WHEN DO EFFECT SIZES *ACTUALLY* CAPTURE EFFECTS? A META-ANALYTIC REVIEW

by

Paul Matthew Amari

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Arts in
Industrial/Organizational Psychology

Charlotte

2023

Approved by:

_____
Dr. George Banks

_____
Dr. Janaki Gooty

_____
Dr. Eric Heggestad

ABSTRACT

PAUL MATTHEW AMARI. When Do Effect Sizes *Actually* Capture Effects? A Meta-Analytic Review. (Under the direction of DR. GEORGE BANKS)

Effect size benchmarks are used as guidelines for conducting power analysis and Bayesian analysis, guiding theory, interpreting practical significance, and reviewing scientific progress. However, effect size estimates that are correlational directly violate the definition of an "effect", as they do not capture a cause-and-effect relationship. The current work begins with a review of the current state of the literature and presents a continuum of causal-inference strength. Next, to demonstrate this conceptualization, a comprehensive review was conducted of the leadership literature: (1) a second-order meta-analysis of leader individual differences (total k = 1,829; total N = 640,388), (2) meta-analyzed lab and field experiments (total k = 110; total N = 18,402), and (3) a narrative review of effect sizes from quasi-experimental and non-traditional experimental designs. This work concludes with implications for theory and practice, future directions for research, and methodological best practices (e.g., experimental design).

Key words: Leadership, effect sizes, endogeneity bias, meta-analysis, second-order meta-analysis, individual differences, experiments, causal inference

ACKNOWLEDGEMENTS

I would like to thank my committee chair and advisor, Dr. George Banks, for his guidance, support, and invaluable experience and expertise throughout the duration of this project. I am honored and immensely grateful to have gotten the opportunity to work with and learn from such a great scholar that will continue to chart the path forward for the field of leadership. I also would like to thank my committee members, Dr. Janaki Gooty and Dr. Eric Heggestad, as well as Dr. Ernest O'Boyle. Their knowledge and expertise strengthened this project into being the best version possible, and for that I am grateful. Lastly, I would like to thank the members of my research team, Leah Bourque, Holly Holladay, Bushra Zaidi, and Rachel Liang, for their support and strong work ethic that they displayed throughout various phases of the project that played an integral role in its successful execution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PRISMA an acronym for Preferred Reporting Items for Systematic Reviews and Meta-Analyses

CHAPTER 1: INTRODUCTION

Traditionally scholars have relied on significance testing to advance theory and practice (Williams et al., 2020). However, the use of effect sizes has become an alternative solution that has grown in popularity in the past several decades in leadership research (American Psychological Association Task Force on Statistical Inference, 1999) and beyond (Kelley, 2013; Aguinis & Pierce, 2006). The use of Cohen's (1962, 1988) effect size benchmark estimates have been regarded with the utmost importance for conducting power analyses (Kruschke et al., 2012), interpreting practical significance (Aguinis et al., 2010; Ellis, 2010; Brooks et al., 2014), and reviewing scientific progress (Cohen, 1988; Cumming, 2012). A number of efforts have later advanced the utility of these original benchmarks (Bosco et al., 2015; Gignac & Szodorai, 2016). However, there are a host of limitations that render the majority of primary and meta-analytic effect sizes estimates as not meaningful to scholars because they are not causally identified, and thus do not truly capture effects due to endogeneity bias (e.g., Banks et al., 2016; 2018; Hoch et al., 2018; Judge & Piccolo, 2004; Rockstuhl et al., 2012).

Typical correlational benchmarks are still acceptable quantifications of relationships among variables, as long as they are interpreted and described as purely relational, and causality is not implied. That is, the term effect size should not be included as is commonly done in meta-analytic work (Borenstein et al., 2009; Schmidt & Hunter, 2015). Furthermore, these associations between variables only demonstrate early preliminary evidence to suggest a causal relationship. As such, there are many criteria needed for establishing a causal inference (Cook et al., 1979; Podsakoff and Podsakoff, 2019). In addition, endogeneity bias has rendered the correlation effect size benchmarks available to scholars as highly limited or often misleading (Antonakis et al., 2010). When correlations are influenced by sources of endogeneity bias, such as omitted

variables, common-method bias, or selection effects, they also include the effects of unmeasured causes (Hill et al., 2021). This means that the observed correlations are not causally identified (Wulff et al., 2023; Kennedy, 2008).

To address this gap, I present the first purely causal leadership effect size estimates to date. First, I begin with a review of the effect size literature, past work on benchmarks, and a number of methodological issues that present problems for interpreting effect sizes and their ability to adequately capture effects ($x \rightarrow y$). Then, a complementary visualization of a continuum is presented to demonstrate this as well as the strength of causal claims. Third, three systematic reviews of the leadership literature were conducted and presented to illustrate different types of causally identified effect sizes.

The leadership literature was chosen specifically due to its high level of suitability to serve as a strong illustrative example for the continuum of causal strength. That is, the leadership literature was deemed an area ripe for investigation in that it had a high number of field, lab, and quasi experiments, with many of these experimental designs naturally demonstrating a cause-and-effect relationship. This cause-and-effect nature is inherently assumed through the social influence process that can be seen when researchers manipulate leader behaviors and then measure follower outcomes (Yukl & Falbe, 1990; Furst & Cable, 2008; Oc & Bashshur, 2013). Furthermore, the leadership literature has shown a consistent increase in emphasis around making stronger causal inferences to aid in better theory building and refinement (Antonakis et al., 2010; Day & Antonakis, 2013; Podsakoff & Podsakoff, 2019; Güntner et al., 2020; Eden, 2021). Lastly, the results and implications from this paper can better inform thought leaders in organizations who directly influence or contribute to the minimum of $200 billion dollars spent annually on leader selection, assessment, and development in corporate America alone

(Moldoveanu & Narayandas, 2019). However, it is also important to note that the conversation put forth throughout this paper is far-reaching and extends well-beyond leadership. That is, many other research domains in organizational behavior could also benefit from establishing more robust effect size benchmarks based on our recommendations.

First, a second-order meta-analysis of leader individual differences was conducted and their effects on various outcomes (total k = 21,829; total N = 640,388). The decision to select individual differences was due to their exogenous nature as well the long history of leader individual differences investigated as major predictors of leader effectiveness (Lord et al., 1986; Judge et al., 2002; Hoffman et al., 2012; Derue et al., 2011). Second, a meta-analysis of effect sizes from lab and field experiments was completed to demonstrate effect size estimates that meet traditional "gold standards" (total k = 110; total N = 18,402). Lastly, a third review was completed in a narrative fashion to present examples of effect sizes from quasi-experimental and non-traditional experimental designs. This work concludes with implications for theory and practice, future directions for research, and methodological best practices (e.g., experimental design).

1.1 Defining Effect Sizes

Effect sizes are one of the primary quantifications of empirical results in the social sciences (Tukey, 1969). Cohen (1988) defined effect size as, "the degree to which the null hypothesis is false" (p. 9). Later, Kelley and Preacher (2012) broke effect sizes into three distinct facets (dimension, measure/index, value) and included ten resulting propositions that follow to make effect sizes more inclusive and adaptable. They explained that there are certain properties and types of effect sizes that may make them more or less important for specific circumstances. Examples of some consequences are that effect sizes can represent different values (samples or

populations), dimensions (unidimensional or multidimensional), or standardization (standardized vs. unstandardized). Flora's (2020) recent definition of effect size builds off of Kelley and Preacher's (2012) interpretation and is classified as the magnitude of the association between two or more variables.

Oxford's English dictionary defines an effect as, "a change which is a result or consequence of an action or other cause." This proposed definition of an effect implies that there is a causal mechanism (*x* causes *y*) in regard to effect sizes. Furthermore, Cook and colleagues (2002) stated in their seminal piece on drawing causal inferences that the terms "cause" and "effect" are partly dependent on another due to the causal nature embedded within their inherent definitions. Hence, the term "effect size" should include some degree of certainty regarding the magnitude of the influence and thus, should be causally identified.

1.2 Current Effect Size Benchmark Standards

Cohen (1988). Although Cohen's (1962, 1988) benchmarks for the field are useful and served as a major accomplishment in scientific progress at the time, researchers should bring into question the extent to which leadership research and organizational behavior more broadly, are *actually* capturing effects in the estimates reported. For instance, Hemphill (2003) argued that the minimum cutoff value of 0.30 is unrealistically high. Aguinis and colleagues (2010) even stated that Cohen's (1988) effect size benchmarks provide a false sense of security through objectivity and standardization for academics because they were mostly determined by his own subjective opinion. We also lack the information to discern whether his interpretation of a small effect size may be more meaningful and impactful for specific circumstances.

Some have also claimed that these benchmarks are commonly misjudged to the point where researchers underestimate the true value of effects (Cortina & Landis, 2009; McCartney &

Rosenthal, 2000; Rosenthal, 1990, Götz et al., 2022). One example of this can be seen with Ernst et al.'s (2021) prospective meta-analysis on charismatic leadership. The authors found an effect for charismatic leadership tactics (CLTs) on follower task performance. Although this was considered a "small effect," comparing this to the effect for CLTs on an outcome like follower vaccination rates is incomparable despite having the same magnitude ($r = 0.10$). A "small effect" on outcomes like follower vaccination rates or mortality rates may be more meaningful, thus potentially making it a "large effect" (McCartney & Rosenthal, 2000; Martin et al., 2021; Flora, 2020, Götz et al., 2022). A firsthand example of this can be seen with Phan (2021), where their effect size estimates pertaining to COVID-19 infection rates were deemed "small" despite the critical implications of their intervention. Another everyday example that Götz et al. (2022) discuss is the correlation between ibuprofen intake and pain alleviation ($r = .14$; Meyer et al., 2001). And although Cohen may never have created his benchmarks with the intention of accounting for context, they demonstrate that additional context can better inform practical value and that Cohen's (1962, 1988) benchmarks do not sufficiently represent meaningfulness for every effect size in question.

Bosco et al. (2015). Although there has been heavy reliance on the estimates first derived from Cohen in 1962, the typical small, medium, and large classification of effect sizes determined by Cohen's (1962, 1988) benchmarks has little resemblance to findings in the field (Götz et al., 2022). To ameliorate these concerns, Bosco et al (2015) derived evidence-based benchmarks and incorporated context. Their findings demonstrated that context is critical to the interpretation of effects. They examined 20 broad-level bivariate relationship types (also known as coarse relations) from the years 1980-2010 in an effort to distinguish and contextualize varying types of effects. This marked one of the first major efforts to establish evidence-based

benchmarks founded on a large amount of data and can be considered a major advancement in increasing the utility of effect size benchmarks for organizational behavior scholars.

Despite this improvement to benchmark standards, there are still some limitations. Firstly, the majority of studies included in their review utilized more passive observational designs with little causal estimates. This issue increases endogeneity concerns and can impact the key insights being drawn from their effect size classifications (Antonakis et al., 2010; Schmidt & Pohler, 2018). Second, their review coded for effect sizes exclusively in the form of *r*, suggesting that potentially higher magnitude differences were overlooked without the inclusion of effect sizes in the form of Cohen's *d* from experimental designs. The authors also used only two journals (*Journal of Applied Psychology* and *Personnel Psychology*) for their systematic search. This introduces the possibility of publication bias because of the potential for a systematic difference of correlations published versus those not published (Kepes et al., 2012).

Gignac and Szodorai (2016). Another relied upon source for effect size benchmarks are Gignac and Szodorai's (2016). Similar to Bosco et al. (2015), their benchmarks are data derived and also go beyond Cohen's (1962, 1988) by accounting for key contextual differences. Gignac and Szodorai's (2016) benchmarks are much more refined in scope though, as they chose to center their focus on individual differences specifically. Based on over 700 meta-analytically derived correlations, they found that less than 3% of correlations were deemed large (0.50) according to Cohen's (1988) standards. Thus, by contextualizing the respective effects included in their systematic review, they suggest magnitudes of .10 for small, .20 for moderate, and .30 for large as suitable benchmarks for individual difference researchers. Despite this advancement towards more meaningful effect size benchmarks, endogeneity concerns can still greatly impact the key insights being drawn from their effect size classifications (Antonakis et al., 2010;

Schmidt & Pohler, 2018). To address this gap, the future sections outline a continuum of causal strength followed by the first purely causal leadership effect size estimates to date beginning first with a review of the various sources of endogeneity bias that limit causal inferences.

1.3 Forms of Endogeneity Bias

Relying on the magnitude of any effect size estimates can be problematic when considering endogeneity bias in the following forms: common-method bias, omitted variables, measurement error, omitted selection, simultaneity, inconsistent inference, and model misspecification. This is because these estimates include the effects of unmeasured causes (Hill et al., 2021). This can negate any causal assumptions being made and serve as serious threats to internal validity (Antonakis et al., 2010; Hill et al., 2021) while also resulting in false recommendations for theory and practice (Güntner et al., 2020; Schmidt & Pohler, 2018).

Common-method bias. Common-method bias can result when independent and dependent variables are gathered from the same rating source, thus inflating effect sizes (Podsakoff et al., 2003). It is important to consider whether effect sizes may be influenced by common-method bias given that many studies in the leadership field solely rely on percept–percept data in the form of self-report measures (Fischer et al., 2020; Baumeister et al., 2007). This can result in inflated predictor-criterion relationships (e.g., extra-role behaviors and performance) (Van Dyne & LePine, 1998; Ng et al., 2005). Another example may be judging men vs. women leaders on leadership effectiveness when the selection systems were different, thus causing biased correlations and erroneous comparisons (Antonakis et al., 2010). Other variables that have been found to be easily susceptible to common-method bias are job satisfaction, turnover, and performance appraisal (Crampton & Wagner, 1994; Spector, 1987). One likely circumstance highly susceptible for common-method bias in a leadership context may

be measuring perceptions of abusive supervision (Wu & Hu, 2013; Henle & Gross, 2014). Another likely circumstance would be testing hypotheses for job satisfaction and a leadership style using questionnaires distributed to the same sample (Jensen & Luthans, 2006; Darto et al., 2015).

Omitted variables. Omitted variables are a result of not including a variable in a model when it is related to *x* and *y* (Hill et al., 2021). This may occur in leadership research by not controlling for other similar leadership styles (e.g., charismatic vs. transformational) or differences in groups (Antonakis et al., 2010). One example of this would be not controlling for authentic leadership in a study designed to measure transformational leadership due to their high overlap (Banks et al., 2016; Gardner et al., 2011). Omitted causes can also be common particularly when testing the effects of leader individual differences (MacLaren et al., 2020; Crampton & Wagner, 1994). Although failing to account for omitted variables can lead to significantly biased estimates, it can also be mitigated through statistical techniques like sensitivity analysis and instrumental variables (Hill et al., 2021).

Measurement error. Measurement error results from the failure of the researcher to account for error in a variable within their model, thus introducing bias by altering the variable estimate and other independent variable estimates within the same model that correlate with *x* (Antonakis & Dietz, 2011; Kennedy, 2008). Unlike common-method bias which inflates effect sizes, measurement error attenuates them when considering bivariate relations (Blake & Gangestad, 2020; Fern & Monroe, 1996). For example, Brunell et al. (2008) examined the predictive ability of several individual differences for leader emergence and found the strongest predictor to be narcissism. If the researchers did not account for error in narcissism, it is likely that this would then taint the model and other key predictors in the model such as extraversion.

Another example can be seen with the meta-analytic review by Judge et al. (2004) on the relationship between intelligence and leadership, which resulted in a stronger magnitude of effects after correcting for measurement error.

Simultaneity. Endogeneity bias in the form of simultaneity occurs when two variables simultaneously cause one another (Güntner et al., 2020). An example of this could be any select leadership style as *x* and follower performance as *y*, where the style may cause significant changes in follower performance. However, the follower performance may simultaneously cause significant changes in leadership style thus suggesting reverse causality. For instance, follower behavior can influence leader behavior (e.g., Ashkanasy & Paulsen, 2013; Ahmad et al., 2020; Güntner et al., 2020) or organizational level factors can influence important outcomes as well (e.g., Desmet et al., 2015; Lee & Kray, 2021). In addition, leadership style may not explain follower performance in full because other factors like an ongoing historical event (e.g., COVID-19 pandemic) may result in a random decrease in performance, thus meaning *e* correlates with *x* making β1 inconsistent (Antonakis et al., 2010; Kennedy, 2008).

