2009

A meta-analytic review of leadership impact research: Experimental and quasi-experimental studies

Bruce J. Avolio
University of Washington, bavolio@u.washington.edu

Rebecca J. Reichard
Claremont Graduate University

Sean T. Hannah
West Point – United States Military Academy

Fred O. Walumbwa
Arizona State University

Adrian Chan
University of Washington

The Leadership Quarterly 20 (2009) 764–784

Contents lists available at ScienceDirect

The Leadership Quarterly

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

A meta-analytic review of leadership impact research: Experimental and quasi-experimental studies

Bruce J. Avolio [a,*], Rebecca J. Reichard [b], Sean T. Hannah [c], Fred O. Walumbwa [d], Adrian Chan [a,e]

[a] University of Washington, United States
[b] Claremont Graduate University, United States
[c] West Point – United States Military Academy, United States
[d] Arizona State University, United States
[e] Singapore Ministry of Defense, Singapore

ARTICLE INFO

Keywords:
Leadership
Experiment
Quasi-experiment
Development
Meta-analysis

ABSTRACT

In this study we set out to conduct a comprehensive quantitative research analysis of literature reporting results on the causal impact of leadership by focusing on examining what we refer to as 'leadership interventions.' We defined leadership interventions as those studies where the researcher overtly manipulated leadership as the independent variable through training, assignment, scenario or other means. Our focus included both examining experimental and quasi-experimental as well as lab and field studies conducted in public and private organizations. Our goal was to address a simple question: *do leadership interventions have the intended impact and if so to what degree?* We conducted a comprehensive review of the published and unpublished literature and uncovered 200 lab and field studies that met our criterion as leadership intervention studies. We report here the findings of a series of meta-analyzed effects comparing the relative impact of leadership interventions across intervention types, leadership theories, and several common dependent variables. Overall, leadership interventions produced a 66% probability of achieving a positive outcome versus a 50–50 random effect for treatment participants, but this effect varied significantly when assessing moderators such as type of leadership theory.

1. Proposing an integrative view for evaluating leadership impact research

Numerous theories of leadership offer a broad array of explanations regarding how leadership impacts follower motivation, thinking, behaviors, and performance. Yet, we know that much of the accumulated research has been based on field survey research versus using experimental designs. Consequently, the conclusions that can be drawn from this literature are limited in terms of being able to verify and validate the cause and effect relationships proposed in the various theories of leadership (Yukl, 1998, 2002).

Another common criticism of leadership research has focused on the research methods used to examine leadership impact on follower performance. For example, Bass (1990) noted that much of the accumulated research is beset with the over use of small convenience samples with cross-sectional designs. Lord and Hall (1992, p. 153) noted, "too much research in the past has attempted to probe the complex issues of leadership using simple bi-variate correlations." Again, these designs limit the conclusions that can be drawn from the accumulated leadership literature for a variety of reasons such as having a high degree of sampling error, lacking temporal precedence, and/or a failure to manipulate leadership as an independent variable in order to examine its impact on performance outcomes.

* Corresponding author. Center for Leadership & Strategic Thinking, Foster School of Business, University of Washington, USA. Tel.: +1 206 543 7908.
E-mail address: bavolio@u.washington.edu (B.J. Avolio).

1048-9843/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.leaqua.2009.06.006

Beyond research design and methods limitations, the limited focus used in the majority of prior meta-analyses reported in the literature has restricted the development of a more integrated summary of the leadership literature. These prior meta-analyses have typically limited their reviews to a single leadership theory when examining leadership impact on various outcomes.

As a foundation for building an integrated research base for the current study, we went back to 1981 where we identified the first leadership study using meta-analytical techniques. This study provided a quantitative review of Fiedler's contingency theory of leadership (see Strube & Garcia, 1981). Over the next 25-year time span we identified 32 meta-analyses conducted in the field of leadership.[1] All of these previous meta-analyses examined one theory, and in many cases one independent variable compared with a limited set of dependent variables/outcomes. For example, Eagly and Karau (1991) focused on gender effects and their relationship to leadership emergence, while for transactional and transformational leadership, prior meta-analyses focused on linking these different leadership styles to outcomes such as performance, rated effectiveness and satisfaction (Dumdum, Lowe & Avolio, 2002; Judge, Bono, Ilies & Gerhardt, 2002; Judge & Piccolo, 2004; Lowe, Kroeck & Sivasubramaniam, 1996).

Consequently, while prior meta-analyses have consistently provided evidence showing the positive relationships that various leadership styles or orientations have on a range of outcomes, they were typically limited to one theory of leadership. Moreover, further limiting conclusions from most prior meta-analyses is the fact that they did not examine experimental interventions where cause and effect could be determined. The current meta-analytic study adds to the body of leadership literature by examining the effects of leadership experimental research, across a number of leadership theories, while incorporating a range of dependent variables as described below.

1.1. A quick look back

Leadership research can be traced as far back as the early part of the twentieth century. Yet, an organized social–scientific approach to studying leadership did not fully emerge until the early 1930s (House & Aditya, 1997). Systematic investigations into what constitutes leadership effectiveness began in earnest over the next several decades at various centers for research on leadership such as at Iowa during the 1930s and at Michigan and Ohio States in the 1940s to 1950s.

Over the next fifty years, models of leadership evolved focusing on traits and personalities of leaders, the nature of leader–follower interactions, what constituted leader and follower cognitions, perceptions and attribution processes, situational factors and contingencies that moderated the effects of leadership, leader behaviors/styles, task and goal orientations, team and shared leadership, and transformational/charismatic leadership. And as the twentieth century drew to a close, there were also attempts to integrate the various theories and models of leadership into a broader more integrative framework called a "full range theory of leadership" (Avolio, 1999).

Along with the growth of research on what constitutes leadership style and effectiveness, a number of authors have called for a renewed emphasis on conducting experimental studies to broaden our understanding of the causal effects of leadership on follower affect, cognition, ability, motivation, and performance (Day, Zaccaro & Halpin, 2004; Day & O'Connor, 2003). Here we define a *leadership intervention* study as one in which the researcher overtly manipulated leadership to examine its impact on some specific intermediate process variables or outcomes. For example, leadership may have been manipulated in either lab or field settings through the use of actors trained to portray specific leadership styles, via leadership scenarios, or as a result of leadership training.

As we note in more detail below, out of the nearly 500 leadership studies initially identified in our review of the leadership literature, less then half ended up meeting our criteria for inclusion in the meta-analysis as they were not interventions testing cause and effect. Most of the approximately 300 research papers dropped from further consideration were either not empirically based or were based on a correlation design without any intervention.

In sum, the guiding purpose for this meta-analytic study was to provide a comprehensive assessment and best estimate of impacts of leadership interventions by obtaining and examining all experimental and quasi-experimental research that have attempted to change and/or develop leadership. Our secondary goal was to then examine more closely the theoretical frameworks, methods, and dependent variables under which experimental and quasi-experimental leadership research varied in its impact in order to develop a base of comparison upon which future leadership research, theory, and practice could build upon. We also identified several research questions that were based on core assumptions in prior leadership models that set the stage for comparing the impacts of different leadership models on various follower outcomes.

1.2. Evidence for higher and lower-order effects of leadership impact

Numerous primary studies and meta-analyses have reported positive associations between a range of leadership orientations including initiation of structure, consideration, and transactional and transformational leadership, all with respect to a very broad range of follower outcomes (Bono & Judge, 2004; Dumdum et al., 2002; Judge et al., 2002; Judge & Piccolo, 2004; Lowe et al., 1996). Yet, as the field of leadership has evolved over the last 20 years, there has been greater attention paid to examining what might be considered higher impact theories or models of leadership, such as charismatic, inspiring and transformational leadership. Indeed, Bass (1985) signaled this shift in terms of a focus on leadership impact, entitling his book, "Leadership and Performance Beyond Expectations." Bass argued that leadership defined as transformational would augment transactional forms of leadership in predicting performance, thus the choice of the term 'performance beyond expectations'. This set the stage for distinguishing what Bryman (1992) and Antonakis and House (2002) referred to as more traditional theories of leadership versus the newer leadership theories.

[1] References of all other leadership meta-analyses found and used in meta-analysis are available from the first author.

1.2.1. Newer versus traditional leadership

Bryman (1992) commented that "there was considerable disillusionment with leadership theory and research in the early 1980s. Out of this pessimism emerged a number of alternative approaches, which shared some common features …collectively referred to as the new leadership" (Bryman, 1992, p. 21). Unlike the "traditional" leadership models, which described leader behavior in terms of leader–follower exchange relationships, providing direction and support, and reinforcement behaviors; or what Bass (1985) has called "economic cost–benefit assumptions" (p. 5); the new leadership approaches emphasized symbolic leader behavior, visionary, inspirational messages, emotional feelings, ideological and moral values, individualized attention, and intellectual stimulation. Bass (1985) summarized well the need for a new approach to leadership in commenting that, "subordinate motivation to work cannot be fully accounted for by any notion of a simple swap of desired material and psychic payments from a superior in exchange for satisfactory services rendered by subordinates" (p. 9). Emerging from these early works, charismatic and transformational leadership theories have turned out to be the most frequently researched theories over the last fifteen years (Bass & Riggio, 2006; Judge & Piccolo, 2004) with a particular emphasis on determining how they affect or contribute to performance beyond the standard range of expectations.

The theory of charismatic/transformational leadership suggests that such leaders raise followers' aspirations and activate their higher-order values such that followers identify with the leader and his or her mission/vision, feel better about their work, and perform beyond expectations (Avolio, 1999; Bass, 1985, 1998; Burns, 1978; Conger & Kanungo, 1987, 1998; House, 1976; Shamir, House & Arthur, 1993). The accumulated research has reported that charismatic/transformational leadership was positively associated with leadership effectiveness and a number of important individual, group and organizational outcomes across many different types of organizations, situations, levels of analyses, and cultures (see Avolio, Bass, Walumbwa, and Zhu (2004) for a summary of this literature). Indeed, as Bono and Judge (2004) concluded, "there is little controversy regarding the positive associations between such leadership [charismatic/transformational] and follower attitudes, such as trust, job satisfaction, and organizational commitment, and behaviors, such as job performance at the individual, group, and organizational levels" (p. 554).

Although the pattern of findings regarding newer theories of leadership has confirmed positive relationships with a range of outcomes, we are not aware of any empirical research that has specifically examined the relative causal effects of the newer theories of leadership, nor compared those causal effects to those of earlier traditional models of leadership. Although there have been several comprehensive meta-analyses of this literature, none of those prior quantitative and qualitative reviews focused on assessing interventions where cause and effect could be determined.

Based on the accumulated effects reported in prior literature, we expect that charismatic/transformational leadership would account for more variance than any of the traditional leadership models due to its focus on a higher order of exchange and affective and emotional needs and responses of followers (Bass, 1985, 1998; Yukl & Van Fleet, 1992). Drawing on previous literature, Bass' (1985) original argument, and three recent meta-analyses (e.g., Bono & Judge, 2004; Dumdum et al., 2002; Judge & Piccolo, 2004), we suggest that the newer leadership theories should account for a greater share of the variance in follower attitudes, behaviors, and performance when compared with more traditional transactional models of leadership. Our predictions are based on the central assumption underlying newer leadership models that such leaders work to transform their followers to higher levels of motivation, ability, and performance. We build on this assumption in terms of our overview of the current study below.

2. Study overview

2.1. Impact of leadership research

The purpose in conducting the meta-analytic study described in this paper was to address the question commonly asked about leadership research: 'do leadership interventions or leadership development initiatives make a difference, and if so, by what models or methods and with which outcomes?' Additionally, we were interested in looking at the effects of the various forms of 'researcher-manipulated' leadership versus training or 'developed' leadership to see how they differ in their causal impact as a foundation for stimulating future work in both of these respective broad areas of inquiry.

Most prior meta-analyses have focused on the relationship between a limited subset of independent and dependent variables. The current study adopted a broader strategy from the outset by aggregating the effects from all identified experimental and quasi-experimental leadership studies across newer and traditional theories. We examined how the causal impact of leadership varied across the most commonly researched theoretical frameworks and how effects for each theory category differed by comparing three types of outcomes that commonly appear in the literature: affective, behavioral, and cognitive outcomes.

