Consulting Psychology Journal: Practice and Research
2016, Vol. 68, No. 1, 72–98

1065-9293/16/$12.00 http://dx.doi.org/10.1037/cpb0000059

GENDER BIAS AND SUBSTANTIVE DIFFERENCES IN RATINGS OF LEADERSHIP BEHAVIOR: TOWARD A NEW NARRATIVE

Robert B. Kaiser
Kaiser Leadership Solutions, Greensboro, North Carolina

Wanda T. Wallace
Leadership Forum, Inc., Durham, North Carolina

Women make up about half of the U.S. workforce but fewer than 15% of corporate officers and 5% of CEOs. Much popular discussion about this disparity focuses on unconscious bias against women in leadership roles. The current study analyzed the presence of gender bias as distinct from substantive gender differences in ratings of the leadership behavior of a matched sample of 857 women and 857 men from upper-levels of management in 6 companies, representing 5 industries, based in the United States, Western Europe, and Australia. The results found virtually no evidence of bias against women leaders and some evidence of bias *in favor* of them. Further, the analysis of substantive gender differences in behavior indicated that women used a more forceful-operational style associated with the tactical management of execution whereas men used a more strategic-enabling style associated with senior organizational leadership. The findings are interpreted in terms of changing attitudes toward women in leadership and how current approaches to achieving parity in upper management may overlook the importance of the broadening career experiences women need to develop strategic leadership skills.

Keywords: gender bias, glass ceiling, leadership, sex differences, women and leadership

There has been much spirited discussion lately about women in leadership. The subject is frequently in the media, from best-selling books like *Lean In* (Sandberg, 2013) and *The Confidence Code* (Kay & Shipman, 2014) to regular articles in popular magazines, newspapers, and websites such as a recent special section in the *Wall Street Journal*, "What's Holding Women Back" (Waller & Lublin, 2015). The subject is also of great corporate concern. Most large companies have made a priority

Editor's Note. Larry Norton served as the action editor for this article.

Robert B. Kaiser, Kaiser Leadership Solutions, Greensboro, North Carolina; Wanda T. Wallace, Leadership Forum, Inc., Durham, North Carolina.

Robert B. Kaiser has a commercial interest in the *Leadership Versatility Index* described in this article.

Correspondence concerning this article should be addressed to Robert B. Kaiser, Kaiser Leadership Solutions, 1903-G Ashwood Court, Greensboro, NC 27455. E-mail: rob@kaiserleadership.com or Wanda.Wallace@leadershipforuminc.com

out of increasing the number of women in leadership roles and many have established diversity and inclusion offices to focus on it.

A lot of this interest comes from a glaring disparity: although women make up nearly half of the American workforce (United States Department of Labor, 2014), among S&P 500 companies only one in seven corporate officers and fewer than one in twenty chief executive officers is a woman (Catalyst, 2015). These demographic trends are not just an American problem; the figures are similar throughout Western societies across Europe and Oceania, although some notable variations exist (e.g., Scandinavian countries have higher rates of women participating in corporate and public leadership). The gender agenda is a global issue.

The Popular Narrative

Discussions about why women are underrepresented in senior leadership seem to rely on a common narrative to tell a story that provides cultural significance and a shared understanding of the problem (Dailey & Browning, 2014). Consequently, this narrative also guides problem-solving efforts and shapes the kinds of solutions that get considered and implemented.

A Paradox

The popular narrative about women in contemporary leadership invokes a paradox. First, whereas the 20th century workplace was hierarchical, the modern workplace is more collaborative. Flatter structures, team-based work designs, increased demographic diversity, greater cross-cultural interaction, enhanced communication technologies, and a faster pace of change have put a premium on the ability of people to work together (Mohrman & Cohen, 1995; National Research Council, Committee on Techniques for Enhancement of Human Performance, 2001). The levelling of hierarchy also means that employees are motivated more by influence and relationships than by the command and control model that dominated the industrial age (Leonard, 2003).

This more collaborative work environment, continues the narrative, creates a female leadership advantage because the naturally more relational, empathic, and inclusive style of women makes them better at working with others and fostering a more cooperative atmosphere (Eagly & Carli, 2003; Maccoby, 2007; Yukl, 2010). Although this claim may seem like an overgeneralization, it is often stated as a matter of fact in the popular media. For example, a headline in the New York Times read "No Doubt: Women are Better Managers" (Smith, 2009) and one in the Daily Mail read "Women in Top Jobs are Viewed as 'Better Leaders' Than Men" (2010). A *Psychology Today* article explained "why women may be better leaders than men" because their leadership style is "more suited to modern organizations" (Williams, 2012).