Omitted selection. Omitted selection refers to the problem of not randomly assigning individuals resulting in an endogenous treatment group (Certo et al., 2016). The issue lies in the fact that *x* may be explained by other factors (Clougherty et al., 2016; Connelly et al., 2013). For instance, imagine if twenty managers were to volunteer for one of two training session types. The first session involves ethical leadership training, and the other is a more standardized training. Since the managers self-selected themselves into the prospective training session type, the researcher would be unable to determine the true effect of the treatment group on the managers. As a result, any research that does not correct for omitted causes is likely to lead to inaccurate parameter estimates (Kennedy, 2008). One example of this can be seen with a meta-

analysis conducted by Reyes et al. (2019), which provided estimates to evaluate the effectiveness of leadership development programs. They highlighted the importance of accounting for selection effects that can be a common concern when training vs. non-training act as the treatment and control group (Martin et al., 2021).

Inconsistent inference. The sixth common form of endogeneity is inconsistent inference, which posits that the standard error across groups is inconsistent (Antonakis et al., 2010). This can lead to false *p*-value estimates, thus potentially impacting overall significance (Antonakis & Deitz, 2011). In a leadership context this may occur when evaluating the effect of a district manager's behavior on the followers of a particular team in one department. If the standard error of this group is inconsistent with the other groups, then any comparisons made by inferences from the researcher would be deemed unreliable (Schmidt & Pohler, 2018). These endogeneity concerns can render past effect size benchmarks as misleading, which is why the current review searches for evidence where forms of this bias have been minimized.

1.4 A Continuum of Causal-Inference Strength

Typical correlational benchmarks are still acceptable quantifications of relationships among variables, as long as they are interpreted and described as purely relational, and causality is not implied. That is, the term effect size should not be included as is commonly done in meta-analytic work (Borenstein et al., 2009; Schmidt & Hunter, 2015). However, these associations between variables only demonstrate early preliminary evidence to suggest a causal relationship. As such, there are many criteria needed for establishing a causal inference (Cook et al., 1979; Podsakoff and Podsakoff, 2019). Furthermore, endogeneity bias has rendered the correlation effect size benchmarks available to leadership scholars as highly limited or often misleading (Antonakis et al., 2010). To visualize this, a continuum is presented in Figure 1 Panels A and B

to illustrate the strength of causal claims. A visual representation is needed since causal inferences are a matter of relative degree (Cook et al., 2002). Hence the fluidity of the continuum. This may also suggest that a balance of high internal and external validity is needed through triangulation, making it extremely difficult to establish exact effect size magnitudes. The design of the continuum is based on the guidelines of experimental work proposed by past scholars (Cook et al., 1979; Cook et al., 2002; Podsakoff & Podsakoff, 2019; Lonati et al., 2018).

Figure 1 Panel A. *Causal Inference Continuum*

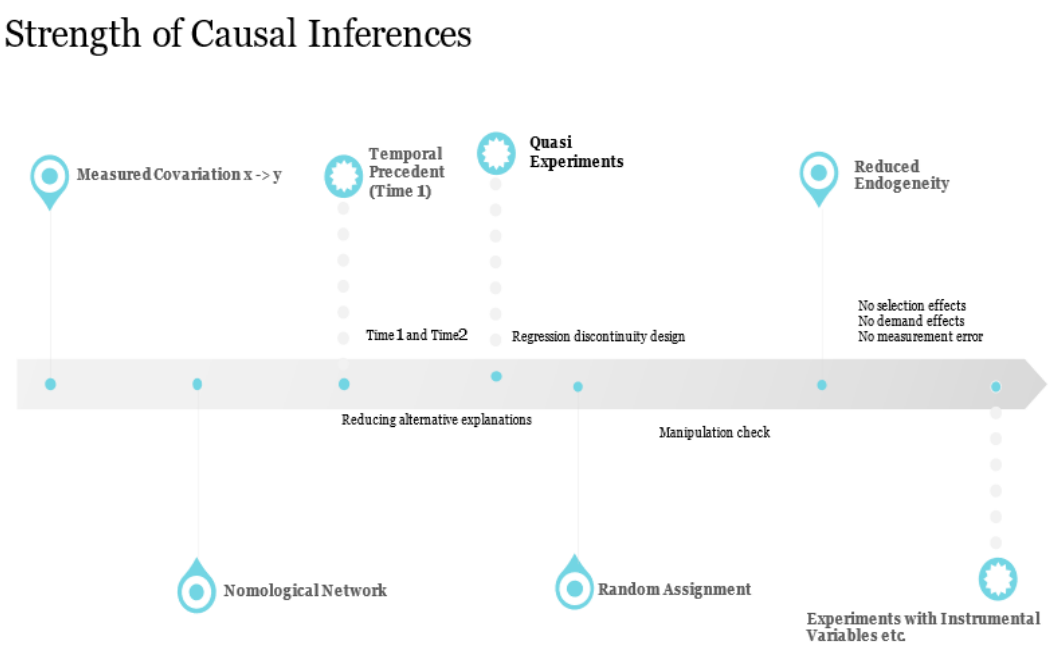
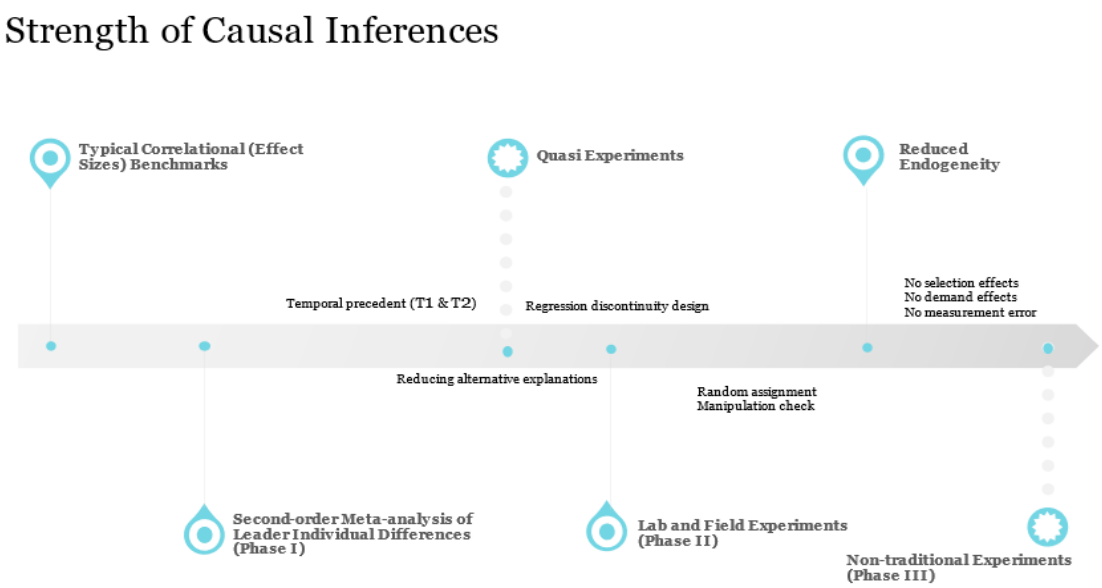


Figure 1 Panel B. *Causal Inference Continuum with Current Work*



Prerequisites to causal inferences. Moving from left to right, measured covariation is represented at the beginning of the continuum followed by developing a nomological network and establishing temporal precedent, with multiple time points adding additional strength. Each of these are prerequisites and deemed necessary to establish a casual inference (Podsakoff and Podsakoff, 2019; Byrne, 1984; Aguinis & Vandenberg, 2014). Understanding the relations between a concept and other concepts in its network helps to establish a case of how x might cause y. Furthermore, establishing a nomological network reduces the likelihood of alternative explanations (Aguinis & Edwards, 2014; Podsakoff and Podsakoff, 2019). Although these conditions for establishing casual claims may seem straightforward, the reliance of cross-sectional designs quickly negates evidence of temporal precedence (Bowen & Wiersema, 1999), and reducing alternative explanations is rarely satisfied by scholars (Aguinis & Edwards, 2014).

Temporal precedence is often misunderstood by leadership scholars. Simply because *x* is measured at Time 1 and *y* is measured at Time 2 and there is a relation does not mean *x* caused *y*.

Again, a supervisor may give an employee a positive performance evaluation commonly depicted as *y*. A researcher asks employees to evaluate a leader which is commonly depicted as *x* (and noted in studies as Time 1) and later the supervisor again evaluates the followers which is commonly depicted as *y* (and noted in studies as Time 2). In reality what is happening is something more akin to x_1 (Time 1)→ y_1 (Time 2) → x_2 (Time 3)→ y_2 (Time 4) → x_3 (Time 5)→ y_3 (Time 6) and the researcher is only capturing x_3 (Time 5)→ y_3 (Time 6). As such, our theories suggest that leadership is a dynamic social influence process (Yukl & Falbe, 1990; Furst & Cable, 2008; Oc & Bashshur, 2013; Day & Antonakis, 2013) and our measurements rarely reflect this (Yammarino et al., 2005).

Since the majority of studies included in Bosco et al.'s (2015) review utilized more passive observational designs with little causal estimates, their benchmarks are situated near the far-left end of the continuum previously illustrated. However, since individual differences serve as an example for the effects of exogeneous variables, Gignac and Szodorai's (2016) benchmarks are slightly increased in strength despite their effects still being correlational. This is because individual differences are stable and correlations which include them are less likely to be affected by omitted variables that influence both *x* and *y*.

"Gold standard" experiments. The middle of the continuum is characterized by quasi, field, and laboratory experimental designs. The usage of regression discontinuity designs has been highly supported by Cook and colleagues (2002) for mitigating biases and various threats to internal validity. From there, demonstrating random assignment, a typical advantage expressed via traditional lab or field experiments, is deemed higher in strength for drawing causal-inferences (Cook et al., 2002; Podsakoff and Podsakoff, 2019). Additional displays of rigor such as careful design of priming techniques or manipulation checks and avoiding unfair comparisons

by comparing a treatment group to a suitable control group are several practices that should be held paramount that further increase the strength of a causal claim (Lonati et al., 2018; Wulff et al., 2023).

Optimal causal-inference strength. Endogeneity bias can make correlations uninterpretable when the variable being manipulated (x) is correlated with error (e) as it directly violates the principles of ordinary least squares regression (OLS) and bivariate correlations, thus resulting in an effect size estimate that also includes the effects of unmeasured causes (Hill et al., 2021). This can greatly influence the ability to make causal claims (Podsakoff & Podsakoff, 2019). To illustrate this, the far right of the continuum are experiments that use advanced methods and analyses such as a two-stage least squares regression (2SLS) analysis for instrumental variables (Angrist & Imbens, 1995; Wooldridge, 1997), such as MacLaren et al. (2020). An instrumental variable estimator allows the researcher to go beyond typical OLS and ANOVA analyses by mitigating the risk of omitted causes predicting endogenous variables (Lonati et al, 2018). Omitted causes, such as the third variable problem, can be common particularly when testing the effects of individual differences (MacLaren et al., 2020; Crampton & Wagner, 1994). Thus, experiments that effectively mitigate endogeneity risks reduce the likelihood of producing inconsistent and biased estimates and increase the strength of a causal claim (Lonati et al., 2018; Antonakis et al., 2010). Note that the continuum continues, thus it is not posited that 2SLS or regression discontinuity experiments are the "one-size-fits-all" solution to making causal claims.

CHAPTER 2: METHOD

2.1 Overview

Three separate searches with unique inclusion criteria were delineated into three comprehensive phases for the overall systematic review of the leadership literature and are mapped onto the continuum illustrated in Figure 2. The purpose of the three searches are to identify different types of casually identified effect sizes depicted at different points in the continuum. The first review involved a second-order meta-analytic search with common individual differences of leaders and their effects on various outcomes to serve as an example for the effects of exogeneous variables, thus situating these estimates on the left side of the continuum but stronger than causal inferences made from typical correlational benchmarks (See Figure 2). The decision to select individual differences was due to their exogenous nature as well the long history of leader individual differences investigated as major predictors of leader effectiveness (Lord et al., 1986; Judge et al., 2002; Hoffman et al., 2012; Derue et al., 2011).

The second review required a separate search examining only field and lab experiments to demonstrate effect size estimates that are purely causal in nature, thus situating these estimates much further down the continuum. More specifically, this review will focus on the bivariate relationships between prominent variables in the leadership literature similar to Bosco et al.'s (2015) approach. The last unique review was required to report the results of quasi-experiments and non-traditional experimental designs, but in a selective narrative fashion. The effects of the non-traditional experimental designs are illustrated on the far-right end of the causal strength continuum due to the significant decrease in potential forms of endogeneity bias.

The effects of this review are presented in the form of Cohen's *d* to account for potentially higher magnitude differences being overlooked, and accounts for publication bias (Kepes et al., 2012). This will help evaluate the current landscape of effect size distributions and their effectiveness at capturing generality while providing key insights for the leadership field. It

also increases the overall rigor and robustness of the review and adds stronger support for the use of causal effect size benchmarks as it pertains to the leadership domain. Additionally, the current review is more refined in scope than past reviews of effect sizes by examining effect sizes for particularly the leadership domain. Each of the three comprehensive systematic searches adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2020). This was done to provide a complete and transparent account of the searches that were conducted. Studies that met sufficient criteria in any phase during the initial screening process that were published in a different language were translated to English and assessed for full eligibility for inclusion when necessary (e.g., Yamuara et al., 2013; Hirota, 1953).

2.2 Systematic Search and Inclusion Criteria

#1: Second-order Meta-analysis of Leader Individual Differences. For the first systematic review, I performed a second-order meta-analytic search using ABI/INFORM to obtain previously published meta-analyses that reported effect sizes between eleven commonly studied individual differences of leaders and various outcomes. Unpublished meta-analyses were also included in an effort to mitigate publication bias (Kepes et al., 2012). The database was searched from 1980-2022 using *leadership* and *meta* as a key word within abstracts in combination with any key terms situated in the entire text that may be associated with the eleven individual differences selected appearing anywhere in the text. The following individual differences were included: general mental ability, emotional stability, agreeableness, conscientiousness, extraversion, openness, Machiavellianism, psychopathy, narcissism, gender, and emotional intelligence. For a full list of key terms used see Appendix A. Each meta-analysis had to include primary studies exclusively comprised of the individual differences of leaders or

managers, not followers, to avoid any potential confounding of effect size estimates for outcomes associated with common leadership styles, leader emergence, etc. Similar to previous second-order meta-analyses involving personality (e.g., Wilmot et al., 2019; Chiaburu et al., 2011), they also needed to provide an effect size estimate that could be converted to a correlation coefficient for at least one of the individual differences mentioned.

The initial search resulted in 254 records identified (See Figure 2 Panel A). Each of these records underwent the full screening process by reading each abstract and ensuring that all sufficient criteria for inclusion were met. While assessing for eligibility, 89 records were eliminated because they did not examine leader individual differences with direct effects on followers, 22 records were omitted because they did not contain an all-leader sample, 12 records had insufficient data, and 110 were falsely identified in the search and thus were deemed false positives (e.g., book chapters, primary studies, etc.). The final sample comprised of 21 meta-analyses and 82 effects. The meta-analytic findings of this review further support and justify the need for creating a new set of effect size benchmarks for the leadership field, one that contains estimates solely causal in nature.