3. Core research questions

3.1. Type of intervention

The studies in our sample reflected various types of leadership manipulations used by the original researchers. For example, leadership may have been manipulated by the researcher through the use of actors trained to portray specific leadership styles, by participant role-play, through presenting participants with scenarios or vignettes of specific leadership actions or styles, by training or developing leaders, by assigning specific types of leaders to groups or assigning leaders in a particular way, or by altering leader expectations of their followers in an effort to alter their behaviors toward those followers. We felt it was important to determine when a particular leadership theory is more (or less) effective. Table 1 provides a breakdown of the frequency of

different types of leadership interventions found in studies going back to the early part of the 20th century. It was encouraging to note that the number of studies that manipulated leadership has increased over the decades. In particular, the 1990s has shown a surge in the number of experimental and quasi-experimental leadership studies compared to previous decades. This trend looks set to continue into this current decade.

The most common form of manipulation in these studies was leader training/development (62 of the 200 studies). Developmental studies likely offer the greatest practical implications for organizations, and we were therefore keenly interested in how the effects of this form of manipulation may differ from the other forms of interventions.

The first meta-analysis that investigated managerial training studies reported moderately positive effects. In their meta-analytic review of 70 studies conducted from 1952 to 1982 on managerial training effectiveness, Burke and Day (1986) concluded that while managerial training was moderately effective, more empirical research was still needed before any firm conclusions could be drawn.

In a follow-up meta-analysis of 83 studies reported in the literature from 1982 to 2001, Collins and Holton (2004) replicated earlier findings showing that managerial training produced positive outcomes, with effect sizes ranging from .96 to 1.37 for knowledge outcomes; from .35 to 1.01 for expertise outcomes; and around .39 for performance outcomes. In addition, Collins and Holton echoed the need for more clarity on the effectiveness of managerial training, especially in terms of training impact on organizational performance outcomes. They also remarked that there was little done to determine which theories were more or less appropriate for implementing in the various organizations where training was conducted.

To our knowledge, no other study has attempted to conduct a quantitative review of leadership training/development studies nor compared the effects of those studies to other forms of manipulation, thus our first research question is:

Research question 1. Does the impact of experimental/quasi-experimental leadership interventions differ comparing training or developmental versus other types of leadership interventions?

3.2. Leadership theory

After determining at an aggregate level the extent to which leadership interventions have an impact across various forms of manipulation, we were also interested in determining if experimental/quasi-experimental research based on different leadership theories have more (less) of an impact, and how those effects vary across four primary outcomes commonly reported in the leadership literature: affective, cognitive, and behavioral/performance. By affective we mean how individuals feel about their leader, work, colleagues and so forth. Cognitive outcomes represent how individuals perceive information, process information and make decisions. Finally, behavioral outcomes represent the actions taken by individuals that are observable, for which we have also included individual and group performance. Further, we assessed organization-level performance (e.g. profit) in another category.

Authors of some theories of leadership have suggested that certain leadership styles such as transformational leadership are expected to have a greater impact than other styles, and will affect followers at a higher level of development in terms of emotions, cognitions, behavior and performance (Bass, 1990, 1998). We therefore wanted to determine whether and in what way different theories of leadership had differing impacts on affective, cognitive or behavior/performance outcomes. As far as we can ascertain, there has been no comprehensive quantitative review that directly examined and compared the effects of the various major leadership theories, nor has any assessed how leadership interventions guided by these respective theories differentially impact various outcomes.

3.3. Classification of leadership theories

To focus our review and analysis, we classified leadership theories into two broad categories based on the theoretical origins of the study: traditional leadership and newer leadership theories. By 'traditional,' we refer to theories that dominated leadership research up to the late 1970s, including behavioral, and contingency approaches to leadership (Yukl, 2002, 2006). By 'newer' we refer to theories that have dominated leadership research in the 1980s forward, including charismatic, inspirational, transformational, and visionary leadership (Bass, 1998; Bryman, 1992; Peterson & Hunt, 1997). In addition, we discovered that a sizeable number of experimental intervention studies were based on Pygmalion based leadership (such as self-fulfilling prophecy) yielding an additional category of potential interest because of their use of experimental research. Finally, included in the overall analysis, but excluded from theory-specific analyses due to the small numbers of studies found in our search meeting

Table 1
Type of leadership manipulation by decade.

	Scenario or vignette	Actor or role play	Leader trained or developed	Leader appointed or assigned	Leader expectation	Others	Total
Post World War I till end World War II	0	1	0	0	0	0	1
1950 till 1959	0	0	3	0	0	0	3
1960 till 1969	1	1	4	2	0	0	8
1970 till 1979	4	4	12	10	3	6	39
1980 till 1989	7	4	12	9	4	1	37
1990 till 1999	8	17	14	12	6	3	60
2000 onwards	15	12	17	1	1	6	52
Total	35	39	62	34	14	16	200

the inclusion criteria, were intervention studies based on cognitive information processing/individual differences, attribution theory, leader-member exchange, and team leadership.[2]

In sum, we are not aware of any empirical research that has specifically examined the relative effects of newer forms of leadership as compared to the traditional models of leadership. On the basis of existing theory and research (e.g., Bono & Judge, 2004) and the interest in higher order of exchange and affective and emotional needs and responses of followers proposed in the newer leadership theories (Bass, 1985, 1990, 1998), we explored the following research questions:

Research question 2a. Does the impact of experimental/quasi-experimental leadership interventions differ as a function of whether it was based on newer leadership theory versus traditional leadership theory?

Research question 2b. Does the impact of experimental/quasi-experimental leadership interventions differ as a function of whether it was based on newer leadership theory versus Pygmalion leadership theory?

3.4. Examining leadership theory by dependent outcomes

The differences in the various theoretical frameworks discussed above led us to establish research questions to assess the differing impact of traditional versus newer leadership theories on four primary outcomes. For example, we could expect that transformational leadership, which makes up the majority of the newer research, to be more positively linked to affective outcome measures such as liking and trust due to transformational leaders exhibiting higher levels of individualized consideration and idealized influence behavior, while the impact on how individuals think and address challenges and problems may be more positively impacted by the greater levels of intellectual stimulation associated with transformational versus more traditional models of leadership (Bass & Riggio, 2006). Conversely, research on path-goal and transactional leadership theories have focused more on behavioral change and outcomes, as opposed to affective or cognitive change (Bass, 1990). Thus we examined how the various leadership theories differentially impacted outcomes depending on the core propositions in those respective theories.

We set out to examine how different leadership theories were moderated in their impact by exploring the following four types of dependent variables included in prior leadership research: affective (e.g., liking, satisfaction or enjoyment), behavioral/performance (e.g., leader emergence, participation, or performance measures of behaviors), and cognitive (e.g., level of idea generation or confidence) and organizational performance (e.g., profit or productivity increases). The latter category, unit/organizational performance, includes those outcomes that although likely result from individual or group behaviors, those behaviors were typically not directly measured.

Research question 3. Does the impact of experimental/quasi-experimental leadership interventions based on newer, traditional, or Pygmalion theories differ for affective, cognitive, behavior and organizational performance outcomes?

In sum, we have attempted to provide an appropriate rationale and conceptual basis for examining three categories of leadership theories to determine how developing or manipulating the leadership style/orientation associated with each theoretical category impacted four different outcomes. The theoretical literature comprised of newer research repeatedly discusses what has been commonly referred to as the 'higher order' impact of these leadership styles on follower emotions or affect, cognitions, behavior and performance. Thus, we feel there is a sufficient base to expect that there will be differences comparing the more traditional theoretically-based studies to those stemming from newer theories.

With regards to comparing Pygmalion research with the remaining two categories, evidence is less clear at least in terms of comparing this line of research to other newer theories. Specifically, we know that Pygmalion research explicitly sets out to manipulate how participants think about the expectations set by the leader within the experiment (Eden et al., 2000). Thus, we expect there to be a greater impact on cognitions than perhaps more traditional theories of leadership, but not necessarily greater than that of other newer theories. We also know from past Pygmalion research that both behaviors and performance have been shown to have been impacted by the Pygmalion manipulations. Yet, we have no reason to believe the impact will be more or less than other newer models and research, but should be greater than traditional. Thus, in our research question, we simply explore whether differences exist and make no specific predictions about direction.

4. Method

4.1. Inclusion criteria

The inclusion criteria set for this study spanned all experimental and quasi-experimental leadership studies. This definition required that 1) the phenomenon under investigation was in fact a leadership phenomenon, 2) that the researcher or trainer attempted to manipulate the independent variable representing this phenomenon, and 3) that the effects of that leadership

[2] Due to space limitations, readers interested in specific leadership theories are directed to numerous qualitative reviews (see for example, Avolio, Sosik, Jung & Berson, 2003; Bass, 1990; House & Aditya, 1997; Lowe & Gardner, 2000; Yukl, 2002; Yukl & Van Fleet, 1992). In terms of classifying studies, we need to point out, that some theoretical models have evolved over time, and have been revised. Nevertheless, we based our assignment of studies using the root theoretical framework upon which the study was based.

intervention were measured. As stated, both true and quasi-experimental designs were accepted, although in the final tally quasi-experiments (those interventions that did not use random assignment) comprised less than 25% of the identified studies.

Our goal was to achieve a full population of all experimental and quasi-experimental leadership studies; therefore our search spanned all sources of studies, published and unpublished. Consistent with prior criticisms of the leadership literature, the outcome of our search process indicated the majority of leadership studies were correlational. Thus, many interesting studies from a broad range of disciplines, although informative about leadership were excluded if the researcher did not directly manipulate leadership as an independent variable. The following list represented our exclusion criteria: (1) correlational studies, (2) studies that failed to report data on leadership effects, and (3) studies that failed to manipulate leadership as an independent variable.

4.2. Literature search

Given our goal of achieving a full population of studies meeting our criteria, we conducted an exhaustive four-phase search process. In phase I, a team of twelve doctoral associates conducted a literature review and developed a list of 18 major research streams found in the leadership literature. A list of 124 search words and phrases related to each of these streams was developed for use in electronic searches. The team then conducted an initial search of 18 electronic databases, including *Academic Search Elite*, *Business Source Premier*, *PsychInfo*, *Sociological Abstracts*, *ERIC*, *Web of Science*, *Annual Reviews*, *Anthropological Index/Royal Anthropological Society*, *Anthropological Literature Index*, *Dissertation Abstracts Online*, *Worldwide Political Science*, *RAND*, *WorldCat*, *Proceedings First*, *Army Research Institute*, *National Security Archive*, and *NTIS*.

Phase II comprised a validation search conducted approximately one year later. A second team of eight doctoral associates replicated the phase I search to see if 1) any articles were missed in phase I, or 2) any new studies were completed within the intervening year in which the first wave of studies were coded. Additionally, with the assumption that authors of experimental and quasi-experimental leadership studies would largely draw from other such studies for theoretical support, the reference lists of all articles found in phase I were reviewed to see if any referenced articles that met our inclusion criteria. Lastly, the reference lists of 32 other published and unpublished leadership meta-analyses were screened to ensure inclusion.

In phase III, email letters were sent to 670 scholars in leadership and related fields from the list of members on leadership journal editorial boards, leadership networks, prominent scholars in the field, and lists of individuals who had attended leadership conferences, asking them to review and validate the proposed listing of studies for inclusiveness, with all returned recommendations investigated. We received approximately a 90% response rate providing further guidance to potential research articles that could be included in our analysis of the literature. After search phases I–III failed to produce any sizeable groups of studies prior to 1970, we conducted a final phase IV search, manually searching leadership handbooks, other related books, and academic journals known to publish leadership studies.

The four search phases combined netted more than 500 studies that were determined as possibilities for inclusion in the current study. A two-person team trained for this purpose reviewed each study to ensure it met all inclusion criteria. When less than 100% agreement between the two raters occurred regarding the suitability of the study, it was sent to a third expert rater who made the final decision. Based on the above detailed inclusion and exclusion criteria, a total of 200 studies passed this review (about 16% unpublished) and were then coded to assess our research questions.[3] Most of the studies excluded from subsequent analyses were excluded because they were not intervention studies with empirical data included.