Despite the advantageous conditions for women, the narrative concludes, few of them get to lead because there is an invisible barrier of bias preventing women from advancing into top jobs. Thirty years ago this phenomenon was called the "glass ceiling" by the former editor of Working Woman magazine, Gay Bryant (Frenkiel, 1984), a term that resonated and became a popular cultural shorthand. More recently this barrier has been described in terms of "unconscious bias," following the way influential companies in Silicon Valley have adopted modern cognitive theory to explain the mechanisms behind the lack of women in science, technology, engineering, and mathematics careers as well as leadership tracks ("Exposing Hidden Bias at Google," 2014). This argument relies on such theories as the "think manager, think male" paradigm (Schein, 2001), the lack of fit model (Heilman, 2001), and role-congruity theory (Eagly & Karau, 2002) along with constructs such as the double-standard (Foschi, 1996; Lyness & Thompson, 2000) and the double-bind (Ibarra, Ely, & Kolb, 2013) to explain commonly held prejudicial beliefs that women are not as capable of succeeding in masculine-oriented leadership roles. These beliefs are thought to operate below the level of awareness and influence the way women are perceived and evaluated by important organizational gate-keepers and thereby determine what kinds of career opportunities they are offered—or denied.

A Question of Utility

Although the popular narrative on why women are not better represented in senior leadership is compelling and consistent with some research findings, we question how helpful it is in achieving gender parity. One reason for the persistence of this narrative is that it too relies on easily activated stereotypes in providing a simple explanation for a complex problem. But these stereotypes may not apply. According to a meta-analytic summary of dozens of published studies, laboratory research with student samples does show that women use more people-oriented leadership behaviors and men use more task-oriented leadership behaviors; however, field research with actual managers finds no such gender differences (Eagly & Johnson, 1990). The leadership behavior of men and women in upper-management jobs is not stereotypical like the argument for a female leadership advantage assumes (Vecchio, 2002).

By contrast, a review of the literature on gender differences in leadership noted that there are differences on nongender-stereotyped behaviors (Eagly & Carli, 2003). For instance, a meta-analysis of transformational and transactional leadership styles (Eagly, Johannesen-Schmidt, & van Engen, 2003) found that women were rated higher on the transactional dimension of contingent reward (providing reinforcement for satisfactory performance) whereas men were rated higher on the transactional dimension of management by exception (reacting to problems rather than proactively addressing them) and the nonleadership dimension, *laissez faire* (an uninvolved, disengaged style). Women also were rated somewhat higher on transformational leadership. The strongest effects were the higher ratings for men on the more reactive and passive dimensions suggesting that the bigger gender difference in leadership behavior is the greater hands-on and involved style of women. However, this theme is not part of the popular narrative.

Moreover, the popular narrative is disempowering: what are aspiring female managers to do about unconscious biases that exist in someone else's mind? Unlike the voice in a positive story arc (Yost, Yoder, Chung, & Voetmann, 2015), this one takes the locus of control away from women and places it in other people and the surrounding cultural milieu. There is also the pragmatic question of whether the popular narrative has reached its limits in terms of influencing policy solutions to achieve parity. Despite enormous effort, the number of women officers and CEOs in the S&P 500 has increased less than 5% since the turn of the century (Catalyst, 2015). Few committed to gender equality are satisfied with this level of progress.

In this article, we report analyses of a unique data set involving a matched sample of upper-level women and men managers from six different corporations. The findings are provocative and in some ways are inconsistent with, or at least not part of, the popular narrative on why so few women are in senior leadership. We suspect the findings may reflect how we have entered a new phase in the process of social change, one that possibly requires an updated narrative that better represents how far we have come and, more importantly, what work remains to facilitate the ascendancy of more qualified women into leadership roles.

The Present Study

This project began as applied research utilizing existing assessment data gathered for various leadership development programs to determine whether it could also provide secondary information to help companies understand the challenges facing their women managers. Each of the six companies in the present sample is a customer of Kaiser Leadership Solutions, a provider of tools for the assessment and development of managers and executives. One of these tools is the *Leadership Versatility Index* (see Kaiser, Overfield, & Kaplan, 2010), a 360-degree feedback instrument that measures a model of complementary leadership behaviors using a unique rating scale. Kaiser Leadership provides a service of strategic talent analytics for large-volume customers, where assessment data from many participants are aggregated and systematically studied to understand the organization's leadership culture.