#2: Lab and field experiments. The literature was systematically searched for "gold standard" experiments using the ABI/INFORM and PsychINFO databases while adhering to best practice guidelines (Kepes et al., 2013). For an experiment to be considered "gold standard," it had to be a field or lab experiment with successful manipulation checks and random assignment. To minimize the threat of publication bias, unpublished works in the form of dissertations and conference papers were also included by searching ProQuest Dissertations and Theses (Rothstein & Hopewell, 2009). The databases were searched from 1940-2022 using *leader* as a key word in

combination with the following key words within abstracts: *experiment*, *random*, and *trial*. Detection heuristics proposed by Wood (2008) allowed for efficient screening of duplicates.

The classifications for what studies constituted as a lab, field, or quasi-experimental design are consistent with the guidelines proposed by Podsakoff and Podsakoff (2019). Any of the papers retrieved from the search also underwent an additional check for robustness and rigor in their experimental design as a form of quality control and to avoid the potential pitfall associated with the "garbage in-garbage out" perspective (Borenstein et al., 2009). This was done by adhering to the experimental standards for determining causality (Cook et al., 1979; Cook et al., 2002; Lonati et al., 2018). Only traditional laboratory and field experiments that met these standards were considered in the final sample. Any quasi-experiments or advanced methodologies (e.g., instrumental variables) that appeared in the search were saved to be assessed for eligibility in the third review in the event that they met sufficient inclusion criteria. It was also required that studies report statistics that could be used to create an effect size (e.g., correlation coefficients, Cohen's *d*, means and standard deviations of treatment and control groups).

The experiments included in the final sample required that the manipulated variable be a leader behavior and the variable measured have a direct effect on follower outcomes. Furthermore, the relationship had to reach a "critical mass" to be included. That is, enough samples had to be present for a meta-analytic estimate to be obtained. For instance, Phillips (2002) experimentally examined the relationship between leader procedural fairness and team task performance, but no other experiments met sufficient criteria to be included for this relationship. Similarly, Boulu-Reshef and colleagues (2020) designed the only experiment from

the review to meet sufficient criteria examining the effect of empowering leader behavior and follower prosocial behaviors.

The focus on leader-to-follower direct effects aligns with the framing of leadership as a social influence process (Yukl & Falbe, 1990; Furst & Cable, 2008; Oc & Bashshur, 2013; Day & Antonakis, 2013). Because of this theoretical framing, studies that measured the effects of leadership training on leaders were omitted (see Martin et al. (2021) for an extensive review on leadership training experiments and causal claims). Other instances may be uniquely designed studies like Spark et al.'s (2021), in which they experimentally manipulated state extraversion to measure leader emergence and affect. No "paper people" in the form of vignette or scenario studies were included in the final sample. This decision is supported by the notion that the hypothetical nature of the situations posed by these studies can increase the risk of social desirability bias and demand effects, resulting in potentially less consequential outcomes (Lonati et al., 2018). In addition, only bivariate effect size estimates were used due to potential issues with combining multivariate or beta related effect sizes with bivariate effects when deriving meta-analytic estimates (Steel & Kammeyer-Mueller, 2002; Roth et al., 2018).

The initial search resulted in 26,142 records identified (See Figure 2 Panel B). 8,581 of these records were screened by scanning each abstract for sufficient criteria for further examination. Approximately 1,464 papers were assessed for full eligibility, with 213 records screened out as duplicates, resulting in 1,251 papers. While assessing for eligibility, 351 records were eliminated because they did not manipulate an exogenous leadership style or behavior as the independent variable and measure the direct effect of these variables on followers, 189 records were omitted because they included vignettes or scenarios as their manipulation, 209 records had insufficient data (e.g., only included the means of the treatment and control group,

contained only multivariate estimates), 57 failed the additional checks for experimental rigor, and 74 were falsely identified in the search and thus were deemed false positives (e.g., book chapters, non-experiments, etc.). A total of 342 of the records identified were quasi-experiments or utilized methods like instrumental variables, and thus were saved to be re-screened in the third review.

The sample included in the review was initially higher, but since many of the direct effects examined did not have at least a *k* of 2, the sample was reduced even further to a final 29 papers comprised of 91 unique samples. Due to the low yield of studies in the final sample, a secondary search was performed by using the reference sections of previous reviews in respective sub-domains of leadership research such as destructive or abusive leadership (e.g., Mackey et al., 2021; Fischer et al., 2021), charismatic leadership (e.g., Ernst et al., 2021; Banks et al., 2017), leader prototypicality (e.g., Barreto & Hogg, 2017), or affect and emotions (e.g., Clarkson et al., 2020; Gooty et al., 2010).

Quasi and non-traditional experiments. The same databases and year range used for identifying the lab and field experiments were used in this search. However, instead of focusing on traditional experimental designs the databases were searched using *leader* as a key word within abstracts in combination with the term *quasi* or potential solutions suggested by Hill et al. (2021) and Antonakis et al. (2010) for remedying endogeneity bias. Some of these key words include *propensity score, Heckman treatment, regression discontinuity, 2SLS.* For a full list of key words used see Appendix A. The initial search resulted in 19,424 records identified (See Figure 2 Panel C). A highly selective approach was adopted for this review.

2.3 Coding Procedures

For all three reviews, I independently coded a subsample of studies that met the inclusion criteria. I then met with a team of collaborators before each of the coding processes to develop a protocol that ensures consistency in judgements (Geyskens et al., 2009). Two of the coders and myself independently coded each study that met the initial inclusion criteria. All coders underwent extensive training regarding expectations for rigor in experimental designs, the appropriateness of variables and potential proxies, necessary effect size information, etc. Interrater reliability was calculated using percent agreement and Cohen's kappa (Cohen, 1960) across 132 coding decisions. The Cohen's kappa estimates for rater 1 and raters 2 and 3 were both .88, thus demonstrating strong consistency (Fleiss, 1981). I double-coded all of the papers used in the final sample for each of three reviews, and any disagreements were discussed among the coders.

During the lab and field experiment coding phase, a select few variables were considered proxies. For example, follower extra effort was coded as a proxy variable for extra-role performance (quantity) and task productivity was coded as a proxy variable for task performance (quantity). When necessary, decisions for variable classifications were based on the measures used in the study. For example, if an experiment tested the effect of leader emotional or affective displays and used the PANAS scale to measure followers' reactions (Watson et al., 1988), then the variable was classified as follower affect. A full list of the independent and dependent leadership variables included in the final review, as well as any proxy variables, can be found in Appendix B. For studies with longitudinal designs, the effect size for Time 1 was used to reduce the potential confounds associated with treatment effects across multiple time points (Onwuegbuzie & Levin, 2003). All papers were coded for level of analysis (individual vs. team/group) and are expressed as separate effect size estimates. Two potential moderators were

also coded for (leader gender and virtual vs. in-person leader manipulation). An example of the additional coding for experimental rigor and various exclusion criteria can be seen in Appendix C.

2.4 Meta-analytic Procedures

The analyses for the lab and field experiment phase were conducted using the R package "metafor" (Viechtbauer, 2010). The observed effect sizes were corrected for artifact distributions of measurement unreliability as outlined by Schmidt and Hunter (2015) where sufficient information exists. The effects for the quasi and non-traditional experiments were summarized and discussed in a narrative fashion.

CHAPTER 3: RESULTS

3.1 Second-order Meta-analysis of Leader Individual Differences

The second-order meta-analytic review resulted in a total of eighty-two meta-analytic effect sizes. A full list of the reported number of samples (*k*), sample size (*n*), and corrected effect size estimates and the corresponding meta-analytic relations can be found in Table 1. Any instance when these were not available by the authors "not reported" (NR) was placed in the table instead. The confidence intervals, credibility intervals, and any test for heterogeneity were also provided when possible. Below is a review of the second-order meta-analytic findings of the eleven individual differences of leaders and their correlation to various outcomes.

Table 1. *Summary of Leader Individual Differences Effect Sizes*

Empirical relationship	*K*	*N*	*P*	95% CI	Evidence for heterogeneity
General Mental Ability					
General mental ability and leader effectiveness	99	15, 985	.17 [g]	.16 to .19	▪ 80% CV: .04 to .31 ▪ *Q*= 340.56
General mental ability and group performance	6	291	.04 [b]	-.07 to .16*	▪ NR
General mental ability and leader emergence	65	NR	.19 [f]	.24 to .30	▪ 80% CV: .05 to .48 ▪ SDρ = .10
General mental ability and transformational leadership	6	826	.16 [b]	.10 to .23*	▪ NR
General mental ability and laissez-faire leadership	2	371	-.16 [b]	-.26 to -.06*	▪ NR
General mental ability and contingent-reward leadership	2	371	.11 [b]	.01 to .21*	▪ NR
General mental ability and initiating structure	2	184	.14 [b]	-.24 to .52*	▪ NR
General mental ability and consideration	1	68	-.07 [b]	NR	▪ NR
General mental ability and inspirational motivation	3	566	.14 [a]	-.05 to .32	▪ 80% CV: -.05 to .33 ▪ SDρ = .17
Emotional stability					
Emotional stability and leader effectiveness	51	8,960	.24 [b]	0.23 to 0.26*	▪ NR
Emotional stability and group performance	1	50	-.03 [b]	NR	▪ NR
Emotional stability and leader emergence	30	277	-.24 [c]	-.30 to -.18	▪ NR
Emotional stability and follower job satisfaction	2	300	.02 [b]	-.10 to .13*	▪ NR
Emotional stability and satisfaction with leader	3	1,078	.08 [b]	.02 to .15*	▪ NR
Emotional stability and transformational leadership	18	3,380	-.17 [h]	-.21 to -.14	▪ 80% CV: -.20 to -.15

Empirical relationship	*K*	*N*	*P*	95% CI	Evidence for heterogeneity
Emotional stability and destructive leadership	28	7,948	.02 [i]	.13 to .27	▪ 80% CV: -.03 to .44 ▪ SDρ = .18
Emotional stability and ethical behavior	28	3,496	.17 [j]	.06 to .18	▪ 80% CV: -.05 to .28 ▪ SDρ = .2
Emotional stability and idealized influence	6	1,310	-.13 [a]	-.19 to -.07	▪ 80% CV: -.17 to -.09
Emotional stability and inspirational motivation	11	2,038	-.14 [a]	-.20 to -.07	▪ SDρ: .09 ▪ 80% CV: -.25 to -.03 ▪ SDρ: .12
Emotional stability and counterproductive knowledge behaviors	2	447	.31 [q] (Corrected r)	NR	▪ NR
Agreeableness					
Agreeableness and leader effectiveness	67	19,670	.14 [k]	.10 to .19	▪ 80% CV: -.10 to .38 ▪ SDρ = .19
Agreeableness and group performance	2	84	.20 [b]	-.02 to .41*	▪ NR
Agreeableness and leader emergence	27	4,174	.24 [k]	.15 to .34	▪ 80% CV: -.06 to .54 ▪ SDρ = .23
Agreeableness and follower job satisfaction	2	300	.01 [b]	-.15 to .17*	▪ NR
Agreeableness and satisfaction with leader	3	1,078	-.03 [b]	.11 to .33*	▪ NR
Agreeableness and transformational leadership	20	3,916	.14 [h]	.06 to .21	▪ 80% CV: -.07 to .34
Agreeableness and destructive leadership	23	6,911	-.15 [i]	-.19 to -.11	▪ 80% CV: -.26 to -.04 ▪ SDρ = .09
Agreeableness and ethical behavior	28	3,496	.08 [j]	.00 to .11	▪ 80% CV: -.11 to .22 ▪ SDρ = .21
Agreeableness and idealized influence	5	1,180	.26 [a]	.19 to .33	▪ 80% CV: .19 to .32 ▪ SDρ = .09
Agreeableness and inspirational motivation	10	1,863	.15 [a]	.09 to .21	▪ 80% CV: .06 to .24 ▪ SDρ = .11
Agreeableness and LMX	4	859	.18 [e]	.11 to .27	▪ 80% CV: .18 to .18 ▪ SDρ = .00 ▪ *Q* = 2.29
Conscientiousness					
Conscientiousness and leader effectiveness	39	10,056	.28 [b]	.26 to .29*	▪ NR
Conscientiousness and group performance	5	203	.31 [b]	.02 to .15*	▪ NR
Conscientiousness and leader emergence	17	277	.33 [c]	.22 to .34	▪ 80% CV: .06 to .51
Conscientiousness and follower job satisfaction	2	300	-.08 [b]	.02 to .15*	▪ NR
Conscientiousness and satisfaction with leader	3	400	-.03 [b]	.02 to .15*	▪ NR
Conscientiousness and transformational leadership	18	3,516	.13 [h]	.06 to .19	▪ 80% CV: -.02 to .28

Empirical relationship	*K*	*N*	*P*	95% CI	Evidence for heterogeneity
Conscientiousness and destructive leadership	27	7,779	-.18 [i]	-.23 to -.13	▪ 80% CV: -.33 to -.02 ▪ SDρ = .12
Conscientiousness and ethical behavior	28	3,496	.14 [j]	.07 to .14	▪ 80% CV: .09 to .12 ▪ SDρ = .12
Conscientiousness and idealized influence	5	1,180	.07 [a]	.01 to .13	▪ 80% CV: .07 to .07 ▪ SDρ = .07
Conscientiousness and inspirational motivation	9	1,760	.10 [a]	.01 to .12	▪ 80% CV: .00 to .13 ▪ SDρ = .10
Extraversion					
Extraversion and leader effectiveness	63	12,640	.31 [c]	.30 to .32*	▪ NR
Extraversion and group performance	3	135	.00 [b]	.02 to .15*	▪ NR
Extraversion and leader emergence	37	277	.33 [c]	.26 to .36	▪ 80% CV: .09 to .53
Extraversion and follower job satisfaction	2	300	.07 [b]	.02 to .15*	▪ NR
Extraversion and satisfaction with leader	3	1,078	.03 [b]	.02 to .15*	▪ NR
Extraversion and transformational leadership	20	3,692	.24 [h]	.21 to .28	▪ 80% CV: .18 to .31
Extraversion and destructive leadership	18	5,409	-.03 [i]	-.08 to .01	▪ 80% CV: -.13 to .06 ▪ SDρ = .08
Extraversion and idealized influence	4	958	.23 [a]	.14 to .32	▪ 80% CV: .22 to .22 ▪ SDρ = .10
Extraversion and inspirational motivation	10	1,908	.34 [a]	.27 to .41	▪ 80% CV: .34 to .34 ▪ SDρ = .03
Extraversion and LMX	4	859	.18 [e]	.04 to .34	▪ 80% CV: .01 to .36 ▪ SDρ = .14 ▪ *Q* = 16.06
Openness					
Openness and leader effectiveness	39	7,762	.24 [c]	.22 to .26*	▪ NR ▪
Openness and group performance	2	117	.13 [b]	.02 to .15*	▪ NR
Openness and leader emergence	20	277	.24 [c]	.19 to .28	▪ 80% CV: .09 to .38
Openness and follower job satisfaction	2	300	.00 [b]	.02 to .15*	▪ NR
Openness and satisfaction with leader	3	400	.03 [b]	.02 to .15*	▪ NR
Openness and transformational leadership	19	3,887	.15 [h]	.08 to .23	▪ 80% CV: -.04 to .35
Openness and destructive leadership	12	4,150	-.08 [i]	-.15 to -.01	▪ 80% CV: -.22 to .06 ▪ SDρ = .11
Openness and idealized influence	6	1,356	.13 [a]	.07 to .19	▪ 80% CV: .08 to .17 ▪ SDρ = .09
Openness and inspirational motivation	11	1,993	.16 [a]	.08 to .17	▪ 80% CV: .16 to .16 ▪ SDρ = .08
Machiavellianism					
Machiavellianism and abusive supervision	2	292	.07 [l]	-.08 to .48	▪ *Q* = 3.19