4.3. Variable coding

We will first describe how we coded for each of the moderator categories discussed in our research questions. A coding team of twelve doctoral associates was formed and trained to conduct data extraction and coding from the studies. This same group of research assistants was used throughout the duration of the project for various coding tasks. This group was broken down into six two-person teams each assigned a group of articles. For non-quantitative data, one person conducted initial coding. Both a primary coder and a secondary rater coded quantitative data independently through a blind review process. Quantitative data was then shared between team members to calculate inter-rater reliability. Initial percent agreement among raters for quantitative data was 91%. All discrepancies were then reinvestigated and consensus reached by team members or referred to a third expert coder, ultimately resulting in 100% agreement.

4.3.1. Type of intervention

We created two mutually exclusive categories of manipulations. The 37 studies that had usable effects (out of the 62 listed) that manipulated leadership through training or development of the leader were placed in category one. The 101 studies that had usable effects (of the remaining 138 studies) that used one of the other five forms of manipulation listed in Table 1 were placed in a second category for comparison. We defined leadership training or development as an attempt by the investigators to enhance an individual's knowledge, skills, ability, motivation, and/or perceived self-concept to enable them to exercise positive influence in the domain of leadership. Non-developmental interventions involved studies where the experimenter was manipulating the leader′s behavior through role plays, scripts, assignment and so forth. We provide examples of the types of interventions uncovered in our review in Table 2.

[3] A reference list of all the 200 studies and search terms included in this study is available from the first author and will be set up on a public web site.

Table 2
Interventions and outcomes of a sample of primary articles.

Author	Year	Theory	Description of leadership intervention	Outcome
Ben-Yoav, Hollander, and Carnevale	1983	Attribution	Leaders were either appointed by experimenter, elected by the group, or not formally designated	Leader responsiveness; leader interest; leader competence; future leadership
Cammalleri	1973	Traditional	Leader confederates were instructed to show behavioral styles associated with either authoritarian or democratic style	Group accuracy
Choi and Mai-Dalton	1999	Newer	Scenario method of having subjects respond to hypothetical settings manipulating self-sacrificial leader behavior, organizational uncertainty, and leader competence	Attributions and effects of charisma; attributions of legitimacy; Intention of reciprocity
Dvir, Eden, and Banjo	1995	Pygmalion	Military psychologist and experimenter raised squad leader expectations of their followers capabilities, focusing on self fulfilling prophecy	Safety performance; light arms performance; topography performance; first aid performance; document security performance; command potential; specialty proficiency performance; platoon leader leadership; equity
Howell and Frost	1989	Newer	Trained actors to act out charismatic versus initiation of structure behavior	Task performance; adjustment to leader; adjustment to group; task adjustment; job satisfaction
Towler	2003	Newer	Trained leaders in charismatic communication style and visionary content	Task satisfaction; task motivation; self efficacy; perceived effective delivery; charismatic leadership; persuasiveness of leader; extra effort; performance accuracy

4.3.2. Theory category

To test research question two, each independent effect size was coded to represent one of the three leadership theory categories: traditional ($n = 41$), Pygmalion ($n = 19$), or newer theories of leadership ($n = 40$). The remaining intervention studies out of the total of 138 did not fall into one of these three theoretical categories, but were included in the overall meta-analysis.

4.3.3. Focus of dependent variable

The focus of each dependent variable from each primary study was coded. Dependent variables were categorized based on coders′ assessment as to their intended focus into four categories: affective ($n = 31$), behavioral ($n = 86$), cognitive ($n = 88$), and organizational performance ($n = 2$). The reader will note that these numbers exceed the actual number of studies noted above, as there were multiple dependent variables coded for specific studies.

Researchers use of criterion varied widely and they infrequently referred to measures using these categorizations, requiring subjective assignment based on inter-rater agreement. Coding some of the dependent variables was problematic as in some instances the selected variables spanned more than one category. For instance, attitudes may often include affective, behavioral, and cognitive dimensions (Kane, Zaccaro, Tremble & Masuda, 2002; Kirkpatrick & Locke, 1996). Variables were thus coded through rater convergence as to the *primary* focus they targeted in a particular study.

4.4. Calculating the effect size statistics (d)

Our first analysis of the literature provided a comparison of all studies that fit our intervention criteria. Following this overall analysis, we then proceeded to examine a series of exploratory non-hierarchical analyses to assess our research questions, in terms of the effect sizes within each category including: study setting (experimental vs. field), target leader level (low, middle and senior level), type of organization (for profit, profit and military), leadership theory category type (Newer, Traditional and Pygmalion), type of intervention (training/development vs. other), quality of the study (low vs. high) and the type of dependent variable (affective, behavioral, cognitive and organizational performance). These analyses were added to provide a more comprehensive examination of leadership intervention impact for future reference.

We also conducted hierarchical analyses of studies by combining the respective theoretical categories and outcome type, as well as the type of intervention and outcome type. The purpose of these hierarchical analyses was to go further into examining some of the potentially unique and interesting differences across the three theoretical categories employed in the current study.

4.4.1. Procedures

During all phases of research synthesis and analysis, we used the methodology recommended by Hunter and Schmidt (2004) as the primary reference source. Given the focus on experimental and quasi-experimental studies, Cohen′s *d* (Cohen, 1977) was chosen as the effect statistic, with all *F*, *t*, chi-square and other statistics transformed into the *d* statistic. Additionally, to calculate and present utility analyses, we drew from Rosenthal and Rubin (1982) for the calculation and interpretation of binomial effect size display (BESD) statistics, and from Hedges and Olkin (1985) for homogeneity analysis (*Q*). All meta-analysis calculations were computed using a spreadsheet specifically designed by the authors for this study and based on Hunter and Schmidt formulas.

4.4.2. Coding of effect sizes

For each study, all available cell-level statistics, simple, main, and interaction effects were coded for each separate dependent variable. In this way, we could extract all possible leadership effects from each study. This provided for a range of effect sizes that

could be pulled from each dependent variable, enabling us to match the most appropriate effect from each study specifically to each of the research questions. For example, we could examine how the main effect of transformational leadership (newer category) impacted the level of extra effort exhibited by employees in a particular study. When cell sample sizes, means, and standard deviations were available, we used these cell statistics to calculate *d* values. When any of the cell statistics were missing, we looked to other effects (e.g., *t*, *F*) that were reported and calculated the *d* value using standard conversion formulas.

For example, in a simple leadership training experiment with a control group, the leadership effect could be the *t* statistic, comparing the experimental (leadership training) and control (no training) group differences on a leadership outcome, such as follower engagement. This *t* statistic would then be converted into a *d* statistic for meta-analysis.

4.5. Data assumptions and decision rules

Meta-analyses require numerous decision rules during coding and analysis, which ultimately affect the quality of the methodology and the interpretability of the findings. We established firm criteria for such judgment calls to enable informed interpretation of our findings (Wanous, Sullivan & Malinak, 1989). Judgment calls in this meta-analysis were related to the following problems: dealing with missing data, maintaining the assumption of independence, correcting for study artifacts, and handling extreme data points.

4.5.1. Missing data

In order to maximize the number of leadership effects calculated, two assumptions were made to minimize unusable data due to missing information. The first missing data problem involved missing cell sample sizes. When cell sample size was missing, it was imputed by dividing the overall sample size for the study by the number of cells, based on the assumption of equal sample size per cell. These imputed cell sample sizes were then used to calculate *d* values. In total, only four such cases were found.

To impute cell sizes, for each study noted below, we took the overall n size of say 100 across two conditions, and then split the value equally for each of the two conditions to calculate the effect size. Two of the samples in which we did this adjustment came from the same study.

The second and more frequent missing data problem occurred due to under-reporting of main and interaction effects, resulting in difficulty calculating *d* values from two-way or three-way ANOVAs (Lipsey & Wilson, 1996). For example, in 76 cases, the primary authors did not report interaction effects and oftentimes noted it was due to non-significant results. Excluding these effects would result in an unrepresentative sample and therefore, we implemented a decision rule that if the interaction effects were not reported, then we would assume a null finding. This assumption allowed us to use the one-way *F* conversion formula (ignoring interaction effects). We felt that disregarding these interaction effects would bias our results towards larger effect size differences. Together, these two assumptions regarding missing data increased the number of possible useable effects from 533 to a total of 612.

4.5.2. Assumption of independence

Using the procedures described above, multiple leadership effects were often extracted from the same study and sample, for a total of 612 non-independent effect sizes. To maintain sample independence, effect sizes from the same sample were averaged yielding one independent effect per study as recommended by Hunter and Schmidt (2004). For research questions one and two which did not require hierarchical moderator analysis, the total number of independent effect sizes equaled 140 with an overall sample size of 13,656 unique participants. Of these 140 independent effects sizes, 40 effects came from the newer theories, 19 from Pygmalion, 41 from traditional leadership, 9 from various individual difference studies, and the remainder from a variety of other various theoretical approaches (i.e., attribution, LMX) (note the last two categories had insufficient numbers of effects to meta-analyze). The total number of effect sizes (*k*) calculated for each research question varied depending on the number of effect sizes that could be calculated to provide the data for the specific research question being analyzed. Consequently, each research question can be viewed as a separate 'study,' with each unique sample providing only one independent effect size used to calculate the overall effect of leadership interventions specific to the parameters of that specific research question.

4.5.3. Correction for study artifacts

According to Hunter and Schmidt (2004), the overall effect size is attenuated due to various study artifacts. We corrected for two of these artifacts throughout our meta-analysis: sampling error and measurement error. The most commonly accepted correction is for sampling error, based on the statistical principle that effects from larger samples are more accurate. By weighting effects from larger samples more heavily, the corrected overall effect size is expected to become closer to the true effect of leadership interventions (Hunter & Schmidt, 2004). For example, we found that the raw, uncorrected effect of leadership interventions on all outcomes represented by *d* was .59. After correcting for study artifacts described above, the true combined effect of all leadership interventions in this study equaled .65 with a standard deviation of .80.

Another issue with primary research that attenuates the overall effect size is measurement error or unreliability in the dependent measure (Hunter & Schmidt, 2004). Of the 48% of the studies that reported reliability estimates, the alpha coefficients ranged from a low of .59 to a high of .97. As less than half of the authors reported reliability estimates, we selected the correction method recommended by Hunter and Schmidt (2004). This method involves the calculation of an attenuation factor, using the square root of the available reliabilities for the dependent measures. The sample weighted effect size was corrected by dividing by the mean attenuation factor. Thus, in the remainder of the paper, we use 'corrected' effect size to refer to the effect size corrected

for *both* sampling error and unreliability in the dependent measure. For comparison purposes both the raw and corrected effects are reported in the tables.

4.5.4. Outlier analysis

Outlier analyses, based on both effect size magnitude and sample size, were conducted on the overall set of data. According to Hunter and Schmidt (2004), extreme values may cause significant within-group heterogeneity of individual effect sizes that may not exist in reality. Furthermore, the weighted averages given to large sample size studies may cause the overall effect size to be influenced by relatively few studies. Thus, effects were examined and are reported both with (data Set 1) and without (data Set 2) these extreme values for each research question as shown in the data tables.

The first step in the search for extreme values was to compute histograms of the effect size and sample size values. From visual inspection of histograms, it was clear that some extreme values were present. We then implemented the three-sigma rule based on the recommendations for setting the statistical standard for selecting outliers using both Kline (1998) as well as Champ and Woodall (1987) recommendations for cutoffs. Specifically, we treated those values more than three standard deviations above the mean as outliers. The cutoff value for effect sizes was 3.11, indicating three outliers ($d = 3.17$, 3.28, and 3.47). With regard to extreme sample sizes, the cutoff value of three standard deviations above the mean was 452 indicating three additional outliers ($n = 499$, 520, and 975). "Set 2" reflects data with these six outliers removed, and thus will always be calculated on slightly fewer effects than the full data set. Small to moderate extreme values were retained in the analysis following Hunter and Schmidt's (2004) suggestion that these values may be simply due to large sampling errors, which we had previously corrected.

