The Leadership Quarterly xxx (xxxx) xxxx

Contents lists available at ScienceDirect

The Leadership Quarterly

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

The state of higher education leadership development program evaluation: A meta-analysis, critical review, and recommendations

Denise L. Reyes[a], Julie Dinh[a], Christina N. Lacerenza[b], Shannon L. Marlow[c], Dana L. Joseph[d], Eduardo Salas[e,*]

[a] Department of Psychological Sciences, 6100 Main St., Rice University, Houston, TX 77005, United States of America
[b] Leeds School of Business, University of Colorado Boulder, 995 Regent Dr., Boulder, CO 80309, United States of America
[c] College of Business, 1 UTSA Circle, The University of Texas at San Antonio, San Antonio, TX 78249, United States of America
[d] Department of Management, University of Central Florida, College of Business, 12744 Pegasus Dr., Orlando, FL 32816, United States of America
[e] Department of Psychological Sciences, Rice University, 6100 Main Street, Sewall Hall 429C, Houston, TX 77005, United States of America

A B S T R A C T

There is a widespread use of leadership development (LD) for students in higher education; however, less is known about the effectiveness of such practices. We provide a summative and meta-analytic review to identify the state of LD programs for students in higher education (i.e., undergraduate and graduate students). The overall objective is to demonstrate whether LD programs are implementing the most effective strategies with any discrepancy revealing a gap between management science and higher education practice. Our results suggest that LD programs within higher education work, but evaluation studies need to more effectively address endogeneity concerns. As a way moving forward, we provide recommendations for conducting a LD program evaluation study and for conducting a meta-analysis on evaluation studies. This meta-analysis can be used as a starting point for the discussion on these issues. We hope that our findings can guide the future development of LD programs.

Introduction

Over the last three decades, a large body of knowledge has amassed regarding how to develop leaders, advancing beyond the notion that individuals can only be born as leaders. Simultaneously, the field of higher education has increasingly recognized the value in allocating resources to train future generations of leaders. A search of the top 50 universities as ranked by U.S. News and World Report (2018) showed that every school on the list offers some form of leadership development (LD) for both undergraduate and graduate students. However, despite the apparent widespread use of LD, little is known about the effectiveness of such practices.

Although it has been argued that leadership is a dispositional trait (Judge, Bono, Ilies, & Gerhardt, 2002), research suggests that it can also be developed through experience (Day, 2000; Day, Fleenor, Atwater, Sturm, & McKee, 2014; Day, Harrison, & Halpin, 2009; Lacerenza, Reyes, Marlow, Joseph, & Salas, 2017). In particular, LD programs can increase leadership knowledge, skills, and abilities (KSAs), which, in turn, can produce other positive downstream effects (Arvey, Rotundo, Johnson, Zhang, & McGue, 2006; Day et al., 2009). Accordingly, the literature yields promising evidence for LD programs, with several meta-analyses linking training to desirable outcomes (Avolio, Reichard, Hannah, Walumbwa, & Chan, 2009; Burke & Day, 1986; Collins & Holton, 2004; Lacerenza et al., 2017; Powell & Yalcin, 2010). Collins and Holton's (2004) meta-analysis found that "organizations should feel comfortable that their managerial leadership development programs will produce substantial results, especially if they offer the right development programs for the right people at the right time" (p. 240). Indeed, training design may enhance individual knowledge and behavior, as well as organizational results (Collins & Holton, 2004), particularly if it takes the specific needs of the trainees into account. For example, LD programs for school administrators have been shown to improve performance, including improvements in student achievement (Darling-Hammond, LaPointe, Meyerson, Orr, & Cohen, 2007). LD programs can bolster these outcomes by drawing upon the hands-on nature of the role and implementing mentoring, coaching, and problem-based frameworks (Darling-Hammond et al., 2007). Notably, this is only one such example of optimizing LD programs for an audience; training developers should consider all delivery and design factors that may influence overall program effectiveness.

Most recently, Lacerenza et al. (2017) found support for the effectiveness of leadership programs across 335 independent employee samples. Their meta-analysis tested 15 moderators and found support for the use of needs analysis, feedback, multiple delivery methods,

* Corresponding author at: Department of Psychological Sciences, Rice University, 6100 Main Street, Houston, TX 77995, United States of America.
E-mail addresses: Denise.L.Reyes@rice.edu (D.L. Reyes), Julie.Dinh@rice.edu (J. Dinh), Christina.Lacerenza@colorado.edu (C.N. Lacerenza), Shannon.Marlow@utsa.edu (S.L. Marlow), Dana.Joseph@ucf.edu (D.L. Joseph), Eduardo.Salas@rice.edu (E. Salas).

https://doi.org/10.1016/j.leaqua.2019.101311
Received 6 July 2018; Received in revised form 23 August 2019; Accepted 24 August 2019

spaced training sessions, on-site training, and face-to-face delivery in improving a host of outcomes. However, mixed findings on the impact of attendance policy and content (i.e., hard skills versus soft skills) indicated that the effectiveness of these features is contingent on the desired outcome type. For example, voluntary programs led to greater transfer (i.e., use of trained skills on-the-job) of training, while mandatory attendance yielded greater results (i.e., organizational outcomes). This underscores the differential impact of LD program characteristics based on the overarching goal of the organization.

Although the aforementioned investigations contribute to our understanding of LD within organizational settings (with employees as trainees), the literature has not reviewed these programs in higher education. However, a substantial number of studies have developed, implemented, and evaluated LD programs within these contexts. The quantity of work in this area suggests that it is a topic not only of interest to educational researchers, but also one of relevance to organizational investment. Despite this, it is unclear how these interventions are being built and assessed. Thus, the current meta-analysis and review expands on previous work by examining the effectiveness of LD within higher education, thus identifying whether training characteristics shown to affect organizational outcomes can generalize to student populations. In doing so, we are also able to characterize the nature of LD research in this context, identifying and synthesizing important areas for future research and practice.

The present work comprehensively identifies the state of LD programs for higher education students. Specifically, the purpose of this paper is threefold: (1) to meta-analytically uncover the design and delivery methods that are best suited to develop students as leaders, (2) to provide a summative review of the design and delivery methods most commonly used in student LD programs, and (3) to assess the state of higher education LD program evaluation studies. Together, these objectives serve an overarching goal of demonstrating whether LD programs are implementing the most effective strategies, where any discrepancies between effectiveness and popularity of use suggest a gap between management science and higher education practice. The identification of these lapses in translation are key, as scholars have called for the rapprochement between the sciences of leadership and training to increase LD effectiveness (Barling, Christie, & Hoption, 2010).

Overall, the current investigation offers several contributions to the literature. First, we provide a meta-analytic evaluation of LD programs over a wide span of years (1951 to 2018), focusing exclusively on higher education programs for students. Second, we use updated meta-analytic techniques that account for different types of primary study designs (i.e., repeated measures, independent groups, independent groups with repeated measures; Morris & DeShon, 2002). Third, we supplement our current meta-analytic findings with a qualitative review to provide further insight into our samples and to get a better idea of the state of the field.

Lastly, and perhaps most importantly, we explain the state of the science of LD program evaluations. Specifically, we discuss endogeneity issues in LD program evaluations that potentially harm the science and provide suggestions on how to overcome these concerns. As a starting point, we provide our meta-analysis as an example of quality issues of primary studies. We provide recommendations for researchers conducting evaluation studies to proactively address this issue in future research, noting that the reduction of endogeneity concerns is of utmost importance to advance the science of leadership development in education.

Outcomes of leadership development programs in higher education

The most widely used training evaluation framework by Kirkpatrick (1959) identifies four types of desired outcomes that a training program may aim to accomplish: trainee reactions, learning, transfer of training, and results. Trainee reactions include the trainee's opinion of the attractiveness of the program as well as perceived utility. Learning refers to an increase in the trainee's level of knowledge pertaining to a specific KSA that was targeted in the training (i.e., the “can do” outcome). Transfer of training, or behavior, is how well the trainee took what he or she learned and applied it to the workplace (i.e., the “will do” outcome). Results of training include bottom-line organizational outcomes, such as amount of sales or other financial outcomes and subordinate outcomes (e.g., turnover).

Within the LD literature, other meta-analyses have used this framework to examine the impact that LD programs have on producing desirable outcomes (e.g., Burke & Day, 1986; Lacerenza et al., 2017); the current investigation follows suit. Specifically, we are interested in testing whether the same conditions apply in LD programs conducted in higher education. We define LD programs in higher education as any program that has been systematically designed to enhance leader KSAs and other components (Day, 2000) for either undergraduate or graduate students. However, in educational contexts, the results criterion is less relevant (i.e., there are no financial outcomes to assess and only one primary sample we are aware of assessed this criterion; Benischek, 1996).

Reactions constitute a particularly important outcome type because they serve as a precursor to trainee learning and may indicate how motivated or interested the trainee is in receiving training (Hughes et al., 2016). Previous research indicates that trainees generally have positive reactions following training, universally (Brown, 2005), potentially due to perceptions of training as a form of support (Dugan & Komives, 2007; Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008). However, a limited number of evaluation studies provide pre-versus post-test or control versus treatment data for reactions.[1]

Similar to reactions, learning (Kirkpatrick, 1956, 1967) has also been found to increase as a natural function of training at large (Hughes et al., 2016). It can be divided into the following categories: affective (i.e., attitudinal change), cognitive (i.e., acquired knowledge change), and skill-based (i.e., acquired technical or motor skill change; Kraiger, Ford, & Salas, 1993). Learning from training is posited to occur for several reasons. In the context of healthcare team training, Hughes et al. (2016) suggests that such increases are due to the perceived importance of the skills being targeted; this value signaling can motivate trainees to ensure they acquire knowledge during training. In the case of LD programs, trainees may be aware of the widespread importance of the types of interpersonal skills (e.g., communication; Hogan & Warrenfeltz, 2003) generally targeted with this form of training; this may consequently motivate trainees to engage in processes conducive to learning.

Drawing from adult learning theory, training can change preexisting ideas or assumptions about the world to produce a change in knowledge, which can explain the positive impact LD programs have on learning (Mezirow & Taylor, 2009). Similarly, a link between expectations and behaviors has been established in the greater social and organizational psychology literature (Ajzen, 1985; Armitage & Conner, 2001). Jernigan (2004) argues that in the context of general education, students may be expecting to learn as a function of being provided educational materials, predisposing them to acquire KSAs. Indeed, higher education contexts inherently prioritize and evaluate learning. For example, Fullerton (2010) required students to self-assess the degree to which they understood leadership competencies such as delegation before and immediately after a LD program. Many studies in higher education and elsewhere use similar testing methods to assess the extent to which students have gained new knowledge.

Finally, since the overarching goal of LD programs is to produce

[1] Due to a small number of primary samples, we were unable to meta-analytically test the effect of LD programs on reactions and result outcomes ($k = 2$ and 1, respectively). They were appropriately excluded as specific criteria of investigation, but were included in an overall evaluation criterion.

positive changes in behavior on-the-job or in applied situations, it is unsurprising that previous evidence indicates a positive effect of training on transfer. Indeed, LD programs have been found to increase transfer among organizational employees (Lacerenza et al., 2017). Transfer has also been assessed in higher education contexts. For example, Muyia and Kacirek (2009) demonstrated transfer measurement by administering a self-report measure one year after training and asking students to rate themselves on competencies such as adaptability. Training-related enhancement in skills may be due to programs' bolstering of cognitive resources available to the trainees on-the-job (Hughes et al., 2016). In other words, Hughes et al. (2016) posit that the skills taught to trainees during training allow them to better cope with the demands of the job. For example, LD programs often enhance communication skills (Hogan & Warrenfeltz, 2003), an intrinsic requirement of many jobs. Consequently, trainees may heavily rely on using this newly-trained skill to cope with other job demands and increase overall performance (Payne, 2005; Pincus, 1986), thus allowing its transfer to the job and enhancement of leadership effectiveness. Within the context of education, these skills may be instrumental in successfully collaborating with other students and completing assignments (Schulz, 2008), leading to a higher reliance on the newly trained abilities and, ultimately, improved skills. Based on the aforementioned evidence, we hypothesize the following:

Hypothesis 1. Leadership development programs have a positive effect on trainee learning outcomes (H1a) and transfer (H1b).