Empirical relationship	*K*	*N*	*P*	95% CI	Evidence for heterogeneity
			(Uncorrected r)		▪ I^2 = 68.66
Psychopathy					
Psychopathy and leader effectiveness	42	6,838	-.04 [m]	-.09 to -.00	▪ 80% CV: -.16 to .07 ▪ SDρ = .09
Psychopathy and leader emergence	46	32,680	.07 [m]	.04 to .10	▪ 80% CV: -.04 to .19 ▪ SDρ = .09
Psychopathy and transformational leadership	13	1,220	-.18 [m]	-.35 to -.01	▪ 80% CV: -.54 to .17 ▪ SDρ = .28
Narcissism					
Narcissism and firm size	30	11,948	.08 [n]	.04 to .13	▪ *Q* = 165.68 ▪ SDρ = .02
Narcissism and firm financial performance	19	8,307	.06 [n]	.03 to .09	▪ *Q* = 32.66 ▪ SDρ = .02
Narcissism and firm innovation/growth	19	9,776	.09 [n]	.05 to .13	▪ *Q* = 58.61 ▪ SDρ = .02
Narcissism and risk taking	7	3,663	.07 [n]	-.03 to .16	▪ *Q* = 53.08 ▪ SDρ = .05
Narcissism and financial leverage	7	1,672	.03 [n]	-.04 to .10	▪ *Q* = 14.86 ▪ SDρ = .04
Narcissism and CEO duality	11	2,759	.09 [n]	.05 to .15	▪ *Q* = 23.19 ▪ SDρ = .03
Narcissism and counterproductive knowledge behaviors	3	581	.27 [q] (Corrected r)	NR	▪ NR
Gender					
Gender and leader effectiveness	99	101,676	-.02 [o]	-.10 to .00	▪ *Q* = 415.3
Gender and leader emergence	136	19,073	.09 [p]	.13 to .29	▪ 80% CV: -.27 to .71
Gender and follower job satisfaction	8	3,824	-.04 [b]	-.03 to .03*	▪ NR
Gender and satisfaction with leader	7	NR	.00 [d]	-.08 to .07	▪ NR
Gender and transformational leadership	44	29,806	-.06 [d]	-.13 to -.08	▪ *Q* = 152.94
Gender and idealized influence	23	16,851	-.05 [a]	-.07 to -.02	▪ 80% CV: -.10 to .00
Gender and inspirational motivation	26	22,802	-.02 [a]	-.05 to .00	▪ SDρ = .06 ▪ 80% CV: -.09 to .04 ▪ SDρ = .06
Gender and laissez-faire leadership	16	12,947	.09 [d]	.14 to .19	▪ *Q* = 18.74
Gender and firm financial performance	78	117,639	.02 [r]	.01 to .04	▪ 80% CV: -.07 to .12 ▪ *Q* = 345.51
Gender and financial leverage	18	56,119	.06 [r]	-.00 to .05	▪ 80% CV: -.11 to .13 ▪ *Q* = 119.56
Gender and destructive leadership	35	7,561	.06 [i]	-.11 to -.02	▪ 80% CV: -.22 to .09

Empirical relationship	*K*	*N*	*P*	95% CI	Evidence for heterogeneity
					▪ SDρ = .12
Emotional intelligence					
Emotional intelligence and leader effectiveness	52	6,052	.25 [s]	NR	▪ 80% CV: .13 to .38 ▪ SDρ = .09
Emotional intelligence and transformational leadership	38	4,519	.37 [s]	NR	▪ 80% CV: .16 to .58 ▪ SDρ = .16
Emotional intelligence and transactional leadership	18	2,141	.10 [s]	NR	▪ 80% CV: .10 to .10 ▪ SDρ = .00
Emotional intelligence and LMX	10	880	.27 [s]	NR	▪ 80% CV: .04 to .50 ▪ SDρ = .18
Emotional intelligence and abusive supervision	5	889	-.43 [l] (Uncorrected r)	-.56 to -.27	▪ *Q* = 28.31 ▪ I^2 = 85.87
Emotional intelligence and authentic leadership	11	3,507	.49 [t]	.35 to .64	▪ 80% CV: .19 to .80 ▪ SDρ = .24
Emotional intelligence and servant leadership	18	2,409	.57 [u]	.48 to .66	▪ 80% CV: .34 to .80 ▪ *Q* = 184.20

Note. * = 90% confidence interval, *NR* = not reported, ρ = rho estimate, CI = confidence interval, CV = credibility interval, SDρ = standard deviation of population correlation estimates across studies, *Q* = statistic that assesses the heterogeneity in effect sizes across studies, I^2 = statistic that assess heterogeneity in effect sizes across studies. Gender was coded as 1 = male and 0 = female. [a.] Banks et al. (2017); [b.] Derue et al. (2011); [c.] Judge et al. (2002); [d.] Eagly et al. (2003); [e.] Dulebohn (2012); [f.] Judge et al. (2004); [g.] Hoffman et al. (2014); [h.] Bono & Judge (2004); [i.] Mackey et al. (2021); [j.] Nei et al. (2018); [k.] Blake et al. (2022);); [l.] Zhang & Bednall (2016); [m.] Harms & Credé (2018); [n.] Cragun et al. (2020); [o.] Paustian-Underdahl et al. (2014); [p.] Badura et al. (2018); [q.] Afshar-Jalili et al. (2021); [r.] Hoobler et al. (2018); [s.] Whitman et al. (2009); [t.] Miao et al. (2018); [u.] Miao et al. (2021).

General Mental Ability. Three outcomes related to leader general mental ability are reported. There is a positive relation between general mental ability of the leader and evaluations of leader effectiveness (ρ = .17, *k* = 99, n = 15,985, 95% CI [.16, .19], leader emergence (ρ = .19, k = 65, n = NR, 95% CI [.24, .30]), and group performance (ρ = .04, k = 6, n = 291, 90% CI [-.07, .16]). There is a relative amount of heterogeneity of the samples for leader general mental ability and leader emergence in particular (80% CV [.05 to .48], SDρ = .10), thus suggesting moderation and a need for more research exploring this relationship.

The Big Five. A total of twelve different outcomes related to a leader's Big Five traits are reported. A leader's emotional stability is associated with follower reported leader effectiveness (ρ = .24, k = 51, n = 8,960, 90% CI [.23, .26]), leader emergence (ρ = -.24, k = 30, n = 277 95% CI [-.30, -.18]), group performance (ρ = -.03, k = 1, n = 50, NR), satisfaction with leader (ρ = .08, k = 3, n = 1,078, 90% CI [.02, .15]), and follower job satisfaction (ρ = .02, k = 2, n = 300, 90% CI [-.10, .13]). The correlations between leader agreeableness and outcomes are mostly

positive. For example, agreeableness is associated with several commonly studied leadership styles and their related constructs such as follower subjective evaluations of transformational leadership (ρ = .14, k = 20, n = 3,196, 95% CI [.06, .21]), destructive leadership (ρ = -.15, k = 23, n = 6,911, 95% CI [-.19, -.11]), and ethical leadership (ρ = .08, k = 28, n = 3,496, 95% CI [.19, .33]). Leader conscientiousness is positively associated with group performance (ρ = .31, k = 5, n = 203, 90% CI [.02, .15]), leader emergence (ρ = .33, k = 17, n = NR, 95% CI [.22, .34]). A leader's extraversion is positively associated with follower evaluations of transformational leadership (ρ = .24 k = 20, n = 3,692, 95% CI [.21, .28]), leader effectiveness (ρ = .31, k = 63, n = 12,640, 90% CI [.30, .32]), and leader emergence (ρ = .33, k = 37, n = NR, 95% CI [.26, .36]). Finally, a leader's openness is positively associated with follower ratings of idealized influence (ρ = .13, k = 6, n = 1,356, 95% CI [.07, .19]) and leader effectiveness (ρ = .24, k = 39, n = 7,762, 90% CI [.22, .26]), but negatively associated with follower evaluations of destructive leadership (ρ = -.08, k = 12, n = 4,150, 90% CI [-.15, -.01]).

Leaders' conscientiousness and openness illustrate a more consistent and larger magnitude for effects in general than those for emotional stability, agreeableness, and extraversion. There are mixed results whether the outcomes have a positive or negative relationship as well as the level of heterogeneity across samples. For instance, the level of heterogeneity of the samples for leader conscientiousness and leader ethical behavior is relatively low (80% CV [.09 to .12], SDρ = .12), but the heterogeneity of the sample for leader agreeableness and leader emergence is relatively high (80% CV [-.06 to .54], SDρ = .23). In addition to these mixed results across samples, no tests for heterogeneity were provided for many of the effects.

The Dark Triad. Eleven outcomes related to leaders' Dark Triad are reported. Leaders' Machiavellianism is positively related to follower evaluations of abusive supervision (ρ = .07, k = 42, n = 6,838, 95% CI [-.09, .00]). Leaders' psychopathy is mostly negative and around the same magnitude as Machiavellianism in relation to the outcomes of leader effectiveness (ρ = -.04, k = 99, n = 15,985, 95% CI [.16, .19]) and leader emergence (ρ = .07, k = 46, n = 32,680, 95% CI [.04, .10]). The relationship between leaders' psychopathy and follower evaluations of transformational leadership (ρ = -.18, k = 13, n = 1,220, 95% CI [-.35, -.01]) displayed a slightly higher magnitude. Seven different outcomes related to narcissism alone are reported (e.g., firm size (ρ = .09, k = 30, n = 11,948, 95% CI [.04, .13]); firm financial performance (ρ = .06, k = 19, n = 8,307, 95% CI [.03, .09]). Nearly all of the effects for leader psychopathy demonstrate a moderate to high level of heterogeneity from their respective samples. For example, the study that reported the meta-analytic estimate for psychopathy and transformational leadership (Harms & Credé, 2018) had highly heterogeneous samples (80% CV [-.54 to .17], SDρ = .28), thus suggesting there is need to explore more potential moderators that could be confounding this relationship.

Gender. A total of ten different outcomes related to a leader's gender are reported, with a positive sign indicating that men were higher than women, and a negative sign indicating that women were higher than men on a particular outcome. Leaders' gender is positively associated with leader emergence (ρ = .09, k = 136, n = 19,073, 95% CI [.13, .29]), laissez-faire leadership (ρ = .09, k = 16, n = 12,947, 95% CI [.14, .19]), and firm financial performance (ρ = .02, k = 78, n = 117,639, 95% CI [.01, .04]). Leaders' gender is negatively associated with follower evaluations of idealized influence (ρ = -.09, k = 23, n = 16,851, 95% CI [-.07, -.02]), inspirational motivation (ρ = -.05, k = 23, n = 16,851, 95% CI [-.13, -.08]), and destructive

leadership (ρ = -.02, k = 26, n = 22,802, 95% CI [-.05, .00]). All of the effects related to leader gender are relatively close in magnitude. The credibility intervals are fairly wide, thus suggesting a relatively high level of heterogeneity across the samples used for leader gender and the outcomes mentioned and a high likelihood of potential moderators for these relationships. One example of this are the samples for leader gender and leader emergence in particular (80% CV [-.27 to .71]).

Emotional intelligence. Seven different outcomes related to a leader's emotional intelligence are reported. The findings illustrate that the correlation between leader emotional intelligence as a stable individual difference and the seven outcomes are mostly positive. Leaders' emotional intelligence is associated with evaluations of leader effectiveness (ρ = .25, k = 52, n = 6,052, NR), transformational leadership (ρ = .37, k = 38, n = 4,519, NR), transactional leadership (ρ = .10, k = 18, n = 2,141, NR), LMX (ρ = .27, k = 10, n = 880, NR), abusive supervision (ρ = -.43, k = 5, n = 889, 95% CI [-.56, -.27]), authentic leadership (ρ = .49, k = 11, n = 3,507, 95% CI [.35, .64]), and servant leadership (ρ = .57, k = 18, n = 2,409, 95% CI [.48, .66]). A relatively medium to high level of heterogeneity was present across these samples, with nearly all of the 80% credibility intervals provided being relatively wide.

3.2 Lab and field experiments

This meta-analytic search resulted in a final total of thirty meta-analytic effect sizes. A full list of the reported number of samples (*k*), sample size (*n*), and corrected effect size estimates in terms of Cohen's *d* and the corresponding meta-analytic relations can be found in Table 2. The standard error (SE), 95% confidence intervals, and tests for heterogeneity in the form of the I^2 statistic were also provided. Below is a review of the meta-analytic findings of eight experimentally manipulated leader behaviors and their effects on followers. Surprisingly, only

charismatic leadership had more than four samples included for most estimates. The implications of this finding are discussed in more detail in a future section.

Table 2. *Summary of Leader-to-Follower Causal Effect Sizes*

Leadership style/behavior and follower outcome	*k*	*N*	*d* (SE)	I^2	*p* value	95% CI
Charismatic leadership						
Task performance (quantity)	11	1769	.20 (.10)	.95	.07	-.02 to .41
Task performance (quality)	8	1942	.07 (.06)	.80	.25	-.05 to .18
Extra-role performance (quantity)	6	772	-.07 (.10)	.87	.52	-.26 to .13
Extra-role performance (quality)	6	772	-.13 (.09)	.85	.18	-.31 to .06
Team prosocial behavior	3	221	.27 (.14)	.77	.07	-.02 to .55
Trust in leader	2	489	.01 (.21)	.95	.99	-.42 to .42
Task satisfaction	2	426	.19 (.05)	.00	.92	.10 to .29
Transformational leadership						
Task performance (quality)	2	236	.19 (.16)	.81	.23	-.12 to .50
Team task performance (quality)	3	414	.26 (.28)	.97	.35	-.29 to .82
Extra-role performance (quantity)	3	326	.23 (.10)	.66	.02	.03 to .42
Team evaluation of leader effectiveness	2	231	.87 (.07)	.00	.001*	.74 to .99
Team liking of leader	2	231	.82 (.10)	.48	.001*	.63 to 1.0
Task satisfaction	2	236	-.01 (.25)	.92	.97	-.51 to .49
Team task satisfaction	2	231	.22 (.53)	.99	.68	.72 to 1.0
Empowering leadership						
Team task performance (quantity)	3	603	.43 (.14)	.91	.001*	.17 to .70
Team task performance (quality)	2	342	.18 (.18)	.91	.32	-.17 to .53
Satisfaction with leader	2	298	.54 (.38)	.96	.16	-.22 to 1.2
Team task satisfaction	2	342	.08 (.53)	.99	.89	-.95 to 1.1
Leader vision						
Trust in leader	2	505	.29 (.70)	.99	.67	-1.0 to 1.6
Perceived charisma of leader	3	809	.11 (.57)	.99	.85	-1.0 to 1.2
Destructive leadership						
Task performance (quantity)	3	280	-.62 (.21)	.91	.001*	-1.0 to -.21
Task performance (quality)	2	196	-.41 (.35)	.96	.24	-1.0 to .27
Prosocial behavior	3	352	-1.10 (.45)	.99	.02	-1.9 to -.21
Negative affect	3	280	.23 (.27)	.95	.41	-.31 to .76
Deviance	2	254	.39 (.15)	.78	.01	.09 to .69
Creativity	2	196	-.53 (.07)	.00	.001*	-.67 to -.39
Participative leadership						
Team task performance (quantity)	3	341	.05 (.34)	.97	.88	-.62 to .73
Group orientedness	2	217	.83 (.24)	.91	.001*	.36 to 1.3
Team evaluation of leader effectiveness	2	221	.90 (.26)	.91	.001*	.38 to 1.4
Task satisfaction	2	514	.52 (.18)	.93	.001*	.18 to .87
Leader prototypicality						
Team evaluation of leader effectiveness	3	252	.67 (.33)	.96	.04	.03 to 1.3
Liking of leader	2	314	.68 (.18)	.89	.001*	.31 to 1.0
Leader positive affective display						
Team evaluation of leader effectiveness	2	679	.98 (.27)	.98	.001*	.45 to 1.5
Leader negative affective display						
Evaluation of leader effectiveness	3	996	-.60 (.11)	.92	.001*	-.82 to -.39
Team evaluation of leader effectiveness	2	679	-.51 (.42)	.99	.23	-1.3 to .31
Leader positive emotion						
Team evaluation of leader effectiveness	2	184	1.92 (.28)	.92	.001*	1.3 to 1.9

Leader negative emotion Evaluation of leader effectiveness	2	497	-.78 (.05)	.00	.001*	-.87 to -.69
Leader humility Evaluation of leader effectiveness	2	644	.05(.05)	.37	.31	-.05 to .15

Note: k = number of samples; n = total sample size; *d* = corrected standardized mean difference; SE = standard error; I = total heterogeneity/total variability; **p*-value < .001; 95% CI = 95% confidence internal. Any necessary conversions (e.g., means and standard deviations of treatment and control groups to Cohen's *d*) utilized David Wilson's "Practical Meta-Analysis Effect Size Calculator

Charismatic leadership. Six outcomes on followers are reported relating to charismatic leadership. There is a positive relation between charismatic leadership and follower task performance (quantity) (*d* = .20, *k* = 11, n = 1,769, SE = .10, 95% CI [-.02, .41], I^2 = .95), task performance (quality) (*d* = .07, *k* = 8, n = 1,942, SE = .06, 95% CI [-.05, .18], I^2 = .80), trust in the leader (*d* = .01, *k* = 2, n = 489, SE = .21, 95% CI [-.42, .42], I^2 = .95), team prosocial behavior (*d* = .27, *k* = 3, n = 221, SE = .14, 95% CI [-.02, .55], I^2 = .77), and followers' task satisfaction (*d* = .19, *k* = 2, n = 426, SE = .05, 95% CI [.10, .29], I^2 = .00). There is a slightly negative relation between charismatic leadership and follower extra-role performance (quantity) (*d* = -.07, *k* = 6, n = 772, SE = .10, 95% CI [-.26, .13], I^2 = .87) and follower extra-role performance (quality) (*d* = -.13, *k* = 6, n = 772, SE = .09, 95% CI [-.31, .06], I^2 = .85). All of the effects related to charismatic leadership are relatively close in magnitude. There is a relatively high amount of heterogeneity of across samples for charismatic leadership and its effects as well, except for transformational leadership and followers' task satisfaction.