Many organizations are concerned about their lack of women in upper management, and one type of analysis Kaiser Leadership offers is called a "gender audit." Gender audits use established

research methodologies from the study of bias in psychological tests to identify bias in the ratings of women leaders. Once bias is statistically isolated, substantive differences between women and men leaders can then be analyzed to identify distinct patterns of behavior. In other words, the analyses are designed to distinguish unfair bias in how women leaders are rated from real differences in their leadership behavior compared with men.

Methodology

The gender audit methodology involves first identifying all the women available in the database for a given company, and then carefully selecting a matched sample of men from the same company. There are typically many more men in the database, and men are chosen for inclusion in the study on a case-by-case basis based on their match with an individual woman in terms of hierarchical level, function, age, management experience, and tenure in the present job. This is a crucial methodological design feature because the matched-sample approach effectively controls for these demographic variables and rules them out as competing explanations for apparent evidence of gender bias and gender differences (Lyness & Thompson, 1997; Privitera, 2015). For example, women often have less management experience than men due to taking time off from a career to bear and raise children (Ruderman & Ohlott, 2002), and differences in experience could affect the way women leaders are perceived as well as their leadership skill.

Two unique features of the measurement tool, the *Leadership Versatility Index* (LVI) 360, also are an important part of the research methodology. First, the LVI measures two pairs of opposing but complementary dimensions of behavior. One pair concerns interpersonal style and is "gender stereotypical"; it contrasts a masculine forceful style with a more feminine enabling style. The other pair concerns organizational issues and is "gender neutral"; it contrasts a big-picture focus on strategy with a tactical focus on execution (see Table 1 for further definitions and specificity for the LVI leadership model). Prior research has focused much more on gender differences in the stereotypical aspects of interpersonal style, with little attention given to differences in the nonstereotypical organizational issues women and men focus upon.

Table 1
Dimensions, Subdimensions, and Definitions for the Leadership Model

Dimension/subdimension	Definition
Forceful	Asserting personal and position power
Take charge	Assuming authority and control
Decisive	Taking a position and speaking up
Demanding	Holding people to high standards
Enabling	Creating conditions for others to contribute
Empowering	Giving people autonomy
Participative	Being open to input and influence
Encouraging	Providing support to people
Strategic	Positioning the organization for the future
Direction	Setting the course
Expansion	Growing the organization
Innovation	Supporting change and creativity
Operational	Focusing the organization on short-term results
Execution	Driving implementation
Efficiency	Conserving resources
Order	Using process discipline

Note. From *Leadership Versatility Index: A User Guide for Version 5.0*, by R. B. Kaiser and D. V. Overfield, in press, Greensboro, NC: Kaiser Leadership Solutions. Copyright 2016 by Kaiser Leadership Solutions, LLC. Adapted with permission.

Second, the rating scale on the LVI is different from the typical five-point scale where higher scores are assumed to be "better." The LVI scale, presented in Figure 1, gives respondents the option of rating each behavior as used "too little," "the right amount," or "too much" to various degrees. The distinction between "the right amount" versus "too little" or "too much" is left to the individual rater to decide based on a subjective evaluation of perceived behavior (Kaiser & Kaplan, 2005)—which provides leeway for the operation of unconscious biases to influence the ratings.

Theoretically, one may expect there to be greater evidence of bias on the forceful and enabling dimensions than the strategic and operational dimensions because forceful and enabling behaviors are so closely linked to gender stereotypes. Common prejudicial attitudes and beliefs may affect how much forceful and enabling behavior is considered "the right amount" for women leaders. The double-bind argument (Ibarra et al., 2013) suggests that women leaders are "damned if they do and damned if they don't": if they behave in a masculine manner they are penalized for violating expectations for how a woman should behave, but if they behave in a feminine manner they are penalized for not matching the masculine ideal associated with being "leader-like." Women who adopt a more stereotypically masculine style may be seen as too forceful and not enough enabling whereas women who adopt a more stereotypically feminine style may be seen as too enabling and not forceful enough.

