46***	F(33,373) = 2.14***	F(33,373) = 1.97***	F(33,373) = 3.17***
Observations	408	408	408	408
R-squared	.24	.19	.19	.29

t-statistics in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.
Note. Education and industry are composed of several dummy variables.

the coefficient for the effect of previous success on the evaluative questionnaire differed from that for the average effect of the other leadership measures. The effect of previous success on the evaluative questionnaire was slightly higher: $\chi^2(1) = 3.74,\ p = .053$. However, the significance and direction of the effect were largely the same. We also tested whether the coefficient of the evaluative questionnaire predicting donations differed from the corresponding average effect of the three leadership styles, which it did not: $\chi^2(1) = 1.26,\ p = .26$.

Moreover, in our exploratory analyses, participants' positive trait affect consistently predicted the measures of leadership styles and the EvalQ (see Table 3). Other individual variables did not have significant effects across the models. Concerning the prediction of donations, age consistently increased donations in the low but not in the high previous success condition (see Table 4). Other variables did not have significant effects across the four models.

Discussion of Study 1

In Study 1, we found reliable evidence that the three measures of positive leadership styles (ALQ, ELS, and SL-7) represented nonbehavioral evaluative outcomes that were predicted by information about previous success, supporting Hypothesis 1. In addition, these three measures predicted donations within (i.e., when actual leader behavior was kept constant) and across experimental conditions (i.e., when leader behavior was statistically controlled for), supporting Hypothesis 2. It is obvious that interpreting such predictions as effects of leader behaviors would be baseless because actual leadership behavior was factually or statistically kept constant. Lastly, as evidence for Hypothesis 3, the pattern of results with regards to both antecedents and outcomes was highly similar for the three positive leadership style measures and the nonbehavioral, purely evaluative EvalQ. In particular, there was no difference between the style measures and the EvalQ in predicting donations. The support for the hypotheses contradicts the dominant view of positive leadership styles as behavioral constructs. Given the strong claims we make, and the ensuing repercussions of Study 1, we sought to constructively replicate it with a second study.

Study 2

Unlike in Study1, we did not manipulate charisma in Study 2. Because the Study 1 results were highly similar across the two levels of the manipulation, we exposed all participants to the same level—that is, the baseline condition (i.e., low charisma). Doing so allowed an increase in the effective sample size and simplified the interpretation of results. Not varying charisma also meant that leader behavior was constant across participants. In Study 2, as a variant of the manipulation of information about previous success in Study 1, we manipulated information about the leader's previous ethical achievements as a potential nonbehavioral cause of leadership style ratings. By studying the effects of different types of success, we sought to demonstrate that the evaluative nature of leadership style ratings is not limited to cues about previous productivity. Moreover, the sample for Study 2 (members of a Swiss public university's laboratory subject pool) was different from that for Study 1 (MTurkers). Lastly, the decision to donate was more costly in Study 2 than it was in Study 1 due to the absolute and relative size of the sum involved. We tested the following hypotheses:

Table 4
Illusory Causation in Study 1 (i.e., Propositions 2 & 3).

	(1)	(2)	(3)	(4)
	Donation	Donation	Donation	Donation
Authentic leadership	.06/.10*/.08**			
	(1.44/1.84/2.47)			
Ethical leadership		.06*/.08**/.07***		
		(1.77/2.24/3.23)		
Servant leadership			.06*/.08*/.07***	
			(1.92/1.95/3.07)	
Evaluative questionnaire				.08**/.12**/.08***
				(2.35/2.48/3.41)
Charismatic tactics	.10/.02	.10/.02	.10/.02	.10/.03
	(1.45/31)	(1.43/.33)	(1.44/.26)	(1.37/.40)
Positive affect	.02/-.04	.02/-.04	.02/-.05	.02/-.04
	(.45/.69)	(.38/.75)	(.42/.80)	(.31/.76)
Negative affect	.06/.05	.06/-.04	.06/-.05	.05/-.05
	(.80/.83)	(.82/.74)	(.81/.88)	(.71/.77)
Extraversion	-.02/.02	-.02/.02	-.02/.02	-.02/.02
	(.50/.58)	(.47/.61)	(.53/.65)	(.56/.51)
Agreeableness	.02/.03	.02/.03	.02/.02	.01/.01
	(.51/.64)	(.51/.58)	(.46/.50)	(.29/.16)
Conscientiousness	.06/-.01	.06/.00	.06/-.01	.05/-.02
	(1.23/.23)	(1.19/.08)	(1.24/.14)	(1.06/.28)
Neuroticism	.04/.02	.04/.02	.03/.02	.03/.01
	(.86/.50)	(.89/.44)	(.82/.46)	(.74/.31)
Openness	-.02/-.04	-.03/-.04	-.03/-.04	-.02/-.04
	(.62/1.05)	(.71/.92)	(.68/1.00)	(.52/.93)
Age	.02**/.01	.02**/.01	.02**/.01	.02**/.01
	(2.20/1.31)	(2.17/1.45)	(2.18/1.51)	(2.19/1.26)
Male	-.02/-.05	-.02/-.03	-.02/-.04	-.03/-.03
	(.27/.57)	(.26/.40)	(.32/.45)	(.33/.39)
Work experience	-.01/.00	-.01/.00	-.01/.00	-.01/.00
	(1.65/1.18)	(1.53/.28)	(1.62/.31)	(1.58/.11)
Education	Included	Included	Included	Included
Industry	Included	Included	Included	Included
Constant	-.97**/.05	-.98**/-.01	-.99**/.09	-.80/.13
	(-2.00/.10)	(-2.04/.02)	(-2.06/.18)	(-1.62/.25)
F-test for all controls	F(31,168) = 3.09***/	F(31,168) = 4.08***/	F(31,168) = 4.79***/	F(31,168) = 3.69***/
	F(30,151) = 5.74***	F(30,151) = 4.48***	F(30,151) = 4.56***	F(30,151) = 5.78***
# Observations	221/187/408	221/187/408	221/187/408	221/187/408
R-squared	.17/.21	.17/.22	.17/.21	.18/.22

t-statistics in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.10.
(a) Education and industry are composed of several dummy variables. (b) For each leadership style respectively, the first data entry before the slash refers to the condition with low previous success, the second refers to the condition of high previous success, and the third to the pooled effect across the performance conditions (reported only for the coefficients and the number of observations). (c) The regressors for the leadership style measures are the residuals, as a function of the manipulation of the charismatic leadership manipulation. (d) The *F*-tests are based on the sample with 8 dropped singleton observations; these were dropped because the variance–covariance matrix of the estimators is not of full rank when there are singletons and a robust estimate of the variance is used (see "j_robustsingular" in Stata; see also Baum, Schaffer, and Stillman (2010) and https://www.stata.com/statalist/archive/2005–10/msg00594.html). Note that the point estimates remain precisely the same.

Hypothesis 1. *Information about a leader's previous ethical achievements positively predicts ratings of authentic, ethical, and servant leadership.*

Hypothesis 2. *Ratings of authentic, ethical, and servant leadership positively predict donations, although leader behavior is constant.*

Hypothesis 3. *For the relationships specified in Hypothesis 1 and Hypothesis 2, ratings on the EvalQ behave in the same way as ratings of authentic, ethical, and servant leadership do.*

Method

Participants. We recruited 394 participants via the experimental subject pool of a Swiss public university. Participants received a fixed remuneration of CHF 10, which corresponded to around USD 10 at the time of the study. At the end of the experiment, participants received an additional CHF 5, which they could either keep for themselves or donate to the leader's charity. The study materials included three questions to test whether participants responded seriously. We restricted the final sample to the 367 participants (93.1%) who responded properly to these questions. In the final sample, 201 participants were female (54.8%), and the average age was 22.80 years (SD = 2.79). We also measured participants' highest educational degree, the faculty in which they were currently enrolled, and their English-language skills.

Procedure. The study consisted of five steps. First, participants read the instructions. Second, they responded to demographic questions. Third, they watched a video of a leader who sought to motivate real workers to prepare as many letters as possible for a fundraising campaign to benefit a charity. Fourth, they rated the leader and themselves via multiple questionnaires. Fifth, they decided whether to donate the CHF 5 that they received at the end of the study; this money was an additional and unexpected remuneration (see Online Appendix 3 for more details on instructions and materials).

Manipulation and measures

Manipulation of ethical achievements. Before watching the video of the leader, participants read a text telling them to assume that the leader was either modestly or very successful in making workers feel that the charity respected them as individuals and not only as workers. This manipulation constructively replicates Study 1′s manipulation of information about previous success. Both making workers feel respected and increasing workers' productivity are positive achievements, although on different criteria (employee treatment as an ethical outcome versus productivity as a bottom-line outcome).

Donation. We coded donations as a binary variable (0 = no donation, 1 = donation).

Leadership. We used the same scales as in Study 1 to measure the three positive leadership styles.

Exploratory variables. We again assessed participants' positive and negative trait affect using Watson et al.'s (1988) five-point scale ranging from 1 ("very slightly or not at all") to 5 ("very much"). We gauged participants' personality using Soto and John's (2017) 30-item measure with a five-point scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). Due to its superior psychometric properties, we used this longer measure instead of the 10-item measure by Rammstedt and John (2007) deployed in Study 1. Furthermore, we collected data on the demographic variables participant age, being male, education (from below high school to PhD), faculty membership, and English-language skills.

Results

Descriptive statistics. Table 5 shows the means, standard deviations, reliabilities, and correlations among the variables.

Recall check. We tested the effectiveness of the ethical-achievement manipulation after the donation decision using three items ("He made workers feel respected"; "He made workers feel well"; "He treats others respectfully") on a scale ranging from 1 to 5 (from "strongly disagree" to "strongly agree"). Participants successfully recalled the manipulation (β = 0.55; p <.01).

Hypothesis Testing. We adopted the statistical approach from Study 1, but we replaced the manipulation of information about previous success with the manipulation of information about ethical achievements and used the adjusted set of exploratory variables. In addition, we did not require residualization because leader behavior was constant. We found support for Hypothesis 1 (see Table 6). Information about ethical achievements predicted measures of authentic, ethical, and servant leadership ($\beta = .38,\ p < .01,\ \beta = .54,\ p < .01,\ \beta = .53,\ p < .01$, respectively). Support for Hypothesis 2 was mixed (see Table 7). Measures of authentic, ethical, and servant leadership did not predict donations across conditions. However, ethical and servant leadership predicted donations in the high ethical-achievement condition ($\beta = .08,\ p < .05,\ \beta = .08,\ p < .10$, respectively), and servant leadership negatively predicted donations in the low ethical-achievement condition ($\beta = -.08,\ p < .05$). Furthermore, in line with Hypothesis 3, the independent variable information about ethical achievements predicted the evaluative questionnaire ($\beta = .37,\ p < .05$; Table 6). This effect did not differ from the average effect of the manipulation of information about ethical achievements on the three leadership style measures: $\chi^2(1) = 2.70,\ p > .10$). The evaluative questionnaire, in turn, predicted donations in the low ethical-achievement condition ($\beta = -.14,\ p < .01$, see Table 7). This coefficient did not differ from the average effect of the three leadership styles ($\chi^2(1) = 2.31,\ p > .10$). Although some of the coefficients were not statistically significant, it was still important to run this test (see Gelman & Stern, 2006).

