Citation:

Klonek, F. E., Gerpott, F. H., & Parker, S. K. (2020). A conceptual replication of ambidextrous leadership theory: An experimental approach. The Leadership Quarterly, 101473. https://doi.org/https://doi.org/10.1016/j.leaqua.2020.101473

[1] A postprint is a digital draft of a research journal article after it has been peer reviewed and accepted for publication, but before it has been typeset and formatted by the journal

Abstract

Ambidextrous leadership theory proposes that a leader's interplay between opening behaviors and closing behaviors enhances followers' exploration and exploitation behaviors, which ultimately increases innovative outcomes. Unfortunately, previous research suffers from problems with causal interpretation and endogeneity concerns threatening the validity of the theory. Our aim was to constructively replicate previous research with an experimental design, more rigorous measures, and state-of-the-art data analytical approaches (2SLS). In two randomized experiments (Study 1: $N = 395$, Study 2: $N = 229$), we manipulated four leadership styles (opening, closing, ambidextrous, and transformational leadership) and tested their effects on participants' exploration/exploitation behaviors as well as objective innovation outcomes. We only found partial support for the hypotheses from ambidextrous leadership theory. We discuss implications in terms of refining central concepts of the theory and offering more accurate assumptions about timing. We also elaborate on more general insights from our constructive replication studies for the leadership field.

Keywords: conceptual replication, ambidexterity, exploration, exploitation, innovation

A Conceptual Replication of Ambidextrous Leadership Theory: An Experimental Approach

Innovation—the creation, promotion, and implementation of new ideas (Hughes, Lee, Tian, Newman, & Legood, 2018; West & Farr, 1990)—is pivotal for building and maintaining organizational competitiveness (Artz, Norman, Hatfield, & Cardinal, 2010; Rubera & Kirca, 2012; Tushman & O'Reilly, 1997). Whereas a significant amount of research has emphasized the importance of leadership for motivating followers to achieve innovation success (e.g., Denti & Hemlin, 2012; Hughes et al., 2018; Mumford, Scott, Gaddis, & Strange, 2002; Rosing, Frese, & Bausch, 2011), there is an ongoing scholarly debate concerning the specific set of leader behaviors that are suitable in this process (Mumford et al., 2002; Stock, Zacharias, & Schnellbaecher, 2017). This debate is based on the notion that traditionally studied leadership styles, such as transformational leadership, are too vague to account for the complex requirements of innovation work (Bledow, Frese, Anderson, Erez, & Farr, 2009; Hughes et al., 2018; Rosing, Rosenbusch, & Frese, 2010). In particular, scholars have argued that leaders need to engage in specific behaviors that align with the dynamic nature of the innovation process, which involves both an exploratory search for creative ideas and the exploitative selection and implementation of ideas (Bledow et al., 2009; Rosing et al., 2011).

To address this call for studying leadership behaviors that are theoretically relevant to the innovation process, a recent research stream has focused on the concept of ambidextrous leadership, defined as leaders' complementary engagement in two distinct types of leadership activities (Rosing et al., 2011; for empirical articles testing this theory, see Table 1). Specifically, ambidextrous leaders switch between opening behaviors (i.e., aimed at enhancing variability in follower behaviors) and closing behaviors (i.e., seeking to reduce variability in follower behaviors)

in consideration of the continuously changing task demands (Rosing et al., 2011). On the one hand, leader opening behaviors allow employees to make errors, encourage them to use alternative methods to accomplish their tasks, and motivate them to take risks. That is, opening behaviors stimulate employees to engage in exploratory variance-increasing "search" behaviors (March, 1991; Rosing et al., 2011). On the other hand, leader closing behaviors focus on establishing work routines, goal monitoring, and rule adherence. That is, closing behaviors encourage employees' variance-decreasing "production" behaviors (March, 1991, Rosing et al., 2011). By switching between opening and closing behaviors according to the progress that has been made on a task, leader ambidexterity (i.e., the interaction between leader opening and closing behaviors) is assumed to drive follower ambidexterity, in which followers engage in both exploration and exploitation. This interplay, in which followers engage in both behaviors, is argued to ultimately drive innovation outcomes because innovation requires the development (i.e., exploration) and the implementation (i.e., exploitation) of ideas in equal measure (March, 1991).

This ambidextrous view on leadership for innovation has attracted considerable interest among both scholars and practitioners. To illustrate, there are 723 academic citations (Google Scholar, while Web of Science lists 250)[2] of the work by Rosing et al. (2011). In this article, the authors developed ambidextrous leadership theory after meta-analytically finding that a range of leadership styles (in particular, transformational and transactional leadership) showed heterogeneous associations with innovation. Practitioner journals have also enthusiastically begun to address the topic in articles such as "The ambidextrous CEO" (Tushman, Smith, & Binns, 2011) or "How to become an ambidextrous leader" (Kinni, 2016). Universities even offer programs that claim to train managers to increase their ambidexterity level (e.g., a Certificate in

[2] Retrieved January 2020

Innovative Leadership; David Eccles School of Business, 2018). However, despite the widespread interest, the hypothesized relationships within ambidextrous leadership theory have not been tested by designs that allow causal conclusions. As we discuss in more detail below, scholars have thus far relied on survey-based correlational designs and acknowledged that this approach limits causal interpretation of the results (e.g., Alghamdi, 2018; Zacher et al., 2016; Zacher & Rosing, 2015). Accordingly, these researchers have noted the need to use controlled experimental approaches to test the suggested causal predictions of leadership ambidexterity theory. We respond to this call by conducting a conceptual replication of ambidextrous leadership theory across two experimental studies. This approach involves purposefully adapting previously used methods to test the theory's underlying assumptions (Makel, Plucker, & Hegarty, 2012).

Our research provides two main contributions to the literature: First, from a methodological perspective, we seek to offer a stronger empirical foundation for the necessity of an innovation-specific leadership style, namely ambidextrous leadership. Specifically, the present research uses experimental research designs with external ratings of innovation and an instrumental variable approach and therefore represents an opportunity to conceptually replicate the results of previous studies. Second, on a more general level, we assume that our replication endeavor is important for the advancement of the research field (Köhler & Cortina, 2019). If a theory receives considerable scholarly attention—"as an arbitrary selection, if a publication is cited 100 times" (Makel et al., 2012: 541)—a constructive replication of findings that methodologically improves the approach of previous (and the original) authors becomes important in building theory with stronger confirmatory power (Köhler & Cortina, 2019; Schmidt, 2009). Replications are also important from a practitioner perspective. The popularization of any leadership theory in mainstream media can strongly influence managerial practice, which can be

problematic if the existing evidence for a theory suffers from methodological artifacts, if it cannot be generalized to different populations, and/or if it does not reflect true causal relationships. As others have warned, practitioners tend to rely "on popular ideas and fads without sufficient consideration given to the validity of these ideas" (Zaccaro & Horn, 2003: 779). In the worst case, managers are trained on a leadership style that may be useless or could potentially harm innovation outcomes. Overall, our research helps to assess the causality of previous findings on leadership and innovation and thus ultimately contributes to building a more theoretically integrated and practically relevant leadership field.

In what follows, we first describe the assumptions underlying ambidextrous leadership theory in more detail and explain the rationale of the hypotheses (Figure 1). To establish why a conceptual replication is needed, we then review the methodological shortcomings of existing studies. We subsequently describe the design, measures, and analytical strategies of the present research.

++++++ Insert Figure 1 about here++++++

Theoretical Model of Ambidextrous Leadership for Innovation

Scholars have long sought to understand how organizations can foster employee innovation by studying organizational-level predictors (e.g., Camisón-Zornoza, Lapiedra-Alcamí, Segarra-Cipés, & Boronat-Navarro, 2004) and team-level predictors (e.g., Hülsheger, Anderson, & Salgado, 2009). Particularly, innovation scholars have put forward the idea that the development of innovations involves a set of complementary variance-increasing activities (i.e., exploration of ideas) and variance-decreasing activities (i.e., selection and implementation of ideas) that align with different phases of the innovation process (Hughes et al., 2018; West & Farr, 1990). Recent research has transferred this idea to the interpersonal level. Acknowledging the role of leadership in this process, scholars have suggested that the innovation process can be managed effectively

through the use of a set of complementary ambidextrous leadership behaviors (e.g., Rosing et al., 2011).

Specifically, the development of ambidextrous leadership was motivated by a meta-analysis summarizing associations between innovation and different leadership styles, particularly transformational and transactional leadership (Rosing et al., 2011). In their study, the authors found that the links between innovation and transformational as well as transactional leadership were moderate and varied largely between studies (see Rosing et al., 2011). From this evidence, Rosing et al. (2011) concluded that the existing constructs do not sufficiently capture a leader's focus on increasing and reducing variance in followers' behavior, two aspects that are essential in the innovation process (West & Farr, 1990). Theoretically, adding such a focus should result in a more aligned leadership style that captures the dynamics of innovation tasks such that "a leadership style is positively related to innovation when complemented by another leadership style that focuses on and fosters different aspects of the innovation process" (Rosing et al., 2011: 965). Notably, while other leadership styles, such as transformational or transactional leadership, can also co-occur, they may or may not include the increase or reduction of variance in follower behaviors. As such, these leadership theories do not capture the central element of variability, which is essential for the definition of ambidextrous leadership.

According to the work by Rosing and colleagues (2011), ambidextrous leadership consists of opening behaviors (i.e., activities that increase variance in followers, such as encouraging experimentation and attempts to challenge established approaches or giving room for independent thinking) and closing behaviors (i.e., activities that decrease variance in followers, such as monitoring goal achievement, taking corrective action, or setting specific

guidelines).[3] It is notable that the concept of leadership ambidexterity (i.e., a set of complementary behaviors that trigger respective complementary follower behaviors) is defined by its outcomes. Thus, the clarity of the concept can be criticized in that it does not address what the nature of leadership ambidexterity actually is (MacKenzie, 2003). In other words, leader opening behaviors are *defined* by their consequences (i.e., follower exploration behaviors, e.g., activities such as deviating from routine work, trying out new approaches, or expanding knowledge to complete an innovation task; Alghamdi, 2018; Rosing et al., 2011; Zacher et al., 2016), yet they are also *hypothesized* to stimulate exploration behavior. In a similar way, leader closing behaviors are conceptualized as antecedents of follower exploitation behaviors (e.g., engaging in standardized or routine activities, applying present work knowledge to the task at hand, or focusing on implementing well-defined tasks; Alghamdi, 2018; Rosing et al., 2011; Zacher et al., 2016). Closing behaviors thus reduce variance in follower behavior (and the theory argues that opening is often required in early phases, whereas closing is required in later phases of innovation tasks when ideas need to be implemented). These assumptions concerning the positive association between leader opening and closing behaviors and follower exploration and exploitation behaviors are summarized in the following hypotheses:

H1: Opening leader behaviors positively predict follower explorative behaviors.

H2: Closing leader behaviors positively predict follower exploitative behaviors.

[3] Providing support for the conceptual distinctiveness of ambidextrous leadership, previous research has clearly distinguished leader opening behaviors from transformational leadership and leader closing behaviors from transactional leadership. Specifically, although opening behaviors and transformational leadership are positively correlated (e.g., r = .67** in Zacher et al., 2016; r = .49** in Zacher & Rosing, 2015), CFA have shown that they do not reflect the same underlying construct. Similarly, closing behaviors and transactional behaviors are positively correlated (e.g., r = .48** in Zacher et al., 2016; r = .48** in Zacher & Rosing, 2015), but are factorially distinct.

Two empirical studies have investigated these suggested positive associations between leader and follower behavior (Alghamdi 2018, Zacher et al., 2016[4]). First, Zacher et al. (2016) used a cross-sectional survey design with 388 employees recruited via the online platform MTurk. Participants were asked to think of their leaders and subsequently assess their leaders' opening and closing behaviors, and they were also asked to report on their own exploration and exploitation behaviors at work. Results showed that leader opening behaviors explained a significant amount of variation in follower exploration (supporting H1) and that leader closing behaviors explained a significant variation in follower exploitation (supporting H2). Notably, the authors controlled for transformational leadership, transactional leadership, and employee person-level factors (i.e., openness to experience, conscientiousness, positive trait affect). Second, Algahmdi (2018) used a single-source self-report study with 147 faculty members in Saudi Arabia. Controlling for gender, educational level, and faculty position, the author showed that supervisors' opening behaviors were positively associated with faculty member exploration behavior (supporting H1) and that supervisors' closing behaviors were positively associated with faculty member exploitation behavior (supporting H2). Our study expands on this previous work by ensuring that leader behavior is manipulated, rather than self-reported, as well as taking further steps to test causality, as we elaborate below. It is pertinent to note that, while the theory has proposed that engagement in particular exploitative versus exploration behaviors needs to occur at the right time (that is, aligned with the needs of the task), this aspect of the theory has not been tested in previous research.

[4] We use the word "predict" in our hypotheses to specify causal relationships of the model. However, we acknowledge that previous studies have worded their hypotheses more cautiously and used the term "positively associated" for H1 and H2.

Leader ambidexterity and follower innovation

Together, opening and closing behaviors form the leader ambidexterity construct. Ambidextrous leadership theory states that leaders need to engage in both behaviors in alignment with the requirements of innovation tasks to promote innovation outcomes. Unfortunately, the original theory did not explain precisely when or how these behaviors should be expressed; rather, it more broadly proposed that an innovation task may require different degrees of variability in terms of follower behaviors (i.e., exploration and exploitation). Hence, if leaders encourage followers to exhibit these behaviors at the right point in time, the outcomes of innovation tasks will be improved. Because these complementary behaviors (i.e., opening and closing) are expected to interactively shape innovation, the higher-order construct of ambidextrous leadership is traditionally captured as a multiplicative interaction of these sub-dimensions (Zacher & Wilden, 2014; see also Gibson & Birkinshaw, 2004, for measuring related constructs such as contextual ambidexterity and Mom, Van Den Bosch, Volberda, 2009, for managerial ambidexterity). Stated formally, these arguments translate into the following hypothesis:

> *H3: The interaction of leader opening and closing behaviors (i.e., leader ambidexterity) positively predicts employee innovation insofar that innovation is highest when both leader opening and leader closing behaviors are high.*

Three studies have tested the interaction effect of leader opening and closing behaviors on follower innovation (Alghamdi, 2018; Zacher & Rosing, 2015; Zacher & Wilden, 2014). First, a single-source self-report study by Algahmdi (2018) showed that the multiplicative interaction of leader opening and closing behaviors predicted faculty members' innovation performance above and beyond the main effects of opening and closing behaviors. Second, Zacher and Wilden (2014)

conducted a diary study with 113 employees and found that followers' self-reported daily innovative work behavior was highest when followers perceived both leader opening and closing behaviors to be high on the same day (controlling for their respective main effects, leaders' daily intellectual stimulation, employees' positive trait affect, and general level of job autonomy). Third, Zacher and Rosing (2015) conducted a team-level study by asking 33 team leaders to rate the innovation performance of their teams and asking followers about their respective leaders' ambidextrous behaviors. The findings indicated a significant interaction effect for leader opening and closing behaviors on team innovation (controlling for transformational leadership and general team success). Specifically, leader opening behaviors were only related to team innovation when closing behaviors were also high.

Leader ambidexterity and follower behavior

In more recent models, scholars have extended the original model (e.g., Zacher et al. 2016; see also Rosing & Zacher, 2017) by suggesting that follower ambidexterity (i.e., the interaction of employee exploration and exploitation behaviors) is the more proximal antecedent of follower innovation. That is, instead of focusing on leader opening and closing behaviors and the direct link between these behaviors and employee innovation, these studies focus on the logical consequence of the previously outlined hypotheses, namely that follower exploration and exploitation behaviors should be drivers of employee innovation. Stated formally,

> *H4: The interplay of follower exploration and exploitation (i.e., follower ambidexterity) positively predicts innovation outcomes insofar that innovation is highest when both follower exploration and exploitation behaviors are high.*

The hypothesis that follower exploration and exploitation interactively shape follower innovation has also received support from Zacher et al. (2016), as well as Rosing and Zacher (2017). First, in their cross-sectional study with MTurk participants, Zacher et al. (2016) reported

that self-reported innovation performance (rated by employees) was significantly predicted by self-rated exploration behavior, exploitation behavior, and the multiplicative interaction of these behaviors. The interaction indicated that employee exploitation was more strongly related to innovative performance when employees also showed high levels of exploration behavior. Second, Rosing and Zacher (2017) conducted two diary studies (one weekly and one daily study) that repeatedly asked employees to rate their weekly (Study 1) and daily (Study 2) levels of exploration behavior, exploitation behavior, and innovative work performance. The authors used polynomial regression with response surface analysis, which is an alternative approach to examining the balance between the two conceptually related constructs of follower exploration and exploitation. The results of their studies showed that employees who exhibited high levels of both exploration and exploitation reported the highest levels of innovative work performance (controlling for weekly levels of positive and negative affect, innovation requirements of the job, age, gender, and education; Rosing & Zacher, 2017). Furthermore, the study showed that an imbalance towards follower exploration was less harmful to innovation than an imbalance towards exploitation.

Table 1 summarizes the previous studies on ambidextrous leadership, their design, control variables, and main findings (see also Figure 1 for an illustration of the hypotheses and their empirical support). Next, we summarize the main methodological shortcomings that limit our knowledge regarding the hypothesized relationships within ambidextrous leadership theory. These shortcomings comprise the starting point for our conceptual replication endeavor.