4.6. Moderator and utility analysis

4.6.1. Moderator analysis

There are several techniques that can be used to test for moderators, and the technique used may impact the conclusion of whether a moderator exists. Sagie and Koslowsky (1993) recommend using the *Q* test when there are either a large number of studies in the meta-analysis or a large number of participants per study. Given that both of those characterize the current study, homogeneity tests were conducted utilizing the *Q* significance test statistic to assess the effects of moderators (Hedges & Olkin, 1985). A significant *Q* statistic indicates the observed effect is heterogeneous and that there is a need to search for moderators to explain further variance in the findings. Each *Q* statistic reported on non-hierarchical research questions was computed independently of the others. Beyond assessing for overall homogeneity, the *Q* statistic allows for examining nested chi-square tests to determine changes in the level of variance-explained after moderator analysis.

4.6.2. Utility analysis

Utility analysis gives meaning to the effect size by translating it into practical terms and, thereby, increases the ease of interpretation. The method used in this study is the Binomial Effect Size Display (*BESD*) (Rosenthal & Rubin, 1982). The *BESD* compares the likelihood of those in the 'treatment' or leadership intervention group experiencing 'success' with the treatment versus the likelihood of those in the control or comparison group experiencing similar success. For example, using a *BESD* one can ask, "What would the effect of the treatment be if 50% of the participants had the occurrence and 50% did not and 50% received treatment and 50% did not?" More specifically, in a study where a leadership manipulation leads to higher efficacy in followers, a *BESD* value of .70 in the treatment group implies a *BESD* value of .30 in the control group since the aggregated *BESD* values of the treatment and control group always adds up to 1. The difference between the two values is .40, which means that participants in the treatment group are 40% more likely than participants in the control group to achieve higher levels of efficacy in followers.

It is worth noting that there are no specific cut-offs for examining what is a good or a bad *BESD* value as the costs and benefits of an intervention vary depending on what is being examined. For example, Rosenthal and Rubin (1982) note that an effect size where just 1% of variance is explained can change the success rate of treatment from 45% to 55%. Even such a small effect size applied to a reduction in employee deaths using some safety intervention program would probably be considered practically significant. While the *BESD* provides an easy to understand and somewhat intuitive effect size, it can only suggest the practical value of any treatment effect and we use it here for that purpose, as well as to support our discussion of what might be the expected return on investment in leadership interventions that we provide as supplemental analyses at the end of the results section.

5. Results

The overall effect of experimental and quasi-experimental leadership research was based on 140 independent effect sizes and 13,656 unique participants. The raw, uncorrected effect of leadership interventions on all outcomes represented by *d* was .59. After correcting for study artifacts described above, the true combined effect of all leadership interventions in this study equaled .65 with a standard deviation of .80. After removing outliers, data Set 2 based on 134 studies and 11,552 unique participants shows the overall effect of leadership interventions yielded an uncorrected effect size of .61 and a corrected effect size of .67 with a standard deviation of .80 (upper credibility interval = 2.24; lower credibility interval = .90). Credibility intervals refer to a potential distribution of parameter values, and is calculated using the standard deviation versus the use of the standard error as used in confidence intervals, which attempts to identify a specific estimate that represents the population value.

Using confidence intervals, our results suggest that there is a 95% probability that the true effect of leadership interventions falls between .26 and 1.08. Because zero does not fall within the 95% confidence interval, we are 95% confident that when assessing the entire

breadth of included studies, experimental and quasi-experimental leadership research had at minimum a small, positive effect on all outcomes. Finally, the *BESD* values for Set 1 were .65 for experimental versus .35 for comparison/control groups. In Set 2 the values were .66 and .34 respectively. This indicates a distinct advantage for success in outcomes for leaders included in the experimental conditions.

5.1. A priori non-hierarchical moderators

Although an overall effect of leadership interventions was found, there was large within-group heterogeneity shown by statistically significant *Q* values for each set of data (Set 1: $Q = 4519.57$; Set 2: $Q = 3494.12$), suggesting the existence of moderators. To further explore this variance in leadership effects, we examined both the type of intervention as well as the form of leadership theory used as moderators.

Table 3
Non-hierarchical effect sizes.

						95% Confidence interval			BESD	
Sample	*k*	*n*	*d*	Corr-*d*	SD	Lower	Upper	*Q*	Exp	Cont/Comp
Intervention type I										
Developmental	37	4423	.55	.60	.57	.23	.97	1587.54	.64	.36
Developmental - Set 2	35	3389	.60	.65	.51	.25	1.06	1025.67	.66	.34
Other	101	8679	.63	.69	.53	.25	1.13	2830.20	.66	.34
Other - Set 2	97	7658	.65	.71	.50	.26	1.17	2335.99	.67	.33
Leadership theory										
Newer	40	3847	.54	.60	.41	.19	1.01	814.85	.64	.36
Pygmalion	19	2516	.70	.78	.74	.42	1.13	1453.35	.68	.32
Pygmalion - Set 2	17	1021	1.24	1.38	.92	.82	1.95	922.83	.79	.21
Traditional	41	3223	.62	.67	.72	.21	1.12	1850.58	.66	.34
Traditional - Set 2	37	2614	.58	.63	.56	.15	1.10	968.96	.65	.35
Outcome type										
Affective	31	3350	.42	.46	.26	.08	.85	340.33	.61	.39
Affective - Set 2	30	2851	.46	.50	.26	.09	.91	316.70	.62	.38
Behavioral	86	7945	.59	.64	.59	.22	1.06	3146.52	.65	.35
Behavioral - Set 2	82	6308	.62	.67	.56	.21	1.13	2328.53	.66	.34
Cognitive	88	9663	.56	.62	.48	.23	1.00	2599.19	.65	.35
Cognitive - Set 2	85	8190	.59	.65	.49	.24	1.06	2330.15	.66	.34
Performance	2	45	.97	.97	.18	.09	1.84	2.24	.72	.28
Experimental setting										
Field	50	5375	.54	.59	.57	.21	.97	1951.40	.64	.36
Field - Set 2	47	3861	.60	.67	.63	.22	1.11	1697.82	.66	.34
Lab	86	7643	.63	.69	.52	.26	1.12	2432.45	.66	.34
Lab - Set 2	83	7054	.63	.69	.44	.25	1.12	1685.01	.66	.34
Type of organization										
For profit	19	1593	.31	.34	.34	.12	.34	263.81	.59	.41
Not for profit	95	8362	.68	.74	.64	.31	1.17	3799.75	.68	.32
Not for profit - Set 2	91	7915	.66	.72	.58	.28	1.15	2980.65	.67	.33
Military	21	2841	.47	.53	.33	.19	.88	392.22	.63	.37
Military - Set 2	19	1346	.63	.73	.41	.24	1.21	293.05	.68	.32
Leadership level										
High	15	1295	.46	.51	.31	.08	.93	181.62	.63	.37
Middle	13	1311	.41	.46	.31	.07	.88	178.82	.62	.38
Middle - Set 2	12	974	.45	.51	.36	.07	.95	172.11	.63	.37
Lower	107	10,421	.63	.69	.59	.28	1.10	4081.77	.66	.34
Lower - Set 2	102	8817	.65	.71	.55	.28	1.15	3114.28	.67	.33
Study quality										
Low	56	5100	.61	.67	.63	.61	.72	1999.92	.66	.34
Low - Set 2	55	4763	.63	.69	.64	.63	.75	1961.58	.67	.33
High	61	5265	.58	.64	.54	.58	.69	1516.78	.65	.35
Intervention type II										
Active	44	4995	.50	.55	.60	.18	.93	1947.45	.64	.36
Active - Set 2	41	3623	.56	.61	.58	.18	1.04	1377.97	.65	.35
Passive	52	5046	.57	.62	.45	.21	1.02	1238.10	.65	.35
Passive - Set 2	50	4496	.55	.60	.42	.17	1.02	983.83	.65	.35

Note. Set 2 is based on data with outliers removed.

In order to explain this variance, we next partitioned the sample of studies into groupings based on the research questions that we stated in our introduction. In some instances, we report the results for only one sample comparison such as newer versus traditional theories. In this case, there were no outliers in the newer grouping, and therefore we compared the newer theory group to the traditional theory group that had outliers removed.

5.1.1. Type of intervention

Research question one addressed the difference in the effects of experimental and quasi-experimental leadership research based on the type of the intervention (training/developmental versus other). As shown in Table 3, for the full data set, the corrected effect size for training/developmental interventions was .60 ($k=37$; $n=4423$) and for Set 2 the corrected effect size was .65 ($k=35$; $n=3389$); versus the other category for Set 1 the corrected effect size was .69 ($k=101$; $n=8679$) and for Set 2 it was .71 ($k=97$; $n=7658$). For both data sets the effect sizes were slightly lower for training and developmental interventions.

5.1.2. Leadership theory

To examine research question 2a, independent effect sizes were analyzed by type of leadership theory including newer, Pygmalion and traditional. Research question 2a contrasted the effects of leadership interventions based on newer theories and those based on traditional theories (i.e., contingency, behavioral). The effect sizes of interventions based on traditional theories (Set 1: corrected $d=.67$; $k=41$; $n=3223$; Set 2: corrected $d=.63$; $k=37$; $n=2614$) versus newer theories (corrected $d=.60$; $k=40$; $n=3847$) did not appear to differ from each other after outliers were removed.

Research question 2b compared interventions based on newer leadership theories and those based on Pygmalion leadership. The largest effect size computed for theory-based leadership interventions was for Pygmalion theory based leadership. As shown in Table 3, for the full data set, the corrected effect size for Pygmalion interventions was .78 ($k=19$; $n=2516$) in Set 1, and for Set 2 the corrected effect size was 1.38 ($k=17$; $n=1021$) versus the .67 for traditional theories and .60 for newer theories noted above. This large change in effect size for Pygmalion interventions from Set 1 to Set 2 was due to the elimination of one study, which had a relatively small effect (.22) with an extremely large sample size ($n=975$). The theoretical implications of this outlier effect/study may be informative and will be addressed in the discussion.

To determine the extent to which leadership theory type might operate as a moderator of the leadership intervention, we performed within and between Q statistic analyses, independent of previous Q statistics reported above. The Q statistic between theories was significant ($Q_B=449.90$, $df=6$), suggesting that the three different leadership theory categories we investigated explain partial variance in the overall leadership effect. However, due to the statistically significant Q statistics still remaining for the effects of interventions based on each theory as shown in Table 4, additional moderators appear to be operating. We next explored types of dependent variables as possible hierarchical moderators operating within the leadership theory categories.

To address research question three, which focused on how the impacts of various theories may be moderated by the type of outcome criteria used in a given study, we divided data from each leadership theory based on the nature of the dependent variable. To examine this question, we created a new data set to maximize the use of the 612 non-independent effect sizes. As described earlier, the assumption of independence was upheld such that only unique samples were used in the calculation of each overall effect. A total of 207 independent effects were calculated and dependent variables coded for an affective focus ($k=31$), behavioral focus ($k=86$), cognitive focus ($k=88$), and organizational performance focus ($k=2$).

Using the designations and coding for affective, behavioral and cognitive dependent variables, we first ran an overall comparison of outcome types before partitioning by theory type (see Table 3). The smallest effects were observed for affective outcomes. Specifically, corrected effects for affective outcomes for Set 1 was .46 ($k=31$; $n=3350$); while for Set 2 the corrected effect was .50 ($k=30$; $n=2851$). With behavioral type outcomes for Set 1 the corrected effect was .64 ($k=86$; $n=7945$); while for Set 2 the corrected effect was .67; ($k=82$; $n=6308$). Finally, for cognitive outcomes for Set 1 the corrected effect was .62 ($k=88$; $n=9663$); while for Set 2 the corrected effect was .65 ($k=85$; $n=8190$). Overall, the largest corrected effect size was for leadership studies with outcomes using an organizational performance focus (e.g., group profit) corrected $d=.97$; ($k=2$; $n=45$). However, these results are based on a sample of only two studies and should be interpreted with caution. As noted above, this was followed in decreasing magnitude by studies assessed by outcomes with behavioral and cognitive focuses, which were each larger when compared to leadership interventions that assessed affective outcomes.