In the exploratory analyses, with regard to predicting measures of leadership styles and the EvalQ, we did not find that the individual

Table 5
Means, Standard Deviations, and Intercorrelations of Study 2.

	M	SD	Alpha	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Donation	.64	.48		1.00														
2. Authentic leadership (ALQ)	3.20	.68	.88	.02	1.00													
3. Ethical leadership (ELS)	4.69	1.01	.87	.04	.74	1.00												
4. Servant leadership (SL-7)	4.48	.99	.79	-.01	.67	.74	1.00											
5. Evaluative questionnaire (EvalQ)	3.06	.72	.80	-.04	.52	.49	.56	1.00										
6. Previous ethical success	.55	.50		-.08	.28	.27	.27	.25	1.00									
7. Positive trait affect	3.63	.59	.81	.02	.13	.18	.12	.07	.05	1.00								
8. Negative trait affect	2.31	.70	.84	-.05	-.03	.03	.05	.04	-.09	-.24	1.00							
9. Extraversion	3.29	.68	.75	-.03	.15	.16	.14	.03	.04	.63	-.26	1.00						
10. Agreeableness	3.70	.62	.71	.16	.15	.11	.17	.10	.09	.28	-.22	.21	1.00					
11. Conscientiousness	3.45	.67	.72	.04	.02	.07	.06	.06	.11	.49	-.27	.35	.24	1.00				
12. Neuroticism	2.88		.83	.03	-.04	.01	-.01	.00	-.05	-.34	.71	-.36	-.11	-.29	1.00			
13. Openness	3.48	.62	.62	-.04	.01	-.01	-.01	-.09	.04	.17	.07	.08	.17	-.06	.11	1.00		
14. Male	.45	.50		-.09	-.01	-.09	-.10	-.05	.01	.16	-.20	.11	-.12	-.02	-.41	-.02	1.00	
15. Age	22.80	2.79		.03	-.14	-.21	-.11	-.06	-.07	-.05	-.10	-.05	.04	.07	-.04	.03	-.01	1.00

Note. For ease of illustration, we do not include the categorical control variables of raters' educational background, faculty, and English-language skills. Male = 1 (else = 0). Correlations greater than |.11| are significant at p <.05; correlations greater than |.14| are significant at p <.01; correlations greater than |.18| are significant at p <.001.

Table 6
Nonbehavioral Prediction of Leadership Styles in Study 2 (i.e., Proposition 1 & 3).

	(1)	(2)	(3)	(4)
	Authentic leadership (ALQ)	Ethical leadership (ELS)	Servant leadership (SL-7)	Evaluative questionnaire (EvalQ)
Ethical achievement	.38***	.54***	.53***	.37**
	(5.56)	(5.34)	(5.26)	(4.99)
Positive trait affect	.06	.23*	.06	.08
	(.70)	(1.87)	(.52)	(.94)
Negative trait affect	.08	.14	.27**	.18**
	(1.16)	(1.37)	(2.53)	(2.30)
Extraversion	.13*	.12	.18*	-.03
	(1.96)	(1.23)	(1.88)	(.41)
Agreeableness	.13**	.11	.23**	.12*
	(2.21)	(1.22)	(2.54)	(1.79)
Conscientiousness	-.11*	-.08	-.08	-.03
	(1.74)	(.90)	(.85)	(.46)
Neuroticism	-.06	-.02	-.16	-.10
	(.89)	(.19)	(1.63)	(1.36)
Openness	-.06	-.12	-.13	-.14**
	(1.02)	(1.43)	(1.47)	(2.28)
Age	-.00	-.04*	-.00	.00
	(.17)	(1.84)	(.21)	(.12)
Male	-.07	-.26**	-.26**	-.11
	(.84)	(2.20)	(2.21)	(1.25)
Education	Included	Included**	Included**	Included
Faculty	Included**	Included*	Included*	Included***
Language	Included	Included	Included	Included**
Constant	2.61***	4.22***	3.24***	2.56***
	(4.32)	(4.77)	(3.68)	(3.99)
F-test for all controls	F(23,342) = 1.79**	F(23,342) = 2.25***	F(23,342) = 1.75**	F(23,342) = 1.80**
# observations	367	367	367	367
R-squared	.18	.20	.17	.17

t-statistics in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Note. Education, faculty, and language are composed of several dummy variables.

exploratory variables had consistent effects (see Table 6). However, jointly these exploratory variables were significant (see *F*-test results in Table 6). In terms of predicting donations, agreeableness had a consistently positive effect agreeableness, and openness had a negative effect in the low ethical-achievement condition (see Table 7). Jointly, the control variables were significant only in the low ethical-achievement condition (see *F*-test results in Table 7).

Discussion of Study 2

In Study 2, we constructively replicated the finding that leadership styles are nonbehavioral outcomes. Although leader behavior did not vary across levels of information about ethical achievements, this information predicted scores on the three leadership style measures, supporting Hypothesis 1. However, regarding Hypothesis 2, we found only partial support for illusory causation. Leadership styles had effects of inconsistent directionality across the two experimental conditions. One plausible explanation for the merely partial support for Hypothesis 2 is the strong main effect of the independent variable information about ethical achievements on donations. This main effect reduced the variance in donations that might be explained by the leadership styles. Finally, regarding Hypothesis 3, the pattern of results was similar for the three leadership styles and the EvalQ. The effect of the EvalQ on donations did not differ from the average effect of the other styles on donations, suggesting that measures of authentic, ethical, and servant leadership are, at least partially, evaluative.

Study 3

As with Studies 1 and 2, Study 3 allowed us to test our three propositions that leadership styles are evaluative, nonbehavioral outcomes that create causal illusions and have empirical properties similar to those of an evaluative questionnaire. Unlike in Studies 1 and 2, however, we conducted Study 3 using U.S. presidents' inaugural speeches instead of videos showing leaders. In addition, instead of manipulating information about a leader's previous success or about ethical achievements, we tested whether leader–follower alignment on political values—another nonbehavioral antecedent—predicted measures of leadership styles and donations. Past research has found values have effects on leadership style assessments (De Luque, Washburn, Waldman, & House, 2008) and leadership outcomes (e.g., Chin, Hambrick, & Treviño, 2013). Thus, value alignment should affect leadership style ratings, and Hypotheses 2 and 3 resemble those from the two previous studies:

Hypothesis 1. *An observer's value alignment with a leader positively predicts ratings of authentic, ethical, and servant leadership.*

Hypothesis 2. *Ratings of authentic, ethical, and servant leadership positively predict donations, even when leader behavior is kept constant.*

Hypothesis 3. *For the relationships specified in Hypothesis 1 and Hypothesis 2, ratings on the EvalQ behave in the same way as ratings of authentic, ethical, and servant leadership do.*

Method

Participants. We recruited 700 U.S.-based participants via Prolific Academic. Participants received a fixed remuneration of GBP 2.50, which corresponded to around USD 3.07 at the time of the experiment. At the end of the experiment, participants received an additional GBP.50, which they could either keep for themselves or donate to the charity of the former president whose inaugural address they read. The study materials included six questions to test whether participants responded seriously. We restricted the final sample to the 689 participants (98.4%) who responded properly to at least four of the six questions. In the final sample, 258 participants were female (37.5%) and the average age was 33.70 years (SD = 11.40). We also collected information on the participants' highest educational degree and their race.

Procedure. The study consisted of four steps. First, participants read

Table 7
Illusory Causation in Study 2 (i.e., Propositions 2 and 3).

	(1)	(2)	(3)	(4)
	Donation	Donation	Donation	Donation
Authentic leadership	-.07/.09/.01			
	(1.29/1.51/.23)			
Ethical leadership		-.05/.08*/.01		
		(1.29/1.88/.48)		
Servant leadership			-.08**/.08*/.00	
			(2.02/1.85/.03)	
EvalQ				-.14***/.06/-.04
				(2.70/.98/1.12)
Positive trait affect	.01/.01	.03/.01	.02/.02	.04/.02
	(.13/.14)	(.33/.10)	(.25/.24)	(.41/.23)
Negative trait affect	-.12/-.01	-.11/-.01	-.09/.02	-.07/-.01
	(1.50/.11)	(1.43/.10)	(1.11/.20)	(.91/.14)
Extraversion	-.04/-.04	-.05/-.04	-.03/-.04	-.07/-.03
	(.59/.55)	(.73/.59)	(.44/.57)	(.93/.49)
Agreeableness	.16**/.10	.16**/.10	.16**/.08	.17**/.10
	(2.16/1.58)	(2.13/1.64)	(2.22/1.31)	(2.40/1.61)
Conscientiousness	-.02/.00	-.02/.00	-.02/.00	-.01/-.01
	(.23/.06)	(.23/.01)	(.33/.06)	(.20/.11)
Neuroticism	.06/.02	.06/.02	.05/.03	.05/.03
	(.84/.32)	(.88/.23)	(.62/.37)	(.67/.39)
Openness	-.17**/.00	-.17**/.01	-.17**/.01	-.18**/.00
	(2.38/.04)	(2.46/.10)	(2.44/.17)	(2.61/.05)
Age	.00/.00	.00/.01	.00/.00	.00/.00
	(.00/.26)	(.06/.48)	(.04/.22)	(.14/.21)
Male	-.01/-.15*	-.02/-.14*	-.03/-.14*	-.02/-.15*
	(.16/1.88)	(.23/1.68)	(.31/1.69)	(.24/1.87)
Education	Included	Included	Included	Included
Faculty	Included	Included	Included	Included
Language	Included	Included	Included	Included
Constant	1.50**/-.17	1.51**/-.29	1.54**/-.19	1.57**/.07
	(1.99/.27)	(2.00/.45)	(2.08/.30)	(2.15/.11)
F-test for all controls	F(22,140) = 1.33/	F(22,140) = 1.33/	F(22,140) = 1.29/	F(22,140) = 1.47*/
	F(22,179) = .85	F(22,179) = .83	F(22,179) = .79	F(22,179) = .86
# Observations	164/203	164/203	164/203	164/203
R-squared	.18/.11	.18/.12	.19/.12	.21/.10

t-statistics in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.
Education, faculty, and language are composed of several dummy variables. In addition, the first entry refers to the condition of low ethical achievements, second entry to the condition of high ethical achievements, and the third entry to across-conditions results (reported only for the coefficients and the number of observations).

the instructions. Second, they read an excerpt of the inaugural address of a former U.S. president. Third, they rated the leader (the U.S. president) and themselves on multiple questionnaires. Fourth, they decided whether to donate GBP.50 to the charity of the respective former president. Participants received this sum at the end of the study; this money was an additional and unexpected remuneration (see Online Appendix 4 for more details on instructions and materials).