++++++ Insert Table 1 about here++++++

Lack of rigorous tests of ambidextrous leadership theory

To test ambidextrous leadership theory, the majority of empirical studies have used a single source (Alghamdi, 2018; Rosing & Zacher, 2017; Zacher et al., 2016; Zacher & Wilden, 2014)

and/or cross-sectional survey field studies (Alghamdi, 2018; Zacher et al., 2016; Zacher & Rosing, 2015). While field research has the advantage of providing high ecological validity, a challenge is that many of these studies were conducted using cross-sectional designs. That is, these studies measured the independent and dependent variables concurrently, which entails that it is unclear whether ambidextrous leadership fosters follower ambidexterity or whether the extent to which followers engage in ambidextrous behaviors increases their leaders' probability of exhibiting ambidextrous behaviors. Furthermore, participants' answer patterns may be driven by several uncaptured variable(s) that correlate with the modeled variables but are not included in the model. This describes an endogeneity problem due to common-method variance (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Podsakoff, Mackenzie, & Podsakoff, 2012; Podsakoff & Organ, 1986), where the effect of a causal predictor (i.e., leader ambidexterity) and/or mediator (follower ambidexterity) on a dependent variable (innovation) cannot be interpreted because the predictor does not vary randomly and its effect on the dependent variable can be explained by other omitted co-variants (Antonakis et al., 2010). If ambidextrous leader or follower behaviors are endogenous regressors and the analytical methods used in a study do not address problems of endogeneity, then previous findings do not help to fully understand the phenomenon; to put it more bluntly, "finding a relationship between an endogenous regressor x—that has not somehow been purged from endogeneity—and y does not help leadership theory one bit" (Antonakis et al., 2014: 94).

To overcome endogeneity in a model, the independent variable and/or mediator must be exogenous, meaning it must be unaffected by any other variable in the model (Antonakis et al., 2014). That is, there must be no variable that would be an antecedent of innovation and would correlate with ambidexterity (i.e., no omitted variable). However, there are reasons to assume that ambidexterity is predicted by certain unmodeled causes that directly affect employee ambidexterity

and/or innovation performance.[5] For instance, empirical evidence indicates that conscientiousness is negatively associated with ambidexterity (Keller & Weibler, 2015), yet conscientiousness is also a direct antecedent of innovation success (Stock, Van Hippel, & Gillert, 2016).

In addition to endogeneity concerns, the exclusive use of survey designs in leadership ambidexterity research can be criticized for reasons of reactivity (i.e., subjects change their responses because they are sensitized to the construct under investigation; Hill, White, & Wallace, 2014) and for self-serving and retrospective biases (Baumeister, Vohs, & Funder, 2007). In other words, perceptions of behavior can differ considerably from actual objective behavior (Behrendt, Matz, & Göritz et al., 2017). However, it is worth noting that previous researchers are well-aware of these challenges, and they have acknowledged these limitations of self-reports in their respective future research sections (e.g., Zacher et al., 2016; Zacher & Wilden, 2015; Rosing & Zacher, 2017).

To summarize, based on previous research that has provided initial support for the application of ambidextrous leadership theory in the field, we consider the time ripe to conduct a conceptual replication that can help address the limitations of previous survey-based and correlational research designs (Clapp-Smith et al., 2018). As others have argued, once empirical data from non-experimental field research "suggest a moderation effect might be present, investigators must devote the additional time, attention, and resources needed to perform more rigorous experimental (…) designs to reveal the true latent model" (Murphy & Russel, 2017: 558).

[5] Previous studies have included multiple control variables that are conceptually argued to predict both ambidextrous leadership and innovation outcomes (for an overview of all control variables, see Table 1). However, a simple estimating system of equations whereby ambidextrous leadership is modeled as an outcome of these controls does not produce the correct estimates if ambidextrous leadership is endogenous and no two-stages-least-squares estimation is used (Antonakis et al., 2010). In other words, including these constructs as control variables does not solve the endogeneity problem because it does not instrumentalize the ambidextrous leadership construct.

Hence, we sought to replicate the results of previous studies on leadership ambidexterity theory by using an experimental design that has significant "confirmatory power" to corroborate the finding that ambidextrous leadership also works in principle and "in isolation" (i.e., without all the other factors present in field environments). Our approach is thus a conceptual replication (which is defined as a repetition of a test of a hypothesis of previous research work using different methods, that is, by using a different material realization, i.e., research design; Schmidt, 2009). In contrast to a direct (or close) replication – which is the most exact possible duplication of previous research procedures (ideally even in the same lab; Brandt et al., 2014; Schmidt, 2009) and which mainly serves to (re)-produce scientific facts – the main purpose of a conceptual replication is to verify the underlying relationships hypothesized in a model (i.e., to extend knowledge and produce understanding; Schmidt, 2009). Given the methodological challenges that often characterize the original studies, conceptual replications therefore regularly need to use a different material realization (i.e., research design, Schmidt, 2009).

Study 1: Randomized Vignette Experiment

Study 1 aims to replicate previous research (i.e., Alghamdi, 2018; Zacher et al., 2016; Rosing & Zacher, 2017; Zacher & Rosing, 2015; Zacher & Wilden, 2014) by testing the causal effects of leader ambidexterity on follower ambidexterity and innovative performance (rated by subject matter experts) through a randomized experimental study with four conditions (i.e., opening, closing, ambidextrous, and a control condition of transformational leadership). An experimental vignette methodology provides a systematic approach to ensure high internal validity (Aguinis & Bradley, 2014). In particular, we aim to test in Study 1 whether leader opening behaviors enhance follower exploration (H1) and whether leader closing behaviors enhance follower exploitation (H2); both hypotheses are conceptual replications of Alghamdi

(2018) and Zacher et al. (2016). Furthermore, we investigate the interactive effect of leader opening and closing (i.e., ambidexterity) on innovation (H3), which aims to replicate Alghamdi (2018), Zacher and Rosing (2015) and Zacher and Wilden (2014). We also assess the interactive effect of follower ambidexterity on innovation (H4), which is a conceptual replication of Rosing and Zacher (2017) and Zacher et al. (2016).

Methods

The study was approved by the Human Research Ethics Committee (HREC) of Curtin University in Australia (HREC approval number HRE2018-0454) prior to conducting the study. All study data and experimental materials are published in anonymized form on the Open Science Framework (https://osf.io/buvq8/files/).

Sample

We used two complementary approaches to determine sample size using an a priori power analysis: First, we reviewed effect sizes reported in the literature and used G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the required sample size (see Appendix A). Second, we also reviewed the sample sizes of previous research (see Appendix B) and compared them with the results of our estimations.

For H1, reported effect sizes have been r = .41 to .42 for the association between leader opening behavior and follower exploration (i.e., H1, Alghamdi, 2018, Zacher et al., 2016). For H2, effect sizes have been r = .45 to .21 for the link between leader closing behavior and follower exploitation (Alghamdi, 2018, Zacher et al., 2016). For H3, effect sizes have ranged between f^2 = .052 (Algahmdi, 2018) to f^2 = .14 (Zacher & Rosing, 2015) for the link between

leader ambidexterity and follower innovation. For H4, effect sizes have been $f^2 = .018$ for the link between follower ambidexterity and follower innovation (Zacher et al., 2016).[6]

We calculated the necessary sample sizes for each of the four hypotheses separately (using $\alpha = .05$ and a high power of $\beta = .95$ based on recommendations from the replication literature; Open Science Collaboration, 2012). Furthermore, when more than one reported effect size was available, we relied on the smaller value (e.g., we used $r = .21$, not $r = .45$ as an effect size measure for H2) to have the most conservative estimate. For H1, the necessary sample size to detect an effect size of $r = .41$ is $N = 71$. For H2, the necessary sample size to detect an effect size of $r = .21$ is $N = 289$. For H3, the necessary sample size to detect a moderation effect $f^2 = .052$ is $N = 252$. For H4, the necessary sample size to detect a moderation effect $f^2 = .05$ is $N = 360$. We considered that some of the participants may not complete the task or may not pass attention check items. Hence, we planned to over-recruit by 10%, resulting in a required sample of $N = 396$ persons for the experiment (hence $n = 100$ per condition).

We also used an alternative approach for determining sample sizes for replication research (Brandt et al., 2014). This approach suggests taking 2.5 times the sample size reported in the original study (Simonsohn, 2013). To obtain this estimate, we calculated the average sample size for those studies that have tested previous hypotheses on ambidextrous leadership theory (i.e., all of the studies listed in Table 1). The average sample from these studies is $N = 129.5$; multiplied by 2.5, this results in a sample size of $N = 324$.

Our final sample of $N = 395$ participants falls in between the suggested sample size determined by the two described sample size calculation approaches. Participants had a mean age

[6] We could not calculate a local effect size f^2 based on the information provided in Rosing and Zacher (2017) and Zacher and Wilden (2014). These studies used more complex hierarchical data structures (i.e., diary studies with observations nested within subjects), which do not allow a straightforward estimation of effect sizes.

of 36 years (*SD* = 9.48), and 44% were female. Participants worked on average 42.3 hours/week and worked in different industries, such as education (13%), information technology (13%), health services (10.4%), manufacturing (8.9%), government (7.1%), business consulting (5.6%), finance (4.1%), human services (3.5%), industry services (3.5%), hospitality (3.3%), building and construction (3.3%), accounting (2.5%), legal (2.3%), military/police/security (1%), and agribusiness (0.8%). On average, participants' work experience was 17.55 years (*SD* = 9.6).

Recruitment

We collected data from working professionals via TurkPrime (Litman, Robinson, & Abberbock, 2017) and prolific.ac[7] (Palan & Schitter, 2018), two versatile crowdsourcing data acquisition platforms for researchers that allowed us to recruit participants using pre-screening requirements (Litman et al., 2017; Keith, Tay, & Harms, 2017). Research comparing the validity of results from participants tested via different channels (i.e., face-to-face laboratory settings, social media posts, and platforms such as TurkPrime or MTurk) indicates that crowdsourcing data can reach equivalent—and occasionally even superior—quality when compared to data collected in in-person settings should researchers make careful decisions about the design and applied analytical procedures (Casler, Bickel, & Hackett, 2013, Cheung, Burns, Sinclair, & Sliter, 2017; Thomas & Clifford, 2017). Therefore, our study was only accessible for participants who had an approval rate of ≥ 90%, and we included multiple measures of attention/ comprehension checks to ensure high data quality (i.e., screener items; cf. Thomas & Clifford, 2017).

[7] We ensured that participants who had both an account with prolific and MTurk could not participate twice in the experiment.

To allow for better comparisons with populations from previous research, we carefully reviewed the sample information (see Table 1, Appendix B) and based our pre-selection criteria on this information. In terms of recruitment, the majority of previous research has conducted online surveys (i.e., Rosing & Zacher, 2017; Zacher et al., 2016; Zacher & Wilden, 2014), and one study used a sample of MTurk participants (Zacher et al., 2016). In terms of nationality, previous research used samples from Australia (Zacher & Rosing, 2015; Rosing & Zacher, 2017), the US (Zacher et al., 2016), Germany (Rosing & Zacher), and Saudi Arabia (Alghamdi, 2018). In terms of industries, previous research has surveyed participants working in public universities (Alghamdi, 2018), creative industries (Rosing & Zacher, 2017), architecture and design firms (Zacher & Rosing, 2015), and a "range of industries" (Rosing & Zacher, 2017), including participants working in non-creative industries, such as administration officers and cleaners (Zacher & Wilden, 2014).

We applied multiple procedures to select an appropriate subject pool that (a) aligned with previous study populations and (b) was adequate in terms of providing high-quality responses. First, we used the platforms' integrated pre-screening procedures, which made our study available only to a pre-defined population of participants (for MTurk, pre-screening criteria were occupational status: employee, hours employed: 35+ hours/week, nationality: AU and US, approval rate: 90%+, number of HITs approved: 100+; for prolific, screening criteria were nationality: AU and US, supervisor: yes, employment status: full-time, and first language: English). Using these criteria, the available participant pool in MTurk was $N < 480$,[8] while the available participant pool in prolific was $N = 1{,}054$.

[8] After specifying these criteria, TurkPrime generated a warning regarding the study's feasibility stating that "your study might not get all the participants that you need." The actual number of available participants was much smaller

We also separated the assessment of the personality and trait affect measures from the main experiment by using a prescreening procedure (for details, see Keith et al., 2017). That is, a week prior to the experiment, participants were invited to complete a survey in which they filled out items on their personality, trait affect, and demographics. In addition, using TurkPrime and prolific's pre-screening procedures, we also asked participants to indicate whether (1) they worked full-time (i.e., 35 hours or more), whether (2) English was their first language, and whether (3) they had a direct supervisor. Furthermore, the first survey included (4) one item screening for bots ("M0st hum4ns ar3 pret7y g0od at read1ng t3xt w1th som3 letters replac3d by numbers. So, wh4t 1s th3 f1rst let7er 0f the m0nth bef0r3 M4rch?") and (5) an English language comprehension test [see Appendix C] to identify potential low-quality respondents (i.e., participants who use server-farms and are not located in the target countries; c.f., Litman et al., 2018). Finally, participants had to pass four data quality items that were placed at the end of the screening survey: (6) one self-report item about data quality, (7) one bogus item, and (8–9) two attention checks. We only invited participants to participate in the main experiment if they passed all of the aforementioned target population screening and data quality measures. Of 1,503 participants, 1,072 answered and passed all of these nine screening measures and were invited to participate in the main experiment.

For the prescreening study, participants were paid USD 1.00/ GBP 1.15, while, for the main experiment, participants were paid USD 8.00/ GBP 6.00. In total, participants were paid USD 9.00 (MTurk)/GBP 7.15 (prolific) for their participation (for prolific, the minimum wage is

than the displayed number, as we had to further exclude participants who had already participated in the validation studies of the email and video manipulations.

GBP 5.00/hour, which entailed that the remuneration was slightly higher than the U.S. minimum wage of USD 7.25/hour).

Instrumental variables. As outlined in more detail below, this study made use of instruments to reduce endogeneity concerns (for the endogenous variables). More specifically, we used the manipulated conditions as instruments (i.e., the opening condition as an instrument for exploration and the closing condition as an instrument for exploitation). Furthermore, we measured participants' conscientiousness (M = 4.03, SD = .60, α = .82) and openness to experience (M = 3.84, SD = .75, α = .84) from the HEXACO inventory (using 10 items for each dimension; Ashton & Lee, 2009) and positive trait affect (M = 3.27, SD = .79, α = .85) and negative trait affect (M = 1.41, SD = .62, α = .90) with five items each from Mackinnon et al. (1999). We display the full measures in Appendix D. Although we acknowledge that based on theory (see Figure 1), the relevance condition, in the form of the F-test of the first-stage regression, is less likely to be met with the personality/affect variables than with the manipulated conditions, we nevertheless decided to include these measures to have some alternative instruments and to reflect previous work. That is, we chose to include these alternative instruments because follower personality measures (i.e., openness to experience, conscientiousness) and positive affect are exogenous and have been theoretically (Rosing et al., 2011) as well as empirically (Zacher et al., 2016) associated with follower ambidexterity.

Experimental task

The experimental task was designed to capture a proxy of workplace innovation from previous research (e.g., Zacher et al., 2016; Zacher & Rosing, 2015). In this previous research on leader ambidexterity, employee innovation was rated using four items that assessed the extent to which employees exhibited one of the following behaviors: "coming up with new ideas,"

"working to implement new ideas," "finding improved ways to do things," and "creating better processes and routines" (e.g., Zacher et al., 2016; Zacher & Rosing, 2015; Zacher & Wilden, 2014). Hence, the focus of this innovation construct is on ideas and improvements concerning practices at the workplace (and less on radical innovations such as the development of the television; see also Axtell et al., 2000). To replicate the original study as closely as possible (cf. Brandt et al., 2014), we designed a task that allowed us to tap into similar aspects of the innovation construct (e.g., Zacher et al., 2016).

In the experimental task, participants had to improve marketing material for an organization that is attempting to promote a 20-year longitudinal study (Appendix E). To carry out this task, participants received an existing marketing document that the organization intends to use for recruitment purposes. This marketing flyer is relatively poorly designed (see Appendix F; e.g., the flyer features typographical mistakes, missing words, poor formatting, very small font size, etc.). Hence, there are various opportunities for improvements. Participants received relatively narrow task instructions, which were to (1) add pictures to the existing marketing flyer and (2) highlight important information within the flyer by using different colors. Participants had 30 minutes to submit the revised version of the flyer.[9]

Some may argue that the task merely measures creativity because participants can focus on adding pictures and colorful design features. However, we intentionally chose narrow task instructions to be able to capture the implementation aspect of innovation. In other words, participants could implement (novel and useful) changes into their work (e.g., reformatting the flyer) that were not part of the work instructions (i.e., "they can find better ways to do things").

[9] To keep the time window constant for all participants, the experiment was programed so that the revised flyer had to be submitted after 30 minutes.

Accordingly, we argue that this task reflects central aspects of innovation as defined by Hughes et al. (2018): Participants need to identify problems (and opportunities) that are associated with poorly (or well) designed marketing material; they can introduce, adopt, or modify new ideas germane to organizational needs (i.e., recruiting and retaining study participants); and they can practically implement these ideas immediately in a revised document. Furthermore, we conducted a small pilot study (without leader manipulations) using a sample of full-time working MTurk participants ($N = 19$) who carried out the experimental task (improving the flyer for marketing purposes). In this pre-test, we found that participants indeed implemented changes that went beyond the actual task instructions (see Appendix G, example 2).

In addition, we checked the validity of the experimental task by correlating a self-report measure of innovative work behaviors (using scales from previous research, i.e., Zacher et al., 2016, Appendix D) with an external rating of the innovation outcome (i.e., two independent raters rated the uploaded flyers with the innovation outcome scale presented in Appendix H[10]). The objective innovation ratings significantly correlated with participants' self-reported innovation behaviors during the task ($r = .60$, $p = .007$). This result tentatively supports that the experimental task taps into the construct of innovation and allows us to capture innovative behaviors.

Procedure

Participants had access to an external link that directed them to the study, which was set up using the Qualtrics survey platform. Upon clicking on the link, participants were informed of their right to opt out of the study at any time. We guaranteed their anonymity and asked participants to give their informed consent before starting the study. In the intake study,

[10] See Appendix G for examples of flyers with a high and a low innovation rating

participants provided information on their demographics (i.e., age, gender, country of residence, English-language proficiency, work hours per week, job role, and industry) as well as on their personality and positive trait affect.

In the main experimental study, which was conducted one week later, all participants first received the general task instructions (see Appendix E) and were provided with access to the material (Appendix F). They were then randomly assigned to one of the four conditions: (1) leader opening, (2) leader closing, (3) leader ambidexterity, and (4) transformational leadership (see Appendix I). In each of the four conditions, participants received two emails from their supervisor: the first one before the task and the second one during the task (i.e., after 15 minutes). We based our decision to divide the task into two periods of 15 minutes, of which the first 15 minutes should be attributed to exploring (via opening behaviors) and the last 15 minutes to exploiting (via closing behaviors), on Farr et al.'s (2003) phase model of innovation. This phase model assumes that exploratory processes are relevant in the early phases of the innovation process (problem identification and idea generation), whereas late phases require exploitative processes (i.e., idea evaluation and implementation).