Transformational leadership. A total of eight different outcomes are related to transformational leadership. It is associated with follower task performance (quality) (*d* = .19, *k* = 2, n = 236, SE = .16, 95% CI [-.12, .50], I^2 = .81, extra-role performance (quantity) (*d* = .23, *k* = 3, n = 326, SE = .10, 95% CI [.03, .42], I^2 = .66), and task satisfaction (*d* = -.01, *k* = 2, n = 236, SE = .25, 95% CI [-.29, .82], I^2 = .97). Transformational leadership is also associated with team

task performance (quality) (d = .26, k = 3, n = 414, SE = .28, 95% CI [-.11, .15], I^2 = .97), team evaluations of leader effectiveness (d = .87, k = 2, n = 231, SE = .07, 95% CI [.74, .99], I^2 = .00), team liking of leader (d = .82, k = 2, n = 231, SE = .10, 95% CI [.63, 1.0], I^2 = .48), and team task satisfaction (d = .22, k = 2, n = 231, SE = .53, 95% CI [.72, 1.0], I^2 = .99). Most of the relationships are positively associated, except for the effects of transformational leadership and followers' task satisfaction. There are moderate to high levels of heterogeneity of across samples with the exception of transformational leadership and team evaluations of leader effectiveness, which is highly homogeneous.

Empowering and participative leadership. Four outcomes related to empowering leadership and four outcomes related to participative leadership are reported. Empowering leadership is positively associated with team task performance (quantity) (d = .43, k = 3, n = 603, SE = .14, 95% CI [.17, .70], I^2 = .91), team task performance (quality) (d = .18, k = 2, n = 342, SE = .18, 95% CI [-.17, .53], I^2 = .91), individual satisfaction with leader (d = .54, k = 2, n = 298, SE = .38, 95% CI [-.22, 1.2], I^2 = .96), and team task satisfaction (d = .08, k = 2, n = 342, SE = .53, 95% CI [-.95, 1.1], I^2 = .99). Participative leadership is positively associated with team task performance (quantity) (d = .05, k = 3, n = 341, SE = .34, 95% CI [-.62, .73], I^2 = .97), team evaluations of leader effectiveness (d = .90, k = 2, n = 221, SE = .26, 95% CI [.38, 1.4], I^2 = .91), individuals' task satisfaction (d = .52, k = 2, n = 514, SE = .18, 95% CI [.18, .87], I^2 = .93) and group-orientedness (d = .83, k = 2, n = 217, SE = .24, 95% CI [.36, 1.3], I^2 = .91). Most of the effect size estimates for participative leadership are relatively larger than the estimates for empowering leadership. However, the relationship between empowering leadership and team task satisfaction and the relationship between participative leadership and team task performance (quantity) are relatively smaller in magnitude than the other six effects presented. There is a

relatively high level of heterogeneity across samples for empowering and participative leadership and their effects.

Destructive leadership. There are six outcomes reported that are related to destructive leadership. Destructive leadership is negatively associated with task performance (quantity) (d = -.62, k = 3, n = 280, SE = .21, 95% CI [-1.1, .43], I^2 = .91), task performance (quality) (d = -.39, k = 2, n = 196, SE = .33, 95% CI [-1.0, -.21], I^2 = .95), follower prosocial behavior (d = -1.10, k = 3, n = 352, SE = .45, 95% CI [-1.9, -.21], I^2 = .99), and creativity (d = -.53, k = 2, n = 196, SE = .07, 95% CI [-.67, -.39], I^2 = .00). Destructive leadership is positively associated with followers' negative affect (d = .23, k = 3, n = 280, SE = .27, 95% CI [-.31, .76], I^2 = .95) and follower deviance (d = .39, k = 2, n = 254, SE = .15, 95% CI [.09, .69], I^2 = .78). The magnitude of the negative effects for destructive leadership are relatively larger in magnitude than the positive effects for destructive leadership and follower deviance and negative affect. There are relatively high levels of heterogeneity across samples with the exception of destructive leadership and creativity, which are highly homogeneous.

Leader vision and prototypicality. Leader vision is positively associated with follower trust in the leader (d = .29, k = 2, n = 505, SE = .70, 95% CI [-1.0, 1.6], I^2 = .99) and follower perceived charisma of the leader (d = .11, k = 3, n = 809, SE = .57, 95% CI [-1.0, 1.2], I^2 = .99). Leader prototypicality is positively associated with team evaluations of leader effectiveness (d = .67, k = 3, n = 252, SE = .33, 95% CI [.03, 1.3], I^2 = .96) and followers' liking of the leader (d = .68, k = 2, n = 314, SE = .18, 95% CI [.31, 1.0], I^2 = .89). The effects reported for leader prototypicality are relatively larger in magnitude than the effects associated with leader vision. Heterogeneity is relatively high across all samples for these effects.

Leader affective displays, emotion, and humility. Leader positive affective displays are positively related to team evaluations of leader effectiveness (d = .98, k = 2, n = 679, SE = .27, 95% CI [.45, 1.5], I^2 = .98) and leader positive emotional displays are positively associated with team evaluations of leader effectiveness (d = 1.92, k = 2, n = 184, SE = .28, 95% CI [1.3, 1.9], I^2 = .92). Leader humility is positively related to individual followers' evaluations of leader effectiveness (d = .05, k = 2, n = 644, SE = .05, 95% CI [-.05, .15], I^2 = .37). Leader negative affective displays are negatively related to evaluations of leader effectiveness at the individual (d = -.60, k = 3, n = 996, SE = .11, 95% CI [-.82, -.39], I^2 = .92) and team level (d = -.51, k = 2, n = 679, SE = .42, 95% CI [-1.3, .31], I^2 = .99). Finally, leader negative emotion is also negatively associated with followers' evaluations of leader effectiveness (d = -.78, k = 2, n = 497, SE = .05, 95% CI [-.87, -.69], I^2 = .00). The estimates associated with leaders' displays of positive affect and emotion are relatively larger in magnitude than the effect of negative leader displays of affect and emotion. The heterogeneity across samples is relatively high, with the exception of the relationship for both leader humility and leader negative emotion and followers' evaluations of leader effectiveness.

Since experiments containing more advanced methodological designs and techniques could not be included in the same meta-analytic estimates reported in Table 2 (e.g., instrumental variables, propensity scoring, etc.), a third systematic search was conducted solely for the purpose of summarizing the findings of these experiments. The results of these studies are reported in a narrative fashion in the following section.

3.3 Quasi and Non-traditional Experiments

For this section I adopt a selective narrative review focused on the effect size estimates of experiments that utilize more advanced analyses and methodologies (e.g., instrumental variables,

propensity scoring, regression discontinuity), and thus could not be integrated in the meta-analytic effect size estimates obtained in the previous search. These strategies were deemed as potential solutions suggested by Hill et al. (2021) and Antonakis et al. (2010) for remedying endogeneity bias. Any quasi-experiments that appeared in the previous search were reassessed for sufficient criteria to be included.

Instrumental variables. An effective way to reduce endogeneity in the form of simultaneity is to perform a two-step equation technique in which you replace the endogenous variable with a predicted value or include a calculated control variable (Hill et al., 20 21). This is typically done with the use of 2SLS (Angrist & Imbens, 1995; Wooldridge, 1997). MacLaren et al. (2020) utilized this approach by testing the "babble hypothesis," which posits that those who speak more are likelier to be perceived as the leader. They selected intelligence and five personality variables from the NEO-FFI scale (Costa & McCrea, 1992) as their primary instrumental variables, since both have been shown to be highly correlated with leader emergence (Zaccaro et al., 2018; Ensari et al., 2011). Thus, they demonstrate a strong causal effect of speaking time on leader emergence that is high in ecological validity while greatly mitigating the threat of endogeneity. Many other experimental designs, both in laboratories and in the field, have successfully implemented an instrumental variable approach to determine causal effects for commonly studied leadership constructs such as transformational leadership (e.g., de Vries, 2012; Artz et al., 2017; Azoulay et al., 2017) and gender (e.g., Chen et al., 2017; Bernile et al., 2018; Amore et al., 2014; Adhikari, 2018).

Regression discontinuity. Regression discontinuity has been highly praised by Cook and colleagues (2002) for mitigating biases and various threats to internal validity, particularly when a quasi-experimental methodology is the most suited to answer the researcher's question. When

paired with quasi-experimental designs, regression discontinuity allows the researcher to use pre-existing environmental conditions as a natural cutoff for determining random assignment (Thistlethwaite & Campbell, 1960). However, the variable used to determine this cutoff must be a continuous variable (Hill et al., 2021). This cutoff point is selected so that a natural partition occurs between the treatment group and the control group. An effect is implied only if the regression equation illustrates a discontinuity (Hahn et al., 2001).

Arvate et al. (2018) leveraged a regression discontinuity to investigate the "queen bee" phenomenon, which posits that women receive less support from female leaders (Staines et al., 1974), and has previously suffered from endogeneity bias in the form of reverse causality and the third variable problem. They critically evaluate this phenomenon by exploring the potential inequality in earnings for women leaders in both public and private firms in municipalities where women were elected as mayors. The cutoff point in this context is the time point before and after the election of a female mayor. They based their findings on a sample of 8.3 million organizations across 5,600 Brazilian municipalities and largely disproved any evidence of the "queen bee" phenomenon. Other studies have used regression discontinuity to examine charismatic leadership (e.g., Bastardoz et al., 2022; Butler, 2009), transformational leadership (e.g., Grönqvist & Lindqvist, 2016), and leadership in teams (e.g., Dunning & Nilekani, 2013; Lechler & McNamee, 2018).

Propensity scoring. The creation of synthetic control groups in the form of propensity score matching allows the researcher to create a new control group that allows for comparisons across groups while reducing the risk of endogeneity through means of highly specified treatment selection (Hill et al., 2021). However, both groups must display similar observable traits or features (Caliendo & Kopeinig, 2008). The goal is to ideally balance the covariate

distribution for both the treatment and control (Stuart, 2010). One example of this can be seen with Li et al. (2020), in which they tested the potential malleability of personality based on a person-environment perspective. To do this, they compared the personalities of those promoted to a leader role (treatment group) to those who did not (control group) across two experiments and multiple time points using propensity score matching. Their results partially supported that an individual's conscientiousness can slightly increase after transitioning into a leader role based on the context and situational pressures involved with assuming a leadership position.

CHAPTER 4: DISCUSSION

4.1 Theoretical Implications

Several overarching theoretical implications can be seen as a result of this comprehensive systematic review. First and foremost, this work presents the first purely causal leadership effect sizes to date. Second, when correlations are influenced by sources of endogeneity bias, such as omitted variables, common-method bias, or selection effects, they also include the effects of unmeasured causes (Hill et al., 2021). This means that the majority of observed correlations and pre-existing benchmarks that have been relied upon by leadership scholars to date are not causally identified (Antonakis et al., 2010; Kennedy, 2008). As such, the continuum of effect sizes provided from the three unique reviews highlight these key differences in effect size strength and interpretability. Third, this systematic review suggests that a multitude of other issues influence the utility of the reported effect sizes that are casually identified, thus resulting in a surprising lack of experiments on leadership processes and phenomena that are high in rigor and quality.

Second-order Meta-analysis of Leader Individual Differences. A few major takeaways can be made from the estimates reported in the second-order meta. Firstly, there is still much more to be learned regarding the stable individual differences of leaders (Antonakis et al., 2012) and biological causes of leader emergence and leader effectiveness, such as genes (Van Vugt & von Rueden, 2020; Götz et al., 2022). The general high-level of heterogeneity displayed across samples for many of the effects demonstrate a need for additional work exploring potential moderators. Second, the leadership field in general is lacking fully reliable updated sources for meta-analytic estimates that relate key individual differences of leaders/managers to effects presented in Table 1, many do not meet modern standards of rigor or

robustness. For example, several of the meta-analyses on emotional intelligence did not provide confidence intervals and nearly half of the estimates for Big Five traits like emotional stability and extraversion are lacking any evidence of homogeneous samples. Third, for those that did include tests of heterogeneity or credibility intervals, most were wide/large. This can prove to be problematic in various ways, such as when researchers need to determine the correct sample size for performing a power analysis.

Fourth, the findings illustrated in the second-order meta-analysis (See Table 1) demonstrate a more wide-ranging and comprehensive overview of previously reported effect size estimates contained within prior meta-analyses, as well as a more critical perspective of the effects demonstrated through the additional reported information of credibility intervals, I^2, etc. when possible. This helps provide a comprehensive look into the reported estimates of commonly studied antecedents and outcomes pertaining to the leadership literature and furthers the conversations introduced by Derue et al. (2011) and Badura et al. (2020) with their respective second-order meta-analyses that examined similar antecedents and outcomes. Additionally, these estimates include antecedents and outcomes overlooked in Derue et al.'s (2011), Badura et al.'s (2020), or both, while simultaneously updating many of the effects they reported for highly relevant leadership variables like gender and leader emergence. Although the effect sizes from the second-order meta presented in Table 1 are an improvement from many previous correlational benchmarks in general, the following sections discuss the effects with increased causal strength. This gets at a closer realistic preview for capturing effects and goes beyond the typical effect size estimates that are reported in many meta-analyses (e.g., Hoch et al., 2018; Banks et al., 2016; Miao et al., 2018) which fail to meet the discussed definition of an effect size (Flora, 2020).

Lab and field experiments. Numerous theoretical implications can be taken from the causal effect size estimates presented. Perhaps the most striking is the surprising dearth of experiments on leadership processes and phenomena that are high in rigor and quality, with some leading scholars recently suggesting the same dismay (e.g., Eden, 2021; Martin et al., 2021). This is also evidenced in previous meta-analytic reviews on leadership from past (Dumdum et al., 2002) to present (Chandler et al., 2022). For example, Martin et al. (2021) recently found in a systematic review of the leadership training literature that the majority of experiments conducted in this area do not meet many of the criteria needed to establish causality. The only exception to this in the current review may be charismatic leadership, which had eleven samples that examined the relationship between charismatic leadership and task performance (quantity). One explanation for this finding is the prospective meta-analysis by Ernst et al. (2021), which included six separate experiments exploring the effect of charismatic leadership on follower task performance and extra-role performance. This may suggest a need for more prospective meta-analyses in the future.