Manipulation and measures

Manipulation of speaker identity. Before reading the excerpt of the speech, in one condition, participants were instructed to presume that George W. Bush had given the speech; in a second condition, participants were instructed to presume that Bill Clinton had been the speaker; and in the neutral condition, we asked participants to presume that a former U.S. president had given the speech. This manipulation served to indicate whether the president's (leader's) values are more (Bush) or less (Clinton) conservative based on the assumption that participants knew these two presidents' party affiliations.

Participants then read the speech. Within each condition, half of the participants received an excerpt of a speech by Bush and the other half an excerpt of a speech by Clinton. We counterbalanced the excerpts across the independent variable presumed speaker identity. The speeches were very similar in length and style, and the actual speech content was not of theoretical interest to us (and it had a weak effect on donations in 2 of 12 cases; see Table 10). However, by using two different speeches, we ruled out that findings would hold only for one particular inaugural speech.

Value alignment. We assessed political-economic conservatism using the three-item measure by Pratto et al. (1994) with the original seven-point agreement scale (1= "very liberal," 7 = "very conservative"). The interaction between this measure and the identity of the president captured the political or value-based alignment between participants and the leader. That is, a participant who scored highly on conservatism was in political alignment with George W. Bush but not with Bill Clinton. As an alternative indicator of conservatism (not reported in the main analyses), we measured social dominance orientation with Pratto et al.'s (1994) 16-item scale.

Donation. We coded donations as a binary variable (0 = no donation, 1 = donation).[5]

Leadership. We measured authentic, ethical, and servant leadership using the same questionnaires as those administered in Studies 1 and 2, and we used the same evaluative questionnaire.

Other variables. We included four control variables: participant age, being male (0 = female, 1 = male), education (from below high

[5] We measured a second DV, namely whether participants are willing to spend 2–5 min on an extra-task. The effects have the hypothesized directionality yet are mostly statistically insignificant. Data for this variable are part of the uploaded dataset. In hindsight, however, we regard the logic for gathering this data conceptually flawed. Whereas making a donation to a charity is both costly for participants and useful for the charity, spending time on the extra task is not necessarily costly (participants can do other stuff in the meantime) and not necessarily useful (depending on perceived task significance). Thus, we do not report relationship with this variable in the manuscript and generally call for caution in interpreting such relationships.

Table 8
Means, Standard Deviations, and Intercorrelations of Study 3.

	M	SD	alpha	1	2	3	4	5	6	7	8	9	10	11	12
1. Donation	.36	.48		1.00											
2. Authentic leadership (ALQ)	3.52	.78	.94	.26	1.00										
3. Ethical leadership (ELS)	5.10	1.14	.94	.25	.81	1.00									
4. Servant leadership (SL-7)	4.91	1.15	.89	.25	.79	.85	1.00								
5. Evaluative questionnaire (EvalQ)	3.76	.82	.89	.21	.68	.72	.70	1.00							
6. Actual speaker Bush	.47	.50		-.02	.09	.08	.11	.06	1.00						
7. Presumed speaker neutral	.30	.46		.01	.10	.11	.09	.03	-.02	1.00					
8. Presumed speaker Bush	.34	.47		.02	-.10	-.07	-.08	-.06	-.02	-.47	1.00				
9. Conservatism	3.43	1.67	.89	.04	.17	.18	.22	.22	.00	.02	-.03	1.00			
10. Social dom. orient. (SDO)	2.47	1.22	.94	-.01	.05	.09	.11	.15	-.02	.04	-.01	.54	1.00		
11. Male	.63	.48		-.05	.02	.04	.09	.12	-.01	-.02	.01	.14	.31	1.00	
12. Age	33.70	11.40		.04	.05	.03	.01	.09	-.02	.03	.01	.07	-.03	-.05	1.00

Note. For ease of illustration, we do not include the categorical control variables of raters' educational background and race. Male = 1 (else = 0). Correlations greater than |.08| are significant at $p < .05$; correlations greater than |.10| are significant at $p < .01$; correlations greater than |.13| are significant at $p < .001$.

Table 9
Nonbehavioral Prediction of Leadership Styles in Study 3 (i.e., Propositions 1 & 3).

	(1)	(2)	(3)	(4)
	Authentic leadership (ALQ)	Ethical leadership (ELS)	Servant leadership (SL-7)	Evaluative questionnaire (EvalQ)
Conservatism	.04	.04	.07	.06*
	(1.33)	(0.90)	(1.62)	(1.85)
Presumed speaker Bush	-.47***	-.74***	-.66***	-.51***
	(-2.78)	(-2.88)	(-2.66)	(-2.74)
Presumed speaker neutral	.17	.18	.06	.09
	(.91)	(.62)	(.12)	(.41)
Interaction: Presumed speaker Bush *x* conservatism	.10**	.18***	.13**	.10**
	(2.15)	(2.62)	(2.09)	(2.18)
Interaction: Presumed speaker neutral *x* conservatism	-.04	-.01	.00	-.05
	(-.83)	(-.23)	(.01)	(-1.02)
Actual speech Bush	.16***	.21***	.29***	.13**
	(2.87)	(2.61)	(3.57)	(2.18)
Male	-.01	-.01	.10	.14**
	(-.11)	(-.10)	(1.15)	(2.31)
Age	.00	-.00	-.00	.00
	(.59)	(-.09)	(-.82)	(1.23)
Education	Included**	Included**	Included***	Included***
Race	Included***	Included***	Included**	Included**
Constant	3.07***	4.66***	4.24***	3.08***
	(12.87)	(15.15)	(13.26)	(14.19)
F-test for all controls	F(11,672) = 6.42***	F(11,672) = 7.77***	F(11,672) = 8.22***	F(11,672) = 8.25***
Observations	689	689	689	689
R-squared	.13	.16	.16	.17

t-statistics in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.
Note. Education and race are multiple dummy variables.

school to PhD), and race (based on the U.S. census categories).

Results

Descriptive statistics. Table 8 shows the means, standard deviations, reliabilities, and correlations among the variables.

Recall check. After participants had made their donation decision, we asked them to identify, from a list of five options ("Bill Clinton," "George W. Bush," "Barack Obama," "Donald Trump," "A former president of the U.S."), the president whom they had been told to presume had given the speech; 598 out of 689 (86.8%) selected the correct name, indicating that our manipulation of the president's identity had been successfully recalled. Afterward, we informed participants that the presumed president might or might not have been the president who actually delivered the inaugural address, and we asked participants whether they were able to identify the actual speaker. Out of 689 participants, 530 indicated that they did not know the identity, and among the 159 participants who indicated that they knew the actual identity, 73 gave a correct response and 86 an incorrect one. Participants, therefore, were not systematically able to determine the actual president's identity.

Hypothesis testing. We used ordinary least squares regressions (OLS) and specified the manipulations as independent variables and the other variables—that is, the two versions of the speech and the demographic variables—as control variables. To examine Hypothesis 1, we tested whether value alignment (i.e., the interaction term between the presumed speaker identity and conservatism) predicted ratings of authentic, ethical, and servant leadership (see Table 9). To do so, we specified three regression models. In each model, the dependent variable was one of the three leadership styles. The independent variables were conservatism, two dummy variables indicating different levels of the manipulated variable presumed identity, and the two interaction terms between conservatism and each dummy variable, as well as the exploratory variables. We found consistent support for Hypothesis 1 (see Table 9). Value alignment (i.e., interaction of Bush as presumed speaker and conservatism) predicted measures of authentic, ethical, and servant

Table 10
Illusory Causation in Study 3 (i.e., Propositions 2 & 3).

	Donation	Donation	Donation	Donation
Authentic leadership	.17***/.13***/.15***/.15*** (3.97/3.49/3.20/6.39)			
Ethical leadership		.11***/.10***/.11***/.11*** (3.79/3.72/3.28/6.42)		
Servant leadership			.13***/.11***/.07**/.10*** (4.43/4.03/2.23/6.40)	
Evaluative questionnaire				.15***/.13***/.04/.12*** (3.87/3.43/.90/5.06)
Age	.00/.00/.00 (.05/1.60/.22)	.00/.01**/.00 (.11/2.01/.37)	.00/.01*/.00 (.40/1.91/.32)	.00/.00/.00 (.10/1.60/.35)
Male	-.04/-.06/-.04 (.54/1.04/.53)	-.03/-.05/-.05 (.49/.87/.68)	-.06/-.06/-.05 (.92/1.06/.70)	-.07/-.08/-.04 (.98/1.41/.59)
Actual speech Bush	-.11*/.09/-.07 (1.76/1.57/1.10)	-.10/.08/-.07 (1.56/1.36/1.07)	-.09/.08/-.08 (-1.52/1.43/-1.15)	-.11*/.08/-.07 (-1.77/1.38/-1.09)
Education	Included	Included	Included	Included
Race	Included	Included	Included	Included
Constant	.27/-.18/.44* (1.00/-.96/1.83)	.13/-.19/.46* (.51/.99/1.88)	.23/-.20/.50** (.89/1.03/2.02)	.21/-.14/.51** (.80/.70/2.03)
F-test for all controls	F(10,222) =.75 F(9,234) = 2.22** F(10,195) = 1.32	F(10,222) =.90 F(9,234) = 2.11** F(10,195) = 1.66*	F(10,222) =.92 F(9,234) = 2.13** F(10,195) = 1.52	F(10,222) =.95 F(9,234) = 2.15** F(10,195) = 1.33
Observations	235/246/208	235/246/208	235/246/208	235/246/208
R-squared	.11/.17/.12	.10/.17/.12	.12/.18/.10	.11/.17/.08

t-statistics in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Note. For ease of illustration, we do not include the categorical control variables of raters' educational background and race. Male = 1 (else = 0). First entry is the condition of presumed identity Bush, second entry is the condition of Clinton, third entry is the neutral condition. The fourth entry refers to the whole sample across conditions (indicated for the leadership styles only).
The regressors for the leadership style measures are the residuals, as a function of the manipulation of speaker identity.

leadership ($\beta = .10$, $p < .05$, $\beta = .18$, $p < .01$, $\beta = .13$, $p < .05$, respectively).