To directly address research question three, we examined the impact of experimental and quasi-experimental leadership studies from each category of leadership theory separately for affective, behavioral, and cognitive outcome variables in a hierarchical analysis. As shown in Table 5, for affective and cognitive outcomes, results show that newer theory affective: corrected $d=.65$; ($k=11$; $n=1028$); cognitive: corrected $d=.63$; ($k=29$; $n=3121$) and Pygmalion affective: corrected $d=.66$; ($k=4$; $n=132$); cognitive: corrected $d=.76$; ($k=16$; $n=2338$) leadership interventions had larger effects than those based on

Table 4
Homogeneity statistics on set 2 data — leadership theory by focus of outcome.

Moderators	k	Q_T	Q_b (df)	Q_w (df)
Newer	58	1341.43	3.50 (2)	1337.93* (54)
Pygmalion	32	1474.40	54.19* (2)	1420.21* (29)
Traditional	48	1173.47	72.58* (2)	1100.89* (44)

Note. k = number of studies, while * denotes significance at the $p=.01$ level. Set 2 is based on data with outliers removed.

traditional leadership theories affective: corrected $d = .32$; ($k = 9$; $n = 1178$) for Set 1 and for Set 2 corrected $d = .37$; ($k = 8$; $n = 679$); cognitive: corrected $d = .40$; ($k = 12$; $n = 1054$) for Set 1 and for Set 2 corrected $d = .34$; ($k = 11$; $n = 1034$).

Conversely, for behavioral outcomes, interventions based on traditional theories for the full data Set 1 corrected $d = .79$; ($k = 31$; $n = 2179$) and Set 2 corrected $d = .67$; ($k = 29$; $n = 2089$) had a larger impact than those based on newer theories corrected $d = .56$; ($k = 18$; $n = 1348$) and Pygmalion full data Set 1 corrected $d = .57$; ($k = 16$; $n = 2242$). Pygmalion Set 2 corrected $d = 1.08$; ($k = 14$; $n = 696$), however, showed the largest effect on behaviors.

We also show in Table 5, comparisons for interventions based on training and development versus those based on other manipulations such as using actors or scripts. For the developmental interventions the effect sizes were similar across the 3 types of dependent variables with a range of corrected d values of .41; ($k = 14$; $n = 1435$) to .48 for both sets of data with behavioral outcomes ($k = 32$; $n = 3630$ set 1; k = 29; $n = 2438$). The one exception was with the cognitive outcomes in set 2, which had a $d = .67$; ($k = 28$; $n = 2575$). With the other intervention types the impact was lowest with affective outcomes with a $d = .54$; ($k = k = 22$; $n = 1777$) and highest with behavioral outcomes with a $d = .76$; ($k = 58$; $n = 4429$) for set 1 and $d = .78$; ($k = 56$, $n = 3879$) for set 2, respectively. The impact on the cognitive variables was in between behavioral and affective for set 1 $d = .62$; $k = 68$; $n = 6568$) and for set 2 $d = .61$; ($k = 66$; $n = 6028$). Overall, effect sizes were generally higher for other versus the training/developmental interventions.

5.2. Post hoc comparisons

There were a number of additional variables that were collected for which we did not pose any specific research questions. In order to provide a more complete picture of our assessment of leadership interventions, we conducted several additional exploratory analyses where we partitioned the sample of leadership studies based on the descriptions made in more detail below.

5.2.1. Experimental setting

The setting was determined based on where the actual study was conducted. The two mutually exclusive categories we used here were studies conducted in the lab versus a field setting.

Table 5
Effect sizes hierarchical.

						95% Confidence interval			BESD	
Sample	*k*	*n*	*d*	Corr-*d*	S.d.	Lower	Upper	*Q*	Exp	Cont/comp
Affective outcomes										
Newer	11	1028	.61	.65	.31	.23	1.06	142.90	.65	.35
Pygmalion	4	132	.55	.66	.55	−.04	1.36	54.62	.66	.34
Traditional	9	1178	.30	.32	.13	−.03	.66	56.26	.58	.42
Traditional - Set 2	8	679	.35	.37	.17	−.05	.80	51.37	.59	.41
Behavioral outcomes										
Newer	18	1348	.53	.56	.42	.10	1.03	306.64	.64	.36
Pygmalion	16	2242	.53	.57	.65	.24	.91	998.90	.64	.36
Pygmalion - Set 2	14	696	1.01	1.08	1.00	.49	1.68	756.25	.74	.26
Traditional	31	2179	.72	.79	.83	.30	1.27	1630.72	.68	.32
Traditional - Set 2	29	2089	.60	.67	.64	.19	1.14	981.89	.66	.34
Cognitive outcomes										
Newer	29	3121	.57	.63	.50	.24	1.02	888.39	.65	.35
Pygmalion	16	2338	.67	.76	.66	.42	1.09	1067.97	.68	.32
Pygmalion - Set 2	14	885	1.20	1.37	.80	.82	1.91	609.34	.78	.22
Traditional	12	1054	.36	.40	.41	−.03	.82	226.88	.60	.4
Traditional - Set 2	11	1034	.31	.34	.15	−.07	.74	66.83	.58	.42
Developmental interventions										
Affective outcome	14	1435	.38	.41	.17	.36	.46	170.34	.61	.39
Affective - Set 2	13	1035	.39	.42	.17	.36	.49	169.88	.61	.39
Behavioral outcome	32	3630	.43	.48	.46	.45	.51	904.09	.62	.38
Behavioral - Set 2	29	2438	.43	.48	.32	.44	.52	377.50	.62	.38
Cognitive outcome	31	5061	.40	.43	.53	.40	.46	1398.66	.61	.39
Cognitive - Set 2	28	2575	.62	.67	.57	.63	.71	1113.10	.66	.34
Other interventions										
Affective outcome	22	1777	.52	.54	.29	.50	.59	410.06	.64	.36
Behavioral outcome	58	4429	.72	.76	.66	.73	.79	2082.99	.68	.32
Behavioral - Set 2	56	3879	.74	.78	.65	.75	.81	1821.44	.68	.32
Cognitive outcome	68	6568	.57	.62	.45	.60	.64	1694.20	.65	.35
Cognitive -Set 2	66	6028	.56	.61	.45	.58	.63	1557.06	.65	.35

Note. Set 2 is based on data with outliers removed.

5.2.2. Type of organization

We used three mutually exclusive categories to code organizational setting. The types of organization included for profit, not for profit and military settings.

5.2.3. Leadership level

We coded the target leader in the study as being at one of three levels. The first level was a direct supervisor. The second level represented middle to a more senior level position, and would represent a leader that had less direct contact with targeted followers on a day to day basis. The third category represented top management, such as CEOs or Presidents.

5.2.4. Study quality

Study quality was coded and split into high versus low quality using the following criteria for high quality: published study; controlled lab study; actual leaders as participants vs. role play; control croup; random selection; random assignment to conditions; if yes to conditions, random assignment of individuals (if not random assignment, were participants matched?); manipulation check (if so, manipulation check supported); experimenter blind to hypotheses; participants blind to hypotheses; control variables or covariates; pre-test; post-test; 2nd post-test; validity of measures cited/documented; multiple raters to assess outcomes; and unit of analysis specified. Studies typically coded as low quality were quasi-experimental designs, lacking in terms of randomization of participants, control groups and so forth.

5.2.5. Active versus passive

Active study interventions represented interventions where the experimenter was attempting to change the target individual's leadership in some way, whereas passive involved participating in a research project where the leader in a scenario exhibited different types of leadership styles.

We provide a breakdown of our meta-analytic findings using the post hoc comparison categories described above in Table 3. Where appropriate, we have provided the results for both data sets including outliers and with outliers removed.

5.2.6. Lab versus field setting

Returning to Table 3, the setting for the intervention had less effects when conducted in the field for Set 1: corrected $d = .59$; ($k = 50$; $n = 5375$) and Set 2: corrected $d = .67$; ($k = 47$; $n = 3861$); versus laboratory-based interventions for Set 1: corrected $d = .69$; ($k = 86$; $n = 7643$) and for Set 2: corrected $d = .69$; ($k = 83$; $n = 7054$).

5.2.7. Organizational setting

Comparing the three organizational settings used for the interventions as shown in Table 3, we found substantial differences in effect sizes. For those studies conducted in a for profit setting the effect size was on average lower: Set 1: corrected $d = .34$; ($k = 19$; $n = 1593$) versus not for profit Set 1: corrected $d = .74$; ($k = 95$; $n = 8362$); Set 2: corrected $d = .72$; ($k = 91$; $n = 7915$) and military organizational settings Set 1: corrected $d = .53$; ($k = .21$; $n = 2841$); Set 2: corrected $d = .73$; ($k = 19$; $n = 1346$). The corrected effect sizes for the not for profit and military organizational settings in the second data set were similar after outliers were removed.

5.2.8. Leader level

As shown, lower level leaders (direct supervisors) had a greater overall effect in both data Set 1: corrected $d = .69$; ($k = 107$; $n = 10{,}421$) and Set 2: corrected $d = .71$; ($k = 102$; $n = 8817$) than either middle level leaders Set 1: corrected $d = .46$; ($k = 13$; $n = 1{,}311$); Set 2: corrected $d = .51$; ($k = 12$; $n = 974$) or high level leaders Set 1: corrected $d = .51$; ($k = 15$; $n = 1295$; no data Set 2), which were both similar in their leadership effects.

5.2.9. Study quality

As shown in Table 3, studies judged to be of lower quality as defined above had the following impact: Set 1: corrected $d = .67$; ($k = 56$; $n = 5100$) and Set 2: corrected $d = .69$; ($k = 55$; $n = 4763$). For studies judged to be of higher quality, we observed a corrected effect of $d = .64$; ($k = 61$; $n = 5265$). We did not have a second set of data here for high quality studies.

5.2.9.1. Active versus passive. The last comparison shown in Table 3 examined active versus passive forms of manipulations used in the primary studies. As defined above for active manipulations the results were as follows: Set 1: corrected $d = .55$; ($k = 44$; $n = 4995$) and Set 2: corrected $d = .61$; ($k = 41$; $n = 3623$). For studies judged to be passive manipulations the results were as follows: Set 1: corrected $d = .52$; ($k = 51$; $n = 5046$) and Set 2: corrected $d = .60$; ($k = 50$; $n = 4496$). Effects were similar across active and passive manipulations except in Set 1 for the active, which had the lowest effect size.

5.3. Return on development investment (R.O.D.I.)

One of the primary goals for the current meta-analysis was to provide a point in time estimate of the effects of leadership interventions on various performance outcome measures. Based on the findings presented above, we would like to show how the return on investment into leadership development can be calculated. Specifically, by knowing the 'average' effect sizes and their ranges, as well as the estimated cost of investment, it is possible to calculate a return on development investment for future leadership training programs.

To calculate the R.O.D.I., we used the range of effect sizes from the meta-analysis, coupled with some standard human resource cost accounting methods to estimate the possible return from a leadership intervention (see Cascio, 1991 for further details). In our example below, we calculated the R.O.D.I. for a 1.5 day training intervention, which would be in the middle to lower range of the time we observed in the developmental interventions reported above.

In addition to examining length of time, we know that interventions can occur on-site at a company, off-site, or mediated by technology (i.e., networking through interactive internet programming). Changes in location influences cost structure (e.g., travel expenses) and thus influence R.O.D.I. Consequently, we also calculated R.O.D.I. for an on-site, off-site and technology mediated leadership intervention program.

5.3.1. Developing a cost structure for estimating return on development investment (R.O.D.I.)

The cost structure included in our analyses was identified by interviewing a small group of leaders from several different organizational levels (e.g., supervisor and executives) drawn from a cross section of organizations including three Fortune 500 organizations. Information gathered from the interviews became the basis for the cost structure and was separated into three groups reflecting three different mediums of intervention: on-site, off-site and on-line. Information for calculating overall cost per program can be found in Table 6. If accurate costs are known for each of these categories by an organization considering or assessing development programs, those costs can readily replace the averages in our formulas. These estimates were based directly on what we gathered from the interviews and may not necessarily generalize to a broader sampling of organizations.