Moderators of leadership development program effectiveness

As previously mentioned, several investigations have pointed to the significance of moderators in explaining leadership training effectiveness (e.g., Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Lacerenza et al., 2017). Given that these moderators are theoretically and empirically derived (e.g., Salas et al., 2012), we herein test those most applicable to the education context. Similar to training programs within other contexts, there are several design and delivery characteristics that may specifically impact LD program outcomes (e.g., Arthur, Bennett, Edens, & Bell, 2003; Baldwin & Ford, 1988). Hypotheses developed from the extant training, learning, and leadership sciences are discussed in the following section, and the relationships tested are depicted in Fig. 1.

Voluntary and mandatory training

Higher education LD programs can be voluntary programs that provide students with the option to attend (e.g., an optional leadership workshop open to all students) or mandatory programs that require students to participate as part of class activities. Whether an individual decides to participate in training or is required to participate has been found to influence motivation (Hicks & Klimoski, 1987). Specifically, Hicks and Klimoski (1987) found that perceived pressure to attend training decreased trainee motivation. Self-determination theory (Ryan & Deci, 2000) may be able to explain this occurrence, as it proposes that autonomy fosters motivation; thus, when students are given the option to participate in training, this need for autonomy is fulfilled (Cohen, 1990).

Training theory highlights the importance of trainee motivation in facilitating outcomes (Baldwin & Ford, 1988). A key finding from the training literature is the importance of motivation, or the decision to pursue some behaviors over others (Tsai & Tai, 2003), in increasing training effectiveness (Blume, Ford, Baldwin, & Huang, 2010; Curado, Henriques, & Ribeiro, 2015; Noe & Schmitt, 1986). Motivation can serve as a buffer against criticism and lack of reinforcement, ultimately leading to greater use of the trained KSAs on-the-job or during class (Noe & Schmitt, 1986; Steers & Porter, 1975). In accordance with this theory and evidence, Curado et al. (2015) found that voluntary training programs were associated with a higher motivation to transfer than mandatory programs. As motivation has been meta-analytically linked with higher rates of transfer (Blume et al., 2010), we expect voluntary programs to be more effective than mandatory programs. Given this link between motivation and voluntary training, we hypothesize:

Hypothesis 2. Voluntary leadership development programs enhance trainee learning (H2a) and transfer (H2b) outcomes to a greater degree than involuntary programs.

Timing structure of training delivery

The typical timing structure for a higher education course is fairly stable, with students meeting weekly over the course of multiple weeks. Similar to the distinction between a lump sum payment of money versus an annuity, this training structure contrasts with “massed” schedules wherein students are exposed to all course materials in a single sitting. The argument for the former approach, referred to as a *spacing effect* or *technique* (Hintzman, 1974), is rooted in learning efficiency theories. For example, cognitive load theory (e.g., van Merrienboer & Sweller, 2005) suggests that an individual's working memory capacity is limited; thus, effective learning occurs only if this mental bandwidth has not been overwhelmed. By temporally spacing out training sessions, learning is more likely to occur because cognitive load is less likely to be exhausted (Janiszewski, Noel, & Sawyer, 2003; Lee & Genovese, 1988). As such, we argue for the positive effects of spaced training programs within the higher education context and hypothesize the following:

Hypothesis 3. Leadership development programs spanning multiple

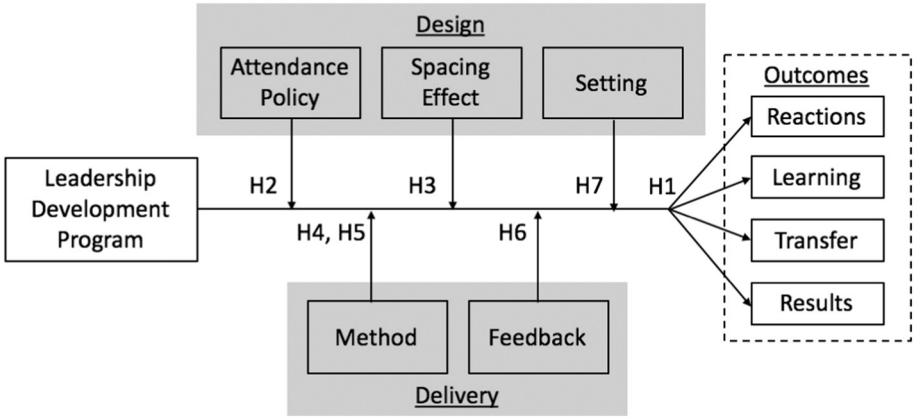

Fig. 1. The design and delivery of leadership development programs. This figure illustrates the relationships tested within the current meta-analytic investigation.

training sessions result in greater effects on learning (H3a) and transfer (H3b) outcomes compared to training programs with one massed training session.

Training program delivery methods

Training programs can leverage the following three primary delivery methods: information, demonstration, and practice. Information-based training provides the trainee with knowledge on a topic and can convey this through mediums such as the instructor providing lectures, presentations, and reading materials. Demonstration-based training offers trainees examples to which they can relate; this may involve watching real-life situations on video or in person. Lastly, practice-based training gives trainees an opportunity to perform what is being taught, including activities such as roleplaying, on-the-job training, in-basket exercises, and simulations.

Of all three methods, theory and evidence suggest that practice is the most effective method for fostering skills (Burke & Day, 1986; Weaver, Rosen, Salas, Baum, & King, 2010). Constructivist learning theory (Piaget, 1952) argues that learning is enriched when the individual reflects and develops understanding and knowledge through their own experiences; in other words, learning by doing. Practice allows trainees to enact needed behavioral skills in a safe environment; in the case of LD, this is especially critical because the majority of such skills relate to interpersonal interaction (e.g., building relationships, communication, team building; Hogan & Warrenfeltz, 2003). We argue that practice, such as roleplaying interpersonal interactions in the classroom, will facilitate these skills more effectively than passively receiving information about them as in the case of information or demonstration (Garavaglia, 1993). In practice, trainees can engage in conversations and scenarios that mirror real-world situations requiring specific interpersonal skills, thus allowing them to build the needed skills. In accordance with this rationale, we hypothesize:

Hypothesis 4. Leadership development programs incorporating only a practice-based method lead to greater effects on trainee learning (H4a) and transfer (H4b) outcomes compared to programs incorporating only information- or demonstration-based methods.

Although one training method may prove to be more effective than the others, the general training literature includes over 30 years of research that suggests training can benefit from using a combination of all three methods (Salas et al., 2015; Salas & Cannon-Bowers, 2001). Given that each has its unique advantages and drawbacks, using all three methods could be a more holistic approach to training. Information can provide trainees with the knowledge and understanding to precede practice opportunities (e.g., Birnbaum, 1984). Demonstration can supplement information by presenting a visible representation and relatable examples of the newly learned knowledge (Salas & Cannon-Bowers, 2000). Finally, practice provides the trainees an opportunity to apply what they learned in a non-threatening environment, such as practicing with other students in order to enhance the learning experience (McCauley & Van Velsor, 2004; Piaget, 1952). Therefore, we hypothesize:

Hypothesis 5. Leadership development programs incorporating information-, demonstration-, and practice-based methods demonstrate greater learning (H5a) and transfer (H5b) effects on trainee outcomes compared to programs implementing only one (e.g., information only) or two methods (e.g., demonstration and information).

Feedback

Feedback theory states that feedback should address both successes and failures to maintain and adjust behavior, respectively (Kluger & DeNisi, 1996); this input can be particularly useful during formative years. Training literature supports the use of feedback because it provides trainees with a better understanding of their ability level and what specific aspects of their performance need to be improved (Brown, Bransford, Ferrara, & Campione, 1983; Komaki, Heinzmann, & Lawson, 1980; Salas et al., 2015). Moreover, it relates the course material to the student in a personalized manner (Hounsell, 2003). Conversely, in the absence of feedback, a student may have a more difficult time gauging his/her ability level and how the material relates to him/herself, consequently hindering any necessary revisions in behavior and reducing the transfer of learning (Ford, Smith, Weissbein, Gully, & Salas, 1998). Notably, some approaches to delivering feedback may work better than others, such as providing both positive and negative feedback as opposed to solely negative feedback (Ellis & Davidi, 2005). All of this considered, we posit that delivering feedback is more effective than not using it. As such, we hypothesize:

Hypothesis 6. Leadership development programs reporting the use of feedback display a greater effect on trainee learning (H6a) and transfer (H6b) outcomes compared to programs that do not report the use of feedback.

Online and face-to-face training

Online training has become a prevalent approach to providing instruction within education (c.f., Clarke, 2004). As such, this self-administered modality (as compared to traditional face-to-face instruction) has particular influence and relevance to students receiving LD in today's educational context. However, online training programs possess features that may make them less effective than face-to-face instruction. First, self-administered, online training may be less adaptive to trainee reactions in real-time, compared to in-person formats (Gerbaud, Gouranton, & Arnaldi, 2009). For example, instruction provided face-to-face is facilitated by a live trainer, who can alter the training content as needed to ensure and maintain the optimal level of difficulty for engagement (Magerko, Wray, Holt, & Stensrud, 2005). Conversely, online platforms are more likely to possess rigid parameters of adaptation, including being pre-programmed with specific content (Appana, 2008). Thus, although technology is becoming more sophisticated, current online training generally does not achieve the same level of flexibility and responsiveness as in-person approaches.

A second concern centers around the depth of education achieved by virtual programs. Trainers have criticized online instruction because they do not believe it conveys difficult teaching and learning problems (Conlon, 1997). Alternatively, in-person techniques such as lectures have evinced positive outcomes (Arthur et al., 2003). This may be attributable to the fact that live facilitators can provide more specific guidance, adapt material, and provide a customized experience to trainees, which can increase engagement and lead to enhanced learning (Merriam, 2001).

Finally, researchers have identified a number of barriers that hinder effective online education, including the quickly-evolving nature of technology, the complexity of networked systems, the lack of stability in online environments, and the limited understanding of student and trainer preparation needed to use such technology (Brandt, 1996). Indeed, not all courses can effectively move from the classroom to computers (Appana, 2008). In-person instruction inherently does not suffer from the same technical issues that can plague online education platforms. In considering each of these factors, we thus hypothesize:

Hypothesis 7. Face-to-face leadership development programs with live facilitators increase positive trainee learning (H7a) and transfer (H7b) outcomes to a greater degree than online, self-administered programs.

Research questions

In addition to these meta-analytic hypotheses, we also pose a number of exploratory research questions that we address through a descriptive review of these samples. These lines of inquiry are designed to better characterize the LD programs being developed, implemented, and evaluated today.

First, Kraiger et al. (1993) note that learning outcomes are multidimensional, indicating that learning can manifest itself through changes in cognitive, affective, or skill capacities. Cognitive learning includes verbal knowledge, knowledge organization, and cognitive strategies. In terms of LD, a training program can train individuals how to develop a leader mental model, which is how leaders view themselves, others, and the environment; this is helpful for shaping behavior, adapting quickly, and forming heuristics for future leadership experiences (Day et al., 2009). Skill-based outcomes include skill compilation and automaticity. Finally, attitudinal outcomes include motivational disposition, self-efficacy, and goal setting; these outcomes also map onto intrapersonal skills that Hogan and Warrenfeltz's (2003) domain model identifies as the earliest stage of development for leaders. Given the lack of theory regarding exactly what types of outcomes are most beneficial to train at the student level, we cannot assert what student LD programs should focus on or evaluate. However, we pose the following exploratory question to understand what occurs most in practice:

Research Question 1: What type of learning outcomes are measured most often in higher education LD programs?

There are a few methods that are commonly used to evaluate training: self-report, observer ratings, objective scores (e.g., declarative knowledge test), and peer evaluations. Evaluations can be completed before and after the training using a repeated measures design. Alternatively, one evaluation can be collected from a trained group and compared with results from a control group (i.e., independent groups design). Some measurement sources are less accurate than others (Wexley & Latham, 2002). In particular, self-report approaches may lead to inflated scores, whereas objective reports are less prone to issues of bias and dishonesty (Blume et al., 2010). The training literature suggests using multiple methods in order to compensate for the strengths and weaknesses of each method (Salas et al., 2015). In practice, we hope to see limited reliance on self-report assessments and, rather, a combination of multiple approaches, which leads us to ask:

Research Question 2: How are the outcomes evaluated most often in higher education LD programs?