To examine Hypothesis 2, we tested whether ratings of authentic, ethical, and servant leadership predicted donations. We specified nine regression models with donations as the dependent variable and the other variables as control variables. Each regression had one leadership style measure as the focal independent variable.; more specifically, we were using the residuals of these leadership styles that are orthogonal to the two versions of the speech (see our section on testing proposition 2). For each of these independent variables, we specified a separate regression model to test for effects within the three different experimental conditions of presumed speaker identity. We found consistent support for Hypothesis 2 (see Table 10). Authentic, ethical, and servant leadership measures predicted donations across conditions ($\beta = .15$, $p < .01$, $\beta = .11$, $p < .01$, $\beta = .10$, $p < .01$, respectively), but also within each of the two coded levels of the interaction between presumed speaker identity and conservatism.

Furthermore, we examined Hypothesis 3 on similar empirical properties by redoing the tests for Hypotheses 1 and 2 using the evaluative questionnaire (EvalQ) instead of the leadership style measures (see the right column of Tables 9 and 10). In line with Hypothesis 3, value alignment predicted the evaluative questionnaire ($\beta = .10$, $p < .05$); this effect did not differ from the average effect of value alignment on the three leader measures ($\chi^2(1) = .12$, $p > .10$). Also in line with Hypothesis 3, the evaluative questionnaire predicted donations within conditions, and this effect was also not significantly different from that of the other styles ($\chi^2(1) = .44$, $p > .10$).

In addition, we found that jointly the control variables were significant predictors of the leadership styles and the EvalQ (Table 9). However, jointly the control variables did not have consistent effects on donations, predicting them only when the presumed speaker was President Clinton (Table 10).

Discussion of Study 3

In Study 3, we found value alignment between participants and a presumed former U.S. president to be a nonbehavioral antecedent of positive leadership style assessments, even when we kept leadership behavior constant. We again showed that leadership styles are, at least partially, an evaluative outcome (Hypothesis 1). Further replicating results from Studies 1 and 2, we showed that the leadership style measures were illusory predictors of donations (Hypothesis 2). Lastly, the pattern of results was similar for the three leadership styles and the EvalQ (Hypothesis 3).

Study 4

Unlike Studies 1 to 3, which tested all three propositions, Study 4 only tested Proposition 2 about causal illusions stemming from positive leadership styles and Proposition 3 about the empirical properties these styles share with a purely evaluative measure. The unique feature of Study 4 is its use of a manipulation that was behavioral and relevant to positive leadership styles or, more precisely speaking, to their behavioral component. We intended to isolate variance in positive leadership style ratings due to this manipulation from variance due to raters' evaluative idiosyncrasies. Thus, unlike Studies 1 to 3, Study 4 allowed us to simultaneously model the causal impact of the behavioral and evaluative components of leadership styles on objective outcomes. Finding an effect of this evaluative component, even in presence of relevant behavioral variation, would further undermine the traditional conception of positive leadership styles as purely behavioral constructs.

Study 4 also differed from Studies 1 to 3 in other ways. First, whereas in Studies 1 to 3 participants were observers of leadership or voters, in Study 4 participants assumed the role of followers who had to execute a task that a leader had given them. This change in participant role helped to check the robustness of our findings across types and levels of participant involvement. Second, the objective outcome variable was lying (not donations as in the previous studies). Third, we also included a measure of transformational leadership to test whether our previous findings generalize to this positive leadership style too. We tested the following hypotheses:

Hypothesis 1. *Rater idiosyncrasies in ratings of authentic, ethical, servant, and transformational leadership predict donations above and beyond variation in these ratings caused by leader behaviors.*

Hypothesis 2. *For the relationships specified in Hypothesis 1, ratings on the EvalQ behave in the same way as ratings of authentic, ethical, servant, and transformational leadership do.*

Method

Participants. We recruited 555 U.S.-based participants via Prolific Academic. Participants received a fixed remuneration of GBP 4.00, which corresponded to around USD 4.80 at the time of the experiment. They could additionally receive up to GBP 2.00, depending on their responses in a real-effort task (a matrix task adapted from Mazar, Amir, & Ariely, 2008). If participants lied, they could increase their payoff, although the leader asked them to not cheat. The study materials included four questions to test whether participants responded seriously. We restricted the final sample to the 543 participants (97.84%) who responded properly to at least three of the four questions. In the final sample, 258 participants were female (47.50%), and the average age was 40.25 years (SD = 13.21). We also collected data on participants' highest educational degree, their race, and their cognitive ability as control variables for participants' performance in the matrix task.

Procedure. The study consisted of four steps. First, participants read the instructions. Second, they were randomly assigned to watch a video of a leader who explained the task and asked participants not to cheat, developed by Banks et al. (2022). Half of the participants watched a version of the video in which the leader used more ethical leadership signals, and the other half watched a version with fewer ethical leadership signals. Third, participants responded to a set of positive leadership style measures to assess the leadership of the speaker in the video. Fourth, they completed the matrix task. In self-reporting their performance, participants could lie by overstating their performance, which would increase their income. Lying was an objective outcome measure.

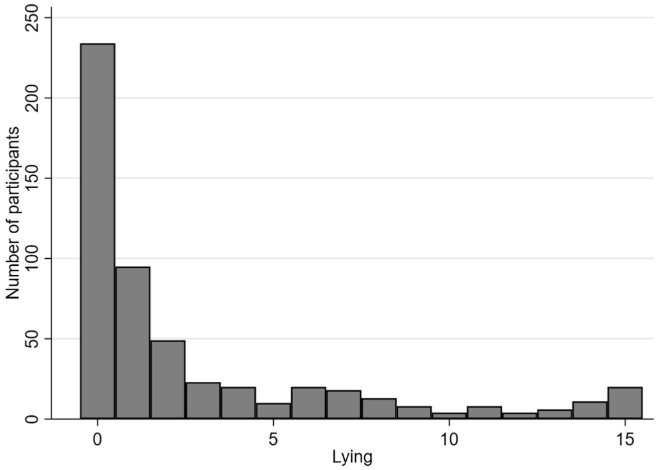

Fig. 2. The Empirical Distribution of the Outcome Variable Lying Behavior in Study 4.

Manipulation and measures

Manipulation of behaviors signaling ethical leadership. The behaviors signaling ethical leadership were embedded in videotaped instructions by a leader to the participants. To ensure that the two versions of the video differed only on the number of ethical leadership signals, we removed the part of the video from the high ethical signaling condition in which the leader suggested that participants could change their response, which could engender a confound (see the limitation noted by Banks et al., 2022; p. 12).

Lying. Participants had to work on 20 matrix tasks. Each matrix consisted of twelve numbers and participants had to indicate whether they found a pair of numbers adding up to exactly 10 or not (Mazar et al., 2008). For each matrix that participants reported as solved, we paid them GBP.10 in addition to their regular payment, meaning they could earn as much as an extra GBP 2.00. Thus, participants had a monetary incentive to report that they solved every matrix; about 3.5 % of participants did so. However, only five matrices were actually solvable. Hence, if participants responded that they had solved one or more of the 15 unsolvable matrices, they were lying or they had miscalculated; miscalculating is a measurement error that should be randomly

Table 11
Means, standard deviations, and intercorrelations of Study 4.

	Mean	S.D.	alpha	1	2	3	4	5	6	7	8	9	10
1. Lying	2.80	4.19		1.000									
2. Ethical behav. manip	0.53	0.50		−0.068 (0.116)	1.000								
3. Authentic leadership	3.72	0.65	.92	−0.062 (0.151)	0.014 (0.751)	1.000							
4. Ethical leadership	5.53	0.81	.90	−0.072 (0.094)	0.054 (0.206)	0.735 (0.000)	1.000						
5. Servant leadership	5.07	0.92	.84	−0.058 (0.181)	−0.073 (0.091)	0.695 (0.000)	0.699 (0.000)	1.000					
6. Transformational leadership	4.28	0.81	.92	−0.008 (0.851)	−0.080 (0.061)	0.707 (0.000)	0.699 (0.000)	0.754 (0.000)	1.000				
7. Evaluative questionnaire	3.63	0.71	.88	0.004 (0.918)	0.002 (0.971)	0.517 (0.000)	0.502 (0.000)	0.538 (0.000)	0.602 (0.000)	1.000			
8. Male	0.52	0.50		−0.032 (0.462)	0.021 (0.620)	−0.018 (0.683)	0.000 (0.992)	0.017 (0.686)	−0.076 (0.075)	−0.037 (0.390)	1.000		
9. Age	40.25	13.22		−0.169 (0.000)	−0.005 (0.899)	0.126 (0.003)	0.011 (0.797)	−0.043 (0.313)	0.000 (0.995)	0.036 (0.406)	−0.083 (0.054)	1.000	
10. American	0.95	0.21		−0.033 (0.439)	−0.057 (0.184)	0.106 (0.014)	0.077 (0.071)	0.082 (0.056)	0.058 (0.177)	0.043 (0.320)	0.046 (0.288)	0.019 (0.654)	1.000

Note. For ease of illustration, we do not include the categorical control variables of raters' cognitive ability, educational background, and race. Male = 1 (else = 0). Correlations greater than |.085| are significant at $p < .05$; correlations greater than |.111| are significant at $p < .01$; correlations greater than |.141| are significant at $p < .001$.

Table 12
Illusory Causation in Study 4 (i.e., Propositions 2 & 3).

	(1)	(2)	(3)	(4)	(5)
	Lying	Lying	Lying	Lying	Lying
Ethical behavior manipulation	−0.26*	−0.25	−0.26	−0.25	−0.24
	(-1.97)	(-1.89)	(-1.92)	(-1.87)	(-1.83)
Authentic leadership	−0.21*				
	(-2.10)				
Ethical leadership		−0.15*			
		(-1.98)			
Servant leadership			−0.16*		
			(-2.37)		
Transformational leadership				−0.12	
				(-1.42)	
Eval. questionnaire					−0.13
					(-1.40)
Age	−0.02**	−0.02**	−0.02**	−0.02**	−0.02**
	(-4.17)	(-4.37)	(-4.54)	(-4.39)	(-4.42)
Male	−0.04	−0.05	−0.03	−0.05	−0.04
	(-0.33)	(-0.35)	(-0.20)	(-0.38)	(-0.28)
American	−0.33	−0.33	−0.37	−0.34	−0.36
	(-1.03)	(-1.02)	(-1.13)	(-1.05)	(-1.11)
Cognitive ability	Included	Included	Included	Included	Included
Education	Included	Included	Included	Included	Included
Race	Included	Included	Included	Included	Included
Constant	3.44**	3.53**	3.61**	3.57**	3.57**
	(5.43)	(5.51)	(5.56)	(5.51)	(5.55)
F-test for all controls	F(15)=105.12***	F(15)=106.18***	F(15)=112.26***	F(15)=107.85***	F(15)=109.23***
Observations	543	543	543	543	543
R-squared	.099	.099	.102	.003	.096

Robust *z*-statistics in parentheses; ** p < 0.01, * p < 0.05.
Note. Cognitive ability, education, and race are composed of several dummy variables. The regressors for the leadership style measures are the residuals, as a function of the ethical behavior manipulation. Note, when bootstrapping standard errors (Cameron & Trivedi, 2009), the *z*-statistic for the residualized rated leader behavior is −1.97, −1.79, −2.13, −1.31, and −1.28 respectively. Thus, substantive inferences remained unchanged.

distributed across the experimental conditions and thus be orthogonal to the treatment effect. The lying score could range from 0 to 15.