The double-standard argument (Foschi, 1996; Lyness & Thompson, 2000) suggests that different levels of these behaviors may be tolerated for men compared to women. For instance, an aggressive man may be regarded as "assertive," but an equally aggressive woman may be seen as "pushy" (Heilman, Wallen, Fuchs, & Tamkins, 2004). A participative man may be lauded as a modern, power-sharing, and inclusive leader whereas a participative woman may be criticized as indecisive and dependent on input from others. In other words, what is "the right amount" of forceful and enabling behavior for a male leader could be regarded as "too much" for a female leader.

Research Questions

Gender audits were conducted one organization at a time, and insights from previous studies were used to refine the design and execution of later studies. In the first few studies the intention of the research team as well as representatives of the organizations who commissioned the research was to find evidence of bias. The working hypothesis was that the ratings of women were influenced by unconscious bias and our goal was to identify the nature and source of that bias. Guided largely by

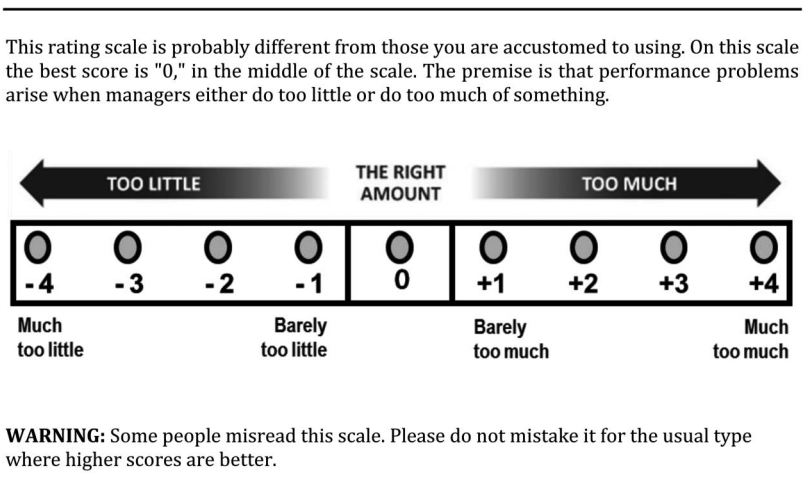

Figure 1. The "Too Little/Too Much" rating scale. Reproduced from *Leadership Versatility Index version 4.1 survey,* Greensboro, NC: Kaiser Leadership Solutions. Copyright, 2010–2016 by Kaiser Leadership Solutions. Used with permission from the publisher. U.S. Patent number 7,121,830.

the popular narrative, we were looking for the "smoking gun" through the use of a rigorous statistical methodology applied to a carefully constructed dataset.

What we found was surprising, and shaped the way we approached later studies by not seeking to prove a particular hypothesis but rather trying to learn about contemporary gender issues in organizational leadership through the systematic and relatively objective observations afforded by our research design. That is, the later analyses were conducted from a problem-oriented perspective as "hypothesis generating" rather than from a theory-oriented perspective as "hypothesis testing" (Lawrence, 1992).

We next report a single, overall analysis combining the data from the six original studies in this same spirit of exploratory research using a broadly representative Western sample of corporate managers. We were interested in two general research questions:

1. Is there evidence of bias against women in ratings of their leadership behavior?
2. Are there substantive differences in the leadership behavior of women and men?

We treat the findings as empirical observations and, combining our experience in working with organizations on diversity goals and women leaders on their development along with recent findings from other applied leadership research on gender, we elaborate an interpretation of the findings and their implications in the Discussion section.

Method

Sample

All data were collected between 2009 and 2014 to provide feedback to participants in various leadership development programs. Participants were from six different organizations, in five different industries, based in the United States, Western Europe, and Australia. The organizations are all large, publicly traded corporations and are described in Table 2 in terms of industry, location, and the number of matched pairs of women and men leaders from each.

A total of 1,714 leaders—857 matched pairs of women and men—were selected for inclusion in the present study based on a number of considerations. Each leader has ratings from the "full circle"—superiors, peers, and subordinates as well as self-ratings. Each leader also has the demographic information needed to pair up a similarly aged, experienced, tenured, and senior male for every female. Descriptive statistics for the demographic variables are presented in Table 3. Participants were mostly executives (with titles like CxO, President, Vice-president) and directors (heads of functional units), and a minority were managers working within a functional area. Women and men were identically represented at these three hierarchical levels of management and there were no significant mean differences on the age, experience, and tenure variables according to a one-way analysis of variance (ANOVA).