5.3.2. Time in participant salary

Number of hours that participants are engaged in the intervention multiplied by their hourly rate (salary). This is used to account for lost direct labor as the leader will be engaged in development and not directly working toward the company's objectives. Our assumptions for this analysis were a salary of $100k for senior level leaders, $70k for mid level leaders and $50k for followers of mid level leaders. Of course, these salaries will vary quite substantially depending on industry, geographical location and the credentials of individual leaders. We use these figures merely to illustrate our strategy for calculating R.O.D.I.

5.3.3. Lost production time

In addition to the hourly rate of participants we included opportunity costs for individuals who directly impact revenue for the company (e.g., sales). For example, if a sales employee attends the training, we include costs as both hourly salary and as lost sales for that time. The conservative assumption for this analysis was twice the salary rate. In this example, if an employee earned $100 per hour, then the 1.5 day training would cost $1200. In addition, we assumed that the employee could earn revenue for the company of up to $2400 in that same time frame. Therefore the labor and opportunity cost for this participant would be $3600. Although this may be an artificially high estimate (e.g., a sales person who generated $2400 in revenue is likely only creating 10%–30% of that in actual profit), it is a conservative estimate in terms of cost structure for this analysis as well as R.O.D.I.

5.3.4. Technology

Includes any technology needed for conducting the interventions. High cost of on-line technology includes hardware, programmers for coding software and initial platform set up.

5.3.5. Mid. versus senior leaders

Based on conversations with large (greater than 1000) organizations from across the United States (e.g., Connecticut, Washington, Nebraska) including multiple Fortune 500 organizations, we assumed an average leadership development

Table 6
Estimated cost structure based on 1.5 day developmental leadership intervention.

Cost of training	On-site [a]	Off-site	On-line
Time in participant salary [a]			
Lost production time [a]			
Instructor	5000	5000	1500
Instructor support staff	1000	1000	5000
Technology [a]	500	500	10,000
Materials	250	250	250
Trainer traveling expenses	2000	2000	0
Travel costs for participants — Sr. [a]	0	3000	0
Travel costs for participants — Mid. [a]	0	10,000	0
Meals — Sr.	3600	3600	0
Meals — Mid.	2000	2000	0
Hotel conference room for training — Sr.	400	500	0
Hotel conference room for training — Mid.	800	1000	0
Hotel stay for participants — Sr.	0	4500	0
Hotel stay for participants — Mid.	0	15000	0

[a] On site is the location of the leadership development program which helps determine costs. These analyses were intended to show the cost structure and effect size of outside providers (e.g. academic/practitioners using a validated leadership model as the base of their intervention). On site in this case does not insinuate internal (e.g. HR) personnel delivering the program.

intervention targets either 30 senior leaders or 100 mid-level leaders. We also calculated returns for their respective followers.

5.4. Calculations for overall return

Overall return was calculated for each level (Sr. level leader and followers, Mid-level leader and followers) and for each type of leadership intervention (on-site, off-site and virtual).

We also chose to compare low, moderate and high effect sizes taken from the meta-analysis results, when examining the R.O.D I. based on the range of effects we noted for developmental interventions, as well as across theory type and outcome measures.

Referring to Table 7, one can see that for low effect sizes, most program interventions would lose money in terms of return on investment. However, even with low effect sizes the return was positive for lower level employees due to a lower cost structure. Yet, as we move to a moderate to strong effect size, in all cases the program interventions showed a positive return that potentially spanned to up to 200% of the investment in development. We would expect such effect sizes would be associated with well-validated evidence-based leadership training interventions.

6. Discussion

6.1. Do leadership interventions matter and in what way?

The overarching goals for this study were three-fold. First, we set out to survey the existing literature on experimental and quasi-experimental leadership research to synthesize what we have learned about leadership impact as a field. In the case of the current meta-analysis, all of the research examined cause and effect relationships. Our second goal was to determine if certain leadership theories and interventions have more or less of a positive impact, and if so, how, when and in what way. Finally, our third goal was to provide an empirical basis for making recommendations to leadership scholars and practitioners to guide a more refined agenda for future experimental interventions in both leadership research and practice. We felt the time was propitious to provide an empirical summary of the causal links between leadership and various organizational outcomes to hopefully promote more frequent lab and field experiments when conducting leadership research in the future.

Our findings affirm that experimental/quasi-experimental leadership interventions had a positive impact across a broad array of interventions, organization types, leadership levels, theories, levels of quality of research, and outcomes. We also found moderately positive effects when we examined leadership interventions including the three general theoretical categories in the sample. In terms of utility, participants in the leadership treatment condition broadly defined, had on average a 66% chance of positive outcomes compared to only a 34% chance of success for the comparison group for the data set corrected for outliers. While we have empirically shown that leadership interventions do make a positive difference, we also found that the ranges of these effects are quite heterogeneous.

Table 7
1.5 Day intervention in U.S. dollars.

	On-site		
	Low return	Average return	High return
Sr. level leader	(94,733.08)	39,534.92	173,802.92
Mid level leader	(160,459.23)	152,832.77	466,124.77
Sr. level follower	14,028.00	116,900.00	215,096.00
Mid level follower	100,200.00	835,000.00	1,536,400.00
ROD %— Sr. ldr	−146.48%	61.13%	200.00%
ROD %— Mid. ldr	−177.66%	169.21%	200.00%
		Off−Site	
Sr. level leader	(102,333.08)	31,934.92	166,202.92
Mid level leader	(185,659.23)	127,632.77	440,924.77
Sr. level follower	14,028.00	116,900.00	215,096.00
Mid level follower	100,200.00	835,000.00	1,536,400.00
ROD %— Sr. ldr	−141.59%	44.19%	200.00[d]
ROD %— Mid. ldr	−87.08%	27.17%	141.43%
		On-Line	
Sr. level leader	(98,733.08)	35,534.92	169,802.92
Mid level leader	(167,659.23)	145,632.77	458,924.77
Sr. level follower	14,028.00	116,900.00	215,096.00
Mid level follower	100,200.00	835,000.00	1,536,400.00
ROD % — Sr. ldr	−143.77%	51.75%	200.00
ROD % — Mid. ldr	−101.24%	36.44%	174.12%

Note: Values in parenthesis represent losses.

6.1.1. Impact by type of intervention

Analysis of research question one sheds some light on the differing impact of experimental and quasi-experimental leadership studies that are developmental versus other types (e.g., leader assignment or scenarios). Results indicated slightly stronger effects for leadership interventions that were not training oriented versus those that were developmental. This finding may reflect in part, that greater levels of intrapersonal change are required in developmental studies versus temporal affective reactions or behavioral adaptations more common in non-developmental studies. Although it is important to point out that the developmental interventions tended to last longer than non-developmental interventions (e.g., from 1 to 7 days on average versus minutes or hours for other types of experimental leadership interventions). Yet, the training interventions were likely targeting real change in organizations, where the other interventions may have had a greater impact because there were less alternative forces attempting to keep the leader and follower from changing. Nevertheless, where the intervention was training or other focused the effects reported in this study require further research to determine how we can enhance the training/developmental interventions.

We also suggest that the outcome criteria used to reflect these developmental changes may be more difficult to effect, or at the very least may vary depending on which outcome is chosen. For example, increasing a leader′s level of self awareness about how they listen to others may be easier to effect as compared to impacting the level of inquiry demonstrated by followers or followers′ willingness to challenge their leader. Also, with all leadership impact variables, the amount of time it takes for the effect to be observed may be longer due to leadership being mediated by other variables prior to impacting performance outcomes. The similarity in effect sizes, however, offers positive evidence of the efficacy of leader training and developmental interventions and their utility to organizations.

6.1.2. Impact by leadership theory

One of the strengths of this study was the inclusion of research conducted across leadership research streams. The inclusion of different theoretical categories contributed to high levels of variance in the overall effect size, but also presented for the first time an opportunity to compare theories and also explore some key moderators across the body of leadership research.

Research question two dealt with the comparative impacts of interventions based on newer leadership theories versus traditional leadership, and Pygmalion. While both newer and traditional leadership interventions had a moderately positive overall impact, their effects do not appear to significantly differ until moderators are included in the analysis as discussed below. Interestingly, our results showed that Pygmalion interventions produced the largest overall effect size for both the full sample as well as after removing one outlier. In fact, we demonstrated through utility analysis that participants receiving Pygmalion leadership interventions had up to a 79% (Set 2 data) chance of success versus 64% chance of success for treatment participants in other newer theory interventions.

For Pygmalion, however, maintaining the one statistical outlier in data set one shows a lower chance of success (68%). Of note, this large sample study (Eden et al., 2000; Study 1 of 7) although a statistical outlier, may actually adequately reflect the phenomenon it is measuring. This study tested the Pygmalion leadership style (PLS). Mirroring our differences found between Set 1 and Set 2 data in this study, the results for PLS has shown only small effects in research as compared to non-PLS Pygmalion studies (Eden et al., 2000). This outlier study was also the most unique one included in the sample in terms of the typical focus of prior Pygmalion research, which has not been on leadership style per se, but leader expectations. Finally, examining the confidence intervals for Pygmalion research as with the other theories shows a considerable range of effects, thus our conclusions regarding impact of Pygmalion interventions must be evaluated in light of this broad range of effects.

These differing findings suggest that for a leader′s expectations to become a self-fulfilling prophecy by increasing followers performance, the leader must truly believe the expectancy and not just try to display they believe it. A cynical interpretation of those studies was that the Pygmalion treatments were only effective if the leader was 'duped' into the expectations toward their followers. Eden, Geller, Gewirtz, Gorden-Terner, Inbar, and Liberman, et al. (2000, pp. 197–198) propose this is partially due to conflicting automatic and nonverbal behaviors the leader may give off when they do not truly believe the expectation, and they therefore question the "application validity" of the PLS theory in organizational practice.

Dov Eden (personal correspondence) suggests that the strong Pygmalion effects may be partially due to its focus on one key variable–leader expectations. Eden also suggests that the communication of positive leader expectancies to the follower may be but a special case of transformational leadership. Specifically, a leader may not be able to be truly transformational without being able to create self-fulfilling prophecies in their followers; as the communication of expectancies is inherent in at least parts of transformational leadership. Importantly, including Pygmalion as a newer form of leadership would have raised the effects for that category and thereby its primacy of effects over traditional. We feel it is important to separate Pygmalion in our analysis, however, to illuminate the power of that unique theoretical construct and its important implications concerning the impact of perceived leader authenticity of heightened expectations. Foremost, holding the largest corrected effect sizes in this meta-analysis, we call for further analysis of Pygmalion′s underlying mechanisms and relationships to perceived leader authenticity and as a possible integral element or mediator of newer theory leader behaviors.

While the field of leadership has been drawn heavily to examining the newer models of transformational, visionary, and charismatic leadership (Avolio, 1999; House & Aditya, 1997; Lowe & Gardner, 2000; Yukl & Van Fleet, 1992), our findings suggest that separating leader expectancies from this category, and without taking outcome criteria into account, interventions based on newer theory approaches, while positive, do not necessarily have a greater impact than other theoretical approaches and that results may depend on the outcome one is striving to influence as discussed in the next section.

It is also interesting to note, that Lowe, Kroeck, and Sivasubramaniam (1996) reported an average effect of .64 for the three transformational scales (.71 for charisma, .62 for individualized consideration and .60 for intellectual stimulation) with rated effectiveness as a criterion. This effect was comparable to a corrected effect of .60 for newer studies in the current study. In an

updated meta-analysis of this literature, Dumdum, Lowe, and Avolio (2002) reported an average corrected effect size of .62 for the transformational scales that was comparable to the effect size reported in the current study using rated effectiveness as a criterion. However, we should note that such direct comparisons are difficult as each study above used different scales and different criteria, as well as being subject to more or less single source/method effects. In addition, in the current study beyond transformational, our newer theory category also included charismatic, visionary and inspirational leadership, where these additions may influence the findings.

In sum, our BESD utility analysis using Set 2 data shows that for traditional theories those in the treatment condition have a 65% chance of success, versus 64% for newer theories, and 79% for Pygmalion (68% for Pygmalion Set 1 data). However, this story changes when the type of outcome variable is taken into consideration.