Leadership. We measured authentic, ethical, and servant leadership using the same questionnaires as in Studies 1 to 3, and we used the same evaluative questionnaire. In addition, we measured transformational leadership with the transformational leadership inventory (TLI) on its original seven-point Likert scale ranging from strongly disagree to strongly agree (Podsakoff, MacKenzie, Moorman, & Fetter, 1990) to test if an additional leadership style followed the same pattern of results.

Control variables. We controlled for participant age, education (from below high school to Ph.D.), race (based on the U.S. census categories), gender, and citizenship (U.S. or other). We also controlled for cognitive ability because it might affect participants' responses in the matrix task. To do so, we used five exercises adapted from the cognitive reflection test of Primi, Morsanyi, Chiesi, Donati, and Hamilton (2016).

Results

Descriptive statistics. Table 11 shows the means, standard deviations, reliabilities, and correlations among the variables. In addition, Fig. 2 shows the distribution of lying as our dependent variable, which had an unexpectedly high number of zeros (i.e., participants who did not lie). To ensure that the models' distributional assumptions were in line with the actual data structure, we used negative binomial regression.

Recall check. We checked whether the manipulation of behaviors signaling ethical leadership (Banks et al., 2022) increased responses on the ethical leadership scale (Brown et al., 2005). The effect was insignificant ($\beta = .10,\ p > .10$). The effect of the manipulation was marginally significant for transformational leadership (Podsakoff et al., 1990) and insignificant for the other leadership measures. Normally a failed recall check would be a cause for alarm. However, the failed check is not a concern because our interest is in simultaneously studying the causal impact of the behavioral and evaluative components of leadership styles on objective outcomes, and not whether the ethical leadership manipulation changes leadership style ratings.

Hypothesis testing. Given the very high dispersion in the count variable with a variance much higher than the mean, we used negative binomial regression (Long & Freese, 2006). We found support for Hypothesis 1 (i.e., causal illusions) for authentic, ethical, and servant leadership, but not for transformational leadership (see Table 12; $\beta = -.21,\ p < .05,\ \beta = -.15,\ p < .05,\ \beta = .16,\ p < .05,\ \beta = -.12,\ p > .10$ respectively). That is, the parts of authentic, ethical, and servant leadership styles that are unrelated to the manipulation and thus unrelated to any behavioral variation spuriously predicted lying. However, we did not find support for Hypothesis 2 because the evaluative questionnaire did not—unlike authentic, ethical, and servant leadership—predict lying (see Table 12; $\beta = -.13,\ p > .10$).

In addition, we found that the manipulation of behaviors signaling ethical leadership significantly reduced lying only in the model in which authentic leadership was the independent variable ($\beta = -.26,\ p < .05$), marginally significantly reduced lying in the models in which ethical ($\beta = -.25, p < .10$) and servant leadership ($\beta = -.25,\ p < .10$) were the independent variable, and did not significantly reduce lying in the models in which transformational leadership ($\beta = -.12,\ p < .10$) and the EvalQ ($\beta = -.13,\ p < .10$) was the independent variable.

Discussion of Study 4

Study 4 extends the findings of Studies 1 to 3 in two ways. First, even in the presence of explicitly manipulating and modeling variation in leader behavior, we find that the relationship between authentic, ethical, and servant leadership style ratings and objective outcomes is largely driven by rater-level idiosyncratic evaluations. Second, Study 4 largely constructively replicates the findings of causal illusion (Proposition 2) from Studies 1 to 3 with a different dependent variable (lying instead of making donations) and with a different participant role (follower instead observer). However, a replication of the findings on Proposition 3 from Studies 1 to 3 was not possible because the EvalQ did not predict lying.

It is notable that in Study 4 the effect of the evaluative component is clearer than that of the behavioral component. The behavioral manipulation was only marginally significant for ethical and servant leadership. Hence, we do not replicate the results of Banks et al. (2022), possibly in part due to the slight adjustments we made to the manipulation, as mentioned in the methods section. The evaluative component of the authentic, ethical, and servant leadership styles, however, significantly predicted lying. The effects for transformational leadership and for the evaluative questionnaire had the same directionality but were nonsignificant. We can only speculate about these nonsignificant effects. For example, null findings for the evaluative questionnaire might have resulted from Study 4′s setting that differed considerably from those of Studies 1 to 3. More importantly, however, Study 4 reaffirmed that authentic, ethical, and servant leadership measures create causal illusions.

General discussion

Our four empirical studies offer two related insights (see Table 1 for an overview of the studies and their findings): positive leadership styles are conflated constructs, and these styles produce causal illusions. Regarding the first insight, positive leadership styles are conflated constructs that might be partially behavioral, but in large part represent positive summary evaluations of leader behaviors and other leader properties. Stated differently, leadership styles are likely affected by leader behaviors, but whether leaders have, for example, an ethical leadership style is not an objective leader property but a subjective evaluation through the eye of the beholder. We found that evaluative, nonbehavioral antecedents cause meaningful variation in positive leadership styles even when leader behaviors do not change. Moreover, measures of positive leadership styles have empirical properties that are in large part similar to those of an entirely evaluative (and entirely nonbehavioral) questionnaire. These findings validate Fischer and Sitkin's (2023) concerns about description-evaluation conflation and cause-effect mingling, as well as our observation that positive leadership styles are mixed leader-rater constructs.

Our second insight serves as an answer to an argumentation that supporters of leadership styles may adopt: if leadership is in the eye of the beholder, then using observer ratings is simply the correct measured choice and we ought to continue "business as usual." However, because subjective evaluations of leadership styles are outcomes themselves, using these styles as independent variables to predict other outcomes is a futile exercise that can only produce causal illusions. These causal illusions stem from two sources: misinterpreting leadership styles as purely behavioral constructs, and ignoring third variables (e.g., information about a leader's past performance or leader–follower value alignment) that can causally affect both positive leadership styles and other outcomes.

Taken together, our research refutes the key assumption of past research that positive leadership styles are behavioral constructs. Instead, past research has unknowingly produced and leveraged conceptual conflation of behavioral descriptions and evaluations that goes beyond perception-based measurement bias (see also Fischer, 2023). Our four studies also invalidate evidence suggesting that positive leadership styles have meaningful causal effects. At first glance, our findings and insights are sobering, because they point to fundamental conceptual and methodological weaknesses in past research, thus knocking past wisdom about positive leadership styles off its foundation. At second glance, however, our findings are a starting point for advancing knowledge about leadership styles. For Popper (1959), scientific progress rested on treating theories as preliminarily true until their refutation. Refutations of theories are instances of learning that lead to the generation and testing of new conjectures and, eventually, to the construction of revised theories. Having provided evidence for refuting past knowledge on positive leadership styles, we now turn to discussing implications for improving future research.

Implications for future research

To address the conflated nature of positive leadership style constructs and disentangle the effects of leader behaviors and their evaluations, future research could proceed along three lines: (1) leader behaviors as unique constructs; (2) follower evaluations as unique outcome constructs; and (3) leadership styles as configurations. For each of these lines, we offer exemplary *meta*-theoretical frameworks and past studies as methodological best practices.

Studying leader behaviors. For decades, the search for the most effective leader behaviors has been a dominant research question (Fleishman et al., 1991; Yukl, 2012). In these efforts, signaling theory has recently emerged as a theoretical foundation (Antonakis, Bastardoz, Jacquart, & Shamir, 2016; Banks et al., 2021a; Ernst et al., 2021; Fest, Kvaløy, Nieken, & Schöttner, 2021; Meslec et al., 2020). Signaling theory is a *meta*-theoretical framework with a long tradition in various natural and social science disciplines, and it is now increasingly making inroads into management research (for an overview, see Connelly, Certo, Ireland, & Reutzel, 2011), in particular at the organizational level (e.g., Dorobantu, Henisz, & Nartey, 2017; Kovács & Sharkey, 2014), but recently also at the individual level (e.g., Antonakis et al., 2016; Banks et al., 2021a).

At the individual level, a signaling theory lens can address the question of how leader behaviors influence follower evaluations—that is, how different behaviors and features of leaders' signals influence the interpretation of these signals by followers (e.g., Westphal, Park, McDonald, & Hayward, 2012). Most management applications focus on signals that are costly, be it directly (e.g., in recruiting) or indirectly (e.g., acquiring certain communication skills), and credibility is essential for the effectiveness of signals (Connelly et al., 2011). Costly signals increase the likelihood that receivers will take them seriously. A line of leadership research that has adopted a signaling lens is charismatic signaling. According to Antonakis et al. (2016), communicating charismatically requires the use of symbolic tools such as metaphors, which are not easy to produce, and requires intelligence and expertise. Thus, the signal (e.g., metaphor) carries information about the signaler (e.g., skills that are costly or hard to acquire); evidence shows indeed that charismatic signaling is correlated with general intelligence (Akstinaite, Jensen, Vlachos, Erne, & Antonakis, 2022). Further features that matter for interpretation are, for instance, the signaler's credibility and the receiver's attention (Connelly et al., 2011). That is, costly signals from credible signalers to attentive receivers have a high likelihood of being effective.