The appropriate time scale that is necessary to elicit the proposed positive consequences of opening and closing behaviors is currently unclear from theory. However, three starting points offer evidence for fine-grained variations of opening and closing behaviors over time: First, research has suggested that ambidextrous leadership can vary and exert influence on employees on a daily level (Zacher et al., 2016), providing support for a dynamic fluctuation on the day level. Second, the originators of the theory noted that employee ambidexterity may co-vary with innovation "across even shorter time scales, for example, on an hourly basis" (Rosing & Zacher, 2017, p. 706). Finally, research in the field of economics (e.g., Ederer & Manso, 2013) has

provided evidence that short experiments can be suitable to create variations in exploration behavior and innovation. For these reasons, we assume that a timeframe of 30 minutes is suitable for an experimental test of the theory.

The content of the email contained the manipulations of the different leadership behaviors (Appendix I). Such vignette-based experiments have proven to be as effective as laboratory experiments at evoking responses and can be used to zoom in on a "snapshot" of a daily situation, such as leader-follower collaboration on innovation tasks (Aguinis & Bradley, 2014; for a similar approach, see Farh & Chen, 2014).

Following the first email, participants started working on the task. After half of the allocated time, they received the second email from their supervisor. When the time was over, participants were asked to upload the revised file. After uploading the work output file (i.e., the flyer), participants reported on their exploration and exploitation behaviors during the task. We included multiple attention and comprehension checks (i.e., instructional manipulation checks and comprehension questions about the experimental materials; cf. Thomas & Clifford, 2017) at the end of the survey to ensure that participants had carefully engaged with this task. $N = 436$ participants started and finished the main experiment. We excluded 21 participants because they failed to correctly respond to any of four response quality items [(1) "some of the questions referred to a supervisor," (2) "I am answering questions on a website currently," (3) self-reported careful responding, (4) correct identification of both emails that they received from their supervisor]. We also excluded another 12 participants who uploaded documents that could not be opened (e.g., damaged PDF or Word files) and/or documents that could not be rated (e.g., pictures/ documents that were unrelated to the task). This resulted in a final sample of $N = 395$ participants for our analyses.

Experimental conditions

Appendix I presents the full material that we used for each leadership condition.

Leader opening condition. Leader opening has been defined as behaviors that increase variance in followers' behavior. However, scholars have criticized that constructs should not be defined by their outcomes and that construct definitions should not contain (often incomplete) examples of what is included in a construct (MacKenzie, 2003). Nevertheless, in the absence of a more accurate definition of ambidextrous leadership, we had to rely on items from published scales to experimentally manipulate opening and closing behaviors as expressed in different supervisor emails (i.e., conditions). Specifically, we created these emails based on the list of opening behaviors provided by Rosing et al. (2011; the items identified by Rosing et al. were also used as survey items by by Zacher et al., 2016; Zacher & Rosing, 2015; Zacher & Wilden, 2014). Both emails from the supervisor expressed opening behaviors. Specifically, the supervisor encouraged experimentation with different ideas ("I want to strongly encourage you to play with different ideas"), motivated taking risks ("So I encourage you to take a risk"), and encouraged different ways of accomplishing a task ("think about different ways and methods to make the most of this").

Leader closing condition. Leader closing as expressed in the supervisor email focuses on decreasing variance in the participants' behaviors based on the list of closing behaviors provided by Rosing et al. (2011; also used as survey items by Zacher et al., 2016; Zacher & Wilden, 2014; Zacher & Rosing, 2015). Thus, both emails from the supervisor expressed closing behaviors. Specifically, the supervisor stressed the importance of variance reduction ("I am here to provide you with specific instructions for this task"), monitored goal attainment ("I regularly check the progress of those workers that I have to supervise"), established routines ("If you have

established a work routine, I would recommend sticking to it now"), and noted the importance of adhering to rules ("I believe that close adherence to the rules and the formal task requirements is the best way to be successful here").

Leader ambidexterity condition. Leader ambidexterity as expressed in the supervisor email is a combination of the opening and closing behaviors identified by Rosing et al. (2011). Specifically, in line with the requirements of the innovation task, the supervisor attempts to increase variance in the participants' behavior (i.e., opening) in the first email (e.g., "There is no right or wrong way in doing this…") and then shifts towards variance reduction (i.e., closing) in the second email (e.g., "If you have established a work routine, I would recommend sticking to it now").

Transformational leadership condition (control). To allow for a fair comparison of our experimental conditions (Cooper & Richardson, 1986), we used transformational leadership as a control condition. We made this choice for two reasons: First, ambidextrous leadership theory was explicitly introduced as an innovation-specific leadership style that should more accurately predict employee innovation than the heterogeneous findings derived from studies on transformational leadership and innovation (Rosing et al., 2011). Second, two of the studies that we sought to replicate also used transformational leadership as a control variable (Zacher & Rosing, 2015; Zacher et al., 2016).

To create supervisor emails that express high levels of transformational leadership, we first carefully reviewed published research using transformational leadership vignettes (e.g., Christie, Barlin, & Turner, 2011; Felfe & Schyns, 2006; Hentschel, Braun, Peus, & Frey, 2018). We then created the transformational emails by adapting existing transformational leadership vignettes (Christie, Barlin, & Turner, 2011) to our context and by rewording items from

published transformational leadership scales (e.g., Rafferty & Griffin, 2004). Both emails from the supervisor expressed transformational leadership. Specifically, the leader expresses a strong *vision* ("My goal is that our research should create significantly more insights compared to studies from other institutions"), attempts to motivate by *inspirational communication* ("This project makes me very proud. I sincerely hope that I can inspire you to feel proud of it as well"), engages in *intellectual stimulation* ("I am encouraging you to think about these problems in new ways"), expresses *support* ("My goal as a leader is to encourage your personal development and to pay attention to your individual needs"), and emphasizes *personal recognition* ("I will acknowledge your accomplishments when I see outstanding work").

Validity of leadership manipulations (pilot study)

To ensure the validity of the experimental manipulation, we conducted a pilot study containing only the email vignettes and the manipulation check items. We separated this study from the main experiment in acknowledgement of the recent discussion concerning a manipulation check in itself being an intervention that potentially influences participants' subsequent behaviors in an experimental task (e.g., Bless & Burger, 2016; Fayant, Sigall, Lemonnier, Retsin, & Alexopoulos, 2017). We recruited participants in exchange for financial compensation (USD 3.00 for about 15 minutes) using MTurk workers who had an approval rate of more than 90% (Keith et al., 2017) and were located in English-speaking countries (mostly the US). Upon providing their informed consent, participants were randomly assigned to one of the four conditions.

Participants read the emails from the supervisor and were able to respond via email (e.g., "You can write an email to shortly discuss the task with your supervisor J.P."). We included this interactive element to make the task more immersive and also to serve as an implicit measure to capture participants' task engagement. Additionally, we employed four attention check measures

and excluded participants who did not answer these checks correctly to ensure that all participants in our final sample had paid careful attention to (and comprehended) the leadership manipulations (see Appendix J for details). Participants (N = 75) had a mean age of 36 years (SD = 10.2) and worked on average 37 hours per week; 52% were female.

After reading the supervisor emails, participants rated the extent to which they perceived opening and closing behavior during the experiment ("To which extent did the supervisor (J.P.) from this task mention or show the following aspects in his/her email?"). We used seven items for each scale (items from Zacher et al., 2016; see also Zacher & Rosing, 2015; Zacher & Wilden, 2014). Examples of items used to measure opening behaviors are "J.P…," "…allows different ways of accomplishing a task," "…encourages experimentation with different ideas," and "…motivates to take risks" (α = .95). Examples of closing items are "…monitors and controls goal attainment," "…takes corrective action," and "…controls adherence to rules" (α = .91). Following the approach adopted by Zacher et al. (2016), we calculated the multiplicative interaction term between the opening and the closing scales to obtain a measure of leader ambidexterity.

Participants also rated the extent to which the leader displayed a transformational leadership style (using items adapted from Rafferty and Griffin, 2004). These items assessed core aspects of transformational leadership, such as *vision* (e.g., "J.P. mentioned that he/she had a clear understanding of where we are going"), *inspirational communication* (e.g., "…wants to make me proud to be a part of this project"), *intellectual stimulation* (e.g., "… wants to challenge me to rethink some of my basic assumptions about this task"), supportive leadership ("… wants to consider my personal feelings"), and *personal recognition* ("he/she would commend me when I do a better than average job"). To separate the effects of transformational leadership on followers (e.g., "this leader inspires me") from actual behavioral elements of leadership, we added a stem to each item that

highlighted the behavioral component of the leadership style (i.e., "J.P. mentioned in the email that s/he wants to…,") followed by the items. All items were answered using a 5-point Likert response format (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, and 5 = a great deal; $\alpha = .92$).

We conducted analyses of variance (ANOVAs) to test whether the four conditions resulted in significantly different perceptions of leadership styles. The ANOVAs showed that perceptions of opening behaviors differed significantly among the leadership conditions ($F(3, 71) = 31.91$, $p < .001$), perceptions of closing behavior differed significantly among the leadership conditions ($F(3, 71) = 19.28$, $p < .001$), measures of ambidexterity differed weakly among the conditions ($F(3, 71) = 2.69$, $p = .052$), and perceptions of transformational leadership behaviors also differed significantly among the conditions ($F(3, 71) = 7.45$, $p < .001$).

Furthermore, we tested specific contrasts to determine whether the conditions were sufficiently distinct from each other. The contrasts showed that perceptions of opening behaviors were highest in the opening condition ($M = 4.48$, $SD = 0.52$); that is, they were significantly higher than in the closing condition ($M = 2.06$, $SD = 1.09$, $t(24) = 8.5$, $p < .001$), higher than in the transformational condition ($M = 2.87$, $SD = 1.01$, $t(36) = 4.31$; $p < .001$), and also higher than in the ambidexterity condition ($M = 4.06$, SD = 0.68, t(39)= 2.16, $p = .037$).

Perceptions of closing behaviors were highest in the closing condition ($M = 3.63$, $SD = 0.68$); that is, they were significantly higher than in the opening condition ($M = 1.66$, SD = 0.73, t(35) = 8.40; $p < .001$), significantly higher than in the transformational condition ($M = 1.95$, $SD = 0.84$, $t(32) = 6.39$, $p < .001$), and also significantly higher than in the ambidexterity condition ($M = 2.36$, SD = 1.02, $t(38)= 4.5$, $p < .001$).

Perceptions of transformational leadership were highest in the transformational condition ($M = 3.49$, $SD = 0.72$); that is, they were significantly higher than in the closing condition ($M =$

2.12, *SD* = 1.08, *t*(32)= 4.29; *p* < .001), significantly higher than in the opening condition (*M* = 2.72, *SD* = 0.79; t(33)= 2.98; *p* = .005), and also significantly higher than in the ambidexterity condition (*M* = 2.32, *SD* = 0.98; *t*(36) = 4, *p* < .001).

Finally, the ambidexterity score was highest in the ambidexterity leadership condition (*M* = 9.61, *SD* = 4.92); that is, it was significantly higher than in the transformational leadership condition (*M* = 5.9, *SD* = 3.6, *t(36)* = 2.56; *p* = .015) and higher (albeit non-significantly) than in the opening condition (*M* = 7.37, *SD* = 3.24; *t*(39) = 1.70; *p* = .098) and the closing condition (*M* = 7.41, *SD* = 4.20, *t(38)* = 1.5; *p* = .141).

These results not only show that the different leadership conditions were effective manipulations of the different leadership styles but also indicate that the conditions differentiate in nuanced ways among leadership styles that share conceptual overlap (e.g., the ambidexterity and opening conditions shared 50% of content overlap; opening behaviors and transformational leadership shared conceptual elements such as intellectual stimulation and opening behaviors; and opening behaviors and transformational leadership have also been shown to correlate in past research, e.g., *r* = .67 in Zacher et al., 2016; *r* = .49 in Zacher & Rosing, 2015).

Finally, we assessed whether the scenarios were perceived as realistic using two questions from Farh and Chen (2014) that were rated on a 5-point Likert scale (1 = completely disagree, 5 = completely agree; α = .88). The items were "It is realistic that I might experience a supervisor like J. P.," and "At some point during my career, I will probably encounter a situation like the one described above." Participants generally agreed that the scenario was realistic (*M* = 3.66, 72% scored ≥ 3), and this perception was not affected by different conditions (*F*(3,71) = .40, *p* = .751). The email responses from participants also indicated that participants were immersed in the scenario (average length per email in characters: *M* = 431; 97% of all participants wrote > 77

characters). In the general feedback section of the survey, participants also commented about J.P. (e.g., "I feel like this is the type of supervisor I would prefer to work with. They give you enough room to hang yourself but sound like they would encourage you to learn from whatever mistake you make" or "The contents were realistic. JP was very focused on the results and consistency"). Participants also commented on the interactive elements ("it was an interactive survey where you had to pay attention even to the little details you thought that didn't matter. it was interesting"). These results further indicate that the experimental conditions were immersive and increase the external validity of this study.

Measures

Follower exploration, exploitation, and ambidexterity. After uploading their work output, participants rated the extent to which they had engaged in exploration and exploitation with 14 items adapted from Zacher et al. (2016) and Mom et al. (2009) using a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). As these original scales were used within an organizational field context, we had to reword the items to better align the constructs with the context of the experimental task (see Table 2). In doing so, we carefully paid attention to capture core conceptual features from the definition of individual (non-managerial) ambidexterity. According to this definition, exploration encompasses "behaviors related to experimentation, searching for alternative ways to accomplish tasks, and learning from errors, […] deviat[ing] from routines, trying out something new, and […] not rely[ing] on established knowledge" (Rosing & Zacher, 2017, p. 696). Hence, exploration was measured using items such as "During the task, I focused on strong renewal of the flyer" (original item: "Focusing on strong renewal of products/services or processes) and "I searched for novel ways to make the flyer more interesting" (original item: "Searching for new possibilities with respect to my work").

Exploitation is defined as "relying on previous experience, putting things into action, and incrementally improving well-learned actions (…), doing things as they have always been done and relying on existing rules and routines" (Rosing & Zacher, 2017, p. 696). Accordingly, examples of exploitation items include "During the task, I focused mainly on carrying out those task activities that were provided in the task description (i.e., adding pictures and colors)" (original item: "Activities which I clearly know how to conduct") or "During the task, I focused on getting the task done as quickly as possible" (original item: "Activities primarily focused on achieving short-term goals"). Following Mom et al. (2009) and Zacher et al. (2016), ambidexterity was operationalized by calculating the multiplicative term of the exploration and exploitation scores.

We ran confirmatory factor analyses (CFAs) to check the robustness of the measurement models underlying the adapted measures of follower exploration and exploitation. We used Mplus (Muthén & Muthén, 2010) to run the CFAs and relied on the χ^2 and its associated p-value to assess model fit. Results indicated that both the two-factor model (i.e., lower chi-square value, χ^2 (64) = 342.61, $p < 0.001$) and the one-factor model (χ^2 (65) = 680.67, $p < 0.001$) did not fit the data well. Since the χ^2 suggested to reject both the one and two-factorial model, we calculated modification indices to check for sources of potential misfit (details on the CFAs and the modification indices can be found in Appendix K). Inspection of the largest modification indices from the two-factor model suggested that model fit could be increased by allowing items from one factor (i.e., exploration) to cross-load on both factors (e.g., allowing exploration item 6 to also load on the exploitation factor; see Table 7 for the full item texts). Furthermore, model fit could have been increased by allowing correlations across items (for exploration, e.g., item 6 with item 7; for exploration, e.g., item 3 with item 4; item 6 with item 1). Due to the replication

nature of this study, we did not apply any of these suggested modifications. However, we report them to allow future researchers to develop better scales.

++++++ Insert Table 2 about here++++++

Innovation outcome. The degree of innovation of the revised marketing flyer was assessed by four raters: two subject matter experts (i.e., marketing experts who are working on the marketing of the actual 20-year longitudinal study) and two research assistants with a background in psychology. All raters were blind to the experimental conditions and independently rated the outcomes (i.e., flyer) submitted by participants using four items (see Appendix H) adapted from Rosing et al. (2018) on a 5-point scale. These items focus on the product facet of innovation (i.e., the "innovative outputs implemented") rather than the process facet of innovation (i.e., "behaviors, actions, and cognitive processes that a person (…) engages in when attempting to generate and implement creative ideas"; Hughes et al., 2018, p. 13). Furthermore, this rating scale captures both the creative dimension ("this task outcome is completely novel and does not at all rely on conventional solutions"; "this task outcome is very creative") and the implementation dimension of innovation ("this task outcome can readily be applied in the 'real world'"; "this task outcome exceeds the quality standards").[11] We used a fully crossed rating design (Hallgren, 2012); that is, each of the four raters rated the entire corpus of 395 flyers, which resulted in 16 measurements (4 items x 4 raters) for each flyer.

We performed an interrater reliability analysis using the innovation outcome ratings from all four raters to calculate the intraclass correlations (ICC) values. The ICC is a statistical index

[11] We acknowledge that another conceptual feature of workplace innovation is the promotion/selling of ideas (cf. Hughes et al., 2018). However, this dimension has not been assessed by previous studies on leader ambidexterity theory (e.g., Zacher et al., 2016; Zacher & Rosing, 2015; Zacher & Wilden, 2014). To keep the replication closer to original studies (Brandt et al., 2014), we decided to not include it in our innovation measure.

commonly used to estimate reliability because it adjusts for chance agreement and systematic differences between raters (Fleiss & Shrout, 1978); it is therefore a more conservative estimate than the Pearson product-moment correlation. Furthermore, we computed the absolute agreement ICCs between raters, which is stricter than using the consistency-based ICC (Hallgren, 2012; McGraw & Wong, 1996). Following Cicchetti (1994), we classified ICCs by cut-off criteria; that is, a value below .40 is considered as poor, between .40–.59 as fair, between .60–.74 as good, and equal or greater than .75 as excellent. To ensure consistently high-quality standards, we regularly monitored the reliability during the process of rating the flyers. When the ICC for a subsample of flyers was ≤ .70, the first author discussed misalignments for all flyers that had "considerable" differences in innovation ratings (e.g., when differences between raters were ≥ 1 point on the rating scale). It is worth noting that, while we discussed the best rating solutions for flyers that were rated differently to ensure a continuous learning process on the part of the raters, we did not correct (i.e., clean) flyer ratings retrospectively, meaning that all ratings still reflect *independent* ratings. In other words, we used discrepancy discussions to inform raters for ongoing ratings (of future flyer batches) and to avoid rater drifts (i.e., to maintain "calibration").