6.1.3. Impact of theory type by focus of the dependent variable

In order to further break down the variance found in leadership effects and to address research question three, the focus of each dependent variable was coded. While only two effect sizes relating to organizational performance outcomes could be calculated with available data, these interventions showed the largest effect size. However, one must be very cautious in interpreting these findings given the limited data available for our meta-analysis. Behavioral and cognitive outcomes also had a larger impact than did those leadership interventions measured by an affective outcome. However, investigation of affect and emotions in the leadership process is an area that clearly needs further research (Lord, Klimoski & Kanfer, 2002).

By analyzing each leadership theory hierarchically and by focusing on the dependent variables measured, additional light can be shed on the unexpected findings related to research question two, which focused on comparing the different leadership theories. In fact, hierarchical analyses indicated that experimental and quasi-experimental studies based on newer forms of leadership theories did have appreciably *larger* effects than those based on traditional leadership theories for both affective and cognitive dependent variables. On the other hand, interventions based on traditional theories had a larger effect on behavioral outcomes than did the newer theory based interventions. Pygmalion leadership, beyond having the largest overall effects, additionally showed in hierarchical analysis to have the greatest effects among the theory categories on both the behavioral, and cognitive categories, and was similar to newer theory effects on affective outcomes.

Given the core focus of these respective theories, it may not be surprising that theories such as transformational leadership had a larger impact on followers' feelings and thinking, while traditional approaches had a great impact on more proximal target behaviors. These findings underscore the importance of Yukl and Van Fleet's (1992) summary of the main methodological problems with leadership studies including the arbitrary manner in which measurement criteria for key variables and constructs are determined. We suggest that future leadership research give greater consideration to the choice of dependent variables with respect to the theoretical framework that is being tested.

Future leadership intervention research should now consider identifying appropriate criteria based on the core propositions in leadership theories. For example, with respect to transformational leadership theory, we would expect future research to try to obtain performance measures that get at individuals or groups 'performing beyond expectations'. Most of the prior intervention research on transformational leadership has not focused on extreme performance, potentially reducing the validities obtained in prior research (also see Korman, 1966; Kroeck, Lowe & Brown, 2004 for additional discussion on linking theory to outcomes measured).

Our suggestions do not preclude using a broader set of measures, but rather a determination of the outcome measures that are theoretically most appropriate to be impacted by changing the leader's style or behavior given the logic theorized in the theory under investigation. In field research in particular, researchers will oftentimes accept whatever performance or other outcomes they receive from the host organization. Unfortunately, as our findings suggest, such arbitrary choices of dependent measures may have a distinct impact on the magnitude of the observed effects.

We also suggest that the differential effects observed across the theoretical groupings used in this study should be interpreted with some degree of caution in that we have yet to fully explore some of the mediating mechanisms that could have produced these results. For instance, although traditional theories appear to yield greater intervention effects with behavioral outcomes, it is possible that these effects may be due in part to mediating cognitive and affective processes that drive those behaviors (e.g., Lord & Brown, 2004; Wofford, Goodwin & Whittington, 1998). Further, greater attention needs to be paid to how cognition and affect interact in the exercise of leadership. Some researchers suggest that affect and cognition may be inextricably intertwined (e.g., Bower, 1983; Cacioppo & Gardner, 1999; Smith & Lazarus, 1990). For example, job satisfaction is purported to have both cognitive and affective components (Organ & Near, 1985). Our findings, however, should help guide researchers to choose outcome variables in future studies that best represent the effects of each theory, and that by doing so, begin to also investigate in a more focused manner the mediating and moderating variables in each theory's nomological network. Practically, however, any non-behavioral criterion should be ultimately linked to increased human and organizational performance.

6.2. Investigating temporal effects

It is important to highlight the impact of time in leadership theory and intervention research and we call for much more focus in this area. Pygmalion interventions, and to a lesser extent traditional interventions may be effective over a shorter time frame, while the impact of newer theory interventions require the development of a transformational relationship in which deep and veritable changes in followers emerge over an extended period of time. Burns (1978) and Bass (1985) suggested that what transforming leaders do is to transform and develop followers into leaders over time, and some researchers (e.g., Lord & Brown, 2004; Wofford et al., 1998) have suggested this requires lasting cognitive and schema changes in individuals.

If the main criterion that distinguishes a newer theoretical approach from traditional approaches is the depth of intra or interpersonal change theorized, then we have not yet investigated how long such deep change takes at supervisory, middle and more senior levels of management. For example, it is very possible that the effects at more senior levels of management may take a greater amount of time to manifest than what could be observed at lower and more direct levels of leadership.

It is important to reiterate that only 9% of studies included in this meta-analysis exceeded seven days in duration. The median intervention length was 3 to 6 h, while the modal intervention length was under 1 h. Given the condensed nature of most of these studies, as a discipline, we must question what we really know about the temporal effects of leadership interventions and of the lasting power of those effects across different leadership levels, theories and dependent variables. For example, only 34% of studies in the sample used a longitudinal design to assess within-group effects, whereas 56% were between-group experiments and 10% were of mixed design. As a source of comparison, in their 10 year analysis of *The Leadership Quarterly*, Lowe and Gardner (2000) reported that 82% of published studies (including non-intervention studies) in that journal were cross-sectional and 18% longitudinal.

To assume that the shorter of these studies truly assesses deep or lasting leadership effects may be erroneous. We feel that it is important for future leadership research to better match temporal designs to theoretical frameworks, and to increase attention to using extended longitudinal studies and repeated measures. This is particularly true for leader development studies and for those theories (i.e., transformational) that propose deep, lasting change in followers. We again call for matching criteria to not only theory, but also study length. If a criterion, such as trust or self-concept change requires time to affect, researchers might temporally plan their research programs accordingly.

6.3. Post-hoc analyses

Surprisingly, our post hoc analyses did not necessarily reveal any striking patterns that would change the interpretation of our findings. Generally speaking, after correcting for outliers, there was no appreciable difference between lab and field research in terms of the magnitude of effect sizes. This is due in part to the fact that we found considerable variation across theory types and dependent variables, which may be more useful than the broader category of lab versus field. We have noted above that there were also some differences for study quality, indicating that higher quality studies produced somewhat lower effect sizes, which were all still in the moderately positive range. Study quality did produce what we might expect in terms of more rigorously controlled studies producing more conservative effect sizes.

We did observe some interesting differences for organizational setting and leadership level. For example, with organizational setting, using data set 2 we found that the largest effects or differences were found in military samples. This may be due to a variety of factors including the level of attention placed on leadership development in the military, as has been noted by previous authors (e.g., Bass & Riggio, 2006). Based on the current results, we suggest that future research needs to focus on examining the effects of leadership interventions in a broader array of organizational settings. For example, we would expect some of the interventions based more on traditional leadership theories to have less impact in organizations where there is dramatic change constantly confronting leaders and followers, as opposed to an intervention based on transformational leadership theory. With this example, we suggest that the effects of interventions, may well depend on the characteristics of the organizational context in which that intervention is implemented. In addition, future research should also focus on the conditions or nature of the intervention that might lead to stronger differences in the military versus for profit and not for profit organizations. Such conditions might include how much focus is placed on doing what is good for the individual as opposed to the causes and mission of the organization.

The effects for leadership level showed that interventions had a greater impact on lower versus middle to higher level leaders. We might speculate that lower level leaders have more direct interactions with their followers, and oftentimes deal with less abstraction and complexity as compared with senior level leaders (Jacobs & Jacques, 1987). The level of complexity associated with higher level leadership positions may be one explanation for our results. Yet, the differences across leadership levels may be due to a variety of other factors such as the nature of the intervention, the organizational support for lower versus middle to senior leadership interventions, as well as a number of other possibilities that deserve closer scrutiny.

Finally, for the passive and active interventions we found that the corrected effects were similar for each intervention type. We suggest that these categories may have been too broad and that future research needs to take a much closer look at the characteristics of the intervention before any firm conclusions can be drawn.

6.4. Limitations

Although we attempted to build a comprehensive data base from all sources, the final data sample included 84% from published sources. Due to the inherent bias of journals against publishing null or negative findings, all synthesis research, including this study, may be upwardly biased. For example, Lipsey and Wilson (1996) showed that published studies had larger average effects (.53) than non-published studies (.39). Also, going back to the early part of the last century made it difficult to uncover unpublished studies, which were likely discarded over the years. However, countering these potential inflation effects, effect sizes may conversely be suppressed as some effects used to calculate (d) for this study came from assessing the difference between an intervention and a comparison or control group in the primary research study. For example, in the Dvir, Eden, Avolio and Shamir (2002) study, the authors compared a more focused transformational leadership development intervention to a less focused 'eclectic' leadership intervention instead of a pure control group, likely reducing the effects of between group differences. Using this more conservative comparison could diminish the differences in effect sizes observed in our study results.

As in all research syntheses, it is also likely that we did not find all studies that met the inclusion criteria during the search phase. However, the rigorous search process should have reduced this potential problem, and netted if not a full population sample, hopefully a representative sample. One other limitation inherent in meta-analysis is the lack of significance testing to compare effect sizes. Hunter and Schmidt (2004) discourage the use of significance testing due to the overemphasis of sample size in estimating power. This prevents absolute tests of differences between effect size statistics, leaving interpretation up to the reader. Although we have occasionally referred to apparent differences, direct and absolute tests of differences were not conducted.

Beyond the practice of researchers who oftentimes do not report null findings, another limitation is the lack of detailed reporting (such as all cell data, simple, main, and interaction effects) in primary articles. This practice forced us to assume null findings of non-reported two and three-way interaction effects in conversion of F to d statistics, which may have further suppressed our overall estimate of effects. However, we did find an increase in reporting this type of data over time, with more recent articles tending to report more complete statistical output from their analyses. Along the same lines, we suggest that future research attempt to determine the extent to which data collected from single versus multiple sources may have affected the pattern of results reported from this study. Indeed, the effect of single source/method bias should be explored in future leadership intervention research particularly given the problems of interpreting common source data and differing levels of self-other agreement (Cheung, 1999; Dionne, Yammarino, Atwater & James, 2002). For example, findings reported by Lowe, Kroeck, and Sivasubramaniam (1996) showed substantially higher relationships for data collected from a single source with ratings of transformational leadership and performance outcome ratings.

Lastly, we were forced to make subjective decisions when coding select data as outlined in the methods section. For example, we decided to code all of the studies based on their root theory and its assumptions. We did so realizing that many theories have been modified over time. These procedures were driven by strict rules and assumptions as outlined to inform the reader. Regardless of these efforts, however, missing data prevented complete representation of the 200 intervention studies in any one of the separate analyses.

6.5. Suggestions for future research

To support theory building, in future synthesis studies we hope to explore the interactions between types of manipulation focus (affective, behavioral, or cognitive) and the type of outcome measurement (affective, cognitive, behavioral, performance) to test the variance of effects at different levels of analysis. For example it would be informative for research synthesis to assess how behavioral manipulations influence cognitive outcomes or vice versa. Such analyses may begin to inform us of important mediating and other leadership processes. Surprisingly, organizational performance outcomes were particularly lacking ($n=2$) and would be important to be included in future intervention studies. Related to this is the need for leadership research to be conducted in more naturalistic settings. Interactions between types of manipulation and level of analysis (follower, leader, group, etc.) would also be informative in guiding future research to assess within and cross-level leadership effects.

Assessing demographics is another area of important needed research. Although the frequency of reporting gender in the sample was just over 50%, analyzing the overall effects of gender of both followers and leaders, and the possible interactions between leader and follower gender would be of great interest. Further analysis of nationality is also needed. Specific leadership theories and other factors may produce varying effects when examining different nationalities. Indeed, Bass (1990) pointed out that the recent advent of leadership as a discipline has focused on a Western, U.S.-centric, post-industrial approach. In our data set, approximately two thirds of the studies were conducted in the U.S. Further compounding this problem, even though studies may be conducted in other cultures, there has been a lack of cross-culture theory-building, whereby Western frameworks are merely tested in other cultural settings and samples. This led Peterson and Hunt (1997) to suggest that international research needs to be investigated through new globally oriented theoretical lenses.

Finally, there is a clear need for future research to explore different intervention types beyond those discussed in the current study. For example, leadership interventions that are rolled out from top management down to the bottom may have more support in terms of positive impact as compared to those begun at a middle level manager level. In addition, with the advent of new information technology, future leadership intervention research ought to examine how the effects of training can be boosted over time by sending text messages or emails to reinforce lessons learned in a leadership training intervention. Clearly, there is much to be done given the limited amount of time typically allocated to leadership interventions that we reported above. Thus, future research can also examine time as a factor in terms of the type and nature of leadership intervention.