More can be revealed about the effectiveness of training depending on the timing of the evaluation (Phillips, 2012). Although it is easier to evaluate training immediately after the training is complete (Phillips, 2012), this can only explain how much was learned from the program. Including a delay between the training and evaluation or sustaining measurement longitudinally can demonstrate how much knowledge was actually retained from the program and provide insight into transfer (Baldwin & Ford, 1988). In practice, an evidence-based approach would involve the implementation of evaluations at multiple time points, including long after training is completed. Therefore, we ask:

Research Question 3: When are the outcomes typically evaluated in higher education LD programs?

Although our meta-analytic investigation can determine which of the three main delivery methods is most effective, there are many specific instructional strategies within each delivery method that can be used. For example, a practice-based strategy known as roleplay allows trainees to act out a scenario related to leadership. Another practice-based strategy is the use of case studies, which are detailed examinations of scenarios that trainees read to identify and solve a problem. Information-based approaches include lectures, reading materials, and discussion. Demonstration-based strategies include watching video examples and observing live re-enactments. Given the multitude of possible approaches, we would like to know the specific instructional strategies that are being used in practice and how many strategies a single LD program typically uses. This leads to our final exploratory question:

Research Question 4: What specific instructional strategies are used most often in higher education LD programs?

Method

Literature search and inclusion criteria

We employed several approaches to identify relevant articles. To begin, studies were collected through an electronic search of the following databases: PsycINFO (1886-August 2018), Business Source Premiere (1886-August 2018), and ProQuest Dissertations and Theses (1886-August 2018). Although the search dated back to 1886, the earliest primary study was published in 1955 (Barnlund, 1955). The following search terms were used to facilitate these searches: *leadership*, *leader*, *manag** (the use of the asterisk denotes a wildcard operator, which returns articles with keywords beginning with the root, e.g., manager), *executive*, *supervisory*, *training*, and *development*. We also included searches adding the terms: *charisma**, *transformational*, *authentic*, *ideological*, and *pragmatic* in order to include any relevant training evaluations related to main leadership styles covered in the literature (Day et al., 2014; Gooty, Connelly, Griffith, & Gupta, 2010; Griffith, Connelly, Thiel, & Johnson, 2015). Finally, we reviewed the reference lists from relevant meta-analyses to identify additional articles (Arthur et al., 2003; Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Keith & Frese, 2008; Lacerenza et al., 2017; Powell & Yalcin, 2010; Taylor, Russ-Eft, & Taylor, 2009).

To be considered appropriate for inclusion, the articles had to meet the following criteria: (a) trainee participants were students in higher education (i.e., undergraduate or graduate students); (b) the study included an empirical assessment of a leadership, leader, managerial, supervisory, or executive training (i.e., development or coaching) program; (c) the study employed a repeated measures, independent groups, or an independent groups with repeated measures design; (d) the study included an adult sample (i.e., over 18 years of age, which excludes K-12 education contexts); (e) the article was written in English; and (f) the article provided sample size and effect size information, or enough information to calculate an effect size.

This resulted in a final sample of 73 independent samples with 5654 participants; 56 undergraduate samples, 12 graduate samples, and five samples that were mixed. The programs lasted anywhere between 30 min to three semesters. Samples reported the length of the program differently (i.e., hours, weeks, months, and semesters). Using as much information as we could draw from the samples, we provide the ranges of time spent in training based on how the length was reported. A total of 27 (37.0%) samples reported duration in minutes and hours, ranging in length from 30 min to 45 h. The average for such programs was 18.44 h. Duration was reported in weeks for 20 (27.4%) programs and these ranged between one week to 18 weeks with an average of 10.7 weeks. Length of training was reported in months for 3 (4.1%) training programs, which were three, four, and six months long. Additionally, nine (12.3%) programs reported duration in school semesters, six of which were a semester long, and three programs that were three semesters long. Lastly, 14 (19.2%) programs did not report training duration.

Coding procedures

We extracted information relevant to the following for both our frequency and meta-analytic analyses: (a) outcome type, (b) attendance policy (i.e., voluntary vs. mandatory), (c) spacing effect (i.e., massed vs. spaced sessions), (d) delivery methods (i.e., information, demonstration, practice, and feedback), (e) training setting (i.e., online vs. face-to-face programs), and (f) training evaluation (e.g., how the outcomes

were collected). Training evaluation could be obtained through self-report, objective scores (e.g., declarative knowledge tests), or observer ratings. Regarding outcome type, we used the framework presented by Kirkpatrick (1959). This classifies evaluations into the following categories: (a) reactions (i.e., the extent to which trainees enjoyed or perceived the training as useful relative to how they initially believed they would enjoy or perceive it as useful), (b) learning (i.e., the extent to which trainees acquired new KSAs due to training), (c) transfer (i.e., the extent to which trained KSAs are demonstrated on the job), and (d) results (i.e., the extent to which the training impacted organizational outcomes). As previously mentioned, due to a small number of primary samples, we were unable to meta-analytically test the effect of LD programs on reactions and result outcomes.

Each article included was double-coded by two of four authors, who had all undergone rater training. Inter-coder agreement, calculated as the percentage of training features coded in agreement out of 9601 decisions, was 94.0%. To account for the possibility of agreement occurring by chance, Cohen's kappa was calculated for all variables with binary categories. These nine variables ranged from kappas of 0.83 to 0.99, with an average of 0.90. Any discrepancies were resolved through discussion. Coded information from the primary samples is presented in Table 1.

Analyses

For the meta-analytic methods, the effect size used was a Cohen's *d*. If a Cohen's *d* was not directly reported, it was calculated from other statistics when possible (e.g., means and standard deviations, *t*-values). If multiple, non-independent effect sizes were reported, we used the intercorrelations among the measures to create a linear composite as described by Nunnally (1978). Similar to procedures used in previous meta-analyses (e.g., DeChurch & Mesmer-Magnus, 2010; Litwiller, Snyder, Taylor, & Steele, 2017), if intercorrelations were not reported, we averaged the effect sizes.

As noted, three types of study designs were included within our analyses (i.e., repeated measures, independent groups, and independent groups with repeated measures). It is necessary to convert effect sizes from multiple study designs to a single study design effect size because the estimated population parameters are influenced by the design type of each study (c.f., Ray & Shadish, 1996). We used the procedures described by Morris and DeShon (2002) to convert the effect sizes into a common metric, the repeated measures effect size. Before making this adjustment, we tested the meta-analytic effect sizes to assess whether they differed as a function of design type and found no statistically significant differences. Therefore, we proceeded to convert all effect sizes into the repeated measures effect size.

The equations we applied are reported in Morris and DeShon (2002). To apply these procedures, it is necessary to calculate $r_{pre.post}$ (i.e., the correlation between pre- and post-training scores). Given that few primary samples reported this metric, we followed recommendations from Morris and Deshon (2002) to use the inverse sampling error variance-weighted average $r_{pre.post}$ across repeated measures samples. The $r_{pre.post}$ was 0.46. Following these procedures, we used a random effects meta-analysis, as suggested by Morris and DeShon (2002), that weights effect sizes by the reciprocal of the sampling variance to account for sampling error variance (Hedges & Olkin, 1985). We also followed procedures outlined by Hunter and Schmidt (1990, 2004) to correct for criterion-related unreliability.

To test for significant moderators, we used *t*-tests of the mean effect sizes (c.f., Hunter & Schmidt, 2004). Per recommendations from Hunter and Schmidt (2004), we corrected for unreliability in the effect sizes using artifact distributions which were created from averaging the internal consistency estimates reported in the primary samples. The mean reliability of an overall criterion for use in all moderator analyses and the overall effectiveness analysis was 0.93. For the analyses examining effectiveness for learning and transfer outcomes, the mean reliabilities were 0.88 and 0.95, respectively.

Results

Table 2 presents the results of the overall meta-analytic *d* effect size combined across evaluation types (i.e., overall). Tables 3 and 4 list meta-analytic results for each moderator category for learning and transfer outcomes, respectively. We report both the corrected average *d* value (corrected for unreliability in the criterion and sampling error variance; Hunter & Schmidt, 2004) and the observed *d* value. Statistical significance of the effect size was tested by interpreting the 95% confidence interval.

To assess for publication bias, we conducted a trim and fill analysis based on procedures identified by Duval and Tweedie (2000). Results from a fixed effects model on the overall effect suggest that zero samples were imputed to the left of the mean, indicating that publication bias is most likely not present. We also supplemented the trim and fill analysis with an a priori selection model based on procedures from Vevea and Woods (2005), which also indicated that publication bias likely did not occur. No outliers were revealed using the one-sample removed analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009).

We found full support for Hypotheses 1a and b as results suggest LD programs within the student (i.e., undergraduate and graduate) context are effective at producing learning (corrected $d = 0.50$, 95% CI [0.37, 0.63]) and transfer (corrected $d = 0.36$, 95% CI [0.16, 0.56]) outcomes.

We found mixed support for Hypothesis 2, which stated that voluntary LD programs would lead to greater outcomes compared to involuntary LD programs ($t(55) = 2.95$, $p < .05$). Voluntary programs substantially improved trainee learning more than involuntary programs ($t(45) = 2.74$, $p < .05$), but did not reliably increase transfer more than involuntary programs ($t(12) = 1.38$, $p > .05$).

We did not find support for Hypothesis 3, which predicted that programs with temporally spaced training sessions would lead to greater outcomes compared to those implementing a single massed training session; for example, learning was not differentially affected by temporal design ($t(52) = 1.46$, $p > .05$) and there were not enough primary samples reporting no spacing ($k = 1$) to test the relationship for transfer. Furthermore, out of the 52 samples who reported using spaced training, 6 (11.5%) provided at least a day between sessions, 3 (5.8%) spaced out sessions at least a week apart, and 43 (82.7%) LD programs did not provide details regarding the duration of the spacing that occurred.

We did not find support for our hypotheses concerning delivery method(s). That is, findings for Hypothesis 4 did not support greater outcomes of solely practice-based methods as opposed to information-only methods, both in terms of learning ($t(12) = 0.73$, $p > .05$) and transfer ($t(3) = 1.04$, $p > .05$). Notably, there were not enough primary samples to test the relationship with demonstration methods ($k = 1$). Hypothesis 5 considered the quantity of methods used, anticipating that programs that combined all approaches would be superior to other single and paired methods. This was not found to be true in the case of learning outcomes, when comparing against information only ($t(19) = 0.02$, $p > .05$), practice only ($t(15) = 0.82$, $p > .05$), and information- and practice-based methods ($t(37) = 0.85$, $p > .05$). Although the difference in sizes comparing a combined information, demonstration, and practice approach with information-only ($t(5) = 1.55$, $p > .05$) and practice-only ($t(4) = 0.47$, $p > .05$) were not statistically significant in predicting transfer outcomes, results trended in the hypothesized direction.

Results were similar for Hypothesis 6, which suggested that programs incorporating the use of feedback would be more effective than those that did not. Results trended in the hypothesized direction, but were not statistically significant for both learning ($t(64) = 1.58$, $p > .05$) and transfer ($t(13) = 0.94$, $p > .05$) outcomes.

Hypothesis 7 was not supported. The meta-analytic effect size for face-to-face programs was not reliably different compared to that of

Table 1
Primary sample codes.