Testing such hypotheses requires measures that separate behaviors and their evaluations. Following the suggestion of Fischer (2023), one potential pathway for future research is to improve existing questionnaires by making them more descriptive and less evaluative of leadership behaviors. First, it is feasible to drop both entirely evaluative items such as "can be trusted" (Brown et al., 2005; p. 125) and conflated items that are difficult to reformulate. One such latter item is "Makes fair and balanced decisions" (Brown et al., 2005; p. 125), which specifies the behavior ("Makes […] decisions") by its positive connotation (i.e., "fair and balanced"). Second, other conflated items might be reworded more descriptively. An example would be a reformulation of "Is willing to admit mistakes when they are made" (Walumbwa et al., 2008; p. 121) to "Admits to mistakes." This reformulation removes the judgment call as to whether the focus is on the leader's intention (i.e., willingness) or the act (i.e., admitting mistakes). It is important to note, however, that the exercise of removing evaluative connotations from leadership questionnaire items does not fully eliminate the risk that responses are partially driven by respondent attributes (for overviews, see Fischer, 2023; Hansbrough et al., 2015). For example, respondent attributes can influence the interpretation of the item “admits to mistakes” such that respondents might disagree in what constitutes a mistake. Thus, in studying charismatic signaling, Antonakis et al. (2016) advocate the use of objectively coded charismatic signals, instead of questionnaires; these signals can also be manipulated.

A particularly promising alternative pathway for future research is behavior-based field experimentation with leadership styles. Lewin, Lippitt, and White (1939) in fact pioneered such research more than 80 years ago, and Eden (2020) synthesized recent progress on field experimentation in leadership research, offering guidance for scholars in line with the Lewinian tradition. In addition, scholars can use objective coding schemes for studying leadership. Macro-oriented leadership scholars have used coding schemes for behaviors of high-level leaders (e.g., König, Mammen, Luger, Fehn, & Enders, 2018; Stam, Van Knippenberg, Wisse, & Nederveen Pieterse, 2018), and micro-oriented leadership scholars have used objective coding schemes for interactions (e.g., Gerpott, Lehmann-Willenbrock, Voelpel, & Van Vugt, 2019). Measuring physiologically grounded behaviors—for example, leaders’ eye movements—is possible too (e.g., Maran, Furtner, Liegl, Kraus, & Sachse, 2019).

Studying follower evaluations. This topic is important in its own right. Although leader behaviors are a key component of leadership, the effectiveness of such behaviors rests on how followers evaluate these behaviors (Banks et al., 2021a; Meindl, 1995). There are many examples of variation in evaluations of the same behaviors. For instance, research on gender role theory has shown that women are evaluated less favorably than men when they display the same assertive behaviors (Eagly & Karau, 2002). Research using information processing perspectives has demonstrated variability in evaluations of jobs (Salancik & Pfeffer, 1978) and behaviors depending on various cues and outcomes (Lord & Maher, 1994). Scholars in the romance of leadership tradition have studied how media evaluations construe leadership images beyond actual behaviors (Chen & Meindl, 1991).

Attributional or evaluative perspectives therefore have a long tradition (Calder, 1977; Meindl, Ehrlich, & Dukerich, 1985) and remain prominent in leadership research (Bligh, Kohles, & Pillai, 2011). Attributional views have mainly been relatively mechanistic (e.g., Kelley & Michela, 1980), because they attribute the most highly co-varying causes to behaviors (Martinko, Harvey, & Douglas, 2007). A more recent version of attribution theory, by contrast, includes evaluators’ intentionality as part of the attribution process (Malle & Knobe, 1997). Malle’s (1999, 2021) general framework offers *meta*-theoretical guidance: behaviors are not only evaluated as good or bad because they are associated with good or bad outcomes but also because people ascribe good or bad underlying motives. This framework allows scholars to examine how observers come to call a person an effective or ethical leader—a question that has long been at the heart of leadership research (e.g., Meindl & Ehrlich, 1987; Treviño, Hartman, & Brown, 2000). Methodologically, such an approach can use questionnaires, as long as these questionnaires only measure attributions or evaluations as outcomes, and not behaviors as predictors of leadership outcomes (Fischer, 2023). Moreover, these evaluations should not be modeled as causes of other variables unless their endogeneity can be fully accounted for.

Studying configurations of leadership styles. Disentangling leadership styles into leader behaviors and follower evaluations and studying them separately is likely to generate many insights. Advancing knowledge about leadership, however, also requires the interplay of behaviors and evaluations to be understood. To this end, one potential *meta*-theoretical framework is configurational theory, which has a long tradition in organization theory and strategic management research (e.g., Miller, 1987, 1996). Katz and Kahn (1978) identified a configurational approach as a promising avenue for studying behaviors in organizations, including leadership. More recently, Van Knippenberg and Sitkin (2013) and Fischer and Sitkin (2023) suggested configurational theorizing to advance leadership research.

At the heart of configurational theorizing is causal complexity, which manifests itself in two ways: (a) certain behaviors or practices have joint rather than separate effects; and (b) certain behaviors thus co-occur much more frequently with each other than do others (cf. Miller, 1987; Miller, 1996). Consequently, there might be certain sets, or patterns, of leadership behaviors that lead to more favorable follower evaluations if these behaviors co-occur. Because leadership styles are defined as patterns of behaviors (Bass & Bass, 2008), these styles can be seen as configurations. However, to become configurational, positive leadership style research would have to specify (a) what the single constituent components (including behaviors, evaluations, and outcomes) are to avoid conflation; and (b) how these components jointly form a meaningful pattern, or configuration, to ensure conceptual integration (Fischer & Sitkin, 2023; for an example see Cardinal, Sitkin, & Long, 2010). Moreover, future configurational research on leader behaviors might benefit from recent progress in studying patterns of leader traits (Foti, Bray, Thompson, & Allgood, 2012) and from the previously mentioned advances in cleanly measuring behaviors and evaluations (Eden, 2020; Tur, Harstad, & Antonakis, 2021). Finite mixture modeling (McLachlan & Peel, 2004) or machine learning (LeCun, Bengio, & Hinton, 2015) could help to uncover these patterns.

Practical implications

Our studies draw attention to three practically relevant themes. First, past research’s claims that authentic, ethical, and servant-like behaviors lead to positive outcomes lack a solid empirical foundation, rendering these claims speculative. Second, leadership styles are not leader behaviors per se but rather a mix of what leaders do and how followers evaluate leadership. Thus, previous research has not been able to isolate concrete behaviors that practitioners could learn in training and then enact to convey an authentic, ethical, or servant leadership style. Third, our studies do not examine the logic that good deeds lead to good outcomes (“do-good logic”; Fischer & Sitkin, 2023). This logic might be valid (as might be its opposite version), but neither our studies nor past questionnaire-based leadership style research speaks to that. Taken together, our studies warn practitioners that lots of evidence on the effectiveness of positive leadership styles is likely driven by causal illusions and thus unwarranted, casting doubt also on large parts of presumed wisdom about evidence-based leadership development (Leroy et al., 2022).

Limitations and generalizability of our work

We identify five important limitations of our work. First, our four experiments could not test differences in predictive strength of the behavioral-descriptive and evaluative components. In Studies 1 to 3 we kept leadership behaviors constant, and it would be premature to rely on only one study (i.e., Study 4) and its findings to draw inferences about the relative predictive strength of actual behaviors versus subjective evaluations of these behaviors. However, our experimental study design demonstrates that positive leadership styles conflate behaviors and evaluations and that these styles can produce causal illusions.

Second, although we find very similar results in testing our propositions across the three positive leadership styles, the reported consistency of these findings might be inflated due to our experimental design. In Studies 1, 2, and 4, raters were exposed to rather limited information about the leader to ensure experimental control over participants’ information environment. However, the advantage of experimental control comes with the limitation that participants have comparatively little information about the leader. The relative lack of information, in turn, makes it harder for participants to discriminate whether the style of a leader is authentic, ethical, or servant-like. Stated differently, our study design likely contributed to the low discriminant validity among the leadership style measures, which, in turn, might have strengthened the pattern of findings. Nevertheless, it should be noted that low discriminant validity between leadership styles is not unique to our studies but commonplace in field settings too because leadership styles overlap in their descriptions of behaviors (Banks et al., 2018; Hoch et al., 2018). More importantly, Fischer and Sitkin (2023) outline how low discriminant validity might be further reinforced by an overlap in the evaluative aspects of leadership styles. Even more importantly, our main findings hold irrespective of concerns about discriminant validity, because each style is conflated and produces causal illusions.

Third, testing Proposition 3 (“similar predictive properties of the EvalQ and leadership styles”) came with common method bias due to the use of questionnaires. Regarding the testing of Propositions 1 and 2, however, our study design rules out common method bias because we tested either the effect of manipulations on leadership style ratings (Proposition 1) or the supposed effect of leadership style ratings on objective outcomes (Proposition 2). The manipulations and objective outcomes were not questionnaire based, meaning that independent and dependent variables were measured using different methods.

Fourth, our four experiments are only partial explanations of how raters evaluate leadership. We have established the existence of conflations and causal illusions in leadership style research. Yet our research only touches on “how” raters form evaluations of leadership that correlate with and spuriously predict objective outcomes. For example, in Studies 1 and 2 we found that certain personality traits influence leadership style ratings. The “how” question presents an evident opportunity for future research.

Fifth, our research is only a definitive demonstration of conflated measurement and causal illusions for the three selected leadership styles and their measures, and not for other leadership styles. Nevertheless, the underlying conceptual rationale extends to other positive and also negative leadership style constructs (see, e.g., Fischer, Tian, Lee, & Hughes, 2021), and even to other organizational behavior constructs with an evaluative connotation. The valence of lthem. Hence, we expect that conflations and causal illusions can be found for other positive leadership styles, such as empowering leadership, as well as for negative leadership styles, such as abusive supervision or destructive leadership. We also expect our reasoning to hold for value-laden nonleadership constructs, such as organizational citizenship behavior and counterproductive workplace behavior, when used as predictor variables.

Conclusion

The purpose of our article was to examine whether positive leadership styles such as authentic, ethical, and servant leadership are truthful representations of leader behaviors. We argued and found that these styles conflate leader behaviors with follower evaluations. Therefore, we uncovered a fatal flaw. This fatal flaw renders claims that these styles cause leadership outcomes obsolete because leadership styles conjure false correlations between behaviors and outcomes.

Our findings have important implications for science and practice. Researchers must disentangle conflated positive leadership style concepts, and practitioners are not likely to get what they are promised when investing in allegedly evidence-based authentic, ethical, or servant leadership training. It is clear from our work that current positive leadership style constructs and measures are problematic—if not even damaging—to science and practice because they can produce misleading findings. Clean conceptualizations and purely behavioral measures must be developed. Future leadership style research must separate leader behaviors from their evaluation, and we hope that our article serves as a foundation for such research. As is clear from our findings, a radical reorientation is the order of the day to clean up the current mess of positive leadership style constructs and measures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and replication material can be found here: https://osf.io/hjbqt/?view_only=3892c09a9a224fd2967be10f79be067f.