The ICC values were calculated by comparing the average rating across the four items among the four raters. Based on the ratings of a single rater, results revealed good reliability, ICC(absolute, single measures) = .75. Using the average rating across all four raters revealed excellent reliability, ICC(absolute, average measures) = .92. Hence, for subsequent analyses, we used the average innovation rating across all four raters (M = 1.57, SD = 0.57).

Analytical strategy

Appendix L provides a full overview of the analytical steps that we performed (including a commented syntax for SPSS and MPLUS, as well as key decisions for the results). Only

participants who (a) completed the prescreening survey (matching the target population), (b) participated in the main experiment, and (c) passed all data quality items (Thomas & Clifford, 2017; Ward & Meade, 2018) were included in the final analysis ($N = 395$). To determine whether exclusion criteria affected the results, we re-ran all analyses for the full sample of $N = 436$ (we only report these analyses when results differed with respect to conclusions that affected the hypotheses). Furthermore, we have summarized the exclusion criteria in Appendix M.

To test H1–H3, we ran ANOVAs and contrasts (using leadership as an independent variable with four conditions: opening, closing, ambidextrous leadership, and transformational leadership) for each of the three dependent variables (follower exploration, follower exploitation, and innovation). A detailed description of the predicted results is provided in Table 3. To illustrate how to read Table 3, the first row shows that for Hypothesis 1, we expect a significant F-value for the one-factorial ANOVA (with four leadership conditions) when using exploration as a dependent variable. Furthermore, we expect that contrasts between conditions show significant differences between the opening and closing conditions (with higher exploration values in the opening condition).

++++++ Insert Table 3 about here++++++

To test H4, we ran multiple regression analyses using exploration, exploitation, and their multiplicative interaction (follower ambidexterity) as predictors of innovation. Because follower ambidexterity is assessed via self-reports and not experimentally manipulated, it is possible that the error terms across equations are still correlated due to the omission of common causes of both variables (Shaver, 2005). This could result in an inconsistent[12] estimate of the relationship

[12] Consistent estimates are estimates of a presumed causal relationship that would converge with the correct population parameter estimate as the sample size increases (cf. Antonakis et al., 2010).

between the follower measures (i.e., exploration, exploitation) and innovation (Antonakis, et al., 2010). To avoid this problem, we applied a two-stage least squares (2SLS) regression. 2SLS allows obtaining a consistent estimate of the coefficient (i.e., the relationship between follower ambidexterity and innovation) by using an instrumental variables approach. Instrumental variables must be exogenous (i.e., they should vary randomly in nature or be experimentally manipulated), should satisfy the exclusion condition (i.e., they should not be direct predictors of the outcome variable, innovation, beyond their effect on the endogenous variable, i.e., follower exploration and exploitation behaviors), and they should be strong predictors of follower ambidexterity (i.e., the relevance condition; cf. Antonakis et al., 2010; Ketokivi & McIntosh, 2017).

We intended to use the manipulated conditions as instruments. To do so, we first estimated the relationship between the instrumental variables and the ambidexterity of followers (i.e., exploration, X_1, and exploitation, X_2), which provides predicted values of X_1 and X_2. The appropriateness of these instruments was tested with the F-test of this regression (i.e., > 10, Staiger & Stock, 1997). If the manipulated conditions passed the test, the dependent variables were regressed on the predicted values of ambidextrous follower behaviors based on the estimates of the first stage. Given that the manipulated conditions were found to be insufficient predictors of ambidextrous behaviors (see results), we also tested whether other instruments (i.e., personality, positive trait affect) passed the test.

Results

Table 4 presents the means, standard deviations, and correlations for all study variables. It is worth noting that we created dummy-coded variables for each experimental condition and

that the average for innovation was very low (M = 1.57, SD = 0.57), which suggests a restriction of variance in the outcome variable.

++++++ Insert Table 4 about here++++++

For easier interpretation, Table 5 also shows the means and standard deviations of the focal dependent variables (exploration, exploitation, and innovation) across the four experimental conditions.

++++++ Insert Table 5 about here++++++

H1 stated that *opening leader behaviors positively predict follower explorative behaviors.* Using exploration as the dependent variable, the ANOVA showed no significant differences among the four different leadership conditions (F(3, 391) = 1.37, p = .25, η^2 = .01), thus failing to support H1. Since the F-value of the ANOVA was non-significant, we did not further test planned contrast between different leadership conditions.

H2 stated that *closing leader behaviors positively predict follower exploitative behaviors.* Using exploitation as the dependent variable, the ANOVA showed significant differences among the four different leadership conditions (F(3, 391) = 4.05, p = .008, η^2 = .03),[13] providing initial support for H2. Planned contrast showed that exploitation was significantly higher for closing leadership (M = 5.15, SD = 0.93) in comparison to opening leadership (M = 4.76, SD = .80, t(391) = 3.2, p = .001, η^2 = .026) and in comparison to ambidextrous leadership (M = 4.81, SD = .99, t(391) = 2.68, p = .008, η^2 = .018). However, exploitation did not differ when comparing closing leadership to transformational leadership (M = 4.96, SD = .76, t(391) = 1.60, p = .111, η^2 = .006). These results partially support H2.

[13] For the full sample (N = 436), the F-test was not significant (F(3, 432) = 2.41, p = .067).

H3 stated that *leader ambidexterity positively predicts employee innovation insofar that innovation is highest when both leader opening and leader closing behaviors are high.* Using the objective innovation rating (averaged across four raters) as the dependent variable, the ANOVA showed no significant differences among the four different leadership conditions ($F(3, 391) = 0.684$, $p = .562$, $\eta^2 = .005$), thus providing no support for H3.

H4 stated that *the interplay of follower exploration and exploitation (i.e., follower ambidexterity) positively predicts innovation outcomes insofar that innovation is highest when both follower exploration and exploitation behaviors are high.* First, we calculated an OLS regression with innovation as the dependent variable and then entered the other variables in consecutive steps (step 1: positive trait affect, conscientiousness, openness to experience; step 2: opening leadership, closing leadership, ambidextrous leadership; step 3: exploration, exploitation; step 4: ambidexterity (exploration*exploitation)]. We used this approach merely with the aim of replicating the analytical approach of Zacher et al. (2016).[14] We recognize that such a model is wrong because it correlates endogenous predictors (i.e., followers' reported perceptions of exploration, exploitation, and ambidexterity) with an endogenous dependent variable. However, this is a common way in which researchers would test such a hypothesized model. If we find significant predictors, we cannot draw any causal inferences from these results.

Model 3 included exploration (B = 0.12, SE = 0.02, $p < 0.001$) and exploitation (B = -0.10, $p = 0.002$) as predictors and showed a significant effect on innovation ($F(8, 386) = 4.41$, p

[14] In the study by Zacher et al. (2016), the authors also added transactional leadership (which we did not manipulate in our study). Note also that we used the transformational leadership condition as the control condition in the regression (which is comparable to entering it as an observed variable in a first step). That is, in our analyses, the manipulated leadership conditions were added as k-1 dummy-coded variables that were contrasted against transformational leadership. Finally, Zacher et al. (2016) did not test the interactive effect of opening*closing leadership which we included in our regression (the results did not differ if we included or excluded ambidextrous leadership).

< .001; R^2 = .083). Furthermore, Model 4, which included the multiplicative interaction of exploration and exploitation, also showed a significant F-value (F(9, 385) = 4.43, p < .001; R^2 = .093). However, the coefficient for ambidexterity (B = -0.05, p = 0.038) suggested a negative effect on innovation, which is the opposite of what H4 predicted.

Since follower ambidexterity (including exploration and exploitation) was not experimentally manipulated, it is possible that omitted common causes affect the estimated coefficients between exploitation, exploration, and ambidexterity on the one hand and innovation on the other (Shaver, 2005). This could result in an inconsistent estimate of the relationship between the follower measures (i.e., exploration, exploitation, and ambidexterity) and innovation (Antonakis et al., 2010). To avoid this problem, we applied a 2SLS regression using instrumental variables. In a first step, we used only the experimental conditions as instruments. However, the results of the first-stage regression revealed that the manipulated conditions were very weak instruments (F-values from the first-stage regression were < 4.05).

Hence, in a second step, we used the experimental conditions and the personality variables (which showed a correlation with the endogenous predictors) as instruments. We used the "reg3" command in STATA version 11 with innovation rating as the dependent variable; exploration, exploitation, and ambidexterity as endogenous predictors; and positive trait affect, conscientiousness, openness, and the manipulated leader behaviors manipulation as instruments. The results of this analysis were non-significant for the overall model (F(3,391) = 0.43, p = .73), including the parameter estimate for ambidexterity (b = -.27, SE = 0.44, p = .53), and thus failed to support H4 (see Appendix N for first-stage estimation and fit statistics). The Sargan overidentification test was non-significant (Sargan χ^2(3)= 1.74, p = .63; Basmann χ^2(3) = 1.72, p = .63), which provided further evidence that the instruments used did not correlate with the

residuals of the y equation (i.e., the overidentifying restrictions of the system of equations are viable). To determine whether the results of the *efficient* estimation method (OLS regression) or the *consistent* estimation method (2SLS approach) should be prioritized, we tested the exogeneity of the observed predictor variables (i.e., exploration, exploitation, and ambidexterity). The initial results of the Durbin test ($\chi^2(3)=4.03$, $p = .26$) and Wu-Hausman test ($F(3,388) = 1.32$, $p = .27$) suggested that exploration, exploitation, and ambidexterity were exogenous. However, since the first-stage regression in the 2SLS model also revealed that our instruments were weak—the F-values from the first-stage regression were $F(6,388) = 6.91$ (for instrumenting exploration), $F(6,388) = 2.27$ (for instrumenting exploitation), and $F(6,388) = 4.19$ (for instrumenting ambidexterity)—and critical values by Stock and Yogo (2005) indicate that F > 12.20 is required for estimates to have less than 5% bias relative to OLS (and F > 5.35 to have less than 20% bias relative to OLS, Sajons, 2020; Stock & Yogo, 2005), the Durbin and Wu-Hausman tests were underpowered, and their results should be interpreted with caution. Therefore, we ran separate exploratory analyses to identify which of the predictors could be exogenous.[15] These results suggest treating exploration as probably endogenous (Durbin test: $\chi^2(1) = 5.67$, $p = .017$; Hausman test: ($F(1,392) = 5.71$, $p = .017$), while they suggested treating exploitation as probably exogenous (Durbin test: $\chi^2(1) = 0.003$, $p = .95$; Hausman test ($F(1,392) = 0.003$, $p = .95$).

Taken together, the results both of the traditional OLS approach and the 2SLS approach did not support H4. Furthermore, only H2 received some support; that is, as theorized, leader

[15] To do this, we conducted one 2SLS regressions using innovation (as the dependent variable), exploration (as the endogenous predictor), and positive affect and conscientiousness (as instruments). In a second exploratory 2SLS regression, we used innovation (as the dependent variable), exploitation (as the endogenous predictor), and closing leadership (as instruments). For both analyses, we tested for overidentification and endogeneity of the predictors using the Durbin and Hausman test.

closing behavior was positively associated with exploitative behavior amongst followers. Finally, we noted that the average ratings for innovation were notably low, which suggests a restriction of variance in the outcome variable.

Study 2: Video Experiment

The main purpose of Study 2 was to increase the external validity of our research by using a more complex representation of behavior than that allowed in a vignette study and by testing participants in in a laboratory context to allow for a more controlled environment (e.g., although we instructed participants in Study 1 to only participate while in a quiet environment, we had significantly more control over the experiment in Study 2). Specifically, our second study tested the leader ambidexterity model in an experimental lab context that more realistically mimicked the fine-grained dynamics of leadership ambidexterity theory. To do so, we manipulated leadership behaviors by using a trained actress and assigned participants the same experimental task as in Study 1. To ensure a high standardization of the manipulation, we videotaped the actress (Antonakis, d'Adda, Weber, & Zehnder, 2015). Manipulating leadership style through the behaviors of an actress provides a further opportunity to more realistically depict the nature of leadership styles (Avolio, Reichard, Hannah, & Walumbwa, Chan, 2009; Podsakoff, Podsakoff, MacKenzie, & Klinger, 2013). Furthermore, the design allowed for a more subtle manipulation of different leadership styles and, thus, a stricter test of the model.

Methods

The study was approved by the institutional review board of Curtin University in Australia prior to being conducted (HREC approval number HRE2018-0454). Study data have been published in anonymized form on the Open Science Framework (OSF, https://osf.io/buvq8/files/).

Sample

Since most hypotheses (with the exception of H2) could not be confirmed in Study 1, the sample size estimations for Study 2 were based on the same effect sizes as those used for Study 1. However, the recruitment of working participants for a one-hour study (that could not be undertaken remotely) took significantly more time than recruiting participants via flexible crowd-working platforms (as we did in Study 1). After an extension of the registered report deadline, we ultimately collected $N = 271$ participants in Study 2. This sample size was still significantly larger than those of studies that we sought to replicate (e.g., taking into consideration all six samples of the five studies that we sought to replicate [see Table 1, column 2, or Appendix B], this sample is on average about 2.09 times larger).

We excluded participants who self-reported that they had not paid attention during the experiment (using a self-reported single item indicator[16] adapted from Meade & Craig, 2012; cf. DeSimone & Harms, 2018; n = 22) and participants who were not able to recognize the experimental condition that was assigned to them (i.e., manipulation checks; "Please indicate which of the following messages resembles most accurately the first [second] message that you received from your supervisor?", $n = 28$ identified a leader message that was from a condition that was not assigned to them), which resulted in a final analysis sample of $N = 229$.

The participants in our final analysis sample had a mean age of 27 years ($SD = 8.79$), and 75.5% were female (1 participant identified as 'other gender'). Participants worked an average of 25.7/week (in their current job) and 33 hours/week (in their previous job). Participants worked in

[16] The full instruction for this question was "It is important for our research to only include responses from participants that devoted their full attention and did not answer randomly to the questions or made false claims. In your honest opinion, should we use your data in our analyses in this study?" and the response options were "YES, I carefully read all questions and tried to answer everything honestly" versus "NO, I just clicked randomly through the whole survey."

various industries, such as hospitality (15.3%), education (14.8%), health services (9.6%), government (7.9%), business services (5.7%), industry services (4.8%), business consulting (5.6%), finance (4.1%), human services (3.5%), human services/welfare (3.9%), accounting (3.5%), information technology (2.6%), agribusiness (2.2%), financial services (1.3%), legal services (1.3%), media (0.9%), and construction (0.4%). On average, participants had 9.31 years (SD = 8.30) of work experience and were working in positions of different levels: 17% were working on an intern level, 44.5% were working on entry level, 16.2% were working on an associate level, 14.8% were on a manager level, 2.2% on a senior manager level, and 3.5% were working on a director level. In terms of highest level of education, 28.8% had a senior high school level, 29.7% had a graduate certificate/postgraduate diploma/advanced diploma/other award course, 28.8% had a bachelor's degree/honors, 10.5% had a master's degree, and 2.2% had a PhD.

Design. We used the same experimental design as in Study 1 (four leader conditions: opening, closing, ambidextrous leadership, and transformational leadership). Participants worked on the same task as described in Study 1. However, in contrast to Study 1, participants received the leader messages as video messages. We used an actress to enact the leadership conditions [(a) opening behavior, (b) closing behavior, (c) ambidextrous leadership, and (d) transformational leadership]. We conducted the experiment in two locations (Germany and Australia) and thus selected an actress who was able to portray the leader in two languages (German and English). The actress received the email transcripts used in Study 1 (see Appendix I for the English and Appendix O for the German transcripts) to portray the leader in two video messages (for each condition). To minimize the effects of the actress (i.e., physical appearance, attractiveness, and gender), the same actress was used across all four conditions (Podsakoff et al., 2013). The actress

was instructed to maintain contaminating factors (e.g., non-verbal cues, body movements and orientation, and affective tone) constant across all conditions (see Figure 2). Furthermore, the visual perspective and film editing were kept constant across conditions, and we ensured that the actress used the exact wording of the emails (see Appendices I and M for the transcripts; see the OSF repository for the video messages).

++++++ Insert Figure 2 about here++++++

Procedure. Before arriving, participants had to provide their informed consent to participate in the study and filled out a short survey on their personality, positive and negative trait affect, and demographic information.

Upon arrival, participants were instructed that they would be working on marketing material for a real research project (i.e., a longitudinal study). Thereafter, participants were randomly allocated to one of the four leadership conditions. The paradigm was modeled in line with lab-based experimental leadership research methods using video-based messages (e.g., Damen, van Knippeberg, & van Knippenberg, 2008; Venus, Stam, & Van Knippenberg, 2013). Participants were informed that they were to receive two videotaped messages from a fictional supervisor ("In this simulation, you will be working with a fictional supervisor"). In the video message, the supervisor provided the task instructions using a different leadership style in each of the conditions. After 30 minutes, participants had to submit their final work output (i.e., the flyer). After submission, participants self-reported their exploration and exploitation activities during the task. At the end, participants were debriefed and compensated.

Experimental conditions

Opening, closing, ambidexterity, and transformational leadership conditions. The four different video conditions used transcripts that were identical to the emails used in Study 1 (see Appendix I, M).

Additional validity studies for the video manipulation of leadership styles.

To validate the video manipulation, we conducted a separate validation study containing only the video messages and rating measures of four different leadership styles (opening, closing, ambidextrous, and transformational leadership). We used a sample of TurkPrime participants with an approval rate of more than 90% (Keith et al., 2017) from English-speaking countries, who were randomly assigned to one of the four conditions. This validation study was identical to the approach to validation adopted in Study 1, except that participants watched two video messages in each leadership condition. The video validation study was only accessible to participants who were at least 18 ages old, worked full time (≥ 35 hours/week), and had a direct supervisor as part of their job (N = 108).