Author(s)	Publication status[a]	Design[b]	N1	N2	α	Attendance policy[c]	Delivery method(s)[d]	Feedback[e]	Setting[f]	Dependent variable(s)	d_{RM}
Antonakis, J., Fenley, M., & Liechti, S. (2011)	P	RM	41	41		V	I, D, P	Yes	F	Transfer	1.57
Barnlund, D. C. (1955)	P	IG	41	41	0.96		I, D, P		F	Learning	0.28
Benischek, S. A. (1996)	U	RM, IG	43	46	0.91	V			F	Transfer, results	0.03
Bradley, M. J. (1994)	U	RM	86	86	0.93	V	I, P		F	Learning	0.36
Brown, N. R. P. (2015)	U	RM	15	15		V	D		F	Learning	0.11
Bruck, J. L. (1997)	U	RM	61	61	0.92		I		F	Learning	0.16
Buschlen, E. (2009)	U	RM	108	108		V	I, D, P		F	Transfer	0.38
Cajiao, J., & Burke, M. J. (2016)	P	RM	246	246	0.84	V			F	Learning	−0.05
Chang, S., Bhat, C. S., & Chen, Y. (2017) Sample A	P	RM	31	31		V	I, P		F	Learning	0.41
Chang, S., Bhat, C. S., & Chen, Y. (2017) Sample B	P	RM	46	46		V	I, P	Yes	F	Learning	0.29
Christensen, M. A. (2015)	U	RM, IG	50	73		V	I, P		F	Learning	0.75
Cressman, K. K. (2005)	U	RM	28	28			I, D, P	Yes	F	Learning	0.15
Downs, C. W. (1974) Sample A	P	RM	49	49					F	Reactions, learning	0.39
Downs, C. W. (1974) Sample B	U	RM	73	73			D, P		F	Reactions, learning	0.35
Eddy, C. L. (2012) Sample A	U	RM, IG	19	23		V	I		F	Learning	−0.03
Eddy, C. L. (2012) Sample B	U	RM, IG	37	23		V	I		F	Learning	−0.41
Eddy, C. L. (2012) Sample C	U	RM, IG	26	23		V	I		F	Learning	−0.29
Eddy, C. L. (2012) Sample D	U	RM, IG	23	23		V	I		F	Learning	−0.06
Endress, W. L. (2000)	U	RM, IG	77	93	0.99	V	I, P		F	Learning	−0.06
Ericksen, K. S. (2009)	U	RM	12	12						Learning, transfer	0.32
Facca-Miess, T. M. (2015)	P	IG	158	283		M	I, P		F	Learning	0.17
Farrell, N. A. K. (2003)	U	RM	45	46		M	I, P		Mix	Learning	0.40
Fiedler, F. E., & Mahar, L. (1979)	P	IG	11	33		V	I			Transfer	0.32
Fields, A. R. (2010)	U	RM	15	15	0.95	V	I, P		F	Learning	0.64
Friedman, S. D., & Westring, A. (2015)	P	RM	242	242		V	P		Mix	Learning, transfer	0.70
Fullerton, J. R. (2010)	U	RM	5	5		V	I, D, P		F	Learning	0.50
Gabriel, J. (2015)	U	RM	39	106		V				Learning	−0.14
Goertzen, B. J., & Whitaker, B. L. (2015) Sample A	P	RM	55	118	0.94	V	I, P	Yes	F	Learning	0.25
Goertzen, B. J., & Whitaker, B. L. (2015) Sample B	P	RM	72	145	0.94	V	I, P	Yes	V	Learning	0.37
Goertzen, B. J., & Whitaker, B. L. (2015) Sample C	P	RM	83	490	0.94	V	I, P	Yes	Mix	Learning	0.12
Gonzalez, K. A. (2016) Sample A	U	RM	29	29	0.99	V	I, P		F	Learning	0.42
Gonzalez, K. A. (2016) Sample B	U	RM	51	51	0.99	V	I, P		F	Learning	0.16
Gonzalez, K. A. (2016) Sample C	U	RM	26	26	0.99	V	I, P		F	Learning	0.38
Grantham, S., Pidano, A. E., & Whitcomb, J. M. (2014) Sample A	P	RM	17	17		V	I, P		F	Learning	0.33
Grantham, S., Pidano, A. E., & Whitcomb, J. M. (2014) Sample B	P	RM	17	17		V	I, P		F	Learning	1.02
Hamdani, M. R. (2018)	P	RM	34	34		M	I, D, P	Yes	F	Learning	0.35
Harvill, R., West, J., Jacobs, E. E., & Masson, R. L. (1985)	P	RM, IG	26	10					F	Learning	1.25
Heft, M., & Deni, R. (1984)	P	IG	24	24		M	I		F	Learning	0.52
Horowitz, R. (2012) Sample A	U	RM	100	100	0.95	V				Transfer	0.28
Horowitz, R. (2012) Sample B	U	RM	104	104	0.95	V				Transfer	−0.12
Jeong, K., & Bozkurt, I. (2014)	P	RM	47	47		M	P		V	Learning	0.53
Kruml, S. M., & Yockey, M. D. (2011)	P	RM	78	78			I, D, P	Yes		Learning	0.39
Larsen, J. A. (1997) Sample A	U	RM, IG	120	104		V	I, P		V	Learning	1.20
Larsen, J. A. (1997) Sample B	U	RM, IG	118	104		V	I, P		V	Learning	0.88
Lehnert, A. B. (2009) Sample A	U	RM	48	48	0.95		I, P	Yes	V	Learning	1.22
Lehnert, A. B. (2009) Sample B	U	RM	47	47	0.95		P		V	Learning	−0.39
Litt, S. D. (2010) Sample A	U	RM	32	34	0.79	M	I	Yes		Transfer	0.15
Litt, S. D. (2010) Sample B	U	RM	25	29	0.79	M	I	Yes		Transfer	0.18
Livingston, R. E. (2003)	U	RM	19	21			I, D		F	Learning	0.09
Matsos, C. T. (1997)	U	RM, IG	35	35		V	I		F	Learning	0.35
McCormick, M. J. (1999)	U	IG	140	140	0.89	V	I, D, P	Yes	F	Learning	2.11
McEnrue, M. P., Groves, K. S., & Shen, W. (2009)	P	RM, IG	75	60	0.95		I, D, P	Yes	F	Learning	0.33
Midgett, A., Hausheer, R., & Doumas, D. M. (2016)	P	RM	20	20	0.96	M	I, P		F	Learning	0.39
Miscenko, D., Guenter, H., & Day, D. V. (2017)	P	RM	98	98		M	I, D, P		F	Learning	0
Muyia, H. M., & Kacirek, K. (2009)	P	RM	43	112		V	I, D, P	Yes		Transfer	0.02
Newstrom, J. W. (1971) Sample A	P	RM	24	24			I, P		F	Learning	0.57
Newstrom, J. W. (1971) Sample B	P	RM	21	21			I, P		F	Learning	0.48
Putman (1992)	U	RM	192	192	0.95	V	P		V	Transfer	0.05
Radnitzer, K. D. (2010)	U	RM	11	11		V			F	Transfer	−0.02

(*continued on next page*)

Table 1 (*continued*)

Author(s)	Publication status[a]	Design[b]	N1	N2	α	Attendance policy[c]	Delivery method(s)[d]	Feedback[e]	Setting[f]	Dependent variable(s)	d_{RM}
Rohs, F. R. (1999)	P	RM	30	30	0.98		I, P		F	Learning	0.72
Rosch, D. M., & Stephens, C. M. (2017)	P	RM	226	226		V			F	Learning	0.54
Sadler, T. D. (2015)	U	RM, IG	147	147	0.91	M	I, D, P		F	Learning	0.10
Sampl, J., Maran, T., & Furtner, M. R. (2017)	P	RM, IG	39	41		V	I, P		F	Learning, transfer	0.77
Sidor, S. M. (2007)	U	RM	91	91		V	P	Yes	V	Learning	0.19
Siewiorek, A. et al. (2013)	P	RM	8	8		V	P		F	Learning	−0.08
Singleton, T. M. (1978)	P	RM	34	34		V	I, D, P	Yes	F	Learning	0.39
Stover, S. H. (1988) Sample A	U	RM	38	38		V	I		F	Learning	0.78
Stover, S. H. (1988) Sample B	U	RM	48	48		V	I		F	Learning	2.39
Stover, S. H. (1988) Sample C	U	RM	35	35		V	I, D, P	Yes	F	Learning	2.49
Teckchandani, A., & Schultz, F. C. (2014)	P	RM	93	93		V	I, P		F	Learning	0.50
Towler, A. J. (2003)	P	IG	14	13		V	I, D, P	Yes	F	Learning, transfer	0.88
Wilcox, B. (2004)	U	RM	92	92		V	I, P		F	Learning	0.89
Zwikael, O., Shtub, A., & Chih, Y. (2015)	P	RM	42	42		V	I, P		V	Learning	0.32

[a] Publication is published (P) or unpublished (U).
[b] Design coded as repeated measure (RM) and/or independent groups (IG).
[c] Attendance policy coded as voluntary (V) or mandatory (M).
[d] Delivery method coded as information (I), demonstration (D), and/or practice (P), or a combination.
[e] Feedback implemented (yes).
[f] Setting coded as face-to-face (F) and/or virtual (V).

Table 2
Meta-analytic results: overall.

Variable	*k*	*N*	*d*	Corrected *d*	SD	%Var	95% CI	
							LL	UL
Overall	73	5734	0.42	0.46	0.55	3.21	0.34	0.58
Published	33	2590	0.38	0.41	0.32	9.59	0.30	0.51
Unpublished	40	3144	0.47	0.51	0.71	1.75	0.30	0.72
Study design								
Repeated measures	53	3172	0.39	0.42	0.48	5.67	0.29	0.54
Independent groups	6	922	0.85	0.92	0.94	0.07	0.22	1.63
Independent groups and repeated measures	14	1640	0.43	0.47	0.52	2.37	0.21	0.72

Note. *k* = number of independent samples; *N* = sample size; *d* = repeated measures Cohen's *d*; SD = corrected standard deviation; %Var = percent of variance accounted for by sampling error variance; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 3
Meta-analytic results: learning.

Variable	*k*	*N*	*d*	Corrected *d*	SD	%Var	95% CI	
							LL	UL
Overall learning	62	5016	0.47	0.50	0.56	2.63	0.37	0.63
Attendance								
Voluntary	39	3204	0.56	0.60	0.63	1.58	0.41	0.79
Involuntary	8	1027	0.21	0.23	0.17	22.24	0.10	0.35
Spacing effect								
Yes	43	3777	0.49	0.52	0.58	2.14	0.36	0.96
No	11	499	0.31	0.32	0.25	24.49	0.16	0.49
Delivery method								
Information	9	462	0.70	0.75	0.99	0.58	0.13	1.36
Demonstration	1	15	0.11	0.12	0	–	0.12	0.12
Practice	5	435	0.46	0.49	0.37	5.45	0.17	0.81
Information and demonstration	1	19	0.09	0.09	0	–	0.09	0.09
Information and practice	27	2232	0.49	0.52	0.35	6.11	0.39	0.65
Demonstration and practice	1	73	0.24	0.26	0	–	0.26	0.26
Information, demonstration, and practice	12	1130	0.70	0.74	0.91	0.38	0.25	1.23
Feedback								
Yes	14	1046	0.69	0.74	0.80	0.64	0.34	1.14
No	48	3970	0.42	0.44	0.46	4.24	0.32	0.57
Setting								
Virtual	8	793	0.52	0.55	0.51	2.30	0.22	0.89
Face to face	48	3682	0.46	0.49	0.62	2.32	0.32	0.66
Mixed	3	370	0.55	0.58	0.26	6.41	0.30	0.86

Note. *k* = number of independent samples; *N* = sample size; *d* = repeated measures Cohen's *d*; SD = corrected standard deviation; %Var = percent of variance accounted for by sampling error variance; CI = confidence interval; LL = lower limit; UL = upper limit.

online programs when considering both learning ($t(54) = 0.30$, $p > .05$) and transfer ($t(13) = 1.79$, $p > .05$) outcomes.

Answers to research questions

Considering RQ1, relevant to both learning and transfer criteria, we specified whether the outcome was cognitive, affective, or skill-based using Kraiger et al.'s (1993) classification schemes. Table 5 lists the frequency and percentages of samples that reported each category of outcome. The majority of our samples measured skill-based outcomes ($k = 31$, 43.1%), followed by affective outcomes ($k = 15$, 20.8%), and the fewest samples solely measured cognitive outcomes ($k = 5$, 6.9%). There were 21 (29.2%) samples that measured a combination of outcome types, as depicted in Table 5.

Regarding how the outcomes were evaluated (RQ2), 59 (80.8%) samples used self-report methods. Only one (1.4%) sample used peer-ratings; four (5.5%) samples used observers and three (4.1%) used objective reports. There were six (8.2%) samples that used multiple methods: two used self-report and observer ratings, two that used self-report and objective ratings, one used self-report and peer ratings, and one used self-report, objective, and observer ratings.