Appendix A. Prominent definitions of the three positive leadership styles

	Influential definitions
Authentic leadership	Authentic leadership “draws upon and promotes both positive psychological capacities and a positive ethical climate, to foster greater self-awareness, an internalized moral perspective, balanced processing of information, and relational transparency on the part of leaders working with followers, fostering positive self-development” (Walumbwa et al., 2008; p. 94).
Ethical leadership	Ethical leadership is “the demonstration of normatively appropriate conduct through personal actions and interpersonal relationships, and the promotion of such conduct to followers through two-way communication, reinforcement, and decision-making” (Brown et al., 2005; p. 120).
Servant leadership	Servant leaders focus “on developing employees to their fullest potential in the areas of task effectiveness, community stewardship, self-motivation, and future leadership capabilities” (Liden et al., 2008; p. 162). (Liden et al., 2008) do not offer a more formal definition but put forward seven dimensions that characterize servant leaders: conceptual skills, empowerment, helping subordinates grow and succeed, putting subordinates first, behaving ethically, emotional healing, and creating value for the community.

Remark. The three definitions are in line with the notion that leadership styles are patterns of leader behaviors. The measures used in our three studies operationalize the constructs as defined above.

Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.leaqua.2023.101771.

References

Akstinaite, V., Jensen, U. T., Vlachos, M., Erne, A., & Antonakis, J. (2022). *Charisma is a costly signal*. Mykonos, Greece: Interdisciplinary Perspectives on Leadership Symposium.

Alvesson, M. (2020). Upbeat leadership: A recipe for–or against–“successful” leadership studies. *The Leadership Quarterly, 31*(6), Article 101439.

Alvesson, M., & Einola, K. (2019). Warning for excessive positivity: Authentic leadership and other traps in leadership studies. *The Leadership Quarterly, 30*(4), 383–395.

Angrist, J. D., & Pischke, J.-S. (2008). ***Mostly harmless econometrics: An empiricist’s companion***. Princeton University Press.

Antonakis, J., d’Adda, G., Weber, R. A., & Zehnder, C. (2022). “Just words? Just speeches?” On the economic value of charismatic leadership. ***Management Science**, 68*(9), 6355–6381.

Antonakis, J. (2017). Charisma and the “new leadership”. In J. Antonakis, & D. V. Day (Eds.), ***The nature of leadership*** (3rd ed., pp. 56–81). Thousand Oaks: Sage Publications.

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An Ill-Defined and Ill-Measured Gift. ***Annual Review of Organizational Psychology and Organizational Behavior, 3***(1), 293–319.

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. ***Current Directions in Psychological Science, 26***(3), 270–275.

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can Charisma Be Taught? Tests of Two Interventions. ***The Academy of Management Learning and Education, 10***(3), 374–396.

Ashford, S. J., & Sitkin, S. B. (2019). *From problems to progress: A dialogue on prevailing issues in leadership research*. The Leadership Quarterly.

Avolio, B. J., Griffith, J., Wernsing, T. S., & Walumbwa, F. O. (2010). What is authentic leadership development? ***Journal of Positive Psychology, 4***, 39–50.

Banks, G., Fischer, T., Gooty, J., & Stock, G. (2021a). Ethical leadership: Mapping the terrain for concept cleanup and a future research agenda. ***The Leadership Quarterly, 32***, Article 101471.

Banks, G. C., Gooty, J., Ross, R. L., Williams, C. E., & Harrington, N. T. (2018). Construct redundancy in leader behaviors: A review and agenda for the future. ***The Leadership Quarterly, 29***(1), 236–251.

Banks, G., McCauley, K., Gardner, W., & Guler, C. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly, 27*(4), 634–652.

Banks, G., Woznyj, H. M., & Mansfield, C. A. (2021b). Where is “behavior” in organizational behavior? A call for a revolution in leadership research and beyond. ***The Leadership Quarterly, 101581***.

Banks, G. C., Ross, R., Toth, A. A., Tonidandel, S., Goloujeh, A. M., Dou, W., & Wesslen, R. (2022). The triangulation of ethical leader signals using qualitative, experimental, and data science methods. ***The Leadership Quarterly, 101658***.

Bass, B. M., & Bass, R. (2008). ***The Bass handbook of leadership: Theory, research, and managerial applications***. New York, NY: Simon and Schuster.

Baum, C. F., Schaffer, M. E., & Stillman, S. 2010. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. ***http://ideas.repec.org/c/boc/bocode/s425401.html***.

Bligh, M. C., Kohles, J. C., & Pillai, R. (2011). Romancing leadership: Past, present, and future. ***The Leadership Quarterly, 22***(6), 1058–1077.

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. ***Organizational Behavior and Human Decision Processes, 97***(2), 117–134.

Calder, B. J. (1977). An attribution theory of leadership. ***New Directions in Organizational Behavior, 179***, 204.

Cameron, A. C., & Trivedi, P. K. (2009). ***Microeconometrics Using Stata***. College Station, Tex.: Stata Press.

Cardinal, L. B., Sitkin, S. B., & Long, C. P. (2010). A configurational theory of control. In S. B. Sitkin, L. B. Cardinal, & K. Bijlsma-Frankema (Eds.), ***Organizational control*** (Vol. 51, pp. 85–100). New York: Cambridge University Press.

Caudill, S. B. (1988). An Advantage of the Linear Probability Model over Probit or Logit. ***Oxford Bulletin of Economics and Statistics, 50***(4), 425–427.

Chen, C. C., & Meindl, J. R. (1991). The construction of leadership images in the popular press: The case of donald burr and people express. ***Administrative Science Quarterly, 36***(4), 521–551.

Chin, M. K., Hambrick, D. C., & Treviño, L. K. (2013). Political ideologies of CEOs: The influence of executives’ values on corporate social responsibility. ***Administrative Science Quarterly, 58***(2), 197–232.

Chiniara, M., & Bentein, K. (2016). Linking servant leadership to individual performance: Differentiating the mediating role of autonomy, competence and relatedness need satisfaction. ***The Leadership Quarterly, 27***(1), 124–141.

Chiniara, M., & Bentein, K. (2018). The servant leadership advantage: When perceiving low differentiation in leader-member relationship quality influences team cohesion, team task performance and service OCB. ***The Leadership Quarterly, 29***(2), 333–345.

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.

Cooper, C. D., Scandura, T. A., & Schriesheim, C. A. (2005). Looking forward but learning from our past: Potential challenges to developing authentic leadership theory and authentic leaders. ***The Leadership Quarterly, 16***(3), 475–493.

Davidson, R., & MacKinnon, J. G. 1993. ***Estimation and inference in econometrics***: Oxford New York.

De Luque, M. S., Washburn, N. T., Waldman, D. A., & House, R. J. (2008). Unrequited profit: How stakeholder and economic values relate to subordinates’ perceptions of leadership and firm performance. ***Administrative Science Quarterly, 53***(4), 626–654.

Den Hartog, D. N. (2015). Ethical leadership. ***Annual Review of Organizational Psychology and Organizational Behavior, 2***(1), 409–434.

Dinh, J. E., Lord, R. G., Gardner, W. L., Meuser, J. D., Liden, R. C., & Hu, J. (2014). Leadership theory and research in the new millennium: Current theoretical trends and changing perspectives. ***The Leadership Quarterly, 25***(1), 36–62.

Dorobantu, S., Henisz, W. J., & Nartey, L. (2017). Not all sparks light a fire: Stakeholder and shareholder reactions to critical events in contested markets. ***Administrative Science Quarterly, 62***(3), 561–597.

Eagly, A. H., & Chaiken, S. (1993). ***The psychology of attitudes***. Harcourt Brace Jovanovich College Publishers.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. ***Psychological Review, 109***(3), 573.

Eden, D. (2020). The science of leadership: A journey from survey research to field experimentation. ***The Leadership Quarterly, 101472***.

Ehrhart, M. G. (2004). Leadership and procedural justice climate as antecedents of unit-level organizational citizenship behavior. ***Personnel Psychology, 57***(1), 61–94.

Ernst, B. A., Banks, G. C., Loignon, A. C., Frear, K. A., Williams, C. E., Arciniega, L. M., Gupta, R. K., Kodydek, G., & Subramanian, D. (2021). Virtual charismatic leadership and signaling theory: A prospective meta-analysis in five countries. ***The Leadership Quarterly, 101541***.

Felfe, J., & Schyns, B. (2006). Personality and the perception of transformational leadership: The impact of extraversion, neuroticism, personal need for structure, and occupational self-efficacy. ***Journal of Applied Social Psychology, 36***(3), 708–739.

Fest, S., Kvaløy, O., Nieken, P., & Schöttner, A. (2021). How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. ***The Leadership Quarterly***.

Fischer, T. (2023). Measuring behaviors counterfactually. ***The Leadership Quarterly***.

Fischer, T., & Sitkin, S. B. (2023). Leadership styles: A comprehensive assessment and way forward. ***Academy of Management Annals, 17***(1), 331–372.

Fischer, T., Tian, A. W., Lee, A., & Hughes, D. J. (2021). Abusive supervision: A systematic review and fundamental rethink. ***The Leadership Quarterly, 101540***.

Fleishman, E. A., Mumford, M. D., Zaccaro, S. J., Levin, K. Y., Korotkin, A. L., & Hein, M. B. (1991). Taxonomic efforts in the description of leader behavior: A synthesis and functional interpretation. ***The Leadership Quarterly, 2***(4), 245–287.

Foti, R. J., Bray, B. C., Thompson, N. J., & Allgood, S. F. (2012). Know thy self, know thy leader: Contributions of a pattern-oriented approach to examining leader perceptions. ***The Leadership Quarterly, 23***(4), 702–717.

Gardner, W. L., Avolio, B. J., Luthans, F., May, D. R., & Walumbwa, F. (2005). “Can you see the real me?” A self-based model of authentic leader and follower development. ***The Leadership Quarterly, 16***(3), 343–372.

Gardner, W. L., Cogliser, C. C., Davis, K. M., & Dickens, M. P. (2011). Authentic leadership: A review of the literature and research agenda. ***The Leadership Quarterly, 22***(6), 1120–1145.

Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. ***American Statistician, 60***, 328–331.

Gerpott, F. H., Lehmann-Willenbrock, N., Voelpel, S. C., & Van Vugt, M. (2019). It’s not just what is said, but when it’s said: A temporal account of verbal behaviors and emergent leadership in self-managed teams. ***Academy of Management Journal, 62***(3), 717–738.

Gioia, D. A., & Sims, H. P., Jr (1985). On avoiding the influence of implicit leadership theories in leader behavior descriptions. ***Educational and Psychological Measurement, 45***(2), 217–232.

Hansbrough, T. K., Lord, R. G., & Schyns, B. (2015). Reconsidering the accuracy of follower leadership ratings. ***The Leadership Quarterly, 26***(2), 220–237.