We used five attention check measures (two items asking participants to identify the correct video message from their manipulation, one bogus item, one item measuring careless responding indirectly, one item measuring careless responding directing, and one memory item about the survey) and excluded participants (N = 18) who did not pass any of these checks correctly. Exclusion of participants did not systematically affect sample size within the four experimental conditions ($\chi^2(3) = 4.5$, p = .211).

Participants in the validity study (N = 90) had a mean age of 38 years (SD = 11.01) and were working on average 41 hours/week; 34.5% were female. After watching the supervisor messages, participants rated the behavior using the same scales as described in Study 1 (opening

and closing behavior from Zacher et al., 2016, and transformational leadership items adapted from Rafferty and Griffin, 2004, α = .95). ANOVAs showed that perceptions of opening behaviors differed significantly among the leadership conditions ($F(3, 86) = 58.13$, $p < .001$), perceptions of closing behavior differed significantly among the leadership conditions ($F(3, 86) = 15.95$, $p < .001$), measures of ambidexterity differed significantly among the conditions ($F(3, 86) = 3.62$, $p = .016$), and perceptions of transformational leadership behaviors also differed significantly among the different conditions ($F(3, 86) = 9.64$, $p < .001$).

Furthermore, we tested specific contrasts to determine whether the conditions were sufficiently distinct from each other. Contrasts showed that perceptions of opening behaviors were highest in the opening condition ($M = 4.58$, $SD = 0.45$); that is, they were significantly higher than in the closing condition ($M = 1.88$, $SD = 1.10$, $t(25.3) = 10.49$, $p < .001$), higher than in the transformational condition ($M = 3.00$, $SD = 0.79$, $t(48) = 8.74$, $p < .001$), and also higher than in the ambidexterity condition ($M = 4.15$, $SD = 0.51$, $t(43) = 2.92$, $p = .005$).

Perceptions of closing behaviors were highest in the closing condition ($M = 3.95$, $SD = 0.73$), that is, significantly higher than in the opening condition ($M = 1.80$, $SD = 1.06$, $t(45) = 7.88$, $p < .001$), significantly higher than in the transformational condition ($M = 2.52$, $SD = 1.22$, $t(38.1) = 4.8$, $p < .001$) and also significantly higher than in the ambidexterity condition ($M = 2.81$, $SD = 1.78$, $t(29.3) = 3.62$, $p = .001$).

Perceptions of transformational leadership were highest in the transformational condition ($M = 3.60$, $SD = 0.78$); that is, they were significantly higher than the closing condition ($M = 2.06$, $SD = 1.04$, $t(43) = 5.61$; $p < .001$), significantly higher than in the opening condition ($M = 2.55$, $SD = 1.19$; $t(43.3) = 3.71$, $p = .001$), and also significantly higher than in the ambidexterity condition ($M = 2.28$, $SD = 1.11$, $t(30.9) = 4.38$, $p < .001$).

Finally, the ambidexterity score was highest in the ambidexterity leadership condition (M = 11.62, SD = 4.96); that is, it was significantly higher than in the transformational leadership condition (M = 7.83, SD = 4.37, t(41) = 2.65, p = .011), significantly higher than in the opening condition (M = 8.13, SD = 4.47, t(43) = 2.47, p = .017), and also significantly higher than in the closing condition (M = 7.35, SD = 4.52, t(38) = 2.85, p = .007).

Taken together, the results of the validity study (which we conducted with a separate sample) showed that each focal leadership manipulation had a mean rating of $M \geq 3.5$ and was perceived as significantly different when compared to the other leadership conditions. Overall, these results supported the validity of the different leadership styles portrayed in the video messages.

For the data collection in Germany, we created German versions of the four leadership conditions (we used a backtranslation process and involved five leadership experts, who evaluated the conceptual overlap of the translations with the focal leadership concepts; see Appendix O for the full translation). After producing the video messages in German (using the same actress and filming under identical conditions), we conducted another validation study with N = 192 participants. The German validation study supported the validity of the German video messages (for details, see Appendix P).

Measures

Follower exploration, exploitation, and ambidexterity. Exploration behavior and exploitation behavior were captured using the same measures as in Study 1. CFAs were performed to check the robustness of the measurement models underlying the exploration and exploitation measures (using Mplus; Muthén & Muthén, 2010). Results indicated that both the two-factor model (χ^2 (64) = 264.06, p < 0.001) and the one-factor model (χ^2 (65) = 422.79, p <

0.001) did not fit the data well. To identify sources of potential model misfit, we inspected the modification indices from the two-factor model (see Appendix Q), which suggested that model fit could be increased by allowing cross-factor loadings (for exploration: item 6), by allowing some items from the same factor to be correlated (for exploration: e.g., item 1 with 2; item 2 with 6, item 6 with 7, for exploitation: e.g., item 3 with 4; item 6 with item 1), and/or by allowing some items from different factors to be correlated.

Instrumental variables. Similar to Study 1, we collected additional (potential) instrumental variables, namely conscientiousness (M = 3.76, SD = .56, α = .75) and openness to experience (M = 3.43, SD = 0.66, α = .77), from the HEXACO inventory (Asthon & Lee, 2009), as well as their positive (M = 3.39, SD = .64, α = .82) and negative trait affect (M = 1.67, SD = .61, α = .83) (Mackinnon et al., 1999). These variables were assessed before participants started the experiment.

Innovation outcomes. Similar to Study 1, two subject experts and two research assistants (both of whom held a master's degree in psychology) who were blind to the conditions rated the degree of innovation of the flyers developed by the participants using four items (Rosing et al., 2018; Appendix H).

We performed interrater reliability analyses in the same way as for Study 1. Results showed good interrater reliability, ICC(absolute, single measures) = .71. Using the average rating across all four raters revealed excellent reliability, ICC(absolute, average measures) = .91. We therefore used the average innovation rating across all four raters (M = 2.18, SD = 0.86, Min = 1; Max = 4.81) for subsequent analyses.

Perceived interruptions (control variable). To determine whether the video messages for each of the leadership conditions affected the work process of participants across conditions

differently, we measured perceived interruptions in each condition using a scale adapted from Sonnentag, Reinecke, Mata, and Vorderer (2018). Participants indicated their agreement with the following items using a 5-point Likert scale (1 = fully disagree; 5 = fully agree): "Incoming video message kept me from doing my job," "The video messages have reached me at inconvenient moments," and "The video messages disturbed me in doing my work" (M = 2.76, SD = 1.03, α = .80). An ANOVA suggested that there were indeed significant differences in terms of perceived interruptions among the four leadership conditions (F(3, 225) = 5.73, p = .001). Post hoc Bonferroni-corrected comparisons indicated that the video messages in the closing condition were perceived as significantly more interruptive (M = 3.04, SD = 1.12) than video messages in the opening condition (M = 2.39, SD = .91, p = .003). Participants also perceived higher levels of interruption in the transformational condition (M = 3.03, SD = 1.00) in comparison to the opening condition (p = .007). Due to these results, we included *perceived video interruptions* as a control variable in all further analyses.

Careless responding. As outlined above, we only included participants in the analysis if they correctly recognized the experimental condition (i.e., both messages from their supervisor) and if they self-reported that the researchers should use their data (i.e., they took the study seriously). In addition, we also included three measures of careless responding (the same items as in Study 1: (i) one bogus item, (ii) one instructed response item, and (iii) one comprehension check) at the end of the experiment. We weighted the options to further exclude participants who incorrectly answered these careless responding items (n = 39) against reducing the power of our study. We decided to steer a middle course by creating a careless responding measure for each participant, which we included as a control variable in further analyses. The careless responding

score could range from 0 (i.e., all careless responding items were answered correctly) to 3 (i.e., all three careless responding items were answered incorrectly).

Analytical Strategy

We used the same analytical approach as in Study 1 (see Appendix L). However, we added control variables to the analyses (which were analyzed using ANCOVAs). We controlled for (1) the language (i.e., location) of the experiment, (2) the extent to which video messages were perceived as interruptions, and (3) the extent of careless responding.

To determine whether our exclusion criteria affected the results, we also ran all analyses for the sample without exclusion criteria ($N = 271$) and with exclusion criteria ($N = 229$) (we only report results for those additional analyses when they revealed differences concerning rejection/confirmation of hypotheses).

When the F-test of the ANCOVA was non-significant, we did not test contrasts among the different leadership conditions.

Results

Table 6 presents the means, standard deviations, and correlations for all study variables. As in Study 1, we created dummy-coded variables for each experimental condition.

++++++ Insert Table 6 about here++++++

To ease interpretation, Table 7 also shows the means and standard deviations of the focal dependent variables (exploration, exploitation, and innovation) and the continuous control variables (perceived interruptions and careless responding) across the four experimental conditions. The F-values of the ANOVAs indicated significant differences between the conditions for the innovation outcome ($F(3,225) = 2.77$, $p = .042$), as well as for careless responding ($F(3,225) = 3.76$, $p = .012$) and perceived interruptions of the messages ($F(3,225) =$

5.733, p = .001). We included those control variables in our subsequent testing of the hypotheses.

++++++ Insert Table 7 about here++++++

H1 stated that *opening leader behaviors positively predict follower explorative behaviors.* Using exploration as the dependent variable, the ANOVA showed no significant differences between the four different leadership conditions (F(3, 225) = 0.89, p = .45). However, when controlling for interruptions, language, and careless responding, the ANCOVA showed significant differences between the four different leadership conditions (F(6, 222) = 2.82, p = .011, η^2 = .071), providing initial support for H1. Contrasts showed that exploration was significantly higher for opening leadership in comparison to closing leadership (F(1,222) = 4.85, p = .029, η^2 = .02). For the full sample, this contrast was not significant, F(1,263) = 2.089, p = .15, η^2 = .008). However, exploration did not differ between opening leadership and ambidextrous leadership (F(1,222) = 1.86, p = .17, η^2 = .008) or between opening leadership and transformational leadership (F(1,222) = 0.93, p = .33, η^2 = .004). These results partially support H1.

H2 stated that *closing leader behaviors positively predict follower exploitative behaviors.* Using exploitation as the dependent variable, the ANOVA showed no significant differences among the four different leadership conditions (F(3, 225) = .817, p = .49). Similarly, the ANCOVA showed no significant differences between the four different leadership conditions (F(6, 222) = .694, p = .65, η^2 = .018), thus providing no support for H2.

H3 stated that *the leader ambidexterity positively predicts employee innovation insofar that innovation is highest when both leader opening and leader closing behaviors are high.* Using the objective innovation rating (averaged across four raters) as the dependent variable, we

found significant differences between the four different leadership conditions ($F(3, 225) = 2.77$, $p = .042$). Even after adding the control variables (interruptions, language, and careless responding), the ANCOVA still showed significant differences between the four different leadership conditions ($F(6, 222) = 4.30$, $p = .0004$, $\eta^2 = .104$), providing initial support for H3. Contrasts further revealed that innovation was significantly higher in the ambidextrous leadership condition than in the closing leadership condition ($\Delta = 0.326$, SE = .15, $F(1,222) = 4.56$, $p = .034$, $\eta^2 = .02$). Innovation in the ambidextrous leadership condition was also marginally higher in comparison to the transformational leadership condition ($\Delta = 0.285$, SE = .16, $F(1,222) = 3.167$, $p = .076$, $\eta^2 = .014$), and there was no statistical difference between ambidextrous leadership and opening leadership ($\Delta = -0.105$, SE = .15, $F(1,222) = 0.478$, $p = .49$, $\eta^2 = .002$). These results, and, in particular, the finding that ambidextrous leadership positively impacts innovation in comparison to closing (and transformational) leadership provide some evidence for H3. However, these results also indicate that opening leadership and ambidextrous leadership are equally beneficial for innovation.

H4 stated that *the interplay of follower exploration and exploitation behavior (i.e., follower ambidexterity) positively predicts innovation outcomes insofar that innovation is highest when both follower exploration and exploitation behaviors are high.* First, we calculated an OLS regression with innovation as the dependent variable; thereafter, we entered the other variables in consecutive steps (adopting the analytical approach of Zacher et al., 2016; see Table 8: step 1: openness to experience, conscientiousness, positive trait affect; step 2: opening leadership, closing leadership, ambidextrous leadership,[17] careless responding, language, perceived

[17] To replicate the approach of Zacher et al. (2016), we contrasted these manipulated conditions against the transformational leadership condition.

interruptions; step 3: exploration, exploitation; step 4: ambidexterity). Again, we recognize that such a model is wrong because it correlates endogenous predictors (i.e., exploration, exploitation, and ambidexterity) with an endogenous dependent variable, meaning that we cannot draw any causal inferences from the results of the model. However, the main purpose is to replicate analytical approaches from approaches adopted in previous research. Model 4 indicated that exploration was a significant predictor of innovation (B = 0.16, p = 0.003), whereas exploitation was non-significant (B = -0.10, p = 0.089; F(11, 217) = 3.82, p < .001; R^2 = .16). Adding the multiplicative interaction between exploration and exploitation (i.e., ambidexterity, B = -.04, p = 0.367) did not have a significant effect on innovation (F(12, 216) = 3.56, p < .001; R^2 = .16). Thus, this traditional analytical approach did not provide support for H4.

++++++ Insert Table 8 about here++++++

Because the three variables of exploration, exploitation, and ambidexterity are potentially endogenous, we also conducted a 2SLS regression with instrumental variables. In a first step, we used the only the experimental conditions as instruments. However, the results of the first-stage regression revealed that the manipulated conditions were very weak instruments (F-values from the first-stage regression were < 0.89). Hence, in a second step, we included the personality variables as additional instruments. We used the "reg3" command in STATA with innovation rating as the dependent variable; exploration, exploitation, and ambidexterity as endogenous predictors; and closing leadership, opening leadership, ambidextrous leadership, positive trait affect, and openness to experience as instruments (see Appendix R for first-stage estimation and fit statistics). This analyses did not reveal a significant effect for the overall model (F(3,225) = 0.52, p = .67), including the ambidexterity as an instrumented predictor of innovation (b = -.95, SE = 1.27, p = .46), thus failing to support H4. The result of the Sargan overidentification test

was non-significant (Sargan $\chi^2(2) = 1.52$, $p = .47$; Basmann $\chi^2(2) = 1.49$, $p = .47$), which provided further evidence that the instruments used did not correlate with the residuals of the y equation (i.e., the overidentifying restrictions of the system of questions are viable). To determine whether the results of the *efficient* estimation method (OLS regression) or the *consistent* estimation method (2SLS approach) should be prioritized, we tested the exogeneity of the observed predictor variables (i.e., exploration, exploitation, and ambidexterity). The results of the Durbin test ($\chi^2(3) = 3.09$, $p = .38$) and Wu-Hausman test ($F(3,222) = 1.01$, $p = .39$) suggested that exploration, exploitation, and ambidexterity are exogenous. However, the first-stage regression in the 2SLS model also revealed that our instruments were weak, that is, F-values from the first-stage regression were $F(5, 223) = 4.41$ (for instrumenting exploration), $F(5, 223) = 1.07$ (for instrumenting exploitation), and $F(5, 223) = 2.28$ (for instrumenting ambidexterity); and critical values identified by Stock and Yogo (2005) indicated that $F > 9.53$ is required for estimates to have less than 5% bias relative to OLS (and $F > 4.99$ to have less than 20% bias relative to OLS; Sajons, 2020; Stock & Yogo, 2005). Hence, the Durbin and Wu-Hausman tests are underpowered and should be interpreted with caution. Therefore, we ran separate exploratory analyses to identify which of the predictors should be treated as probably exogenous (this was only possible for exploration and ambidexterity, as none of the instruments were strong enough for exploitation). While these additional exploratory results suggested treating exploration as probably exogenous (Durbin test: $\chi^2(1) = 0.50$, $p = .48$; Hausman test: ($F(1,226) = 0.50$, $p = .48$) and treating ambidexterity as probably exogenous (Durbin test: $\chi^2(1) = 0.01$, $p = .91$; Hausman test: ($F(1,226) = 0.01$, $p = .91$), these results still need to be interpreted with caution, as the instruments were found to be very weak. Taken together, neither the results from the OLS regression nor the results of the 2SLS approach provide support for H4.

Discussion

The aim of this conceptual replication was to test four hypotheses derived from ambidextrous leadership theory. We identified five previous studies with various designs (cross-sectional correlational data, diary studies, and team-level designs) that reported empirical associations supporting the main propositions of ambidextrous leadership theory (see Figure 1). Our replication study was motivated by the methodological shortcomings of this previous research (e.g., use of self-report variables, reliance on self-reported innovation outcomes, no correction for endogeneity). Accordingly, our aim was to constructively test the assumptions of ambidextrous leadership theory by means of an experimental design (Köhler & Cortina, 2019) applied in two different contexts: one crowdsourcing online setting (which is similar to the recruitment and population of one the largest sample studies in previous leadership ambidexterity research, e.g., Zacher et al., 2016) and one video experimental study (conducted in two laboratories in Australia and Germany). Furthermore, we introduced an instrumental variable approach to diminish threats of endogeneity for variables that have not been experimentally manipulated (i.e., exploration and exploitation). It is important to note that our studies constitute an "independent replication"; that is, none of the original authors who developed or tested ambidextrous leadership theory were involved in the data collection or the testing of the hypotheses. Independent replications are important for science to be self-correcting (Vazire, 2018) and to rule out the effect of incentives (i.e., findings that are in favor of the hypotheses).

In the two experimental studies, the data indicated that, consistent with H2, closing behavior was positively associated with exploitation behavior (Study 1), although this effect could not be supported using the full sample available and was also not replicated in Study 2. Furthermore, although not found in Study 1, Study 2 provided partial support for H1 in that

opening leadership was found to promote exploration behaviors (although a full-sample analysis also failed to support H1). In addition, there was partial support for H3 in that ambidextrous leadership (i.e., the combination of opening and closing leadership behavior) was found to be most beneficial for increasing followers' innovation outcomes compared with transformational leadership and closing behaviors (albeit not in comparison with opening behavior). Furthermore, neither study replicated the finding that the interplay of follower exploration and exploitation behaviors is most beneficial for innovation (H4). In the following, we explore methodological reasons for these null results, followed by possible theoretical reasons.