6.6. Promoting the calculation of R.O.D.I.

We illustrated at the end of the results section how one can calculate R.O.D.I. using the effect sizes generated by the meta-analysis. One of our main purposes in doing so was to reinforce the idea that telling decision-makers in organizations that leadership training works or is valid, is only one part of the argument. The other part of the argument of whether to invest time and resources in leadership development is whether there is any substantive pay back on that investment. We believe that beyond the pay back, there is another intangible benefit to dollarizing leadership interventions and that is managers taking them more seriously. Until managers see the costs and benefits of leadership development, they will continue to view it as a nice-to-have versus a must-have intervention. In doing so, they may actually reduce the intervention's potential impact, by not reinforcing its effects over time.

We also realize that there are a number of potential drawbacks to the R.O.D.I. analysis reported above that need to be considered in future work. First, we have to assume that the effect sizes obtained in the meta-analysis will generalize to different organizational contexts, which may not be the case. Second, much of the research included in the meta-analysis was based on U.S. samples potentially invalidating our findings in other cultures. Third, we have made a number of assumptions about the costs of the intervention and return to estimate R.O.D.I., which may have been inaccurate. All of these estimates and assumptions require further replication and validation.

In addition to the concerns raised regarding the R.O.D.I. we think it is important to note that our analysis did not consider a number of factors that would suggest that the positive return in most instances was indeed conservative. For example, we primarily examined the direct impact of the intervention on the leader and follower. Certainly, leadership development interventions can have positive and cascading effects on indirect followers, the unit/organizational climate, sharing of what was learned with peers, as well as positive effects on the culture. We also know that to the extent leadership is connected to employee commitment and in turn, customer engagement, that the return on investment may have underestimated the positive impact on customer retention, repeat purchases and so forth.

Nevertheless, we have opted to present this analysis for two important reasons. First, we hope to help shift most managers' thinking about training interventions being a cost to endure as opposed to a potential investment with solid returns. Second, we simply wanted to demonstrate it is possible for us to estimate a return on leadership developmental interventions because in our view this type of analysis is not much different than other disciplines would use in estimating the return on investment in technology, marketing research and/or financial acquisitions. Why should leadership development interventions be perceived differently in terms of a return on investment? We believe they should be viewed the same and have provided the example to hopefully spur additional attempts at estimating R.O.D.I. with respect to leadership development interventions. Such assessment also brings attention to the use of validated approaches to leader and leadership development and places a greater standard of evidence on potential providers.

7. Conclusion and practical implications

One of the practical implications of our study is that regardless of the type of intervention, it appears that leadership interventions do have an impact on a variety of outcomes. Yet, leadership interventions appear to differ in terms of their impact based on the theoretical focus of the leadership model. Leadership theories that have focused more on behavioral change may indeed have a greater impact on behavior versus theories focusing on emotional or cognitive change. This suggests that future work on leadership development should take into consideration how the leadership model being learned by participants is linked to the specific outcomes that one expects to have the greatest impact on over time. By linking the leadership model to specific outcomes, we can not only demonstrate the effectiveness of the leadership intervention, we may also build confidence in participants that what we say will work and where, may actually work in terms of leadership development, thereby building greater leader efficacy. Yet, we have also observed here that the effects of leadership interventions were quite varied and multi-dimensional, which suggests the need to have a greater degree of sophistication in how we design future experiments, methods, and criteria to assess exactly how leadership matters, for whom it matters, and under what circumstances it matters. We hope that the accumulated results and interpretations presented here will provide a fresh start for accelerating work that examines the cause and effect impact of leadership.

References

Antonakis, J., & House, R. J. (2002). The full-range leadership theory: The way forward. In B. J. Avolio & F. J. Yammarino (Eds.), *Transformational and charismatic leadership the road ahead* (pp. 3—34). Amsterdam: JAI Press.

Avolio, B. J. (1999). *Full leadership development: Building the vital forces in organizations.* Thousand Oaks, CA: Sage.

Avolio, B. J., Bass, B. M., Walumbwa, F., & Zhu, W. (2004). *MLQ Multifactor Leadership Questionnaire: Technical report, leader form, rater form, and scoring key for MLQ Form 5x-short* (3rd ed.). Redwood City, CA: Mind Garden.

Avolio, B. J., Sosik, J. J., Jung, D. I., & Berson, Y. (2003). Leadership models, methods, and applications. In I. B. Weiner (Ed.), *Handbook of psychology: Industrial and organizational psychology* (pp. 277—307). Hoboken, NJ: John Wiley & Sons.

Bass, B. M. (1985). *Leadership and performance beyond expectations.* New York: Free Press.

Bass, B. M. (1990). *Bass and Stogdill's handbook of leadership* (3rd ed.). New York: The Free Press.

Bass, B. M. (1998). *Transformational leadership: Industry, military, and educational impact.* Mahwah, NJ: Lawrence Erlbaum.

Bass, B. M., & Riggio, R. E. (2006). *Transformational leadership* (2nd ed.). Lawrence Erlbaum Associates: Mahwah, New Jersey.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal of Applied Psychology, 89*, 901—910.

Bower, G. H. (1983). Affect and cognition. *Philosophical Transactions of the Royal Society of London Series B, 302*, 387—402.

Bryman, A. (1992). *Charisma and leadership in organizations.* London, U.K.: Sage.

Burke, M. J., & Day, D. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology, 71*, 232—246.

Burns, J. M. (1978). *Leadership.* New York: Harper & Row.

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual Review of Psychology, 50*, 191—214.

Cascio, W. E. (1991). *Costing human resources: The financial impact of behavior in organizations* (3rd ed.). Boston: PWS-Kent.

Champ, C. W., & Woodall, W. H. (1987). Exact results for Shewhard control charts with supplementary runs rules. *Technometrics, 29*(4), 393—399.

Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology, 52*, 1—36.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Collins, D. B., & Holton, E. F. (2004). The effectiveness of managerial leadership development programs: A meta-analysis of studies from 1982 to 2001. *Human Resource Development Quarterly, 15*, 217—248.

Conger, J. A., & Kanungo, R. N. (1987). Toward a behavioral theory of charismatic leadership in organizational settings. *Academy of Management Review, 12*, 637—647.

Conger, J. A., & Kanungo, R. N. (1998). *Charismatic leadership in organizations.* Thousand Oaks, CA: Sage.

Day, D. V., & O'Connor, P. M. G. (2003). Leadership development: Understanding the process. In S. E. Murphy & R.E. Riggio (Eds.), *The future of leadership development* (pp. 11—28). Mahwah, NJ: Lawrence Erlbaum.
Day, D. V., Zaccaro, S. J., & Halpin, S. M. (2004). *Leader development for transforming organizations: Growing leaders for tomorrow.* Mahwah, NJ: Lawrence Erlbaum.
Dionne, S. D., Yammarino, F. J., Atwater, L. E., & James, L. R. (2002). Neutralizing substitutes for leadership theory: Leadership effects and common-source bias. *Journal of Applied Psychology, 87*, 454—464.
Dvir, T., Eden, D., Avolio, B. J., & Shamir, B. (2002). Impact of transformational leadership training on follower development and performance: A field experiment. *Academy of Management Journal*, *45*, 735—744.
Dumdum, U. R., Lowe, K. B., & Avolio, B. J. (2002). A meta-analysis of transformational and transactional leadership correlates of effectiveness and satisfaction: An update and extension. In B. J. Avolio & F.J. Yammarino (Eds.), *Transformational and charismatic leadership: The road ahead* (pp. 35—66). Amsterdam: JAI Press.
Eagly, A. H., & Karau, S. J. (1991). Gender and emergence of leaders: A meta-analysis. *Journal of Personality and Social Psychology*, *60*, 685—710.
Eden, D., Geller, D., Gewirtz, A., Gorden-Terner, R., Inbar, I., Liberman, M., et al. (2000). Implanting Pygmalion leadership style through workshop training: Seven field experiments. *The Leadership Quarterly*, *11*, 170—210.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.
House, R. J. (1976). A 1976 theory of charismatic leadership. In J. G. Hunt & L.L. Larson (Eds.), *Leadership: The cutting edge* (pp. 189—207). Carbondale, IL: Southern Illinois University Press.
House, R. J., & Aditya, R. (1997). The social scientific study of leadership: Quo vadis? *Journal of Management*, *23*, 409—474.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
Jacobs, T. O., & Jacques, E. (1987). Leadership in complex systems. In J. Zeidner (Ed.), *Human productivity enhancement: Organizations, personnel, and decision making*, *vol. 2*. (pp. 7—65). New York: Praeger.
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*, 765—780.
Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: A meta-analytic test of their relative validity. *Journal of Applied Psychology*, *89*, 755—768.
Kane, T. D., Zaccaro, S. J., Tremble, T. R., & Masuda, A. D. (2002). An examination of the leader's regulation of groups. *Small Group research*, *33*, 65—120.
Kirkpatrick, S. A., & Locke, E. A. (1996). Direct and indirect effects of three core charismatic leadership components on performance and attitudes. *Journal of Applied Psychology*, *81*, 36—51.
Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.
Korman, A. (1966). "Consideration," "initiating structure," and organization criteria — A review. *Personnel Psychology*, *19*, 349—361.
Kroeck, K. G., Lowe, K. B., & Brown, K. (2004). The assessment of leadership. In J. Antonakis, A. T. Cianciolo & R.J. Sternberg (Eds.), *The nature of leadership* (pp. 71—97). Sage Publications.
Lipsey, M. W., & Wilson, D. B. (1996). *Toolkit for practical meta-analysis.* Nashville, TN: Vanderbilt University.
Lord, R. G., & Brown, D. J. (2004). *Leadership processes and follower self-identity.* Mahwah, NJ: Erlbaum.
Lord, R. G., & Hall, R. J. (1992). Contemporary views of leadership and individual differences. *The Leadership Quarterly*, *3*, 137—157.
Lord, R. G., Klimoski, R. J., & Kanfer, R. (2002). *Emotions in the workplace: understanding the structure and role of emotions in organizational behavior.* San Francisco: Jossey Bass & Sons.
Lowe, K. B., & Gardner, W. L. (2000). Ten years of The Leadership Quarterly: Contributions and challenges for the future. *The Leadership Quarterly*, *11*, 459—514.
Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformation and transactional leadership: A meta-analytic review of the MLQ literature. *The Leadership Quarterly*, *7*, 385—425.
Organ, D. W., & Near, J. P. (1985). Cognitive vs. affect measures of job satisfaction. *International Journal of Psychology*, *20*, 241—254.
Peterson, M. F., & Hunt, J. G. (1997). International perspectives on international leadership. *The Leadership Quarterly*, *8*, 203—231.
Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500—504.
Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology*, *46*, 629—640.
Shamir, B., House, R. J., & Arthur, M. B. (1993). The motivational effects of charismatic leadership: A self-concept based theory. *Organizational Science*, *4*, 577—594.
Smith, C. A, & Lazarus, R. S. (1990). Emotion and adaptation. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 609—637). New York: Guilford.
Strube, M. J., & Garcia, J. E. (1981). A meta-analytical investigation of Fiedler's model of leadership effectiveness. *Psychological Bulletin*, *90*, 307—321.
Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, *74*, 259—264.
Wofford, J. C., Goodwin, V. L., & Whittington, J. L. (1998). A field study of a cognitive approach to understanding transformational and transactional leadership. *The Leadership Quarterly*, 955—984.
Yukl, G. (1998). *Leadership in organizations* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.
Yukl, G. (2002). *Leadership in organizations* (5th ed.). Upper Saddle Creek, NJ: Prentice-Hall.
Yukl, G. (2006). *Leadership in organizations* (6th ed.). Upper Saddle Creek, NJ: Prentice-Hall.
Yukl, G., & Van Fleet, D. D. (1992). Theory and research on leadership in organizations. In M. D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology*, *vol. 3*. (pp. 147—198) Palo Alto, CA: Consulting Psychologists Press.