Hmieleski, K. M., Cole, M. S., & Baron, R. A. (2012). Shared authentic leadership and new venture performance. ***Journal of Management, 38***(5), 1476–1499.

Hoch, J. E., Bommer, W. H., Dulebohn, J. H., & Wu, D. (2018). Do ethical, authentic, and servant leadership explain variance above and beyond transformational leadership? A meta-analysis. ***Journal of Management***, 0149206316665461.

Hu, J., & Liden, R. C. (2011). Antecedents of team potency and team effectiveness: An examination of goal and process clarity and servant leadership. ***Journal of Applied Psychology, 96***(4), 851–862.

Huang, F. L. (2019). Alternatives to logistic regression models in experimental studies. ***The Journal of Experimental Education***, 1–16.

Hunter, E. M., Neubert, M. J., Perry, S. J., Witt, L., Penney, L. M., & Weinberger, E. (2013). Servant leaders inspire servant followers: Antecedents and outcomes for employees and the organization. ***The Leadership Quarterly, 24***(2), 316–331.

Hunter, S. T., Bedell-Avers, K. E., & Mumford, M. D. (2007). The typical leadership study: Assumptions, implications, and potential remedies. ***The Leadership Quarterly, 18***(5), 435–446.

Kalshoven, K., Den Hartog, D. N., & De Hoogh, A. H. (2011). Ethical leadership at work questionnaire (ELW): Development and validation of a multidimensional measure. ***The Leadership Quarterly, 22***(1), 51–69.

Katz, D., & Kahn, R. L. 1978. ***The social psychology of organizations***: Wiley New York.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. ***Annual Review of Psychology, 31***(1), 457–501.

Kerr, S. (1975). On the folly of rewarding A, while hoping B. ***Academy of Management Journal, 18***(4), 769–783.

König, A., Mammen, J., Luger, J., Fehn, A., & Enders, A. (2018). Silver bullet or ricochet? CEOs’ use of metaphorical communication and infomediaries’ evaluations. ***Academy of Management Journal, 61***(4), 1196–1230.

Kovács, B., & Sharkey, A. J. (2014). The paradox of publicity: How awards can negatively affect the evaluation of quality. ***Administrative Science Quarterly, 59***(1), 1–33.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. ***Nature, 521***, 436–444.

Lehman, D. W., O’Connor, K., Kovács, B., & Newman, G. E. (2019). **Authenticity**. ***Academy of Management Annals, 13***(1), 1–42.

Lemoine, G. J., & Blum, T. C. (2021). Servant leadership, leader gender, and team gender role: Testing a female advantage in a cascading model of performance. ***Personnel Psychology, 74***(1), 3–28.

Lemoine, G. J., Hartnell, C. A., & Leroy, H. (2019). Taking stock of moral approaches to leadership: An integrative review of ethical, authentic, and servant Leadership. ***Academy of Management Annals, 13***(1), 148–187.

Leroy, H. L., Anisman-Razin, M., Avolio, B. J., Bresman, H., Stuart Bunderson, J., Burris, E. R., Claeys, J., Detert, J. R., Dragoni, L., & Giessner, S. R. (2022). Walking our evidence-based talk: The case of leadership development in business schools. ***Journal of Leadership & Organizational Studies, 29***(1), 5–32.

Levitis, D. A., Lidicker, W. Z., Jr, & Freund, G. (2009). Behavioural biologists do not agree on what constitutes behaviour. ***Animal Behaviour, 78***(1), 103–110.

Lewin, K., Lippitt, R., & White, R. K. (1939). Patterns of aggressive behavior in experimentally created "social climates". ***The Journal of Social Psychology, 10***(2), 269–299.

Liden, R. C., Wayne, S. J., Meuser, J. D., Hu, J., Wu, J., & Liao, C. (2015). Servant leadership: Validation of a short form of the SL-28. ***The Leadership Quarterly, 26***(2), 254–269.

Liden, R. C., Wayne, S. J., Zhao, H., & Henderson, D. (2008). Servant leadership: Development of a multidimensional measure and multi-level assessment. ***The Leadership Quarterly, 19***(2), 161–177.

Long, J. S., & Freese, J. (2006). ***Regression models for categorical dependent variables using Stata*** (2nd ed.). College Station, TX: StataCorp LP.

Lord, R. G., Binning, J. F., Rush, M. C., & Thomas, J. C. (1978). The effect of performance cues and leader behavior on questionnaire ratings of leadership behavior. ***Organizational Behavior and Human Performance, 21***(1), 27–39.

Lord, R. G., & Maher, K. J. (1994). ***Leadership and information processing: Linking perceptions and performance***. Psychology Press.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. ***Personality and Social Psychology Review, 3***(1), 23–48.

Malle, B. F. (2021). Moral judgments. ***Annual Review of Psychology, 72***, 293–318.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. ***Journal of Experimental Social Psychology, 33***(2), 101–121.

Maran, T., Furtner, M., Liegl, S., Kraus, S., & Sachse, P. (2019). In the eye of a leader: Eye-directed gazing shapes perceptions of leaders' charisma. ***The Leadership Quarterly, 30***(6), Article 101337.

Martinko, M., Mackey, J., Moss, S., Harvey, P., McAllister, C., & Brees, J. (2018). An exploration of the role of subordinate affect in leader evaluations. ***Journal of Applied Psychology***.

Martinko, M. J., Harvey, P., & Douglas, S. C. (2007). The role, function, and contribution of attribution theory to leadership: A review. ***The Leadership Quarterly, 18***(6), 561–585.

Mayer, D. M., Aquino, K., Greenbaum, R. L., & Kuenzi, M. (2012). Who displays ethical leadership, and why does it matter? An examination of antecedents and consequences of ethical leadership. ***Academy of Management Journal, 55***(1), 151–171.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. ***Journal of Marketing Research, 45***(6), 633–644.

McLachlan, G., & Peel, D. (2004). ***Finite mixture models***. John Wiley & Sons.

Meindl, J. R. (1995). The romance of leadership as a follower-centric theory: A social constructionist approach. ***The Leadership Quarterly, 6***(3), 329–341.

Meindl, J. R., & Ehrlich, S. B. (1987). The romance of leadership and the evaluation of organizational performance. ***Academy of Management Journal, 30***(1), 91–109.

Meindl, J. R., Ehrlich, S. B., & Dukerich, J. M. (1985). The Romance of Leadership. ***Administrative Science Quarterly, 30***(1), 78–102.

Merriam-Webster. 2023a. Style.

Merriam-Webster. 2023b. Behavior.

Merriam-Webster. 2023c. Evaluation, Vol. 2021.

Meslec, N., Curseu, P. L., Fodor, O. C., & Kenda, R. (2020). Effects of charismatic leadership and rewards on individual performance. ***The Leadership Quarterly, 101423***.

Miller, D. (1987). The genesis of configuration. ***Academy of Management Review, 12***(4), 686–701.

Miller, D. (1996). Configurations revisited. ***Strategic Management Journal, 17***(7), 505–512.

Ng, T. W. H., & Feldman, D. C. (2015). Ethical Leadership: Meta-Analytic Evidence of Criterion-Related and Incremental Validity. ***Journal of Applied Psychology, 100***(3), 948–965.

Pfeffer, J. (2015). ***Leadership BS: Fixing workplaces and careers one truth at a time***. New York, NY: Harper Collins.

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. ***The Leadership Quarterly, 1***(2), 107–142.

Popper, K. (1959). ***The logic of scientific discovery***. Routledge.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. ***Journal of Personality and Social Psychology, 67***(4), 741–763.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). ***Journal of Behavioral Decision Making, 29***(5), 453–469.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. ***Journal of Research in Personality, 41***(1), 203–212.

Salancik, G. R., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design. ***Administrative Science Quarterly***, 224–253.

Schriesheim, C. A., Castro, S. L., Zhou, X. T., & Yammarino, F. J. (2001). The folly of theorizing "A" but testing "B": A selective level-of-analysis review of the field and a detailed leader–member exchange illustration. ***The Leadership Quarterly, 12***(4), 515–551.

Shaver, J. M. (2020). Causal identification through a cumulative body of research in the study of strategy and organizations. ***Journal of Management, 46***(7), 1244–1256.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. ***Journal of Personality and Social Psychology, 113***(1), 117.

Stam, D., Van Knippenberg, D., Wisse, B., & Nederveen Pieterse, A. (2018). Motivation in words: Promotion-and prevention-oriented leader communication in times of crisis. ***Journal of Management, 44***(7), 2859–2887.

Todorov, A. (2017). ***Face Value: The Irresistible Influence of First Impressions***. Princeton University Press.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. ***Science, 308***(5728), 1623–1626.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. ***Annual Review of Psychology, 66***, 519–545.

Treviño, L. K., Hartman, L. P., & Brown, M. (2000). Moral person and moral manager: How executives develop a reputation for ethical leadership. ***California Management Review, 42***(4), 128–142.

Tur, B., Harstad, J., & Antonakis, J. (2021). Effect of charismatic signaling in social media settings: Evidence from TED and Twitter. ***The Leadership Quarterly, 101476***.

Van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? ***Academy of Management Annals, 7***(1), 1–60.

Walumbwa, F. O., Avolio, B. J., Gardner, W. L., Wernsing, T. S., & Peterson, S. J. (2008). Authentic leadership: Development and validation of a theory-based measure. ***Journal of Management, 34***(1), 89–126.

Walumbwa, F. O., Morrison, E. W., & Christensen, A. L. (2012). Ethical leadership and group in-role performance: The mediating roles of group conscientiousness and group voice. ***The Leadership Quarterly, 23***(5), 953–964.

Wang, G., Van Iddekinge, C. H., Zhang, L., & Bishoff, J. (2019). Meta-analytic and primary investigations of the role of followers in ratings of leadership behavior in organizations. ***Journal of Applied Psychology, 104***(1), 70–106.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. ***Journal of Personality and Social Psychology, 54***(6), 1063–1070.

Westphal, J. D., Park, S. H., McDonald, M. L., & Hayward, M. L. (2012). Helping other CEOs avoid bad press: Social exchange and impression management support among CEOs in communications with journalists. ***Administrative Science Quarterly, 57***(2), 217–268.

Yukl, G. (1999). An evaluation of conceptual weaknesses in transformational and charismatic leadership theories. ***The Leadership Quarterly, 10***(2), 285–305.

Yukl, G. (2012). Effective leadership behavior: What we know and what questions need more attention. ***Academy of Management Perspectives, 26***(4), 66–85.

Yukl, G. A. (2008). How leaders influence organizational effectiveness. ***The Leadership Quarterly, 19***(6), 708–722.