Table 2
Organizations Represented in the Sample

Industry	Location	No. of matched pairs
Technology	United States	92
Manufacturing	United States	192
Financial services	United States	127
Financial services	Australia	96
Pharmaceuticals	Western Europe	188
Energy	Western Europe	162
		857

Note. All organizations are large, publicly traded corporations.

Table 3
Descriptive Statistics for Demographic Variables

Variable	Women leaders			Men leaders		
	N	*M*	*SD*	*N*	*M*	*SD*
Age		42.28	7.51		42.34	6.76
Management experience		10.88	6.59		11.15	6.04
Tenure in present job		2.79	2.73		2.86	2.46
Organizational level						
Executive	382			382		
Director	268			268		
Manager	207			207		

Note. $N = 857$ women and 857 men leaders.

Measures

A total of 23,066 coworkers—including 4,190 superiors, 8,679 peers, and 10,197 subordinates—provided ratings on the LVI. The LVI is a 360-degree feedback instrument that contains four primary scales concerning Forceful, Enabling, Strategic, and Operational behaviors that are composed of 12 items each. Each primary scale is composed of three 4-item subscales for a total of 12 subscales and 48 items. The LVI has been extensively researched and refined; ample evidence exists for its reliability and validity as a psychometrically sound measure that shows the expected patterns of convergent and discriminant relations with other measures of leader behavior (see reviews in Buros Mental Measurements Yearbook: Staal, 2008; Vassar, 2008).

The four dimensions of behavior measured by the LVI can be understood in terms of their conceptual and empirical relationships with existing constructs (see Kaiser, Overfield, & Kaplan, 2010). For instance, Yukl's (2010) taxonomy distinguishes interpersonally-oriented, task-oriented, and change-oriented categories of leader behavior. The LVI's Forceful and Enabling dimensions map onto Yukl's interpersonally-oriented category because interpersonal behavior is defined in terms of a dominant and assertive dimension juxtaposed against an accommodating and nurturing dimension (Wiggins & Trapnell, 1996). Forceful leadership is also correlated with the initiating component of the initiating structure construct and Enabling is correlated with the consideration construct in the classic two-factor model of leader behavior (Stogdill & Coons, 1957). The LVI Operational dimension maps onto Yukl's task-oriented category which concerns organizing and planning for the execution of initiatives, and is correlated with the structuring component of the initiating structure construct. And the LVI Strategic dimension maps onto Yukl's change-oriented category which concerns adapting to shifting environmental demands, establishing new directions, and introducing new organizational structures and processes.

Construct validity evidence is further provided in relationships with theoretically relevant personality traits; for example, in terms of the Hogan Personality Inventory (HPI; Hogan & Hogan, 2007), HPI Ambition correlates most highly with Forceful, HPI Interpersonal Sensitivity with Enabling, HPI Inquisitive with Strategic, and HPI Prudence with Operational behaviors; the seven HPI scales together account for about a quarter of the variance in coworker ratings on the four LVI behavior scales (Kaiser & Hogan, 2011). Further, ratings on the LVI are related to a broad range of leadership outcomes and effectiveness indicators, including employee engagement, motivation, and commitment; team cohesiveness and collective confidence; and business unit productivity (see review in Kaiser, Overfield, & Kaplan, 2010; see also Kaiser, McGinnis, & Overfield, 2012; Kaiser & Overfield, 2010).

Descriptive statistics and internal consistency reliability estimates for the four primary LVI scales and the 12 subscales based on individual, rater-level data are presented separately for self-ratings and coworker ratings in Table 4.

Table 4
Descriptive Statistics and Reliabilities for Leader Behavior Scales and Subscales

	Self-ratings			Coworker ratings		
Scales/subscales	α	*M*	*SD*	α	*M*	*SD*
Forceful	.83	+.09	.65	.87	−.05	.64
Take charge	.69	+.25	.76	.75	+.02	.76
Decisive	.70	+.10	.83	.74	+.01	.75
Demanding	.74	−.09	.87	.80	−.18	.80
Enabling	.79	−.03	.53	.85	−.19	.51
Empowering	.75	−.06	.67	.83	−.18	.68
Participative	.61	+.02	.59	.73	−.25	.62
Encouraging	.70	−.04	.79	.76	−.15	.64
Strategic	.86	−.20	.61	.87	−.27	.52
Direction	.81	−.36	.75	.84	−.32	.67
Expansion	.75	−.29	.83	.78	−.31	.71
Innovation	.72	+.04	.64	.74	−.20	.54
Operational	.70	−.16	.47	.70	−.09	.40
Execution	.67	+.08	.71	.69	−.07	.62
Efficiency	.54	−.20	.57	.62	−.07	.55
Order	.62	−.35	.69	.64	−.11	.56

Note. $N = 1{,}714$ self-raters and 23,066 coworkers. Computed at the individual rater-level of analysis, as used in the Phase 1 analysis of bias.