Observing RQ3, there were 57 (78%) samples that specified collecting outcome data immediately after training, eight (11%) that gathered their outcome data after a delay, and eight (11%) that were unclear. Of the delayed outcomes, they ranged from two weeks to a one year delay, with an average of 16.72 days.

RQ4 aimed to uncover the instructional strategies used in LD programs. Within the three main instructional strategies, there was a wide

Table 4
Meta-analytic results: transfer.

Variable	*k*	*N*	*d*	Corrected *d*	SD	%Var	95% CI	
							LL	UL
Overall transfer	15	1150	0.32	0.36	0.44	6.59	0.16	0.56
Attendance								
Voluntary	12	1081	0.33	0.37	0.46	5.20	0.14	0.61
Involuntary	2	57	0.16	0.19	0	100	0.16	0.21
Spacing effect								
Yes	11	875	0.38	0.44	0.48	5.00	0.18	0.69
No	1	27	0.81	0.94	0	–	0.94	0.94
Delivery method								
Information	3	101	0.19	0.21	0	100	0.15	0.28
Demonstration	–	–	–	–	–	–	–	–
Practice	2	434	0.40	0.46	0.35	3.33	0.03	0.90
Information and demonstration	–	–	–	–	–	–		–
Information and practice	1	80	−0.61	0.70	0	–	−0.70	−0.70
Demonstration and practice	–	–	–	–	–	–	–	–
Information, demonstration, and practice	4	219	0.56	0.64	0.60	3.05	0.12	1.17
Feedback								
Yes	5	168	0.55	0.63	0.73	3.48	0.05	1.21
No	10	982	0.28	0.32	0.35	8.38	0.11	0.52
Setting								
Virtual	1	192	0.05	0.06	0	–	0.06	0.06
Face to face	6	356	0.47	0.54	0.70	2.67	0.03	1.04
Mixed	1	242	0.68	0.79	0	–	0.79	0.79

Note. *k* = number of independent samples; *N* = sample size; *d* = repeated measures Cohen's *d*; SD = corrected standard deviation; %Var = percent of variance accounted for by sampling error variance; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 5
Types of learning outcomes evaluated by samples.

	Frequency of samples
Cognitive	5 (6.9%)
Affective	15 (20.8%)
Skill	31 (43.1%)
Cognitive and affective	9 (12.5%)
Cognitive and skill	0 (0%)
Affective and skill	10 (13.9%)
Cognitive, affective, and skill	2 (2.8%)

Note. *k* = 72 samples because one sample did not fit Kraiger, Ford, and Salas' (1993) classification scheme.

Table 6
Frequency of instructional strategies used in LD programs.

Strategy	Frequency of samples
Information-based	
Lecture	44 (69%)
Discussion	40 (63%)
Reading/text-based materials	24 (38%)
Demonstration-based	
Videos/films/audio	12 (19%)
Practice-based	
Project-based work/exercises	30 (47%)
Case studies	14 (22%)
Self-reflection	22 (34%)
Roleplay	15 (23%)
Setting goals	13 (20%)
Coaching/mentoring	9 (14%)
Problem identification and solving	8 (13%)
Games	4 (6%)
Technology-based simulations	5 (8%)
Outdoor course (e.g., rope course)	2 (3%)
Behavioral modeling	2 (3%)
Action learning	2 (3%)
Leader match	1 (2%)

Note. *k* = 64 because 9 samples did not specify instructional strategies used.

variety of specific approaches used across samples. The number of instructional strategies used ranged from one to ten, averaging four per program. Moreover, nine samples did not specify the approaches used. Given the possibility of many combinations of methods, Table 6 lists strategies used in the samples.[2]

Discussion

The majority of universities offer LD programs to students, thereby potentially providing them the opportunity to better prepare for the future. Given this demonstrable investment, we aimed to identify whether LD programs for students were effective, to understand under what training circumstances they were optimally so, and how higher education LD studies were being evaluated. In doing so, we uncovered both research and practical implications. The practical implications of our findings aim to help training developers recognize what to include in LD programs in higher education. Our research implications build theory on LD and provide recommendations to strengthen future meta-analyses.

Summary of meta-analytic findings

The current meta-analysis complements previous work on LD program effectiveness (Avolio et al., 2009; Burke & Day, 1986; Collins & Holton, 2004; Lacerenza et al., 2017; Powell & Yalcin, 2010). Although comprehensive from a workforce perspective, these existing meta-analyses excluded student samples from their meta-analyses, leaving the effectiveness of LD programs in educational contexts unknown. For example: Avolio et al. (2009) specified type of organization as either profit, not for profit, or military; Burke and Day (1986) solely analyzed managerial/supervisory personnel; Collins and Holton (2004) only used employee samples; and Powell and Yalcin's (2010) meta-analysis focused on private sector organizations. We argue that the effectiveness of LD programs in student populations is an important, separate examination from employee populations for several reasons: (1) student participants have less, or even no, previous leadership experience; (2)

[2] Note that multiple methods could have been used in a single sample.

the content of student LD programs may be more general because they are not job-specific; and (3) the goals of higher education LD programs are likely different from others (e.g., the goal in a higher education program could be to become a better leader and/or attain a leadership position, whereas the goals in organizational LD programs might be to be a better leader, increase follower performance, increase follower job satisfaction, lower follower turnover, etc.). Therefore, the current investigation tests moderators of LD programs to assess the conditions in which higher education LD programs are most effective.

Regarding the effectiveness of LD programs in higher education, our meta-analytic findings suggest that, in our samples, substantial learning occurred (corrected $d = 0.50$ translating to a 19% increase [percent increase is equal to Cohen's U3 - 50; Cohen, 1988]), and transfer also occurred, but to a lesser extent (corrected $d = 0.36$ translating to a 14% increase). This supports the notion that these LD programs improve the extent to which students *can* become better leaders (i.e., by exhibiting changes in learning) more than they improve the extent to which they *will* be better leaders (i.e., by transferring the learned behavior to the workplace). Educators may be paying too much attention to learning outcomes and neglecting to effectively teach students how to transfer their newfound skills. This lack of transfer is not uncommon and referred to throughout the science of training as *the transfer problem* (e.g., Ford & Weissbein, 1997). The greater improvement in learning compared to transfer could potentially be due to teachers creating programs that result in the acquisition of knowledge and skills, but not transfer – which is a separate task entirely. Alternatively, current approaches to measuring transfer within this context may not fully capture the extent to which transfer has occurred. There may be constraints associated with measuring student transfer compared to employee transfer. For example, gathering accurate transfer data after a class has ended and students are no longer in contact with the instructor may be more difficult compared to gathering transfer data from employees that continue to work with the organization that has provided training. Thus, the importance of learning itself should not be disregarded as learning is a necessary step toward transfer (Huang, Blume, Ford, & Baldwin, 2015; Hughes et al., 2016).

We found mixed support regarding moderators of LD program effectiveness for students. Results suggest that voluntary programs are more effective than involuntary programs in fostering learning outcomes, thus supporting training theory which highlights the importance of trainee motivation in facilitating outcomes (Baldwin & Ford, 1988). This could be due to attendees of voluntary programs having more intrinsic motivation to change attitudes and behaviors because they self-select into the program. However, the issue of self-selection in a voluntary program also has a strong chance of leading to inflation in self-report outcomes. We discuss this in greater detail in the upcoming section on recommendations for conducting a LD program evaluation study.

Although results trended in the hypothesized direction, we did not find that programs incorporating multiple delivery methods were significantly more effective than those based on a single delivery method. In regard to LD programs in a more general context (i.e., with employees as trainees), Lacerenza et al. (2017) found the relationship between programs and learning outcomes to be fairly stable across moderator categories; the current results mirror these findings. We encourage LD program developers to continue moving in the direction of using a combination of methods. Though we do not have clear findings, theory and previous research have demonstrated support for this approach (Salas & Cannon-Bowers, 2001). To ensure that the selected methods are actually effective in training leadership skills, developers should evaluate the program and adjust accordingly based on the results.

Our hypothesis in favor of using feedback also showed results trending in the hypothesized direction. Though findings did not demonstrate significant relationships, it is possible that some of the samples, which did not report whether or not they provided feedback, did indeed provide trainees with feedback but neglected to document it. Similarly, many articles that reported the use of feedback did not elaborate on how the feedback was delivered. For example, feedback can be delivered by a single source or it can be provided from multiple individuals who know the trainee from different perspectives for a more holistic view (e.g., trainers, classmates, professors). The latter approach is known as 360-degree feedback (Goldsmith, Lyons, & Freas, 2000). The trainee receives a report with a summary of everyone's ratings and some reports also provide the individual with the average rating for other trainees so that they can better understand what areas need more improvement. Future research should investigate whether 360-degree feedback is more effective than single source feedback and who would be the best source to give comments and criticism.

Interestingly, we did not find a difference in outcomes between online and face-to-face programs. However, previous researchers have suggested that face-to-face training is potentially more impactful than self-administered facilitation because facilitators can make real-time adjustments to fit the needs of the participants (Magerko et al., 2005). Although this should be further investigated, it is promising that we did not find differences, as the future of training is moving toward a more virtual world. If virtual programs can prove to be as effective as face-to-face programs, the benefits of this scalability can enable training to reach many more students. Furthermore, advances in technology can allow online programs to more closely mimic real face-to-face interactions and incorporate additional real-time adjustments and feedback to participants.

Summary of frequency analysis

Our primary goal was to identify whether LD programs are beneficial for students in higher education. Our secondary goal was to move beyond this initial aim by taking a deeper look into the prevalence of certain features of training and the evaluation process; this uncovers additional detail that would have been otherwise ignored in a solely meta-analytic approach. The meta-analysis highlights *what works*; the review reveals *what is used* in practice in more specific detail. For example, our meta-analysis compares the primary, scientifically-based delivery methods (i.e., information-based, demonstration-based, practice-based), whereas the review reveals the exact and exemplary types of practice methods that are being used in these programs.

Promisingly, our meta-analysis found support that LD programs lead to learning. Our review showed that in regard to Kraiger et al.'s (1993) classification scheme of learning outcomes, most programs focus on skill-based learning. This includes communicating, persuading others, setting goals, and problem solving (Bruck, 1997; Kruml & Yockey, 2011; Rohs, 1999). Intuitively, skill-based outcomes are important for training because programs are designed to change behaviors. However, although we cannot judge what type of outcome is most important to evaluate, future research can test cognitive and affective outcomes as well, because affect and cognitions are also important for shaping behaviors (Kahle & Berman, 1979).

Our frequency evaluation of training design, delivery, and implementation characteristics revealed that in practice, LD programs generally use approaches that are convenient and inexpensive rather than rooted in science. We noted this in two main areas: instructional strategies and evaluation. First, we note that lecture and discussion were the predominant instructional strategies used (see Table 6 for specific strategies that were used). We do not discredit the value of these strategies; rather, we encourage researchers and training developers to explore approaches that incorporate more practice (e.g., roleplay, goal setting, games). In doing so, it will be possible to determine exactly which strategies are the most effective and if results mirror training theory and current evidence. Potentially, using more practice-based methods like reflective activities and roleplay could convince students that they can and should incorporate their skills in real scenarios. Also, because many students have yet to hold a professional

Table 7
Endogeneity concerns for higher education leadership development program evaluation and mitigation strategies.