Methodological issues and implications

First, in terms of methodological arguments, there may have been an issue with restriction of variance in Study 1, which may have lowered the likelihood of detecting a potential effect. In particular, the innovation scores for participants had a very low mean and a relatively small standard deviation (M = 1.57, SD = 0.57) indicating that most participants showed low innovative work performance. Restricted variance (i.e., reduced range of a variable) can statistically lessen the strength of an association with other variables (i.e., predictors). Hence, this interpretation might offer an explanation as to why we did not find support for the hypotheses that focused on innovativeness as an outcome (i.e., H3 and H4 [Study 1]).

Second, we had to adapt existing measures of exploration and exploitation to align them with our experimental context. While the model fit of a two-factorial model for these two measures was better than that of a single model, the results of our CFAs still suggested that some modifications would be required to achieve a better model fit (i.e., non-significant chi-square values). In other words, the estimates for exploration and exploitation might potentially have been biased (thus affecting H4). For the purpose of the replication, we did not make any

modifications to the CFA model or exclude any items. However, we encourage future researchers to consult our Appendices (K, O) or to use our data (from the OSF) to identify sources of potential misfit.

Third, the lack of specificity in ambidextrous leadership theory may offer an alternative reason why our experiments mostly found no support, as it required us to specify (a) what is meant by opening and closing behaviors (due to a lack of construct definition) and (b) the timing of these behaviors (due to a lack of consideration of time in the theory). To begin with, opening and closing leader behaviors (and thus the ambidexterity construct) have been defined by their outcomes (i.e., opening behaviors increase variance in followers' behaviors, and closing behaviors reduce variance in followers' behaviors; Rosing et al., 2011). This confronted us with the problem that it was conceptually unclear what opening behaviors actually are (see MacKenzie, 2003). Correspondingly, the definition of ambidextrous leadership—i.e., "leaders need to engage in both behaviors in line with the innovation task to increase innovation outcomes" (Rosing et al., 2011, p. 9)—does not precisely explain *how* certain leadership behaviors should be expressed and implies a tautological line of argumentation. Furthermore, this definition does not precisely explain when or how certain leadership behaviors should be expressed and implies some sort of outcome-based categorical imperative (i.e., if it results in innovation, leaders must have been using opening and closing behaviors; thus, they are ambidextrous leaders). To illustrate, a particular behavior (e.g., sanctioning a follower for delivering a poor-quality visual presentation) may at times increase variance in followers' behaviors (e.g., the follower may attempt to include more visual examples in future presentations) or decrease variance in followers' behaviors (e.g., the follower may decide to focus more on relying on existing presentations without making risky choices such as adding

new slides). Due to the lack of a proper construct definition, we had to rely on items from published scales (Zacher et al., 2016) as a source (and inspiration) to experimentally manipulate opening and closing leader behaviors. Since we are the first researchers to attempt to experimentally manipulate leader ambidexterity, a failure to replicate original findings may not necessarily be attributed to poor theory alone, as it could also reflect poor operationalization of the core construct.

Fourth, ambidextrous leadership theory is rather vague concerning the issue of timing. While the theory acknowledges that innovation processes are complex and do not neatly unfold in a linear fashion, it does not specify when leaders should engage in these behaviors. However, competing models of the innovation process have proposed more linear processes to foster innovation (i.e., early exploration vs. late exploitation; Farr et al., 2003). This to-date still missing understanding of the innovation process makes it much more difficult to test the model of ambidextrous leadership. However, we had to make assumptions regarding timing. We relied on linear phase models of innovation (Farr et al., 2003), which propose that opening is better at the beginning of the innovation task, whereas closing leadership is better towards the end of the task. This issue highlights that propositions from ambidextrous leadership only work in conjunction with temporal models of innovation. However, if assumptions from these temporal models of innovation are wrong (or if different temporal models of the innovation process are more appropriate, e.g., complexity perspectives that assume more intertwined and chaotic cycles of exploration and exploitation; Schroeder, Van de Ven, Scudder, & Polley, 1989), researchers will have a hard time properly testing the hypotheses of ambidextrous leadership theory.

Fifth, proponents of leadership ambidexterity theory may argue that our manipulation did not include enough (or included too many) opening and closing behaviors. The problem is that

the theory is not specific in prescribing the degree to which these behaviors need to be exhibited (other than that both behaviors need to be aligned with the requirements of innovation tasks). Furthermore, testing the theory with an experimental design (in contrast to a field study) may have more profound consequences with respect to the misalignment of leadership behaviors with the requirements of an innovation task. In a single (and brief) experimental situation, misalignment might prove more problematic and disruptive than in an ongoing leadership relationship, which provides followers with multitude of situational cues (beyond leader behaviors) that they can react to. Furthermore, "real" leaders in a field context may be better at aligning their behaviors with the requirements of different tasks—although this assumption concerning leader behavior–task requirements alignment remains to be tested by future research.

Taken together, our replication has revealed that the construct is conceptually underdeveloped and that the theory would benefit from providing more accurate predictions as to the degree to which leaders should exhibit opening and closing behaviors and when they should do so to allow researchers to better test the claims of the theory. This point also illustrates the merit of replication, which forces researchers to operationalize and specify the (often vague) predictions of management and leadership theories.

Theoretical issues and implications

Our study also offers at least four new insights concerning ambidextrous leadership theory.

First, our research points to the importance of extending ambidextrous leadership theory to include non-verbal behaviors (Schyns & Mohr, 2004). We used a design with high internal validity and "confirmatory power" to replicate the results of previous studies that tested ambidextrous leadership theory. While developing the stimulus material, we become aware that leader ambidexterity may go beyond "behavior" alone. To elaborate, we used a video approach

(in Study 2) to manipulate leader behavior in a way that was high in both internal and external validity. To ensure high internal validity, we took particular care to ensure that contaminating factors such as the affective tone of the leader, posture, gaze, and body movements were kept constant across all video conditions (which was difficult to realize for our actress). However, one could question whether such "pure" behaviors, while high in internal validity, really reflect what ambidextrous scholars might have intended when introducing the concept. That is, in the field (and in terms of external validity), a leader showing opening behaviors will most likely also show high levels of positive affect and an open body language, whereas a leader showing closing behaviors will most likely exhibit more negative affect and dominating body movements. Unfortunately, ambidextrous leadership theory does not articulate any propositions regarding affective tone, body movements, or gestures (Rosing et al., 2011); however, it is possible these aspects are in fact crucial to the theory, as affective tone has been identified as a crucial mechanism in driving both exploration behavior (Knight, 2015) and innovation (Madrid, Totterdell, Niven, & Barros, 2016.). We recommend that more attention be paid to the role of emotions and other non-verbal aspects in order to revise (and improve) the theory in the future.

Second, our research provides important insights regarding the limitations or potential boundary conditions of ambidextrous leadership theory. Replication studies that find heterogenous effect sizes across relatively comparable study designs have the potential to identify the boundary conditions (and moderators) of psychological theories (Klein et al., 2014). For example, we only found support for closing leadership being positively related to exploitation behavior in one experiment. Interestingly, the association between closing leadership and exploitation behavior was particularly salient in the population of crowd-workers. In that regard, it is interesting to consider the work characteristics of crowd-workers (Brawley &

Pury, 2016), as their work environment features specific elements that may be particularly receptive towards a closing leadership style: Tasks are limited in scope, skill utilization, and time investment (most "contracted jobs" on MTurk are therefore called *micro-tasks*, as they only take a few seconds or minutes; Aguinis & Lawal, 2013). Furthermore, while running the experiment, we noticed that crowd-workers are strongly concerned about their rejection rates (which influences the likelihood of being "hired" for future tasks). Performance on platforms such as MTurk is constantly monitored (for errors or carelessness; Cheung et al., 2017). In this context, a closing leadership style (e.g., sanctioning errors, monitoring behaviors, controlling adherence to rules) might be particularly salient and thus likely to promote exploitation behaviors. Relatedly, the crowd-working context might also be a reason why we found no support for the hypotheses that link opening leadership behavior and follower behavior to innovation outcomes, given that such opening behavior is highly counter-cultural and therefore not salient to the workers. To conclude, ambidextrous leadership theory could profit from including work design-related components to determine under which circumstances ambidextrous leadership will prove effective.

Third, our research offers some novel insights with respect to the potential side effects of opening behaviors. The second experiment showed that participants' careless responding scores were highest in the opening condition (although, overall, absolute values of careless responding were quite low). It is noteworthy that Rosing et al. (2011) articulated differences between closing and opening with respect to managing errors: The opening style incorporates a lenient error management style, whereas closing leadership incorporates a severe error management style (e.g., "sanctioning errors"; Rosing et al., 2011). Although the theory originally made no propositions as to whether an opening leadership style would also promote errors, our study

tentatively indicates that opening behaviors could promote carelessness. While less attention to detail can be beneficial when it comes to generating innovative work outcomes, future research should further explore whether opening behavior in isolation might be problematic under certain circumstances, such as within safety-critical environments.

Fourth, our replication—which was intended to be constructive (Köhler & Cortina, 2019) with respect to previous empirical research on ambidextrous leadership—makes an important contribution to the larger body of the replication literature within the field of management. Köhler and Cortina (2019) defined constructive replications as research that includes at least one methodological improvement, thus allowing a superior test of hypotheses. More specifically, (a) comprehensively constructive replications seek to solve *all* of the most important weaknesses of previous tests; (b) substantially constructive replications seek to solve *some*, but not all, of the previous weaknesses of previous studies; and (c) incrementally constructive replications seek to solve only a *single* (or more minor) flaw(s) in previous studies. Crucially, a recent review of more than 400 top-management publications revealed that only 1.7% of the reviewed replications were independent and incrementally constructive (Köhler & Cortina, 2019). Moreover, none of the reviewed studies were substantially (or comprehensively) constructive *(*Köhler & Cortina, 2019). Since we consider our replication to be at least substantially constructive, our research is unique within the field of management: To illustrate our attempts to be substantial constructive, we used an experimental design that allows for stronger conclusions regarding causality, used objective innovation ratings, reduced common-method bias, and applied a more critical analytical approach (i.e., 2SLS). While conducting this research, our goal was to maintain elements of the original studies (i.e., using a sample with similar characteristics and keeping our measures as close as possible to the original measures of exploration and

exploitation; cf. Köhler & Cortina, 2019). We also took particular care to ensure that not only our data but also the experimental materials are accessible on the OSF framework (https://osf.io/buvq8/files), and we encourage others to reanalyze this data (to ensure reproducibility; Aguinis, Ramani, & Alabduljader, 2018) as well as to reuse our validated experimental materials (i.e., text and video manipulations of leadership styles) for direct/close replication purposes (Köhler & Cortina, 2019). To conclude, we contribute to the replication movement in the leadership field (Clapp-Smith et al., 2018) by providing a rare example of a substantially constructive replication.

Practical implications

In terms of practical implications for leaders and managers, our findings provided some support that ambidextrous leadership might stimulate innovation outcomes (based on Study 2), albeit not necessarily via increasing exploration and exploitation behaviors on the part of followers. Moreover, leader opening behaviors seem to be just as useful for stimulating innovation as using an ambidextrous style (which requires a combination of opening with closing behaviors). When practitioners and organizations are deciding whether to invest in training their leaders to adopt an ambidextrous leadership style or training them in transformational leadership, our research implies that investing in ambidextrous (or just in opening) leadership would be slightly superior (bearing in mind that the differences in achieving better innovation outcomes with opening/ambidextrous leadership as opposed to transformational leadership are only small).

Furthermore, leaders should keep in mind that ambidextrous leadership may not be helpful in all situations (e.g., when followers are working in a low-innovation work environment, such as piecemeal crowd work). Furthermore, leaders who are working in safety-critical

environments (Griffin & Hu, 2013) should consider being cautious in adopting an opening leadership style, as it may stimulate higher levels of carelessness in their followers.

More importantly, our research demonstrates how the translation of research findings into practical recommendations should be articulated (and read) with caution. In other words, the findings of empirical observational research that does not address endogeneity may falsely inform practitioners (Antonakis, 2017). To illustrate, our studies show that addressing potential omitted causes (either by using an experimental design or by using analytical approaches such as 2SLS) can considerably change the estimated causal relationships when testing a theory. A scientific fact such as a correlation (e.g., "researcher XY found an association between follower ambidexterity and innovation") may not reflect a causal relationship. This reported association from researcher XY should thus not be translated into a statement such as "managers should stimulate follower ambidexterity to achieve innovative outcomes."

Limitations

As with all studies, this research has limitations that should be considered and addressed in future research endeavors.

First, by translating the basic tenets of ambidextrous leadership theory into the context of an experiment, we had to make some choices that could be considered as oversimplifying the ambidextrous leadership concept. In particular, ambidextrous leadership theory proposes that frequent and situationally adequate changes between opening and closing behaviors are necessary to promote innovation. Our experiment included only a single change from opening to closing behaviors in the ambidextrous leadership condition, and this change was independent of participants' behavior. Future replications may thus want to use more complex representations of ambidextrous leadership. Studying temporal trajectories in combination with intervention studies

and/or field experiments could provide alternative operationalizations of ambidextrous leadership theory that are methodologically rigorous (McClean, Barnes, Courtright, & Johnson, 2019; Podsakoff & Podsakoff, 2019). Clearly, these experimental field studies are difficult to realize, as it is challenging to (i) find organizations that are willing to run randomized experiments with their leaders, (ii) to exert full control over the research process when working with organizational stakeholders, (iii) to collect the large samples that are necessary to detect significant effects, and (iv) to simultaneously maintain participant engagement over time when using time-intensive repeated measurements (Klonek, Gerpott, Lehmann-Willenbrock, & Parker, 2019; Podsakoff & Podsakoff, 2019). Another choice was that we had to develop an experimental paradigm to study the effect of different leadership behaviors on employee innovation outcomes. We attempted to incorporate central features of innovation identified by Hughes et al. (2018), such as the such as the implementation of ideas into the final flyer document (in contrast to generating ideas concerning how to develop a flyer); a strong utilitarian focus of the flyer that serves real organizational needs; and aspects of interpersonal, social and practical elements (i.e., the interaction with a leader during the process). Some critics may still argue that this paradigm constitutes a task more focused on creativity than innovation, while others may argue that our selected paradigm constitutes a routine task. Overall, the validity of the experimental paradigm is consequential for testing ambidextrous leadership theory. While we did establish the validity of this experimental task (see the section titled “experimental task”), future studies could attempt to develop different types of innovation tasks (borrowing from the economics literature, e.g., Ederer & Manso, 2013) to test the main propositions of ambidextrous leadership theory.

Second, one of our reviewers suggested that the video messages (in Study 2) could be perceived as an interruption to the work process (in particular, for the closing leadership condition). Therefore, we included a measure of perceived interruptions in our post-experimental questionnaire (in Study 2). We indeed found that the messages in the closing conditions were considered more disruptive to the process than messages in the other conditions. While we safeguarded against the negative effects of interruptions by measuring their effect and including them as control variables, we would encourage future research to consider alternative experimental realizations to minimize this effect.

Third, we acknowledge that the sample size for testing H4 (only in Study 2, but not in Study 1) was potentially underpowered—at least for detecting the (allegedly) very small effect identified in previous research. That is, we were unable to gather the intended sample size for Study 2. There were multiple reasons why the data collection for Study 2 was more challenging in comparison to Study 1. First, in the spirit of a constructive replication, we aimed to match our sample with those of previous studies. This entailed that participants had to meet certain criteria. The most significant hurdle was that we were searching for working participants (i.e., professionals), who proved much more difficult to recruit for a study outside of their workplaces than student populations. In that regard, our compensation (which was kept similar in both studies) was not a major incentive for working professionals to participate in Study 2. Second, due to the nature of the registered report, we only had a very limited timeframe of only a few months. Future researchers who embark on similar ventures should bear in mind that some of the proposed study features can be extraordinarily time-intensive and occasionally require multiple iterations. For example, implementing the design and materials for Study 2 involved recruiting a suitable bilingual actress, filming and editing the experimental conditions such that they only

involved a manipulation of the focal independent variable, translation of all study materials and recruitment material, integrating and management of data collection, implementation of the video footage into experimental software, including two validation studies of the video materials (German and English), and setting up research labs in two countries. Finally, researchers should also estimate time budgets conservatively when attempting to obtain non-self-report measures; for example, it took at least two months to coordinate and obtain the external innovation ratings of more than 600 printed flyers from four different raters. Besides the operational challenges in collecting the data for Study 2, it should also be noted that the effect size that was used to determine the intended sample size was based on a somewhat shaky foundation. That is, the sample size estimate for H4 was based on a single study (Zacher et al., 2016), which did not correct for potential threats of endogeneity. As discussed in our introduction, research that does not correct for omitted causes (i.e., threats of endogeneity) can result in biased parameter estimates. In other words, without correcting for omitted causes, the true estimate of an effect size could be smaller, larger, or even reverse (i.e., negative; Kennedy, 2008), and thus the required samples size could be smaller or larger. Hence, this renders the use of this effect size to obtain sample size estimates questionable.

Fourth, even though we used an instrumental variable approach (i.e., 2SLS) to overcome the problems associated with endogenous variables, our analyses showed that our selected instruments were rather weak and lead to substantial bias in the estimation.

References

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

Aguinis, H., & Gottfredson, R. K. (2010). Best-practice recommendations for estimating interaction effects using moderated multiple regression. *Journal of Organizational Behavior, 31*, 776 –786.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, *32*, 1033–1043.

Aguinis, H., & Lawal, S. O. (2013). eLancing: A review and research agenda for bridging the science–practice gap. *Human Resource Management Review*, *23*(1), 6-17.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, *12*(1), 83-110.

Alghamdi, F. (2018). Ambidextrous leadership, ambidextrous employee, and the interaction between ambidextrous leadership and employee innovative performance. *Journal of Innovation and Entrepreneurship*, *7*(1), 1–14.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, *28*(1), 5–21.

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., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity:

Problems and solutions. In D.V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93–117). New York: Oxford University Press.