Many of the bivariate relationships presented in Table 2 exhibit high levels of heterogeneity. This suggests that there is a need to explore additional moderators for many of these relationships. Another takeaway can be seen in relation to the effects associated with destructive leadership. For example, although it is not surprising that destructive leadership was negatively related to follower task performance (quantity) and prosocial behaviors, it is interesting that the magnitude for these effects are so relatively large. This further supports the notion that even though destructive leadership may be a low base rate phenomenon, its effects can be highly detrimental to followers in a number of ways (Schyns & Schilling, 2013; Mackey et al., 2021).

A multitude of issues regarding the quality of casually identified effects from primary studies largely contributed to the low number of studies included in the final sample. For example, a large number of studies failed to compare their treatment group to a suitable control group. Instead, many designed their studies to compare two treatment groups to one another (e.g., leader self-sacrifice vs. leader self-benefitting), which is considered an unfair comparison since it is unknown which condition caused the effect (Lonati et al., 2018). Another reason for the low number of studies was the absence of sufficient effect size information that could be converted to Cohen's *d*. For instance, many of the experiments pre-1960 only reported the means of the treatment and control groups when the standard deviations are necessary as well for conversion purposes. This may be because the reporting standards of published experiments in this time period were vastly different than today (Thompson, 1999; Aguinis et al., 2010). On the other hand, a good number of the experiments in recent years explored mediation and moderation effects without explicitly stating the direct effects or performed solely multivariate analyses, which can increase the risk of jeopardizing statistical validity (Lonati et al., 2018). It is postulated that both of these common occurrences may be due to the increased emphasis on producing novel findings (Collins, 1985; Nosek et al., 2012; Aguinis et al., 2020). The number of studies excluded for these various reasons gives pause for concern. Methodological recommendations are provided in a future section to address this.

Quasi and non-traditional experiments. A main implication that can be taken from the narrative review is the difference in strength across causal claims. This becomes highly evident when comparing the typical correlational effect size estimates to lab study, or a "gold standard" lab experiment to a controlled lab setting with a method that reduces endogeneity risks. Studying the personality traits of leaders is one instance that is highly susceptible to potential confounds in

experimental settings. For example, Bottger (1984) found a strong effect between perceived leader influence and speaking time. However, Bottger's (1984) experiment failed to account for the high covariation between leader influence and leader expertise (the study's other predictor variable). As such, their claim that perceived leader influence is a strong predictor of speaking time is greatly weakened. Many other studies have attempted to experimentally evaluate speaking time and leader emergence, but also succumbed to forms of endogeneity bias, thus misconstruing this relationship. For instance, Morris and Hackman's (1969) results suffered from common-method bias and Kremer and Mack (1983) did not control for intelligence or personality.

Despite previous discrepancies in results, MacLaren et al. (2020) utilized an instrumental variable approach to control for intelligence and five personality variables from the NEO-FFI scale (Costa & McCrea, 1992) to assess the same proposed relationship. Their results indicated a relatively large causal effect of speaking time and leader emergence that is both high in ecological validity while greatly mitigating the threat of simultaneity bias. MacLaren et al.'s (2020) methodological rigor suggests that their strength in claims differ significantly when compared to previous lab studies examining the same effects. Furthermore, leadership scholars should consider the influence endogeneity bias can have when designing studies. If present, these concerns can negate any causal assumptions being made and serve as serious threats to internal validity (Antonakis et al., 2010; Hill et al., 2021). Or even worse, it can impede scientific progress by resulting in false recommendations for theory and practice (Güntner et al., 2020; Schmidt & Pohler, 2018).

4.2 Practical Implications

There are several practical implications that can be drawn from the systematic review. First, the effects reported from the second-order meta and the narrative review cast a wide net of understanding for key outcomes associated with the individual differences of leaders. Focusing on leader-to-follower effects also sheds more light on the relationship between leaders and their followers in organizations as well as the social influence process (Yukl & Falbe, 1990; Furst & Cable, 2008; Day & Antonakis, 2013). Perhaps more importantly, it may suggest what individual differences should be selected when hiring for leadership roles and the effects certain leader behaviors may have on followers' evaluations or perceptions of them. For instance, assuming that speaking time continues to show a causal effect on leader emergence over a series of additional studies, this may be something a hiring manager would want to consider when observing group interactions or a practitioner that is evaluating performance during a leaderless group discussion, a common exercise required in assessment centers (Thornton et al., 2014). It is useful to know which leader personality traits are correlated with potentially important positive outcomes like perceived leader effectiveness and group performance. And it can be equally important to note which leader traits (e.g., Machiavellianism) predict negative outcomes like abusive supervision or which leader behaviors (e.g., destructive leadership) are associated with negative outcomes like a significant decrease in follower task performance.

It is also important to note that some of the effects presented may also be more practically significant despite their proposed relatively small magnitude (Götz et al., 2022; Flora, 2020). For instance, objectively and causally measuring followers' number of helping behaviors in response to leadership behaviors (e.g., Porath and Erez, 2007) may hold more practical significance despite it being deemed "small" in magnitude per current benchmark standards. A "large effect" can be seen as meaningful to a practitioner despite that it might be conflated, and a "small effect"

might be viewed as not meaningful, and thus completely overlooked when making informative decisions (McCartney & Rosenthal, 2000; Martin et al., 2021; Götz et al., 2022). Furthermore, Martin and colleagues (2021) postulated that a large effect size from a poorly designed study is less theoretically and practically meaningful than a smaller effect size from a well-designed study. Bosco et al. (2015) accounted for some of these differences by contextualizing small changes in evlauted behavior and small changes in attitudes with separate benchmarks. One must consider the potential added practical significance that this may indicate.

4.3 Limitations and Future Directions

A number of limitations are present from the systematic review. First, the small number of *k*s for the casual bivariate relationships suggests that more casually identified effects are needed to increase the robustness of the evidence base from random-sampling error. the confidence of the relationships posed in this study. Second, the small number of samples also restricted the testing of any theoretical or methodological moderators. Furthermore, contextual moderators should be explored in the future as well such as leadership behaviors enacted in virtual contexts vs. in-person and leadership in a crisis. These lenses allow for more dynamic interpretations of leadership which can portray a more accurate account of leader social influence and may even result in making leader behaviors more endogenous than typically implied by scholars (Güntner et al., 2020). Third, the selective nature of the review for quasi and non-traditional experiments limits the scope of experimental leadership studies that successfully mitigated forms of endogeneity bias. Fourth, the inclusion criteria for the causal effect sizes were highly stringent, and although many of the decisions were based on widely accepted standards in experimental research (e.g., Cook et al., 1979; Cook et al., 2002; Podsakoff and Podsakoff,

2019), some can be considered more polarizing such as the exclusion of vignette/scenario studies (Aguinis and Bradley, 2014; Fischer et al., 2021).

There are several potential directions for future research to aid in the theoretical and practical advancement of leadership theory. First, the conflation of behaviors with evaluations can lead to faulty comparisons (Fischer et al., 2020; Banks et al., 2021). This may result in inaccurate effect size interpretations. Conflating behavior-evaluation effect sizes (x_1 and y_1) with evaluation-evaluation effect sizes (y_1 and y_2) is akin to comparing "apples and oranges." Second, incorporating context into effect size reporting is critical to their interpretation (Götz et al., 2022), with initial steps in this direction already demonstrated by the benchmarks of Bosco et al (2015) and Gignac and Szodorai (2016) in a broad sense.

Conflation of evaluations and behaviors. Behaviors can be defined as, "the internally coordinated responses (actions or inactions) of whole living organisms (individuals or groups) to internal and/or external stimuli, excluding responses more easily understood as developmental changes." (Levitis et al., 2009; p. 103). Evaluations on the other hand can be characterized as, "appraisals of behavior." (Banks et al., 2021). The conflation of behaviors with evaluations can lead to faulty comparisons (Fischer et al., 2020; Banks et al., 2021), which may result in inaccurate effect size interpretations. Comparing behavior-behavior, behavior-evaluations, or evaluation-evaluation effect sizes in the same meta-analytic estimates can be misleading. Furthermore, evaluations and behaviors are equally important and should be treated uniquely in their interpretation. One example of examining a solely behavior-behavior bivariate relationship can be seen with Chen (2012), in which the effect of rewarding vs. punishing leader behaviors on follower reporting choices were examined. An example of an evaluation-evaluation derived effect size is general leader charisma and follower trust in leader (Rast, 2016).

A major issue with the conflation of behaviors and evaluations is that the overall impact or value of effect sizes are potentially being overestimated or underestimated when interpreted in practice settings. In other words, a "large effect" can be seen as meaningful to a practitioner despite that it might be conflated, and a "small effect" might be viewed as not meaningful, and thus completely overlooked when making informative decisions (McCartney & Rosenthal, 2000; Martin et al., 2021). One potential reason for this may be that the effect sizes that get published tend to be larger than the population effect (perhaps due to the pressures of publication) (Flora, 2020). Regardless, this discrepancy in typical small, medium, and large classification of effect sizes determined by Cohen's (1962, 1988) benchmarks has little resemblance to findings in the field (Bosco et al., 2015), especially in the medical sciences where "small" effects can have substantial implications for individuals' health and safety (Phan, 2021). Dvir et al. (2002) relied on the effect size benchmarks of Cohen (1988) to find that there was a small effect of transformational leadership on follower extra-role performance. However, when considering Bosco et al.'s (2015) context-specific benchmarks, this same effect would be considered a medium effect. This disconnect in interpretability and discrepancy in value is problematic and should continue to be revised.

Context matters. Context can be defined as, "stimuli and phenomena that surround and thus exist in the environment external to the individual, most often at a different level of analysis." (Mowday & Sutton, 1993, p. 198). Johns (2006) took this interpretation one step further by adding that discrete context refers to specific situational variables that can directly influence behavior, and thus illustrates the various roles context can play in organizations. Bosco et al.'s (2015) updated benchmarks improved upon Cohen's (1962, 1988) by addressing major contextual issues. The importance of context in leadership studies has persisted over time (e.g.,

Bryman et al., 1996; Alvesson, 1996; Gardner et al., 2020). And many contextual factors have been found to impact leaders' social influence process on followers (e.g., Wofford, 1994; Vroom & Jago, 2007; Antonakis et al., 2003), thus it must be considered for determining accurate effect size benchmarks in leadership research (Götz et al., 2022). Bosco et al.'s (2015) and Gignac and Szodorai's (2016) context-specific benchmarks take interpretability one step further in a positive direction, but additional contextualization is needed for leadership effect size benchmarks to be accurate and hold theoretical and practical value (e.g., comparing virtual to in-person leadership).

Pillai (1996) and Hunt (1999) both conducted experiments in which they examined the effectiveness of various leadership styles through the manipulation of a crisis vs. non-crisis context, a critical contextual moderator that has grown in popularity. Mackey et al.'s (2017) meta-analytic findings showed that contextual factors involving followers and their work environments (e.g., cross-cultural differences) can influence the relationship with their supervisor as well as their perception of abusive supervision. The contextualization of effect size benchmarks can better inform practical value as well through the use of power analyses (Cashen & Geiger, 2004; Ellis, 2010), Bayesian analysis (Kruschke et al., 2012; Jackman, 2009), or the interpretation of theoretical or practical significance (Aguinis et al., 2010; Brooks et al., 2014). Despite the difficulty it might present to incorporate and account for contextual factors and their influence on behaviors, it must be accounted for in regard to effect size magnitudes (Bosco et al., 2015; Aguinis & Pierce, 2006; Götz et al., 2022; Cumming, 2014).

4.4 Methodological Best Practices

Two strong methodological ways to estimate the practical value of effects are the use of Bayesian analysis (Kruschke et al., 2012; Jackman, 2009) or through the interpretation of practical significance (Aguinis et al., 2010; Ellis, 2010; Brooks et al., 2014). Bayesian analysis,

which originated from computations developed by Bayes and Price (1763), provides researchers with a distribution of potential credible parameter estimates across multiple predictors. This allows them to evaluate the trade-offs for each parameter and make decisions accordingly (Jackman, 2009). Advantages of performing a Bayesian analysis in place of a power analysis are its ability to consider prior knowledge, create joint distributions of parameters, effectively assess the null hypothesis, and evaluate uneven sample sizes across groups or conditions (Kruschke et al., 2012).

Practical significance requires the consumer of research to make a judgment concerning the value of a set of results in regard to its implications for a specific decision (Vaske et al., 2002), and ask whether the results are notable and or substantial enough to truly matter (Armstrong & Henson, 2004). One way that practical significance can be determined in an effective manner according to Aguinis et al. (2010) is through the use of qualitative methodologies post hoc with practitioners selected for the sample. This process could aid in the effort for stronger contextualization of effect sizes and highlight findings that may have been overlooked per traditional standards. Evaluating the practical value of effect sizes can be important for us to be able to contextualize whether a phenomenon is a big effect or not and to fully understand the true magnitude of these effects in general.

There are also a multitude of methodological best practices that should be considered when conducting experimental research that are largely apparent from the systematic review. The number of leadership experiments that displayed high levels of rigor through their methodological designs are scant to say the least. Because of this, it is highly recommended that leadership scholars revisit the guidelines put forth by Cook et al. (1979), Cook et al. (2002), and Lonati et al. (2018) as a guide to their work. For instance, random assignment and avoiding

unfair comparisons by comparing a treatment group to a suitable control group are several practices that should be held paramount. Other specific recommendations may involve the careful design of priming techniques or manipulation checks (Wulff et al., 2023). More experiments should also examine the effects followers have on leaders (e.g., Ashkanasy & Paulsen, 2013; Ahmad et al., 2020) and organizational level factors as predictors (e.g., Desmet et al., 2015; Lee & Kray, 2021) for helping explain leadership as a dynamic process (Oc & Bashshur, 2013).

Researchers should also review previous noteworthy experiments in their respective sub-domains of leadership. Martin et al., (2013) (empowering leadership), Dvir et al. (2002) (transformational leadership), and Porath and Erez (2007) (destructive leadership) are all exemplars of high-quality experiments in their respective areas of leadership research that adhere to experimental standards. For example, Porath and Erez (2007) trained a confederate leader to be rude to participants to elicit the mundane realism for examining treatment effects across three experiments. Other scholars that study destructive leadership or abusive supervision may benefit from utilizing similar strategies in their designs that get away from the overreliance of vignette methods (Fischer et al., 2021). Furthermore, an increase in emphasis around replication studies should be more encouraged by leadership scholars and social science researchers in general to promote a more robust, rigorous, and reproducible science (Eden, 2021; Furchtgott, 1984; Makel et al., 2012). Wiggins and Chrisopherson (2019) recently found that less than half of studies published in psychology journals are replicated, which is alarming, especially when compared to other disciplines like economics (Camerer et al., 2016). In addition, looking across disciplines at experimental work or guidelines proposed in economics (e.g., Frackenphol et al., 2016; Van der Heijden et al., 2013; Caliendo & Kopeinig, 2008) or political science (e.g., Dunning & Nilekani,

2013; Lechler & McNamee, 2018; John, 2017) may be ripe areas for deriving applicable ideas for methodological designs that result in stronger causal inferences. These directions discussed would strengthen experimental work in leadership research, help get scholars away from traditional publication pressures (Aguinis et al., 2020; Nosek et al., 2012), and are directly conducive to building a better science (Wulff et al. 2023).