Endogeneity concern	Exemplar study combatting concern
Main concerns	
Sample suffers from self-selection or is non-representative	- Although Sampl et al. (2017) recruited undergraduate students via e-mail, making it a voluntary training program, they conducted a pre-posttest design with a training group versus a control group who later received the training. They used a longitudinal randomized controlled study design by offering a voluntary program via e-mail that was split into two groups: an intervention group or a waiting list group. "Students had the opportunity to sign up for the study by completing a questionnaire, which coexisted as the first measurement point (T1). …After the first assessment point at T1, participants were randomly assigned to either a training group or a control group (waiting list). During a fixed period of 10 weeks, participants of the [training] group received the training, whereas participants of the control group received no training. The control group was informed that groups were divided due to the high attendance and received the training at a later point when the study was finished. After the completion of the training or waiting period, all participants were invited again to participate in a second assessment (T2) by completing the same questionnaires as used in T1. In order to reflect as closely as possible the critical variables, T2 took place during the examination period at the end of the summer term" (p. 1397).
Dependent variables are gathered from a single-method self-reporting	- In a military college setting, students acting as platoon leaders were rated by their superior officer. The raters were unaware of whether the cadets completed a leadership training program (Fiedler & Mahar, 1979). - Antonakis et al. (2011) videotaped MBA students giving a speech before and 6 weeks after leadership training and had independent assessors rate the speeches for leader charisma.
Omitting selection by comparing a treatment group to a non-equivalent group	- Facca-Miess (2015) compared three groups of business students: (a) students enrolled only in a capstone marketing course (CC) which incorporated leadership training as part of the course, (b) students enrolled only in a market research and analysis course (MR), which did not discuss leadership, and (c) students enrolled simultaneously in MR and CC.
Other concerns	
Omitting a regressor, that is, failing to include important control variables when testing the predictive validity of dispositional or behavioral variables	- Sampl, Maran, and Furtner (2017) included effects of time and group and pre-intervention group differences. Antonakis et al. (2011) used speech performance ratings for their outcome variable, so they controlled for length of speech and measured communication skills to control for other learning effects that were not taught.
Omitting fixed effects	- To test whether trainees improved charismatic behaviors, participants delivered a speech before and 6 weeks after training, using the same content and wearing the same attire. "The advantage of using this type of within-subjects design is to determine whether variation in charisma predicted subjective ratings of leader prototypicality and other outcomes beyond participant constant (i.e., fixed) effects" (Antonakis et al., 2011, p. 384).
Not using cluster-robust standard errors in panel data (i.e., multilevel hierarchical or longitudinal)	- Antonakis et al. (2011) used cluster-robust standard errors at the rater level because each rater rated four trainee leaders.
Not correlating disturbances of potentially endogenous regressors in mediation models	- See Antonakis et al. (2011, Study 2) for the equations with correlated disturbances to test endogenous regressors.

Note. The endogeneity concerns are from Antonakis, Bendahan, Jacquart, and Lalive (2010).

leadership position, LD programs can aim to provide students with these experiences so that they have an opportunity to practice their skills in a real-world setting. In all, LD does not stop in the classroom, so neither should LD programming. Second, regarding evaluation, the majority of our samples only used self-report ratings and only collected data immediately after training. We elaborate on the boundaries of this evaluation approach and provide recommendations for researchers conducting LD program evaluation studies later in the paper.

Roadmap for future research

Although the current meta-analysis found LD programs to be effective in increasing learning and transfer, a majority of the constituent samples included in the meta-analysis had endogeneity concerns, pointing to a larger endogeneity bias in the LD literature. However, there are a handful of exemplar studies that combat these issues and demonstrate how a LD program evaluation study should be conducted to mitigate these potential issues, which we would like to highlight. In this section, we describe endogeneity concerns, offer recommendations regarding how to conduct a LD program evaluation study, and provide study examples, shown in Table 7. Then, we offer best practices for future meta-analyses in this area; we suggest a set of minimum inclusion criteria for LD program research upon which policy and program decisions can be based. Finally, we conclude with the limitations of the current meta-analysis and additional future directions.

We believe this meta-analysis can be used as a starting point for discussing key endogeneity issues in LD program studies, as well as highlighting how to resolve these concerns in future research. Antonakis, Bendahan, Jacquart, and Lalive (2010) outline a number of threats to causal inference, three of which are substantial concerns in our meta-analysis: (1) omitted selection by comparing a treatment group to a non-equivalent group (i.e., not using random assignment), (2) self-selection within voluntary programs, and (3) single-method self-reporting. Potential compounding bias can also be caused by multiple endogeneity concerns in a single sample. In the current meta-analysis, out of 57 samples that reported adequate information, 36 had all three issues (63.2%), 14 had two of the issues (24.6%), and 7 (12.3%) had one issue. These three concerns are by no means an exhaustive set of causes of endogeneity bias, and the magnitude and direction of such bias is uninterpretable; therefore, we did not analyze any further. Below, we address how researchers can avoid these issues when conducting a LD evaluation study.

First and foremost, the gold standard is to use and compare a randomly assigned and representative student sample in the LD program to an equivalent group that serves as a control (Antonakis et al., 2010). For example, Facca-Miess (2015) compared business students in three groups: (a) students enrolled only in a capstone marketing course (CC) that incorporated leadership training as part of the course, (b) students enrolled only in a market research and analysis course (MR), which did not discuss leadership, and (c) students enrolled simultaneously in MR and CC. Similarly, Heft and Deni (1984) used a sample from a more general training program and randomly assigned the trainees into two sections without disclosing condition. Section I completed the leadership portion of the training, and Section II completed the pre- and post-

test prior to receiving the leadership training. In both these studies, students' random assignment to these groups established an experimental design, wherein the control group acted as a baseline to isolate the effects of the LD intervention.

In cases where omitting selection is unavoidable, the treatment group can be compared to a similar control group; in these cases, the process should be explicitly modeled to estimate the counterfactual correctly. Antonakis et al. (2010) detail multiple method approaches that allow researchers to make stronger causal claims when random assignment is not achievable. For example, they point to the Heckman type two-step selection model (Heckman, 1979) to predict the variance from the error term due to selection, which can then be removed to correctly estimate the treatment term.

A second main concern in educational settings is that students typically self-select to be a part of the LD program, as opposed to training programs that many work organizations require of their employees. Although volunteering to participate in a LD program can be related to higher intrinsic motivation (Hicks & Klimoski, 1987), it may also inflate self-report data because the trainees enter the program with a bias in favor of the program. To avoid this issue, evaluation studies can use Sampl, Maran, and Furtner's (2017) approach of implementing a randomized controlled design. Sampl et al. (2017) offered the training program via e-mail to students and then split the interested students into two subgroups: a training group and a waitlist group, which served as the control group. Both groups completed a pretest as part of the intake form and a post-test after the training was facilitated to the first group. The waitlist-control group was then given the training after this initial period. During this process, it is ideal to have substantial proximal separation, such that the groups are unaware of each other (Athanasopoulou & Dopson, 2018). This waitlist-control alternative not only establishes a control group, but also allows educational administrators to provide all students with the opportunity to gain valuable LD experiences. Notably, self-selection is an inherent part of the vast majority of LD programs in higher education – nearly all students self-select into LD programs (even self-selecting into a degree program that requires an LD course is still self-selection) and thus, self-selection is less of an endogeneity "threat" in the educational context and instead, it may be important to use the estimates we provide in the current meta-analysis as evidence of the effectiveness of these programs within active self-selection contexts.

Third, we note that the majority of the samples only evaluated learning outcomes using self-report methods. Self-report is typically saturated with self-serving and social desirability biases (Fisher, 1993). To help mediate these biases, researchers should take care in using high-quality instruments. One strategy is to use psychological separation of content areas when designing the instrument (Athanasopoulou & Dopson, 2018; Podsakoff, MacKenzie, & Podsakoff, 2012). By adding items to the survey that are unrelated to the LD program, it makes it less apparent to the participants that it is LD program-specific. Another technique is to minimize concerns over the purpose of the assessment, in turn potentially reducing social desirability biases when completing the survey. For example, in Heft and Deni's (1984) sample, the facilitator told the students that the questionnaires were "given to me by the School of Business Administration" so that students would not think they were related to the course.

Additionally, researchers should consider mixed methods to generate more insightful findings and enhance the robustness of the study design (Athanasopoulou & Dopson, 2018). Researchers should leverage multiple sources of data when evaluating LD program impact in order to develop a more holistic view (Salas et al., 2015). For example, trainees can be evaluated by the trainers, peers, and themselves. In their military student sample, Fiedler and Mahar (1979) used a composite rating from cadet superiors, cadet peers, and supervising officers to assess leadership performance (i.e., transfer criterion). In other school settings, the composite score can combine ratings from the facilitator, course peers, and even supervisors from extracurricular activities with which the student is involved (e.g., sports team, volunteer position, internship). This suggestion aligns with research demonstrating that others' ratings (rather than self-report) can be more valid predictors of outcomes such as competence and performance (Atkins & Wood, 2002; Greguras & Robie, 1998). Triangulating data sources can thus powerfully augment the accuracy of LD program evaluation.

As opposed to solely relying on trainees to self-report their perceived learning, outcomes should also be measured using more objective approaches. For example, participants can be given a declarative knowledge test as a measure of learning. Researchers may also consider carefully-designed observational methods, which can provide more useful insight and robust measurement. Antonakis, Fenley, and Liechti (2011) serve as an exemplar study in this regard. They asked MBA students to give a speech before and six weeks after leadership training, using independent assessors to rate each speech for markers of leader charisma. To help control for a number of factors, they required that participants provide the same content and wear the same attire at both sessions. This approach helped "determine whether variation in charisma predicted subjective ratings of leader prototypicality and other outcomes beyond participant constant (i.e., fixed) effects" (Antonakis et al., 2011, p. 384). They also controlled for length of speech and measured communication skills to control for other learning effects that were not taught. As Antonakis et al. (2011) demonstrated, LD program evaluations should account for fixed effects and include important control variables when testing the predictive validity of behavioral variables. These measures can also help alleviate other endogeneity concerns in this area of research (as outlined in Table 7).

Furthermore, future studies are needed that include and evaluate other Kirkpatrick (1959) evaluation criteria, namely, reactions, transfer, and results. For example, LD programs completed for course credit likely collected student evaluations, which could easily be used to measure reactions. Evaluating programs holistically would enable more accurate investigations involving the effect of training design, delivery, and implementation characteristics on all outcomes. Additionally, evaluating outcomes after time has passed can determine whether the learned material has been retained and whether the effectiveness of LD programs degrades over time. Temporal delays in evaluation can also provide the researcher with objective data on performance (e.g., transfer) that would not otherwise be obtainable immediately post-training (e.g., students often do not have a performance episode to display transfer until well after the training is over).

Finally, a simple yet imperative practice in conducting LD program evaluation studies is to include as much detail as possible when reporting program design and delivery. This helps other training developers model new programs after effective LD programs. Not only does complete reporting help administrators and trainers use best practices in the field, but it also supports future scholars as they investigate research in this area. For example, a meta-analysis can only make assumptions based on the details provided in each article, in which case there may be undeterminable and unaccounted-for information. When documenting any design or delivery method used in the study, we suggest acknowledging *who*, *what*, *when*, *where*, and *how* each process was accomplished (e.g., explicitly stating that feedback was written down by the instructor and other peers during a practice activity and given to trainees immediately afterward). In order for the science of LD to grow, researchers must ensure that they comprehensively document necessary program information.

Ultimately, in order for the science of LD to grow, future research on LD programs should adhere to the following:

- *Use a comparison group, especially for voluntary programs*
- *Use a comparison group when outcomes are only self-report*
- *Demonstrate how the comparison group is equivalent to the treatment group*
- *Include control variables in design*
- *Include potential confounding variables in analysis*

- *Ensure that any self-reported learning outcome is difficult to fake.*

We hope this list enhances the quality of next-generation LD research, and can subsequently be used as inclusion criteria in a next-generation LD meta-analysis, upon which policy and program decisions can be based.

Limitations and additional future directions

These recommendations draw attention to limitations of the present effort. The first limitation is the lack of details available from the samples included within our study. As previously stated, researchers should carefully detail all aspects of their program such that meta-analyses can investigate as much detail as is provided in the original samples. For example, regarding spacing of sessions, researchers should report more information so we can examine not just whether spacing is important but what spacing (e.g., length of spacing) is best to maximize effectiveness. Another limitation was the small sample size that prevented some of our hypotheses from being tested. Additionally, though we did not find publication bias, it is possible that researchers have not documented unsuccessful LD programs, therefore limiting the amount of accessible data on ineffective programs (however, a large proportion of our samples were unpublished dissertations). Finally, although we consider the discussion of endogeneity concerns as a contribution to the literature, the issue does lend itself as a limitation in our ability to make causal inferences from our meta-analysis. It should also be noted that although the endogeneity-plagued results may be similar to those of samples with a clear causal design, this does not mean that the endogeneity-plagued results should be used to inform policy. Therefore, we encourage caution when interpreting the results, but hope that this study can serve as a guide for future research on the topic.