Antonakis, J., D'adda, G., Weber, R., & Zehnder, C. (2015). *Just Words? Just Speeches?–On the Economic Effects of Charismatic Leadership*. HEC Lausanne.

Artz, K. W., Norman, P. M., Hatfield, D. E., & Cardinal, L. B. (2010). A longitudinal study of the impact of R&D, patents, and product innovation on firm performance. *Journal of Product Innovation Management*, *27*, 725–740.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91*, 340–345.

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.

Axtell, C. M., Holman, D. J., Unsworth, K. L., Wall, T. D., Waterson, P. E., & Harrington, E. (2000). Shopfloor innovation: Facilitating the suggestion and implementation of ideas. *Journal of Occupational and Organizational Psychology*, *73*, 265 –285.

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

Behrendt, P., Matz, S., & Göritz, A. S. (2017). An integrative model of leadership behavior. *The Leadership Quarterly, 28*, 229–244.

Bledow, R., Frese, M., Anderson, N., Erez, M., & Farr, J. (2009). A dialectic perspective on innovation: Conflicting demands, multiple pathways, and ambidexterity. *Industrial and Organizational Psychology*, *2,* 305–337.

Bless, H., & Burger, A. M. (2016). A closer look at social psychologists' silver bullet: Inevitable and evitable side effects of the experimental approach. *Perspectives on Psychological Science, 11,* 296–308.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Brawley, A. M., & Pury, C. L. (2016). Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, *54*, 531–546.

Camisón-Zornoza, C., Lapiedra-Alcamí, R., Segarra-Ciprés, M., & Boronat-Navarro, M. (2004). A meta-analysis of innovation and organizational size. *Organization Studies*, *25*, 331–361.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29,* 2156–2160.

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*(4), 347–361.

Christie, A., Barling, J., & Turner, N. (2011). Pseudo-transformational leadership: Model specification and outcomes. *Journal of Applied Social Psychology*, *41*(12), 2943–2984.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Clapp-Smith, R., Carsten, M., Gooty, J., Connelly, S., Haslam, A., Bastardoz, N., & Spain, S. (2018). Special registered report issue on replication and rigorous retesting of leadership

models. *The Leadership Quarterly*, 88, V–VI.

Cohen, J. E. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71*(2), 179–184.

Damen, F., Van Knippenberg, B., & Van Knippenberg, D. (2008). Affective match in leadership: Leader emotional displays, follower positive affect, and follower performance. *Journal of Applied Social Psychology, 38*, 868–902.

David Eccles School of Business (2018). Certificate in Innovative Leadership. Retrieved from http://eccles.utah.edu/programs/executive-education/certificates/#innovative-leadership on 1st of June 2018

Denti, L., & Hemlin, S. (2012). Leadership and innovation in organizations: A systematic review of factors that mediate or moderate the relationship. *International Journal of Innovation Management*, *16,* 1240007-1–1240007-20

Ederer, F., & Manso, G. (2013). Is pay for performance detrimental to innovation?. *Management Science*, *59*(7), 1496-1513.

Farr, J. L., Sin, H.-P., & Tesluk, P. E. (2003). Knowledge management processes and work group innovation. In L. V. Shavinina (Ed.), *The international handbook on innovation* (pp. 574–586). New York, NY: Elsevier Science.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191

Fayant, M. P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the

limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology*, 30, 125–130.

Farh, C. I., & Chen, Z. (2014). Beyond the individual victim: Multilevel consequences of abusive supervision in teams. *Journal of Applied Psychology*, *99*, 1074–1095.

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

Fleiss, J. L., & Shrout, P. E. (1978). Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, *43*, 259–262.

Gibson, C. B., & Birkinshaw, J. (2004). The antecedents, consequences, and mediating role of organizational ambidexterity. *Academy of Management Journal, 47*, 209–226.

Griffin, M. A., & Hu, X. (2013). How leaders differentially motivate safety compliance and safety participation: The role of monitoring, inspiring, and learning. *Safety science*, *60*, 196–202.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. doi:10.20982/tqmp.08.1.p023

Hentschel, T., Braun, S., Peus, C., & Frey, D. (2018). The communality-bonus effect for male transformational leaders–leadership style, gender, and promotability. *European Journal of Work and Organizational Psychology*, *27*, 112–125.

Hill, A. D., White, M. A., & Wallace, J. C. (2014). Unobtrusive measurement of psychological constructs in organizational research. *Organizational Psychology Review*, *4*, 148–174.

Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership*

Quarterly, 29, 549–569.

Hülsheger, U. R., Anderson, N., & Salgado, J. F. (2009). Team-level predictors of innovation at work: A comprehensive meta-analysis spanning three decades of research. *Journal of Applied Psychology, 94*, 1128–1145.

Keith, M. G., Tay, L., & Harms, P. D. (2017). Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. *Frontiers in Psychology*, *8*, 1359.

Keller, T., & Weibler, J. (2015). What it takes and costs to be an ambidextrous manager: Linking leadership and cognitive strain to balancing exploration and exploitation. *Journal of Leadership & Organizational Studies*, *22*, 54–71.

Kennedy, P. (2008). *A Guide to Econometrics* (6th ed.) Oxford, UK: Wiley-Blackwell.

Ketokivi, M., & McIntosh, C. N. (2017). Addressing the endogeneity dilemma in operations management research: Theoretical, empirical, and pragmatic considerations. *Journal of Operations Management*, *52*, 1–14.

Kinni , T. (2016). How to become an ambidextrous leader. Retrieved from *strategy+business Blog*, https://www.strategy-business.com/blog/How-to-Become-an-Ambidextrous-Leader?gko=91b76 on 1st of June 2018

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152.

Klonek, F., Gerpott, F. H., Lehmann-Willenbrock, N., & Parker, S. K. (2019). Time to go

wild: How to conceptualize and measure process dynamics in real teams with high-resolution. *Organizational Psychology Review. Advanced online publication.* Doi: doi.org/10.1177/2041386619886674

Knight, A. P. (2015). Mood at the midpoint: Affect and change in exploratory search over time in teams that face a deadline. *Organization Science*, *26*(1), 99–118.

Köhler, T., & Cortina, J. M. (2019). Play it again, Sam! An analysis of constructive replication in the organizational sciences. *Journal of Management. Advanced online publication, DOI:* 10.1177/0149206319843985.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, *49*(2), 433-442.

Litman, L., Robinson, Y., Moss, A.J., Gautam, R. (2018, Sept. 18). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Retrieved from https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science, 31*(3), 323-326.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

McClean, S. T., Barnes, C. M., Courtright, S. H., & Johnson, R. E. (2019). Resetting the clock on dynamic leader behaviors: a conceptual integration and agenda for future research. *Academy of Management Annals*, *13*(2), 479–508.

Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Rodgers, B.

(1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. *Personality and Individual differences*, *27*(3), 405–416.

Madrid, H. P., Totterdell, P., Niven, K., & Barros, E. (2016). Leader affective presence and innovation in teams. *Journal of Applied Psychology*, *101*(5), 673–686.

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

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2,* 71–87.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30–46. http://dx.doi.org/10.1037//1082-989X.1.1.30.

Mom, T. J., Van Den Bosch, F. A., & Volberda, H. W. (2009). Understanding variation in managers' ambidexterity: Investigating direct and interaction effects of formal structural and personal coordination mechanisms. *Organization Science*, *20*, 812–828.

Mumford, M. D., Scott, G. M., Gaddis, B., & Strange, J. M. (2002). Leading creative people: Orchestrating expertise and relationships. *The Leadership Quarterly*, *13,* 705–750.

Murphy, K. R., & Russell, C. J. (2017). Mend it or end it: Redirecting the search for interactions in the organizational sciences. *Organizational Research Methods*, *20*, 549–573.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2018). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and*

Psychology, 1–19.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7,* 657–660.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

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,* 539–569.

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management, 12,* 531–544.

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

Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., & Klinger, R. L. (2013). Are we really measuring what we say we're measuring? Using video techniques to supplement traditional construct validation procedures. *Journal of Applied Psychology*, *98*, 99–113.

Rafferty, A. E., & Griffin, M. A. (2004). Dimensions of transformational leadership: Conceptual and empirical extensions. *The Leadership Quarterly*, *15*(3), 329–354.

Rosing, K., Bledow, R., Frese, M., Baytalskaya, N., Johnson Lascano, J., & Farr, J. L. (2018). The temporal pattern of creativity and implementation in teams. *Journal of Occupational and Organizational Psychology*, *91*(4), 798–822. doi:10.1111/joop.12226

Rosing, K., Frese, M., & Bausch, A. (2011). Explaining the heterogeneity of the leadership-

innovation relationship: ambidextrous leadership. *The Leadership Quarterly, 22,* 956–974.

Rosing, K., Rosenbusch, N., & Frese, M. (2010). Ambidextrous leadership in the innovation process. In A. Gerybadze, U. Hommel, H. W. Reiners & D. Thomaschewski (Eds.), *Innovation and International Corporate Growth* (pp. 191–204). Berlin: Springer.

Rosing, K., & Zacher, H. (2017). Individual ambidexterity: The duality of exploration and exploitation and its relationship with innovative performance. *European Journal of Work and Organizational Psychology*, *26*(5), 694–709.

Rubera, G., & Kirca, A. H. (2012). Firm innovativeness and its performance outcomes: A meta-analytic review and theoretical integration. *Journal of Marketing, 76,* 130–147.

Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *The Leadership Quarterly*, 101348. Advanced online publication, https://doi.org/10.1016/j.leaqua.2019.101348.

Schroeder, R. G., Van de Ven, A. H., Scudder, G. D., & Polley, D. (1989). The development of innovation ideas. In A. H. Van de Ven, H. L. Angle & M. S. Poole (Eds.), Research on the management of innovation: The Minnesota studies (pp. 107–134). New York, NY: Oxford University Press.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90-100.

Schyns, B., & Mohr, G. (2004). Nonverbal elements of leadership behaviour. *German Journal of Human Resource Management*, *18*(3), 289–305.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating Cohen's f2, a measure of local effect size, from PROC MIXED. *Frontiers in psychology*, *3*, 111.

Shaver, J. M. (2005). Testing for mediating variables in management research: Concerns, implications, and alternative strategies. *Journal of Management*, *31*(3), 330–353.

Simonsohn, U. (2013). Evaluating replication results. Available at SSRN: http://ssrn.com/abstract=2259879

Sonnentag, S., Reinecke, L., Mata, J., & Vorderer, P. (2018). Feeling interrupted - being responsive: How online messages relate to affect at work. *Journal of Organizational Behaviour, 39*(3), 369-383.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*, 557–586.

Stock, R. M., von Hippel, E., & Gillert, N. L. (2016). Impacts of personality traits on consumer innovation success. *Research Policy*, *45*, 757–769.

Stock, R. M., Zacharias, N. A., & Schnellbaecher, A. (2017). How do strategy and leadership styles jointly affect co-development and its innovation outcomes? *Journal of Product Innovation Management*, *34*, 201–222.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. Andrews, & J. H. Stock (Eds.). *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197.

Tushman, M. L., & O'Reilly, C. A. (1997). *Winning through innovation*. Cambridge: Harvard Business School Press.

Tushman, M. L., Smith, W. K., & Binns, A. (2011). The ambidextrous CEO. *Harvard Business*

Review, 89, 74–80.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13,* 411–417.

Venus, M., Stam, D., & Van Knippenberg, D. (2013). Leader emotion as a catalyst of effective leader communication of visions, value-laden messages, and goals. *Organizational Behavior and Human Decision Processes, 122*, 53-68.

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*, 231–263.

Welbourne, T. M., Johnson, D. E., & Erez, A. (1998). The role-based performance scale: Validity analysis of a theory-based measure. *Academy of Management Journal, 41*, 540-555.

West, M. A., & Farr, J. L. (1990). *Innovation and creativity at work: Psychological and organizational strategies.* Chichester: Wiley.

Zaccaro, S. J., & Horn, Z. N. (2003). Leadership theory and practice: Fostering an effective symbiosis. *The Leadership Quarterly*, *14*, 769–806.

Zacher, H., Robinson, A. J., & Rosing, K. (2016). Ambidextrous leadership and employees' self-reported innovative performance: The role of exploration and exploitation behaviors. *The Journal of Creative Behavior*, *50*, 24–46.

Zacher, H., & Rosing, K. (2015). Ambidextrous leadership and team innovation. *Leadership & Organization Development Journal*, *36*, 54–68.

Zacher, H., & Wilden, R. G. (2014). A daily diary study on ambidextrous leadership and self-reported employee innovation. *Journal Of Occupational & Organizational Psychology, 87,* 813–820.

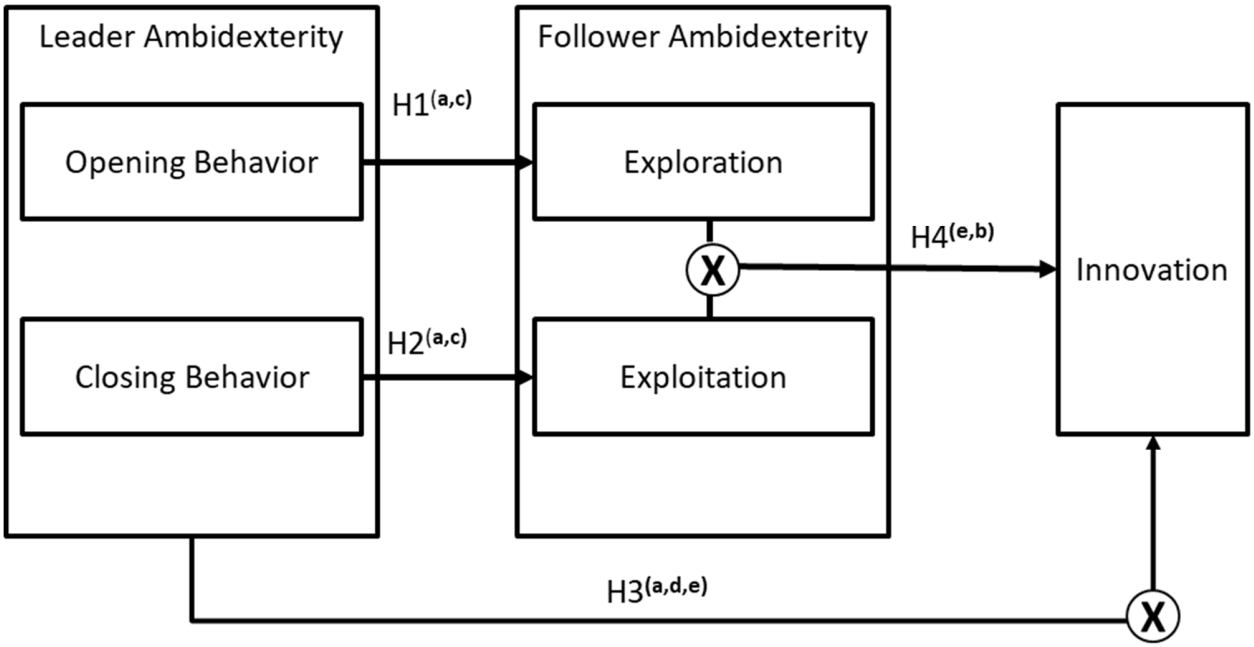
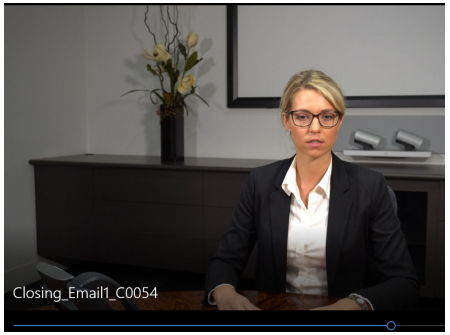


Figure 1. Model of leader ambidexterity for innovation (including empirical support for model paths).

Note: Empirical support for each hypothesis is provided in brackets; superscript letters = direct support
a = Alghamdi (2018)
b = Rosing & Zacher (2017)
c = Zacher et al. (2016)
d = Zacher & Rosing (2015)
e = Zacher & Wilden (2014)

Opening (English)

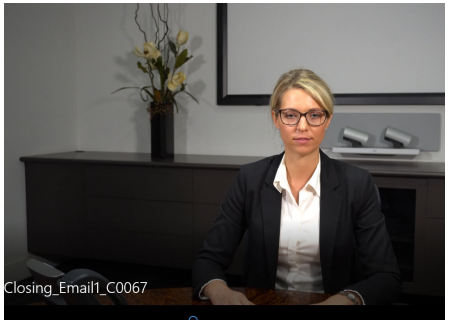
Closing (English)



Transformational (English)

Opening (German)

Closing (German)



Opening (German)

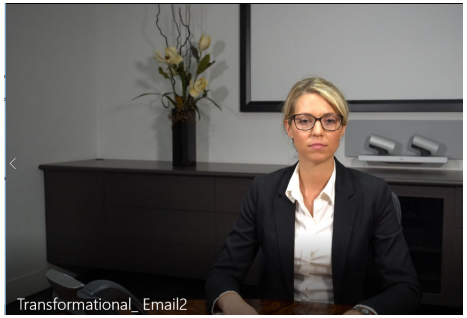


Figure 2. Screenshots of experimental manipulations of leadership styles in Study 2.

Note: Ambidextrous leadership was manipulated by presenting first the opening message and, second, the closing video message within the experiment. The file name (bottom left for each pictures) was not shown during the actual experiment.

Table 1. Literature review outlining the paths of ambidextrous leadership theory that have been tested in previous research.