Procedures

The analyses were conducted in three phases. The first two phases examined bias from two complementary perspectives, and the third examined substantive differences in the leadership behavior of women and men. To make it easier to follow methods and their associated findings, we describe statistical and analytical procedures along with the findings for each phase in the Results section.

Results

Our analytic strategy for identifying gender bias was based on the procedures used to identify racial bias in cognitive tests developed over concerns about adverse impact due to IQ testing for employment (see Jensen, 1980, for a review). According to the Society for Industrial and Organizational Psychology (SIOP, 2003), "In a statistical context, (bias is) a systematic error in a score. In discussing fairness, bias refers to variance due to contamination or deficiency that differentially affects the scores of different groups of individuals" (p. 66). According to this definition, bias is identified statistically when a measure systematically overestimates or underestimates the value of the variable it is intended to measure for members of a particular population (Aguinis, Culpepper, & Pierce, 2005; SIOP, 2003).

Two types of analysis are used to study test bias. The first is to identify *prima facie* evidence of bias: do members of different groups (e.g., Blacks vs. Whites; men vs. women) score differently? In and of itself, *prima facie* evidence does not necessarily indicate bias—it is possible that the two groups really are different on the attribute being measured—but it is used to alert the possibility of unfair bias. Therefore, in the second analysis the scores are studied for evidence of *differential prediction* by examining whether scores for one group are associated with different levels of performance outcomes compared to the other group (Aguinis et al., 2010; Cleary, 1968; SIOP, 2003). For instance, bias is indicated when the job performance of black candidates with certain IQ scores is lower than the job performance of white candidates

with the same scores. In this case, the IQ test *under*predicts the performance of black candidates and is therefore biased against them (Aguinis et al., 2010; Cleary, 1968; SIOP, 2003).

Phase 1: *Prima Facie* Evidence of Bias

We analyzed whether women and men leaders were systematically rated different by certain groups of raters—for instance, do male coworkers rate women leaders more harshly than female coworkers rate them? If so, ratings from male coworkers would seem biased against women leaders. Do superiors rate women leaders more harshly than subordinates rate them? If so, ratings from superiors, who act as gatekeepers to promotions and other advancement opportunities (Lyness & Heilman, 2006), would appear biased. Do male superiors rate women leaders particularly harshly? If so, ratings from male superiors in particular may be biased.

As recommended by Vecchio (2002), these analyses focused on leader-coworker dyads as the unit of analysis. They were conducted on rater-level data with a 2 (Leader gender) × 2 (Rater gender) × 4 (Rater source: self, superior, peer, subordinate) Multiple Analysis of Variance (MANOVA) statistical model. Two MANOVAs were conducted, one using the four primary scales—Forceful, Enabling, Strategic, and Operational—as dependent variables, and the other using the 12 subscales as dependent variables. There were methodological and substantive reasons for conducting two separate MANOVAs. Methodologically, the scales are composed of the subscales and thus including both scales and subscales in one analysis would violate the independence of observations assumption (Tabachnick & Fidell, 2007). Substantively, we wanted to determine whether bias was pervasive and evident in the broad dimensions or whether bias was subtle and localized to more narrowly defined subdimensions.

According to this analytic strategy, *prima facie* evidence of bias may be indicated by a significant two-way interaction effect involving leader gender and either rater gender or rater source or by a significant three-way interaction effect involving leader gender, rater gender, and rater source. A total of 48 possible interactions with leader gender were tested (3 independent variables × 16 dependent variables [4 dimensions + 12 subdimensions]).