Given the limited availability of evaluation studies on LD programs in higher education, our study was not able to assess several moderators that would be helpful to investigate in the future. First, it seems that only brief LD programs have been empirically evaluated. There could be degree programs (i.e., multiple courses) offered to students that would be beneficial to evaluate. Also, future research should continue to evaluate the effectiveness of both modes of training (i.e., face-to-face and online education) and consider the effectiveness of blended learning, which combines both face-to-face and online education (Driscoll, 2002). As technology advances, there may be interesting shifts in modes of training and how participants engage, given that the new generation has been found to prefer technology more than previous generations (Frand, 2000).

Conclusion

Our results suggest that LD programs in higher education work in the studies examined– both learning and transfer increased as a result of these LD programs. However, the samples identified within our meta-analysis also pointed to a concern that appears to be common within this area of literature—endogeneity bias. To this end, we offer a roadmap for future evaluation studies to more effectively address endogeneity concerns. In practice, it appears that LD programs that are being used in education have been following guidelines from scientific research (e.g., the spacing principle, using multiple delivery methods), but there is still room for improvement (e.g., providing feedback, measuring outcomes using a triangulation approach to measurement). Other design, delivery, and implementation elements need further research specific to student leadership development. We hope that our findings can guide the future development of LD programs and their evaluation design.

Acknowledgements

This work was supported in part by contracts NNX16AP96G and NNX16AB08G with National Aeronautics and Space (NASA) to Rice University. This work was also supported, in part, by research grants from the Ann and John Doerr Institute for New Leaders at Rice University. We also thank our reviewers for their helpful comments and suggestions.

References

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl, & J. Beckmann (Eds.). *Action control* (pp. 11–39). Berlin, Heidelberg: Springer.

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

*Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education, 10*(3), 374–396. https://doi.org/10.5465/amle.2010.0014.

Appana, S. (2008). A review of benefits and limitations of online learning in the context of the student, the instructor and the tenured faculty. *International Journal on E-learning, 7*(1), 5–22.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British Journal of Social Psychology, 40*(4), 471–499.

Arthur, W., Jr., Bennett, W., Jr., Edens, P. S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*, 234–245.

Arvey, R. D., Rotundo, M., Johnson, W., Zhang, Z., & McGue, M. (2006). The determinants of leadership role occupancy: Genetic and personality factors. *The Leadership Quarterly, 17*, 1–20. https://doi.org/10.1016/j.leaqua.2005.10.009.

Athanasopoulou, A., & Dopson, S. (2018). A systematic review of executive coaching outcomes: Is it the journey or the destination that matters the most? *The Leadership Quarterly, 29*, 70–88.

Atkins, P. W., & Wood, R. E. (2002). Self-versus others' ratings as predictors of assessment centers ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*, 871–904. https://doi.org/10.1111/j.1744-6570.2002.tb00133.x.

Avolio, B. J., Reichard, R. J., Hannah, S. T., Walumbwa, F. O., & Chan, A. (2009). A meta-analytic review of leadership impact research: Experimental and quasi-experimental studies. *The Leadership Quarterly, 20*, 764–784.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63–105. https://doi.org/10.1111/j.1744-6570.1988.tb00632.x.

Barling, J., Christie, A., & Hoption, C. (2010). Leadership. In S. Zedeck (Ed.). *Handbook of Industrial and Organizational Psychology*(pp. 183–240). Washington, DC: American Psychological Association.

*Barnlund, D. C. (1955). A comparative study of individual, majority, and group judgment. *The Journal of Abnormal and Social Psychology, 58*(1), 55–60. https://doi.org/10.1037/h0040823.

*Benischek, S. A. (1996). *An inquiry into the effect of a management program on perceived locus of control orientation and job satisfaction in managers and non-managers (order no. 9635261).* (Available from ProQuest Dissertations & Theses Global. (304251067)).

Birnbaum, M. L. (1984). The integration of didactic and experiential learning in the teaching of group work. *Journal of Education for Social Work, 20*(1), 50–58.

Blume, B., Ford, J., Baldwin, T., & Huang, J. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* (West Sussex, UK).

*Bradley, M. J. (1994). *The effectiveness of leadership training for traditional-aged college students (order no. 9524815).*

Brandt, D. S. (1996, February). *Teaching the net: Innovative techniques in internet training.* (Paper presented at the 11th Annual Computers in Business Conference, Washington, DC).

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell, & E. M. Markman (Vol. Eds.), *Handbook of child psychology*(4th ed.). *Vol. 3. Handbook of child psychology* (pp. 77–166). New York, NY: Wiley.

Brown, K. G. (2005). An examination of the structure and nomological network of trainee reactions: A closer look at "smile sheets". *Journal of Applied Psychology, 90*, 991–1001.

*Brown, N. R. P. (2015). *College student leadership development participation and emotional intelligence (order no. 10036395).* (Available from ProQuest Dissertations & Theses Global. (1775742279)).

*Bruck, J. L. (1997). *The influence of field dependence on college students' leadership attitudes and self-perceptions (order no. 9729166).* (Available from ProQuest Dissertations & Theses Global. (304405150)).

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology, 71*, 232–265.

*Buschlen, E. (2009). *Can college students learn to lead? An examination of a collegiate leadership course using the social change model of leadership.*

*Cajiao, J., & Burke, M. J. (2016). How instructional methods influence skill development in management education. *Academy of Management Learning & Education, 15*(3), 508–524.

*Chang, S., Bhat, C. S., & Chen, Y. (2017). Experiential group training and group leadership development in Taiwanese school counsellor trainees. *Asia Pacific Journal of Counselling and Psychotherapy, 8*(1), 66–85. https://doi.org/10.1080/21507686.2016.1264442.

*Christensen, M. A. (2015). *Growth in leader efficacy based on participation in a college student leadership development program (order no. 10102193).* (Available from

ProQuest Dissertations & Theses Global. (1785846225)).
Clarke, A. (2004). *e-Learning skills.* Basingstoke, UK: Palgrave Macmillan.
Cohen, D. J. (1990). What motivates trainees? *Training and Development Journal, 44*, 91–94.
Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Hillside, NJ: Erlbaum.
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.
Conlon, T. (1997). The internet is not a panacea. *Scottish Educational Review, 29*(1), 30–38.
*Cressman, K. K. (2005). *The influence of spiritual gifts on effectiveness of leadership development among undergraduate college students.* (Available from ProQuest Dissertations & Theses Global).
Curado, C., Henriques, P. L., & Ribeiro, S. (2015). Voluntary or mandatory enrollment in training and the motivation to transfer training. *International Journal of Training and Development, 19*(2), 98–109.
Darling-Hammond, L., LaPointe, M., Meyerson, D., Orr, M. T., & Cohen, C. (2007). *Preparing school leaders for a changing world: Lessons from exemplary leadership development programs. School leadership study. Final report.* Stanford Educational Leadership Institute.
Day, D., Harrison, M. M., & Halpin, S. M. (2009). *An integrative approach to leader development: Connecting adult development, identity, and expertise* (1st ed.). New York, USA: Taylor & Francis.
Day, D. V. (2000). Leadership development: A review in context. *The Leadership Quarterly, 11*, 581–613. https://doi.org/10.1016/S1048-9843(00)00061-8.
Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., & McKee, R. A. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly, 25*(1), 63–82.
DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology, 95*(1), 32–53.
Downs, C. W. (1974). The impact of laboratory training on leadership orientation, values, and self-image. *Communication Education, 23*(3), 197–205.
Driscoll, M. (2002). Blended learning: Let's get beyond the hype. *E-learning, 1*(4), 1–4.
Dugan, J. P., & Komives, S. R. (2007). *Developing leadership capacity in college students.* College Park, MD: National Clearinghouse for Leadership Programs.
Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x.
*Eddy, C. L. (2012). *Leadership training and behavioral adjustments in undergraduates (order no. 3490336).* (Available from ProQuest Dissertations & Theses Global. (917669019)).
Ellis, S., & Davidi, I. (2005). After-event reviews: drawing lessons from successful and failed experience. *Journal of Applied Psychology, 90*(5), 857–871.
*Endress, W. L. (2000). *An exploratory study of college student self-efficacy for relational leadership: The influence of leadership education, cocurricular involvement, and on -campus employment (order no. 9967894).* (Available from ProQuest Dissertations & Theses Global. (304627339)).
Ericksen, K. S. (2009). *An evaluation of the long-term effectiveness of a women's leadership development program.* (Available from ProQuest Dissertations & Theses Global. (220008796)).
*Facca-Miess, T. M. (2015). Investigating teaching leadership in the capstone marketing course. *Marketing Education Review, 25*(2), 141–157. https://doi.org/10.1080/10528008.2015.1030252.
*Farrell, N. A. K. (2003). *Change agent, mentor, advocate: A study of teachers learning to be leaders (order no. 3079321).* (Available from ProQuest Dissertations & Theses Global. (305240303)).
*Fiedler, F. E., & Mahar, L. (1979). The effectiveness of contingency model training: A review of the validation of leader match. *Personnel Psychology, 32*(1), 45–62.
*Fields, A. R. (2010). *Leadership self-efficacy in university co-curricular programs (order no. 3447772).* (Available from ProQuest Dissertations & Theses Global. (858614938)).
Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2), 303–315. https://doi.org/10.1086/209351.
Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*, 218–233. https://doi.org/10.1037/0021-9010.83.2.218.
Ford, J. K., & Weissbein, D. A. (1997). Transfer of training: An updated review and analysis. *Performance Improvement Quarterly, 10*(2), 22–41.
Frand, J. (2000). The information age mindset: Changes in students and implications for higher education. *Educause Review, 35*(5), 15–24.
*Friedman, S. D., & Westring, A. (2015). Empowering individuals to integrate work and life: Insights for management development. *Journal of Management Development, 34*(3), 299–315. https://doi.org/10.1108/JMD-11-2012-0156.
*Fullerton, J. R. (2010). *Transformative learning in college students: A mixed methods study (order no. 3398454).* (Available from ProQuest Dissertations & Theses Global. (250719387)).
*Gabriel, J. (2015). *Situational leadership awareness development in student outdoor leaders through training versus experience (order no. 3710163).* (Available from ProQuest Dissertations & Theses Global. (1699306455)).
Garavaglia, P. L. (1993). How to ensure transfer of training. *Training & Development, 47*(10), 63–69.
Gerbaud, S., Gouranton, V., & Arnaldi, B. (2009, August). Adaptation in collaborative virtual environments for training. *International conference on technologies for E-learning and digital entertainment* (pp. 316–327). Berlin, Heidelberg: Springer.
*Goertzen, B. J., & Whitaker, B. L. (2015). Development of psychological capital in an academic-based leadership education program. *Journal of Management Development, 34*(7), 773–786. https://doi.org/10.1108/JMD-07-2013-0105.
Goldsmith, M., Lyons, L., & Freas, A. (Eds.). (2000). *Coaching for leadership*. San Francisco, CA: Jossey-Bass.
*Gonzalez, K. A. (2016). *Leadership training for a diverse world: A study of the effectiveness of three interventions with fraternity and sorority leaders (order no. 10103958).* (Available from ProQuest Dissertations & Theses Global. (1790115399)).
Gooty, J., Connelly, S., Griffith, J., & Gupta, A. (2010). Leadership, affect and emotions: A state of the science review. *The Leadership Quarterly, 21*(6), 979–1004.
*Grantham, S., Pidano, A. E., & Whitcomb, J. M. (2014). Female graduate. students' attitudes after leadership training: A case study. *Journal of Leadership Studies, 8*(1), 6–16.
Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*, 960–968. https://doi.org/10.1037/0021-9010.83.6.960.
Griffith, J., Connelly, S., Thiel, C., & Johnson, G. (2015). How outstanding leaders lead with affect: An examination of charismatic, ideological, and pragmatic leaders. *The Leadership Quarterly, 26*(4), 502–517.
*Hamdani, M. R. (2018). Learning how to be a transformational leader through a skill-building, role-play exercise. *International Journal of Management Education (Elsevier Science), 16*(1), 26–36. https://doi.org/10.1016/j.ijme.2017.11.005.
*Harvill, R., West, J., Jacobs, E. E., & Masson, R. L. (1985). Systematic group leader training: Evaluating the effectiveness of the approach. *Journal for Specialists in Group Work, 10*(1), 2–13.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society,* 153–161.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.
*Heft, M., & Deni, R. (1984). Altering preferences for leadership style of men and women undergraduate residence advisors through leadership training. *Psychological Reports, 54*(2), 463–466.
Hicks, W. D., & Klimoski, R. J. (1987). Entry into training programs and its effects on training outcomes: A field experiment. *Academy of Management Journal, 30*(3), 542–552.
Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.). *Theories in cognitive psychology: The Loyola symposium.* Oxford, England: Lawrence Erlbaum.
Hogan, R., & Warrenfeltz, R. (2003). Educating the modern manager. *Academy of Management Learning & Education, 2*, 74–84.
*Horowitz, R. (2012). *An exploratory study of the impact of college student leadership programs upon the construct of mindfulness (order no. 3512267).* (Available from ProQuest Dissertations & Theses Global. (1024287086)).
Hounsell, D. (2003). Student feedback, learning and development. *Higher education and the lifecourse* (pp. 67–78). .
Huang, J. L., Blume, B. D., Ford, K. J., & Baldwin, T. T. (2015). A tale of two transfers: Disentangling maximum and typical transfer and their respective predictors. *Journal of Business and Psychology, 30*(4), 709–732.
Hughes, A. M., Gregory, M. E., Joseph, D. L., Sonesh, S. C., Marlow, S. L., Lacerenza, C., ... Salas, E. (2016). Saving lives: A meta-analysis of team training in healthcare. *Journal of Applied Psychology, 101*, 1266–1304.
Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*, 334–349.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings.* Thousand Oaks, CA: Sage.
Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *The Journal of Consumer Research, 30*, 138–149. https://doi.org/10.1086/374692.
*Jeong, K., & Bozkurt, I. (2014). Evaluating a project management simulation training exercise. *Simulation & Gaming, 45*(2), 183–203. https://doi.org/10.1177/1046878113518481.
Jernigan, C. G. (2004). What do students expect to learn? The role of learner expectancies, beliefs, and attributions for success and failure in student motivation. *Current Issues in Education,* 7–45.
Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780. https://doi.org/10.1037/0021-9010.87.4.765.
Kahle, L. R., & Berman, J. J. (1979). Attitudes cause behaviors: A cross-lagged panel analysis. *Journal of Personality and Social Psychology, 37*(3), 315–321.
Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*, 59–69.
Kirkpatrick, D. (1959). Techniques for evaluating training programs. *Journal of the American Society for Training and Development, 13*, 3–9.
Kirkpatrick, D. L. (1956). How to start an objective evaluation of your training program. *Journal of the American Society of Training Directors, 10*, 18–22.
Kirkpatrick, D. L. (1967). Evaluation of training. In R. Craig, & L. Bittel (Eds.). *Training and development handbook* (pp. 87–98). New York, NY: American Society of Training and Development.
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284. https://doi.org/10.1037/0033-2909.119.2.254.
Komaki, J., Heinzmann, A. T., & Lawson, L. (1980). Effect of training and feedback: Component analysis of a behavioral safety program. *Journal of Applied Psychology, 65*(3), 261–270. https://doi.org/10.1037/0021-9010.65.3.261.
Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal*