Publication	Sample	Design	Control Variables	Measures (IV)	Measures (DV)	Findings	Support of Model Paths
Alghamdi (2018) *Journal of Innovation and Entrepreneurship*	• N = 147 faculty members working in Albaha province, Saudi Arabia	• Cross-sectional • Self-report survey • Single source (from employees) • Correlational	• *Leader:* age, gender, educational level, faculty position	• Employees rated supervisors perceived opening (7 items) and closing behaviors (6 items)[a] • Leader ambidexterity = opening*closing	• Employee-rated exploration (5 items) and exploitation (6 items)[b] • Employee-rated innovation (4 items)[c]	• Opening behaviors → employee exploration • Closing behaviors → employee exploitation • Opening*Closing leader behaviors → Employee innovation	• H1, H2, H3
Rosing & Zacher (2017) *European Journal of Work and Organizational Psychology*	• Study 1 (weekly study): N = 59 employees in Australia • Study 2 (daily study): N = 37 employees working in creative industries in northern Germany	• 2 diary studies: Study 1 (weekly, six waves); Study 2 (daily, 5 waves) • Single-source (all employee-rated)	• *Employee:* Positive weekly affect, negative weekly affect, innovation requirements, age, gender, education	• Employee-rated exploration (5 items) and exploitation (6 items)[b]	• Employee-rated innovative work performance (6 items)	• Employee ambidexterity → employee innovative work performance	• H4
Zacher et al. (2016) *The Journal of Creative Behavior*	• N = 388 employees • Recruited via MTurk (United States)	• Cross-sectional • Single-source survey (from employees) • Between-subjects • Correlational	• *Leader:* Transf. leadership, transact. leadership • *Employee*: open., conscient., trait positive affect	• Perceived leader opening (7 items) and closing behaviors (6 items)[a] • Self-rated exploration (5 items) and	• Self-rated innovation performance (4 items)[c] • Self-rated exploration (5 items) and exploitation (4 items)[b]	• Opening behaviors → Employee exploration • Closing behaviors → Employee exploitation • Employee exploration, exploitation, and exploration*exploitation (controlling for opening and closing behaviors) → self-reported innovation	• H1, H2, H4

				exploitation (4 items)[b]			
Zacher & Rosing (2015) *Leadership & Organization Development Journal*	• N = 33 team leaders and 90 of their employees from Australia	• Cross-sectional • Dual source (ratings from leaders and employees) • Team-level analysis • Between-team comparison • Correlational	• *Leader:* Transf. leadership • *Team:* General team success	• Employees rated perceived leader opening (7 items) and closing behaviors (7 items) • Controls: transf. leadership	• Team leaders rated team innovative performance (4 items)[c] • Team leaders rated team success (1 item)	• Opening behaviors → team innovation • Closing behaviors (→X) team innovation • Opening*Closing leader behaviors → team innovation	• H3
Zacher & Wilden (2014) *Journal of Occupational and Organizational Psychology*	• N = 113 employees • Convenience sampling, recruited through personal contacts and research participant pools (no information about nationality) • Baseline survey & daily survey (five workdays)	• Diary study (within-person and between-person) • Single-source self-report survey (from employees) • Correlational	• *Leader*: Intellectual stimulation • *Employee:* Positive affect, job autonomy	• Perceived leader opening (4 items) and closing behaviors (4 items)[a]	• Self-rated innovation performance (4 items)[c]	• Opening behaviors → Daily self-reported innovation • Closing behaviors (→ X) daily innovation • Opening*Closing behaviors → daily innovative performance	• H3
Current Study	• Study 1 (online experiment): N = 395, recruited via MTurk and prolific • Study 2 (lab experiment): N = 229 recruited in Australia and Germany	• Two randomized experiments (between-person design); online experiment with manipulated emails (Study 1), lab video	• 2SLS with instrumented variables (manipulated conditions, positive trait affect, conscient., openness,)	• Experimentally manipulated leader behaviours (opening, closing, ambidextrous, transf. leadership; see Appendix I and O); validation	• Participant-rated exploration (7 items) and exploitation (6 items) (see also Table 2) • External raters independently rated innovation performance of the	• Opening behaviors → exploration behavior (only partial support in Study 2) • Opening leadership (vs. closing leadership or vs. transf. leadership) → innovation outcomes • Ambidextrous leadership (vs. closing leadership or vs. transf. leadership) → innovation outcomes	• H1 (partial support in Study 2), H2 (only in Study 1), H3 (only partially supported in Study 2), H4 (not supported)

	experiment (Study 2) • Multi-method (self-report exploration/ exploitation behaviour, external ratings)	study for all experimental materials (Appendix P)	submitted outcomes (4 items)[e]

Notes. To reduce the complexity of this table, we only report variables in each study that are central to the aims of this replication study. transf. = transformational, transact. = transactional, open. = openness to experience, conscient. = conscientiousness, IV = predictor variables, DV = dependent variables, → means "positively predicted", → *X* means "did not have a significant effect", * = "Multiplicative interaction between two variables", [a] = items based on behavioral descriptions from Rosing et al. (2011), [b] = items used from Mom et al. (2009), [c] = items used from Welbourne et al. (1998), [d] = see Table 2 (Column 4 for items), [e] = items from Rosing et al. (2018); see also Appendix H.

Table 2. Conceptual definitions of exploration/exploitation, items from field research, and adapted items for the experiment

Construct definition		Managerial (Mom et al., 2009)	Employee (non-managerial) (Zacher et al., 2016).	Items adapted items for the experiment
Individual exploration: "behaviors related to experimentation, searching for alternative ways to accomplish task, and learning from errors. When exploring, individuals deviate from routines, try out something new, and do not rely on established knowledge." (Rosing & Zacher, 2017, p. 696)		*To what extent did you, last year, engage in work-related activities that can be characterized as follows:*	*Rate the extent to which you engaged in the following activities at work:*	*"During the task,*
	1	Searching for new possibilities with respect to products/services, processes, or markets.	Searching for new possibilities with respect to my work	I searched for novel ways to make the flyer more interesting.
	2	Evaluating diverse options with respect to products/services, processes, or markets	Evaluating diverse options with respect to my work	I evaluated diverse options with respect to the flyer.
	3	Focusing on strong renewal of products/services or processes	Focusing on strong renewal of products/services or processes	I focused on strong renewal of the flyer.
	4	Activities requiring quite some adaptability of you	Activities requiring me to be adaptable	I had to be adaptable.
	5	Activities requiring you to learn new skills or knowledge	Activities requiring me to learn new skills or knowledge	I was trying to learn something new.
	6	Activities that are not (yet) clearly existing company policy	--	I engaged in activities that were not formally required by the task description.
	7	Activities of which the associated yields or costs are currently unclear	--	I tried to experiment with different methods to reach the goal.
Individual exploitation: "encompasses relying on previous experience, putting things into action, and incrementally improving well-learned actions. Exploitation involves doing things as they have always been done and relying on existing rules and routines." (Rosing & Zacher, 2017, p. 696)	1	Activities which serve existing (internal) customers with existing services/products	Activities which serve existing customers with existing products/ services	I maintained the existing format and existing text of the flyer draft version.
	2	Activities of which it is clear to you how to conduct them	Activities which I clearly know how to conduct	Focused mainly on carrying out those task activities that were provided in the task description (i.e., adding pictures and colors).
	3	Activities primarily focused on achieving short-term goals	Activities primarily focused on achieving short-term goals	Focused on getting the task done as quickly as possible.
	4	Activities which you can properly conduct by using your present knowledge	Activities I can properly conduct using my existing knowledge	I only conducted those activities which I knew how to conduct.
	5	Activities which clearly fit into existing company policy	Activities which clearly fit into existing company policy	I strictly adhered to the rules and fulfillment of task requirements.
	6	Activities which you carry out as if it were routine	--	Focused on implementing those things that were required.
	7	Activities of which a lot of experience has been accumulated by yourself	Activities in which I have accumulated a lot of experience	

Table 3. Analytical approach and predicted differences for the experimental studies.

Nr.	Hypothesis	Expected results	Expected contrasts
1	*Opening leader behaviors positively predict follower explorative behaviors.*	Significant F-value ($p < .05$) with *follower exploration* as dependent variable and leadership style as independent variable (four conditions: opening, closing, ambidextrous, transformational)	*Dependent variable: Follower exploration* Contrasting specific conditions: Opening condition > Closing condition ($p < .05$) Opening condition > Transformational condition ($p < .05$) Opening condition > or = Ambidexterity condition (p = na)[a]
2	*Closing leader behaviors positively predict follower exploitative behaviors.*	Significant F-value ($p < .05$) with *follower exploitation* as dependent variable and leadership style as independent variable (four conditions: opening, closing, ambidextrous, transformational)	*Dependent variable: Follower exploitation* Contrasting specific conditions: Closing condition > Opening condition ($p < .05$) Closing condition > Transformational condition ($p < .05$) Closing condition > or = Ambidexterity condition (p = na)[a]
3	*The interaction of leader opening and closing behaviors (i.e., leader ambidexterity) positively predicts employee innovation insofar that innovation is highest when both leader opening and leader closing behaviors are high.*	Significant F-value ($p < .05$) with *innovation outcome* as dependent variable and leadership style as independent variable (four conditions: opening, closing, ambidextrous, transformational)	*Dependent variable: External ratings of innovation outcomes* Contrasting specific conditions: Ambidextrous condition > Closing condition ($p < .05$) Ambidextrous condition > Opening condition ($p < .05$) Ambidextrous condition > Transformational condition ($p < .05$)
4	*The interplay of follower exploration and exploitation behavior (i.e., follower ambidexterity) positively predicts innovation outcomes insofar that innovation is highest when both follower exploration and exploitation behaviors are high.*	Significant F-value ($p < .05$) for the 2SLS model with *innovation outcome* as dependent variable and follower activities as endogenous variables (three variables: exploration, exploitation, follower ambidexterity), which are instrumented by the manipulated conditions	*Dependent variable (Y): External ratings of innovation outcomes* X1 = Exploration† X2 = Exploitation† X3 = Follower ambidexterity (exploration*exploitation) 2SLS $Y = B + \beta_1 X1 + \beta_2 X2 + \beta_3 X3$

Note. [a] = Leadership ambidexterity theory (and empirical research) does not allow making specific predictions about this effect; † = in a first stage, both the endogenous variables X1 and X2 were regressed on the instrumental variables to obtain predicted values of X1 and X2 in order to purge endogeneity biases.

Table 4. Correlation matrix all study variables (Study 1)

		M	SD	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1)	Opening Leader Condition	0.26	0.44	-.35**	-.33**	-.35**	0.02	-0.06	0.04	-0.05	-.11*	0.07	-0.01	0.03
(2)	Closing Leader Condition	0.26	0.44	na	-.32**	-.34**	-0.01	0.00	0.04	0.05	.153**	-0.04	0.07	-0.02
(3)	Ambidextrous Leader Condition	0.23	0.42		na	-.32**	-0.01	-0.06	0.02	-0.04	-0.07	0.05	-0.02	0.05
(4)	Transformational Leader Condition	0.25	0.43			na	-0.01	.116*	-.099*	0.04	0.02	-0.07	-0.05	-0.06
(5)	Positive Trait Affect	3.28	0.79				.85[a]	-.225**	.301**	.218**	-0.01	.25**	.20**	-0.05
(6)	Negative Trait Affect	1.41	0.62					.90[a]	-.230**	0.01	0.00	-0.01	-0.01	0.00
(7)	Conscientiousness	4.03	0.60						.82[a]	.185**	-0.05	.21**	.14**	-0.03
(8)	Openness	3.84	0.75							.84[a]	0.04	.15**	.15**	0.00
(9)	Exploitation	4.92	0.88								.66[a]	-0.06	.60**	-.17**
(10)	Exploration	4.89	1.07									.82[a]	.75**	.21**
(11)	Follower Ambidexterity	24.03	6.69										na	.03
(12)	Innovation rating (average across four raters)	1.57	0.57											.93[b]

Note: $N = 395$, [a] = Cronbach's alpha, [b] = Intraclass correlation, Experimental conditions are dummy coded (e.g., 1 = yes, 0 = no); $^{*}p<.05$, $^{*}p<.01$

Table 5. Means, standard deviations and F-statistics across four leadership conditions (Study 1)

	Opening leadership		Closing leadership		Ambidextrous leadership		Transform. leadership		Total		Test statistic	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*
Exploitation	4.76	0.8	5.15	0.93	4.82	0.99	4.96	0.76	4.92	0.88	4.04	0.008
Exploration	5.02	1.08	4.82	1.07	4.98	1.06	4.76	1.05	4.89	1.07	1.37	0.251
Innovation rating	1.6	0.56	1.54	0.56	1.62	0.71	1.51	0.46	1.57	0.57	0.68	0.562

Note: $N = 395$, F-values were calculated based on ANOVA

Table 6. Correlation matrix for study variables (Study 2)

	M	*SD*	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) Opening Leader Condition	0.26	0.44	-.36**	-.36**	-.31**	-0.04	0.10	-0.10	0.01	.18**	-0.01	-.21**	-0.01	0.08	0.06	.13*
(2) Closing Leader Condition	0.26	0.44	na	-.36**	-.31**	0.06	-0.05	.16*	0.01	0.06	0.02	.16*	0.09	-0.08	0.00	-.13*
(3) Ambidextrous Leader Condition	0.27	0.44		na	-.31**	0.00	-0.12	0.04	-0.01	-0.12	-0.01	-0.08	-0.08	-0.03	-0.10	0.08
(4) Transformational Leader Condition	0.21	0.41			na	-0.02	0.08	-0.10	-0.01	-0.12	0.00	.16*	0.00	0.04	0.04	-0.09
(5) Positive Trait Affect	3.39	0.64				.82[a]	-0.03	.29**	.15*	-0.04	-.17*	-0.04	-0.11	.15*	0.05	0.02
(6) Negative Trait Affect	1.67	0.61					.83[a]	-0.05	0.00	0.09	-0.05	-0.04	0.04	0.07	0.12	0.09
(7) Conscientiousness	3.76	0.56						.75[a]	.14*	-0.03	-.15*	0.06	0.03	0.09	0.12	0.02
(8) Openness	3.43	0.66							.77[a]	-0.09	0.07	0.03	-0.03	.25**	.19**	0.00
(9) Careless Responding Score	0.18	0.41								na	0.10	-0.04	0.03	-0.02	-0.01	-0.01
(10) Language (German vs. English)	0.21	0.40									na	0.08	-0.02	.17**	0.12	-.24**
(11) Interruption through messages	2.76	1.03										.80[a]	0.09	.155*	.18**	0.04
(12) Exploitation	4.50	0.97											.67[a]	-.23**	.55**	-.15*
(13) Exploration	4.42	1.07												.78[a]	.66**	.19**
(14) Follower Ambidexterity	19.66	5.62													na	0.03
(15) Innovation Rating (Mean across four raters)	2.18	0.86														.91[b]

Note: $N = 229$, [a] = Cronbach's alpha, [b] = Intraclass correlation; Experimental conditions are dummy coded (e.g., 1 = yes, 0 = no); Coding of Language (0 = English; 1 = German); na = not applicable, *p<.05, *p <.01.

Table 7. Means, standard deviations and F-statistics across four leadership conditions (Study 2)

	Opening leadership		Closing leadership		Ambidextrous leadership		Transform. leadership		Total		Test statistic	
	M	SD	M	SD	M	SD	M	SD	M	SD	*F*	*p*
Exploitation	4.49	0.95	4.65	1.01	4.37	0.98	4.49	0.95	4.50	0.97	0.82	0.485
Exploration	4.56	1.05	4.27	1.14	4.36	1.18	4.50	0.85	4.42	1.07	0.89	0.449
Innovation Rating	2.37	0.94	2.00	0.76	2.29	0.84	2.04	0.84	2.18	0.86	2.77	0.042
Interruption through messages	2.39	0.91	3.04	1.13	2.63	0.95	3.03	1.00	2.76	1.03	5.733	0.001
Careless Responding Score	0.30	0.50	0.22	0.45	0.10	0.30	0.08	0.28	0.18	0.41	3.764	0.012

Note: $N = 229$, F-values were calculated based on ANOVA

Table 8. OLS regression with interaction effects predicting innovation (Study 2)

		Outcome: Innovation Rating (Mean for 4 raters)			
Model	*F*	*B*	*SE(B)*	*t*	*p*
Step 1	0.05				
Positive Trait Affect		0.02	0.09	0.23	0.82
Conscientiousness		0.02	0.11	0.23	0.82
Openness		-0.01	0.09	-0.09	0.93
(Constant)		2.05**	0.49	4.19	0.00
Step 2	2.84*				
Positive Trait Affect		-0.01	0.09	-0.11	0.91
Conscientiousness		-0.01	0.10	-0.11	0.92
Openness		0.02	0.09	0.25	0.81
Opening leader (Condition)		0.39*	0.17	2.33	0.02
Closing leader (Condition)		-0.04	0.16	-0.24	0.81
Ambidextrous leader (Condition)		0.29†	0.16	1.77	0.08
Careless Responding		0.00	0.14	0.01	0.99
Language (German vs. English)		-0.53*	0.14	-3.76	0.00
Interruptions		0.10†	0.06	1.76	0.08
(Constant)		1.86**	0.51	3.65	0.00
Step 3	3.82**				
Positive Trait Affect		-0.07	0.09	-0.78	0.43
Conscientiousness		-0.02	0.10	-0.23	0.82
Openness		-0.03	0.08	-0.40	0.69
Opening leader (Condition)		0.36*	0.16	2.22	0.03
Closing leader (Condition)		0.02	0.16	0.15	0.88
Ambidextrous leader (Condition)		0.29†	0.16	1.85	0.07
Careless Responding		0.01	0.14	0.11	0.92
Language (German vs. English)		-0.62**	0.14	-4.45	0.00
Interruptions		0.08	0.06	1.37	0.17
Exploration		0.16**	0.06	2.95	0.00
Exploitation		-0.10†	0.06	-1.71	0.09
(Constant)		2.08**	0.58	3.60	0.00
Step 4	3.56**				

Positive Trait Affect	-0.07	0.09	-0.82	0.41
Conscientiousness	-0.01	0.10	-0.12	0.90
Openness	-0.03	0.08	-0.40	0.69
Opening leader (Condition)	0.35*	0.16	2.18	0.03
Closing leader (Condition)	0.02	0.16	0.12	0.91
Ambidextrous leader (Condition)	0.28†	0.16	1.75	0.08
Careless Responding	0.01	0.14	0.07	0.94
Language (German vs. English)	-0.62**	0.14	-4.47	0.00
Interruptions	0.07	0.06	1.26	0.21
Exploration	0.34†	0.20	1.70	0.09
Exploitation	0.07	0.20	0.36	0.72
Ambidexterity (Exploration*Exploitation)	-0.04	0.04	-0.90	0.37
(Constant)	1.33	1.01	1.31	0.19

Note. † for p<.10, * for p<.05, ** for p<.01, *** for p<.001