Some other methodological best practices to apply when deemed appropriate may include the use of video-based designs (Podsakoff et al., 2013) or virtual reality tools (Aguinis & Edwards, 2014; Raymondie et al., 2013) to embody more realistic dyadic interactions, utilizing public good or public bad games to examine leader and follower outcomes (Frackenphol et al., 2016; Van der Heijden et al., 2013), or the use of a triangulation approach (Banks et al., 2022) to capture causal effects. Researchers should also consider using more objective measures to study leader and follower behaviors such as examining leader decision making speed (Van de Caleseyde et al., 2021), speaking time (MacLaren et al., 2020), hand gestures (Clarke et al., 2021), or eye gaze (Maran et al., 2019). In terms of additional analyses considerations, one may want to take into account the appropriateness of a 2SLS or a 3SLS and proper reporting standards associated with both (Bastardoz et al., 2023; Sajons, 2020), how causal inferences can be strengthened in mediation analyses (MacKinnon and Pirlott, 2015), interpreting coefficients from regression results when analyzing 2x2 experiments (Wulff et al., 2023), leveraging panel data (Bliese et al., 2020), or ways to increase the precision of estimates like the precise parameter estimation (PPE) approach which can prioritize desirable confidence interval width over statistical significance (Tonidandel et al., 2014).

4.5 Conclusion

Despite the importance of effect sizes for leadership scholars, endogeneity bias has rendered the correlation effect size benchmarks available to leadership scholars as highly limited. To illustrate this, I performed a comprehensive systematic review to introduce a continuum of novel effect size estimates that demonstrate the strength of causal claims. In doing so, this paper made several key contributions by (1) presenting a second-order meta-analysis of leader individual difference variables (total k = 1,829; total N = 640,388), (2) meta-analyzing experiments (total k = 110; total N = 18,402) representing purely causal effects, and (3) providing a selective narrative review of quasi-experimental and non-traditional experimental designs. The findings from the continuum of effect size estimates further supports and justifies the need for creating a new set of effect size benchmarks for the leadership field, one that contains estimates solely causal in nature, increases the level of precision for researchers, and accounts for context. Furthermore, the surprising dearth of rigorous experiments suggests that leadership scholars will need to produce additional high-quality experiments for these casual benchmarks to become a reality. Thus, it is posited that this paper serves as the beginning of a long-series of work that must address each of these challenges moving forward for effect size benchmarks to be more meaningful to scholars and practitioners.

REFERENCES

Adhikari, B. K. (2018). Female executives and corporate cash holdings. *Applied Economics Letters*, *25*(13), 958-963.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *Journal of Applied Psychology*, *90*(1), 94.

Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, *17*(4), 351-371.

Aguinis, H., Cummings, C., Ramani, R. S., & Cummings, T. G. (2020). "An A is an A": The new bottom line for valuing academic research. *Academy of Management Perspectives*, *34*(1), 135-154.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2011). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, *37*(1), 5-38.

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, *51*(1), 143-174.

Aguinis, H., & Pierce, C. A. (2006). Computation of effect size for moderating effects of categorical variables in multiple regression. *Applied Psychological Measurement*, *30*(5), 440-442.

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2008). Comparison of three meta-analytic procedures for estimating moderating effects of categorical variables. *Organizational Research Methods*, *11*(1), 9-34.

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 569-595.

Aguinis, H., Vassar, M., & Wayant, C. (2021). On reporting and interpreting statistical significance and p values in medical research. *BMJ Evidence-Based Medicine*, *26*(2), 39-42.

Aguinis, H., Werner, S., Lanza Abbott, J., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*(3), 515-539.

Ahmad, M. G., Klotz, A. C., & Bolino, M. C. (2021). Can good followers create unethical leaders? How follower citizenship leads to leader moral licensing and unethical behavior. *Journal of Applied Psychology*, *106*(9), 1374.

Amore, M. D., Garofalo, O., & Minichilli, A. (2014). Gender interactions within the family firm. *Management Science*, *60*(5), 1083-1097.

Antonakis, J. (2023). In support of slow science: Robust, open, and multidisciplinary. *The Leadership Quarterly*, 101676.

Antonakis, J., Avolio, B. J., & Sivasubramaniam, N. (2003). Context and leadership: An examination of the nine-factor full-range leadership theory using the Multifactor Leadership Questionnaire. *The Leadership Quarterly*, *14*(3), 261-295.

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*, 293-319.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086-1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. *The Oxford Handbook of Leadership and Organizations*, *1*, 93-117.

Antonakis, J., Day, D. V., & Schyns, B. (2012). Leadership and individual differences: At the cusp of a renaissance. *The Leadership Quarterly*, *23*(4), 643-650.

Antonakis, J., & Dietz, J. (2011). Looking for validity or testing it? The perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. *Personality and Individual Differences*, *50*(3), 409-415.

Armstrong, S. A., & Henson, R. K. (2004). Statistical and Practical Significance in the IJPT: A Research Review from 1993-2003. *International Journal of Play Therapy*, *13*(2), 9.

Artz, B. M., Goodall, A. H., & Oswald, A. J. (2017). Boss competence and worker well-being. *IlR Review*, *70*(2), 419-450.

Arvate, P. R., Galilea, G. W., & Todescat, I. (2018). The queen bee: a myth? the effect of top-level female leadership on subordinate females. *The Leadership Quarterly*, *29*(5), 533–548.

Ashkanasy, N. M., & Paulsen, N. (2013). The influence of follower mood on leader mood and task performance: An affective, follower-centric perspective of leadership. *The Leadership Quarterly*, *24*(4), 496-515.

Azoulay, P., Liu, C. C., & Stuart, T. E. (2017). Social influence given (partially) deliberate matching: Career imprints in the creation of academic entrepreneurs. *American Journal of Sociology*, *122*(4), 1223-1271.

Badura, K. L., Grijalva, E., Newman, D. A., Yan, T. T., & Jeon, G. (2018). Gender and leadership emergence: A meta-analysis and explanatory model. *Personnel Psychology*, *71*(3), 335-367.

Banks, G. C., Engemann, K. N., Williams, C. E., Gooty, J., McCauley, K. D., & Medaugh, M. R. (2017). A meta-analytic review and future research agenda of charismatic leadership. *The Leadership Quarterly*, *28*(4), 508-529.

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. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., ... & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5-20.

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

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323-338.

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.

Banks, G. C., Woznyj, H. M., Kepes, S., Batchelor, J. H., & McDaniel, M. A. (2018). A meta-analytic review of tipping compensation practices: An agency theory perspective. *Personnel Psychology, 71*(3), 457-478.

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

Barreto, N. B., & Hogg, M. A. (2017). Evaluation of and support for group prototypical leaders: A meta-analysis of twenty years of empirical research. *Social Influence*, *12*(1), 41-55.

Bastardoz, N., Jacquart, P., & Antonakis, J. (2022). Effect of crises on charisma signaling: A regression discontinuity design. *The Leadership Quarterly*, 101590.

Bastardoz, N., Matthews, M. J., Sajons, G. B., Ransom, T., Kelemen, T. K., & Matthews, S. H. (2023). Instrumental variables estimation: Assumptions, pitfalls, and guidelines. *The Leadership Quarterly*, 101673.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396-403.

Bernile, G., Bhagwat, V., & Yonker, S. (2018). Board diversity, firm risk, and corporate policies. *Journal of Financial Economics, 127*(3), 588-612.

Blake, K. R., & Gangestad, S. (2020). On attenuated interactions, measurement error, and statistical power: Guidelines for social and personality psychologists. *Personality and Social Psychology Bulletin*, *46*(12), 1702-1711.

Bliese, P. D., Schepker, D. J., Essman, S. M., & Ployhart, R. E. (2020). Bridging methodological divides between macro-and microresearch: Endogeneity and methods for panel data. *Journal of Management*, *46*(1), 70-99.

Bottger, P. C. (1984). Expertise and air time as bases of actual and perceived influence in problem-solving groups. *Journal of Applied Psychology*, *69*(2), 214.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological Methods*, *13*(2), 99.

Borenstein, M., Cooper, H., Hedges, L., & Valentine, J. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, *2*, 221-235.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431.

Boulu-Reshef, B., Holt, C. A., Rodgers, M. S., & Thomas-Hunt, M. C. (2020). The impact of leader communication on free-riding: An incentivized experiment with empowering and directive styles. *The Leadership Quarterly*, *31*(3), 101351.

Bowen, H. P., & Wiersema, M. F. (1999). Matching method to paradigm in strategy research: limitations of cross-sectional analysis and some methodological alternatives. *Strategic Management Journal*, *20*(7), 625-636.

Briner, R. B., & Rousseau, D. M. (2011). Evidence-based I–O psychology: Not there yet. *Industrial and Organizational Psychology*, *4*(1), 3-22.

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*(2), 332.

Brunell, A. B., Gentry, W. A., Campbell, W. K., Hoffman, B. J., Kuhnert, K. W., & DeMarree, K. G. (2008). Leader emergence: The case of the narcissistic leader. *Personality and Social Psychology Bulletin*, *34*(12), 1663-1676.

Bryman, A., Stephens, M., & a Campo, C. (1996). The importance of context: Qualitative research and the study of leadership. *The Leadership Quarterly*, *7*(3), 353-370.

Butler, D. M. (2009). A regression discontinuity design analysis of the incumbency advantage and tenure in the US House. *Electoral Studies*, *28*(1), 123-128.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research*, *54*(3), 427-456.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31-72.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433-1436.

Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal*, *37*(13), 2639-2657.

Chen, M. L. (2012). The effect of leader reward and punishment behaviors on subordinates' budget reports. *The Engineering Economist*, *57*(1), 41-54.

Chen, J., Leung, W. S., & Goergen, M. (2017). The impact of board gender composition on dividend payouts. *Journal of Corporate finance*, *43*, 86-105.

Chiaburu, D. S., Oh, I. S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: a meta-analysis. *Journal of Applied Psychology*, *96*(6), 1140.

Clarkson, B. G., Wagstaff, C. R., Arthur, C. A., & Thelwell, R. C. (2020). Leadership and the contagion of affective phenomena: A systematic review and mini meta-analysis. *European Journal of Social Psychology*, *50*(1), 61-80.

Clougherty, J. A., Duso, T., & Muck, J. (2016). Correcting for self-selection based endogeneity in management research: Review, recommendations and simulations. *Organizational Research Methods*, *19*(2), 286-347.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen-Charash, Y., & Spector, P. E. (2001). The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *86*(2), 278-321.

Connelly, B. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*, *66*(2), 407-442.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings* (Vol. 351). Boston: Houghton Mifflin.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (pp. 103-134). Boston, MA: Houghton Mifflin.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*(2), 161.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*, 287-308.

Cortina, J. M., Markell-Goldstein, H. M., Green, J. P., & Chang, Y. (2021). How are we testing interactions in latent variable models? Surging forward or fighting shy?. *Organizational Research Methods*, *24*(1), 26-54.

Costa Jr, P. T., & McCrae, R. R. (2008). *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc.

Crampton, S. M., & Wagner III, J. A. (1994). Percept-percept inflation in micro-organizational research: An investigation of prevalence and effect. *Journal of Applied Psychology*, *79*(1), 67.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*(3), 138-146.

Darto, M., Setyadi, D., Riadi, S. S., & Hariyadi, S. (2015). The effect of transformational leadership, religiosity, job satisfaction, and organizational culture on organizational citizenship behavior and employee performance in the regional offices of national

institute of public administration, Republic of Indonesia. *European Journal of Business and Management*, *7*(23), 205-219.

Dasborough, M. T., Ashkanasy, N. M., Humphrey, R. H., Harms, P. D., Credé, M., & Wood, D. (2022). Does leadership still not need emotional intelligence? Continuing "The Great EI Debate". *The Leadership Quarterly*, *33*(6), 101539.

Day, D. V., & Antonakis, J. (2013). The future of leadership. *The Wiley-Blackwell Handbook of the Psychology of Leadership, Change, and Organizational Development*, 221-235.

Derue, D. S., Nahrgang, J. D., Wellman, N. E., & Humphrey, S. E. (2011). Trait and behavioral theories of leadership: An integration and meta-analytic test of their relative validity. *Personnel Psychology*, *64*(1), 7-52.

Desmet, P. T., Hoogervorst, N., & Van Dijke, M. (2015). Prophets vs. profits: How market competition influences leaders' disciplining behavior towards ethical transgressions. *The Leadership Quarterly*, *26*(6), 1034-1050.

De Vries, R. E. (2012). Personality predictors of leadership styles and the self–other agreement problem. *The Leadership Quarterly*, *23*(5), 809-821.

Eden, D. (2021). The science of leadership: A journey from survey research to field experimentation. *The Leadership Quarterly*, *32*(3), 101472.

Dunning, T., & Nilekani, J. (2013). Ethnic quotas and political mobilization: caste, parties, and distribution in Indian village councils. *American Political Science Review*, *107*(1), 35-56.

Dvir, T., Eden, D., Avolio, B. J., & Shamir, B. (2002). Impact of transformational leadership on follower development and performance: A field experiment. *Academy of Management Journal*, *45*(4), 735-744.

Ellis, P. D. (2010). Effect sizes and the interpretation of research results in international business. *Journal of International Business Studies, 41*(9), 1581-1588.

Emerson, G. B., Warme, W. J., Wolf, F. M., Heckman, J. D., Brand, R. A., & Leopold, S. S. (2010). Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of Internal Medicine, 170*(21), 1934-1939.

Ensari, N., Riggio, R. E., Christian, J., & Carslaw, G. (2011). Who emerges as a leader? Meta-analyses of individual differences as predictors of leadership emergence. *Personality and Individual Differences*, *51*(4), 532-536.

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

Fern, E. F., & Monroe, K. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, *23*(2), 89-105.

Fischer, T., Hambrick, D. C., Sajons, G. B., & Van Quaquebeke, N. (2020). Beyond the ritualized use of questionnaires: Toward a science of actual behaviors and psychological states. *The Leadership Quarterly*, *31*(4).

Fischer, T., Tian, A. W., Lee, A., & Hughes, D. J. (2021). Abusive supervision: A systematic review and fundamental rethink. *The Leadership Quarterly*, *32*(6), 101540.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement. *Statistical methods for rates and proportions*, *2*(212-236), 22-23.

Flora, D. B. (2020). Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology*, *61*(4), 318.

Frackenpohl, G., Hillenbrand, A., & Kube, S. (2016). Leadership effectiveness and institutional frames. *Experimental Economics*, *19*, 842-863.

Furst, S. A., & Cable, D. M. (2008). Employee resistance to organizational change: managerial influence tactics and leader-member exchange. *Journal of Applied Psychology*, *93*(2), 453.

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.

Gardner, W. L., Lowe, K. B., Meuser, J. D., Noghani, F., Gullifor, D. P., & Cogliser, C. C. (2020). The leadership trilogy: A review of the third decade of the leadership quarterly. *The Leadership Quarterly*, *31*(1), 101-379.

Geyskens, I., Krishnan, R., Steenkamp, J. B. E., & Cunha, P. V. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management*, *35*(2), 393-419.

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205-215.

Grönqvist, E., & Lindqvist, E. (2016). The making of a manager: evidence from military officer training. *Journal of Labor Economics*, *34*(4), 869-898.

Güntner, A. V., Klonek, F. E., Lehmann-Willenbrock, N., & Kauffeld, S. (2020). Follower behavior renders leader behavior endogenous: The simultaneity problem, estimation challenges, and solutions. *The Leadership Quarterly*, *31*(6), 101441.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201-209.

Henle, C. A., & Gross, M. A. (2014). What have I done to deserve this? Effects of employee personality and emotion on abusive supervision. *Journal of Business Ethics*, *122*(3), 461-474.

Hemphill, J. F. (2003). *Interpreting the magnitudes of correlation coefficients*.

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, *47*(1), 105-143.

Hirota, K. (1953). Group problem solving and communication. *Japanese Journal of Psychology*, *24*, 176-177.

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*, *44*(2), 501-529.

Hoffman, B. J., Woehr, D. J., Maldagen-Youngjohn, R., & Lyons, B. D. (2011). Great man or great myth? A quantitative review of the relationship between individual differences and leader effectiveness. *Journal of Occupational and Organizational Psychology*, *84*(2), 347-381.

Hunt, J. G., Boal, K. B., & Dodge, G. E. (1999). The effects of visionary and crisis-responsive charisma on followers: An experimental examination of two kinds of charismatic leadership. *The Leadership Quarterly*, *10*(3), 423-448.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, *8*(1), 3-7.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. John Wiley & Sons.