of Applied Psychology, 78(2), 311–325.
*Kruml, S. M., & Yockey, M. D. (2011). Developing the emotionally intelligent leader: Instructional issues. *Journal of Leadership and Organizational Studies, 18*, 207–215. https://doi.org/10.1177/1548051810372220.
Lacerenza, C. N., Reyes, D. L., Marlow, S. L., Joseph, D. L., & Salas, E. (2017). Leadership training design, delivery, and implementation: A meta-analysis. *Journal of Applied Psychology, 102*(12), 1686–1718.
*Larsen, J. A. (1997). *Application of cognitive, affective, and behavioral theories to measure learning outcomes in management training (order no. 9724012).* (Available from ProQuest Dissertations & Theses Global. (304369176)).
Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport, 59*, 277–287. https://doi.org/10.1080/02701367.1988.10609373.
*Lehnert, A. B. (2009). *The influence of strengths-based development on leadership practices among undergraduate college students.* (Dissertation Abstracts International).
*Litt, S. D. (2010). *Examining behavioral change among supervision and management undergraduates in a selected college (order no. 3416695).* (Available from ProQuest Dissertations & Theses Global. (742515274)).
Litwiller, B., Snyder, L. A., Taylor, W. D., & Steele, L. M. (2017). The relationship between sleep and work: A meta-analysis. *Journal of Applied Psychology, 102*(4), 682–699.
*Livingston, R. E. (2003). *The impact of leadership education on changes in leadership behaviors: An evaluative study of the effectiveness of leadership education as perceived by students in a police executive leadership program (order no. 3098526).* (Available from ProQuest Dissertations & Theses Global. (305224070)).
Magerko, B., Wray, R. E., Holt, L. S., & Stensrud, B. (2005). Improving interactive training through individualized content and increased engagement. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), 2005* (pp. 1–11). .
*Matsos, C. T. (1997). *Student leadership development as a supplement to college fraternity pledge programs (order no. 9735730).* (Available from ProQuest Dissertations & Theses Global. (304331220)).
McCauley, C. D., & Van Velsor, E. (2004). Our view of leadership development. In K. M. Hannum, J. W. Martineau, & C. Reinelt (Eds.). *Handbook of leadership development* (pp. 1–22). (2nd ed.). San Francisco, CA: Jossey-Bass.
*McCormick, M. J. (1999). *The influence of goal-orientation and sex-role identity on the development of leadership self-efficacy during a training intervention. (order no. 9957488).* (Available from ProQuest Dissertations & Theses Global. (304573730)).
*McEnrue, M. P., Groves, K. S., & Shen, W. (2009). Emotional intelligence development: Leveraging individual characteristics. *Journal of Management Development, 28*(2), 150–174. https://doi.org/10.1108/02621710910932106.
Merriam, S. B. (2001). Andragogy and self-directed learning. *New directions for adult and continuing education. 89. New directions for adult and continuing education* (pp. 3–14).
Mezirow, J., & Taylor, E. (Eds.). (2009). *Transformative learning in action: A handbook of practice*. San Francisco, CA: Jossey-Bass.
*Midgett, A., Hausheer, R., & Doumas, D. M. (2016). Training counseling students to develop group leadership self-efficacy and multicultural competence through service learning. *Journal for Specialists in Group Work, 41*(3), 262–282. https://doi.org/10.1080/01933922.2016.1186765.
*Miscenko, D., Guenter, H., & Day, D. V. (2017). Am I a leader? Examining leader identity development over time. *The Leadership Quarterly*. https://doi.org/10.1016/j.leaqua.2017.01.008.
Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125.
*Muyia, H. M., & Kacirek, K. (2009). An empirical study of a leadership development training program and its impact on emotional intelligence quotient (EQ) scores. *Advances in Development Human Resources, 11*, 703–718.
*Newstrom, J. W. (1971). *The assessment of quality and acceptance dimensions of managerial problems (order no. 7205559).* (Available from ProQuest Dissertations & Theses Global. (302627508)).
Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology, 39*(3), 497–523.
Nunnally, J. C. (1978). *Psychometric theory.* New York, NY: McGraw-Hill.
Payne, H. J. (2005). Reconceptualizing social skills in organizations: Exploring the relationship between communication competence, job performance, and supervisory roles. *Journal of Leadership & Organizational Studies, 11*(2), 63–77.
Phillips, J. J. (2012). *Handbook of training evaluation and measurement methods.* Routledge.
Piaget, J. (1952). *The origins of intelligence in children.* NY: Basic Books.
Pincus, J. D. (1986). Communication satisfaction, job satisfaction, and job performance. *Human Communication Research, 12*(3), 395–419.
Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*(1), 539–569.
Powell, K. S., & Yalcin, S. (2010). Managerial training effectiveness: A meta-analysis 1952–2002. *Personnel Review, 39*, 227–241.
Putman, P. G. (2012). *Virtual simulation in leadership development training: The impact of learning styles and conflict management tactics on adult learner performance.* (Available from ProQuest Information & Learning. Dissertation Abstracts International Section A: Humanities and Social Sciences. (2013-99111-269)).
*Radnitzer, K. D. (2010). In K. D. Radnitzer (Ed.). *Emotional intelligence and self -directed learning readiness among college students participating in a leadership development program (order no. 3411453)* (pp. 2010). (Available from ProQuest Dissertations & Theses Global. (577525209)).
Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology, 64*, 1316–1325.
*Rohs, F. R. (1999). Response shift bias: A problem in evaluating leadership development with self-report pretest-posttest measures. *Journal of Agricultural Education, 40*, 28–37.
*Rosch, D. M., & Stephens, C. M. (2017). Campus involvement as a predictor for durable leadership development in conjunction with leadership program participation. *Journal of College Student Development, 58*(7), 1107–1112. https://doi.org/10.1353/csd.2017.0092.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.
*Sadler, T. D. (2015). *The effect of a leadership training program on consideration of future consequences (order no. 1606021).* (Available from ProQuest Dissertations & Theses Global. (1752512036)).
Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest, 13*, 74–101.
Salas, E., Benishek, L., Coultas, C., Diets, A., Grossman, R., Lazzara, E., & Oglesby, J. (2015). *Team training essentials: A research-based guide.* Routledge.
Salas, E., & Cannon-Bowers, J. A. (2000). Design training systematically. In E. A. Locke (Ed.). *The Blackwell handbook of principles of organizational behavior* (pp. 43–59). Malden, MA: Blackwell.
Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*, 471–499.
*Sampl, J., Maran, T., & Furtner, M. R. (2017). A randomized controlled pilot intervention study of a mindfulness-based self-leadership training (MBSLT) on stress and performance. *Mindfulness, 8*(5), 1393–1407. https://doi.org/10.1007/s12671-017-0715-8.
Schulz, B. (2008). The importance of soft skills: Education beyond academic knowledge. *NAWA: Journal of Language and Communication,* 146–154.
*Sidor, S. M. (2007). *The impact of computer based *simulation training on leadership development (order no. 3276388).* (Available from ProQuest Dissertations & Theses Global. (304744504)).
*Siewiorek, A., Gegenfurtner, A., Lainema, T., Saarinen, E., & Lehtinen, E. (2013). The effects of computer-simulation game training on participants' opinions on leadership styles. *British Journal of Educational Technology, 44*(6), 1012–1035.
*Singleton, T. M. (1978). Managerial motivation development: A study of college student leaders. *Academy of Management Journal, 21*(3), 493–498. https://doi.org/10.2307/255730.
Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93*, 280–295.
Steers, R. M., & Porter, L. W. (Eds.). (1975). *Motivation and work behavior*. New York-McCraw-Hill.
*Stover, S. H. (1988). *An investigation of locus of control and its relationship to leadership training (order no. 8901978).* (Available from ProQuest Dissertations & Theses Global. (303722464)).
Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Transfer of management training from alternative perspectives. *Journal of Applied Psychology, 94*, 104–121.
*Teckchandani, A., & Schultz, F. C. (2014). The vision thing: An experiential exercise introducing the key activities performed by leaders. *Journal of Leadership Studies, 8*(1), 63–70. https://doi.org/10.1002/jls.21315.
*Towler, A. J. (2003). Effects of charismatic influence training on attitudes, behavior, and performance. *Personnel Psychology, 56*(2), 363–381.
Tsai, W., & Tai, W. (2003). Perceived importance as a mediator of the relationship between training assignment and training motivation. *Personnel Review, 32*(2), 151–163.
U.S. News and Report. (2018). National University Ranking. (2018). https://www.usnews.com/best-colleges/rankings/national-universities Retrieved from.
Weaver, S. J., Rosen, M. A., Salas, E., Baum, K. D., & King, H. B. (2010). Integrating the science of team training: Guidelines for continuing education. *The Journal of Continuing Education in the Health Professions, 30*, 208–220. https://doi.org/10.1002/chp.20085.
Wexley, K. N., & Latham, G. P. (2002). *Developing and training human resources in organizations* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
*Wilcox, B. (2004). *Leadership development of community college students: Does participation in the phi theta kappa leadership development studies course have an effect on the development of leadership behaviors? (order no. 3152297).* (Available from ProQuest Dissertations & Theses Global. (305113262)).
*Zwikael, O., Shtub, A., & Chih, Y. (2015). Simulation-based training for project management education: Mind the gap, as one size does not fit all. *Journal of Management in Engineering, 31*(2), 1. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000239.
Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177.
Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428–443.