James, E.H., & Wooten, L.P. (2010). *Leading under pressure: From surviving to thriving before, during, and after a crisis.* New York: Routledge Press.

Jensen, S. M., & Luthans, F. (2006). Relationship between entrepreneurs' psychological capital and their authentic leadership. *Journal of Managerial Issues*, 254-273.

John, P. (2017). *Field experiments in political science and public policy: Practical lessons in design and delivery*. Routledge.

Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review*, *31*(2), 386-408.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: a qualitative and quantitative review. *Journal of Applied Psychology*, *87*(4), 765.

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: a meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*(5), 755.

Kelley, K. (2013). Effect Size and Sample Size Planning 11. *The Oxford Handbook of Quantitative Methods in Psychology, Vol. 1*, *1*, 206.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Kennedy, P. (2008). *A guide to econometrics*. John Wiley & Sons.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods*, *15*(4), 624-662.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). *Journal of Business and Psychology*, *28*(2), 123-143.

Kremer, J. M., & Mack, D. (1983). Pre-emptive game behaviour and the emergence of leadership. *British Journal of Social Psychology*, *22*(1), 19-26.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722-752.

Lechler, M., & McNamee, L. (2018). Indirect colonial rule undermines support for democracy: Evidence from a natural experiment in Namibia. *Comparative Political Studies*, *51*(14), 1858-1898.

Lee, M., & Kray, L. J. (2021). A gender gap in managerial span of control: Implications for the gender pay gap. *Organizational Behavior and Human Decision Processes*, *167*, 1-17.

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

Li, W. D., Li, S., Feng, J. J., Wang, M. O., Zhang, H., Frese, M., & Wu, C. H. (2021). Can becoming a leader change your personality? An investigation with two longitudinal studies from a role-based perspective. *Journal of Applied Psychology*, *106*(6), 882.

Littlefield, R. S., & Quenette, A. M. (2007). Crisis leadership and Hurricane Katrina: The portrayal of authority by the media in natural disasters. *Journal of Applied Communication Research*, *35*(1), 26-47.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, *64*, 19-40.

Lord, R. G., De Vader, C. L., & Alliger, G. M. (1986). A meta-analysis of the relation between personality traits and leadership perceptions: An application of validity generalization procedures. *Journal of Applied Psychology*, *71*(3), 402.

Mackey, J. D., Ellen III, B. P., McAllister, C. P., & Alexander, K. C. (2021). The dark side of leadership: A systematic literature review and meta-analysis of destructive leadership research. *Journal of Business Research*, *132*, 705-718.

Mackey, J. D., Frieder, R. E., Brees, J. R., & Martinko, M. J. (2017). Abusive supervision: A meta-analysis and empirical review. *Journal of Management*, *43*(6), 1940-1965.

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, *19*(1), 30-43.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537-542.

Martin, R., Hughes, D. J., Epitropaki, O., & Thomas, G. (2021). In pursuit of causality in leadership training research: A review and pragmatic recommendations. *The Leadership Quarterly*, *32*(5), 101375.

Martin, S. L., Liao, H., & Campbell, E. M. (2013). Directive versus empowering leadership: A field experiment comparing impacts on task proficiency and proactivity. *Academy of Management Journal*, *56*(5), 1372-1395.

McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71, 173–180.

Miao, C., Humphrey, R. H., & Qian, S. (2018). Emotional intelligence and authentic leadership: A meta-analysis. *Leadership & Organization Development Journal*, *39*(5), 679-690.

Miao, C., Humphrey, R. H., & Qian, S. (2021). Emotional intelligence and servant leadership: A meta-analytic review. *Business Ethics, the Environment & Responsibility*, *30*(2), 231-243.

Moldoveanu, M., & Narayandas, D. (2019). The future of leadership development. *Harvard Business Review*, *97*(2), 40-48.

Morris, C. G., & Hackman, J. R. (1969). Behavioral correlates of perceived leadership. *Journal of Personality and Social Psychology*, *13*(4), 350.

Mowday, R. T., & Sutton, R. I. (1993). Organizational behavior: Linking individuals and groups to organizational contexts. *Annual Review of Psychology*, *44*(1), 195-229.

Ng, K. Y., & Van Dyne, L. (2005). Antecedents and performance consequences of helping behavior in work groups: A multilevel analysis. *Group & Organization Management*, *30*(5), 514-540.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205-1226.

Oc, B., & Bashshur, M. R. (2013). Followership, leadership and social influence. *The Leadership Quarterly*, *24*(6), 919-934.

Onwuegbuzie, A. J., & Levin, J. R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods*, *2*(1), 12.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906.

Phan, P. H. (2021). Connecting Research to Policy Is Easier Said Than Done. *Academy of Management Perspectives*, *35*(4), 563-565.

Phillips, J. M. (2002). Antecedents and consequences of procedural justice perceptions in hierarchical decision-making teams. *Small Group Research*, *33*(1), 32-64.

Pillai, R. (1996). Crisis and the emergence of charismatic leadership in groups: An experimental investigation 1. *Journal of Applied Social Psychology*, *26*(6), 543-562.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, *19*(2), 159-203.

Podsakoff, P. M., & Podsakoff, N. P. (2019). Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, *30*(1), 11-33.

Porath, C. L., & Erez, A. (2007). Does rudeness really matter? The effects of rudeness on task performance and helpfulness. *Academy of Management Journal*, *50*(5), 1181-1197.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods, 16,* 93–115.

Raymondie, R. A., & Steiner, D. D. (2022). Backlash against counter-stereotypical leader emotions and the role of follower affect in leader evaluations. *Journal of Applied Social Psychology*, *52*(8), 676-692.

Rast III, D. E., Hogg, M. A., & Giessner, S. R. (2016). Who trusts charismatic leaders who champion change? The role of group identification, membership centrality, and self-uncertainty. *Group Dynamics: Theory, Research, and Practice*, *20*(4), 259.

Reyes, D. L., Dinh, J., Lacerenza, C. N., Marlow, S. L., Joseph, D. L., & Salas, E. (2019). The state of higher education leadership development program evaluation: A meta-analysis, critical review, and recommendations. *The Leadership Quarterly*, *30*(5), 101311.

Rockstuhl, T., Dulebohn, J. H., Ang, S., & Shore, L. M. (2012). Leader–member exchange (LMX) and culture: A meta-analysis of correlates of LMX across 23 countries. *Journal of Applied Psychology*, *97*(6), 1097.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*(4), 1.

Roth, P. L., Le, H., Oh, I. S., Van Iddekinge, C. H., & Bobko, P. (2018). Using beta coefficients *Applied Psychology*, *103*(6), 644.r

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. *The handbook of research synthesis and meta-analysis*, *2*, 103-125.

Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *The Leadership Quarterly, 31*(5), 101348.

Schyns, B., & Schilling, J. (2013). How bad are the effects of bad leaders? A meta-analysis of destructive leadership and its outcomes. *The Leadership Quarterly*, *24*(1), 138-158.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115.

Schmidt, F. L. (2015). History and development of the Schmidt–Hunter meta-analysis methods. *Research Synthesis Methods*, *6*(3), 232-239.

Schmidt, F. L., & Oh, I. S. (2013). Methods for second order meta-analysis and illustrative applications. *Organizational Behavior and Human Decision Processes*, *121*(2), 204-218.

Schmidt, J. A., & Pohler, D. M. (2018). Making stronger causal inferences: Accounting for selection bias in associations between high performance work systems, leadership, and employee and customer satisfaction. *Journal of Applied Psychology*, *103*(9), 1001.

Schriesheim, J. F. (1980). The social context of leader–subordinate relations: An investigation of the effects of group cohesiveness. *Journal of Applied Psychology*, *65*(2), 183.

Schweitzer, M. E., Ordóñez, L., & Douma, B. (2004). Goal setting as a motivator of unethical behavior. *Academy of Management Journal*, *47*(3), 422-432.

Sergent, K., & Stajkovic, A. D. (2020). Women's leadership is associated with fewer deaths during the COVID-19 crisis: Quantitative and qualitative analyses of United States governors. *Journal of Applied Psychology*, *105*(8), 771.

Smithson, M. (2003). *Confidence intervals* (No. 140). Sage.

Sosik, J. J., & Cameron, J. C. (2010). Character and authentic transformational leadership behavior: Expanding the ascetic self toward others. *Consulting Psychology Journal: Practice and Research*, *62*(4), 251.

Spark, A., & O'Connor, P. J. (2021). State extraversion and emergent leadership: Do introverts emerge as leaders when they act like extraverts?. *The Leadership Quarterly*, *32*(3), 101474.

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, *72*(3), 438.

Staines, G., Tavris, C., & Jayaratne, T. E. (1974). *The Queen Bee Syndrome*.

Steel, P. D., & Kammeyer-Mueller, J. D. (2002). Comparing meta-analytic moderator estimation techniques under realistic conditions. *Journal of Applied Psychology*, *87*(1), 96.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal*, *43*(2), 178-190.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them?. *Theory & Psychology*, *9*(2), 165-181.

Thornton III, G. C., Rupp, D. E., & Hoffman, B. J. (2014). *Assessment center perspectives for talent management strategies*. Routledge.

Tonidandel, S., Williams, E. B., & LeBreton, J. M. (2014). Size matters… just not in the way that you think: Myths surrounding sample size requirements for statistical analyses. In *More statistical and methodological myths and urban legends* (pp. 162-183). Routledge.

Van de Calseyde, P. P., Evans, A. M., & Demerouti, E. (2021). Leader decision speed as a signal of honesty. *The Leadership Quarterly*, *32*(2), 101442.

Van der Heijden, E., & Moxnes, E. (2013). Leading by example to protect the environment: Do the costs of leading matter?. *Journal of Conflict Resolution*, *57*(2), 307-326.

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal*, *41*(1), 108-119.

Van Vugt, M., & von Rueden, C. R. (2020). From genes to minds to cultures: Evolutionary approaches to leadership. *The Leadership Quarterly*, *31*(2), 101404.

Vaske, J. J. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife*, *7*(4), 287-300.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

Vroom, V. H., & Jago, A. G. (2007). The role of the situation in leadership. *American Psychologist*, *62*(1), 17.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779-804.

Waldman, D. A., Carter, M. Z., & Hom, P. W. (2015). A multilevel investigation of leadership and turnover behavior. *Journal of Management*, *41*(6), 1724-1744.

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.

Whitman, D. S., Van Rooy, D. L., Viswesvaran, C., & Kraus, E. (2009). Testing the second-order factor structure and measurement equivalence of the Wong and Law Emotional Intelligence Scale across gender and ethnicity. *Educational and Psychological Measurement*, *69*(6), 1059-1074.

Wilkinson, L. (1999). American Psychological Association Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604.

Williams, L. J., O'Boyle, E. H., & Yu, J. (2020). Condition 9 and 10 tests of model confirmation: A review of James, Mulaik, and Brett (1982) and contemporary alternatives. *Organizational Research Methods*, *23*(1), 6-29.

Wilmot, M. P., Wanberg, C. R., Kammeyer-Mueller, J. D., & Ones, D. S. (2019). Extraversion advantages at work: A quantitative review and synthesis of the meta-analytic evidence. *Journal of Applied Psychology*, *104*(12), 1447.

Wilson, D. B. (2016). Formulas Used by the" Practical Meta-Analysis Effect Size Calculator. *Practical meta-analysis*.

Wofford, J. C. (1994). Effects of Situational Variables on Leader's Choice of Behavior. *Psychological Reports*, *75*(3), 1289-1290.

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*, *11*(1), 79-95.

Wu, T. Y., & Hu, C. (2013). Abusive supervision and subordinate emotional labor: The moderating role of openness personality. *Journal of Applied Social Psychology*, *43*(5), 956-970.

Yamaura, K., Horishita, T., & Kanayama, M. (2013). Positive feedback is not fully effective in all situations. *Shinrigaku Kenkyu: The Japanese Journal of Psychology*, *83*(6), 517-525.

Yammarino, F.J., Dionne, S.D., Chun, J.U., & Dansereau, F. (2005). Leadership and levels of analysis: A state-of-the-science review. *The Leadership Quarterly, 16,* 879-919.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association*, *46*(253), 19-34.

Yukl, G., & Falbe, C. M. (1990). Influence tactics and objectives in upward, downward, and lateral influence attempts. *Journal of Applied Psychology*, *75*(2), 13.

Zaccaro, S. J., Green, J. P., Dubrow, S., & Kolze, M. (2018). Leader individual differences, situational parameters, and leadership outcomes: A comprehensive review and integration. *The Leadership Quarterly, 29*(1), 2-43.

Zhang, Y., & Bednall, T. C. (2016). Antecedents of abusive supervision: A meta-analytic review. *Journal of Business Ethics*, *139*, 455-471.

APPENDIX A: Full List of Key Terms Included in Searches

Key words included in search:
Phase 1: Second-order meta-analysis of leader individual differences
"Leadership" in abstracts AND "meta" in abstracts AND "big five" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "extraversion" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "conscientiousness" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "neuroticism" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "emotional stability" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "openness" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "agreeableness" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "mental ability" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "intelligence" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "cognitive ability" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "dark triad" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "narcissism" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "psychopathy" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "Machiavellianism" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "gender" anywhere in text
"Leadership" in abstracts AND "meta" in abstracts AND "emotional intelligence" anywhere in text
Phase 2: Lab and field experiments
"Leader" in abstracts AND "experiment" in abstracts
"Leader" in abstracts AND "random" in abstracts
"Leader" in abstracts AND "trial" in abstracts

Phase 3: Quasi and non-traditional experiments
"Leader" in abstracts AND "quasi" in any field
"Leader" in abstracts AND "instrumental variable" in any field
"Leader" in abstracts AND "instrumental specification test" in any field
"Leader" in abstracts AND "lagged variable" in any field
"Leader" in abstracts AND "Heckman treatment" in any field
"Leader" in abstracts AND "Heckman selection models" in any field
"Leader" in abstracts AND "regression discontinuity" in any field
"Leader" in abstracts AND "synthetic control" in any field
"Leader" in abstracts AND "2SLS" in any field
"Leader" in abstracts AND "3SLS" in any field
"Leader" in abstracts AND "GMM" in any field
"Leader" in abstracts AND "exogenous event" in any field
"Leader" abstracts AND "dynamic panel" in any field
"Leader" in abstracts AND "propensity score" in any field
"Leader" in abstracts AND "simultaneous equation model" in any field
"Leader" in abstracts AND "difference-in-differences model" in any field
"Leader" in abstracts AND "Hausman test" in any field
"Leader" in abstracts AND "Monte Carlo analysis" in any field
"Leader" in abstracts AND "inverse Mills ratio" in any field

APPENDIX B: Leadership Variables Included in Final Coding Phase

IV ID Codes	DV ID Codes	Proxies for DVs:
1 = charismatic leadership	1 = task performance (quantity)	<- task productivity
2 = transformational leadership	2 = task performance (quality)	
3 = empowering leadership	3 = prosocial behavior	<-- helpfulness behavior, follower contributions
4 = leader vision	4 = extra-role performance (quantity)	<- extra effort
5 = destructive leadership	5 = extra-role performance (quality)	
6 = participative leadership	6 = follower trust in leader	
7 = leader prototypcality	7 = perceived leader effectiveness	
8 = leader affective display (+)	8 = follower liking of leader	<- satisfaction with leader, follower support of leader, endorsement of leader
9 = leader affective display (+)	9 = follower group orientedness	<- group relatedness, group cooperation
10 = leader emotions (+)	10 = perceived leader charisma	
11 = leader emotions (-)	11 = follower negative affect	
12 = leader humility	12 = follower creativity	
	13 = follower deviance/unethical behavior	
	14 = follower task satisfaction	
IVs dropped in analysis that failed to reach "critical mass":		
leader procedural fairness/grant voice		
leader behavioral integrity		
leader self-sacrifice		
ethical leadership		
leader confidence		
leader prosocial behaviors		
leader punishing behavior		
leader power		
leader communication style		
transparant leadership		