Results from the multivariate tests are presented in Table 5. Significant main effects were observed for all three independent variables in both MANOVAs.[1] There was also a significant two-way interaction between Rater gender and Rater source at the subscale level of analysis. Follow-up univariate ANOVAs indicated that this effect was due to men leaders rating themselves higher on the Empowering subscale than women leaders rated themselves or than coworkers rated

[1] The main effect for leader gender indicated that women and men leaders were rated differently but consistently across rater genders and rating sources; the interpretation of these results is reserved for the third set of analyses on substantive gender differences. Follow-up univariate ANOVA tests indicated that the main effect for rater gender was due to male raters providing lower mean scores on Strategic leadership [$d = -.07$ standard units lower; $F(1, 13) = 11.40$, $p < .001$; especially the subscales Innovation ($d = -.07$; $F(1, 13) = 13.16$, $p < .01$) and Direction ($d = -.07$; $F(1, 13) = 8.91$, $p < .001$) and to a lesser extent Expansion ($d = -.05$; $F(1, 13) = 4.26$, $p < .05$)]. Female raters provided lower mean scores on Operational leadership [$d = -.06$; $F(1, 13) = 11.21$, $p < .001$; especially the subscales Order ($d = -.08$; $F(1, 13) = 13.65$, $p < .001$) and Efficiency ($d = -.08$; $F(1, 13) = 7.58$, $p < .01$)]. Follow-up univariate ANOVA tests were also used to interpret the main effects for rater source. Significant mean differences were observed for all four scales (and all 12 subscales): Forceful, $F(3, 13) = 26.98$, $p < .001$, Enabling, $F(3, 13) = 21.89$, $p < .001$, Strategic, $F(3, 13) = 34.20$, $p < .001$, Operational, $F(3, 13) = 9.87$, $p < .001$. The overall trend showed that the mean scores for self-ratings were higher than the other groups for Forceful and Enabling (especially the Take charge, Decisive, and Participative subscales which were, respectively, $d = .33$, .15, and .41 standard units higher than the grand mean of the other rater groups). However, self-ratings were lower for Operational (especially the Order subscale, $d = -.46$). The mean superior score was the lowest on Strategic and Forceful (especially the Direction, Expansion, and Demanding subscales, which were, respectively, $d = -.31$, $-.21$, and $-.22$ standard units lower) but highest on Operational (especially the Order subscale, $d = .29$). The effects for rater source were larger in size than the effects for rater gender, accounting for 1.5% of the multivariate variance in scale scores and 3.0% of it in subscale scores (vs. .1% and .2%, respectively, for rater gender). The rater source effects were largely due to differences in self-ratings; mean scores for the superior, peer, and subordinate groups were more similar to one another.

Table 5
Multivariate Tests for Prima Facie Evidence of Gender Bias

Variable	Wilks' λ	*F*	df_1	df_2	*p*
Dependent variables: 4 scales					
Leader gender	.99	43.95	4	24,668	**.000**
Rater gender	1.00	4.89	4	24,668	**.001**
Rater source	.96	93.69	12	65,266	**.000**
Leader gender × Rater gender	1.00	1.02	4	24,668	.397
Leader gender × Rater source	1.00	1.09	8	49,336	.367
Rater gender × Rater source	1.00	1.68	8	49,336	.126
Leader gender × Rater gender × Rater source	1.00	1.44	8	49,336	.172
Dependent variables: 12 subscales					
Leader gender	.99	21.68	12	23,754	**.000**
Rater gender	1.00	4.60	12	23,754	**.000**
Rater source	.91	62.10	36	70,185	**.000**
Leader gender × Rater gender	1.00	1.21	12	23,754	.268
Leader gender × Rater source	1.00	1.03	24	47,508	.423
Rater gender × Rater source	1.00	2.04	24	47,508	**.002**
Leader gender × Rater gender × Rater source	1.00	.71	24	47,508	.842

Note. Bold value indicates significant effect, $p < .05$.

either men or women leaders ($d = .26$ standard units higher compared to the grand mean of all other raters; $F(2, 13) = 4.14$, $p < .05$).

The effects of primary interest were interactions involving leader gender. Neither the Leader gender by Rater gender nor the Leader gender by Rater source two-way interactions nor the three-way Leader gender by Rater gender by Rater source interaction were significant in either MANOVA. These results suggest that there was not a robust trend where women leaders disproportionately were rated less favorably by male raters or a particular rater group.

Nonetheless, the statistical software package we used generated follow-up univariate ANOVAs that indicated a total of four significant interactions with leader gender: one in the scale-level and three in the subscale-level analyses. Four is 8.3% of the 48 total possible gender interactions. Evidently the interaction effects were two few and too small to cumulatively achieve significance in the multivariate analyses (Tabachnick & Fidell, 2007). Each accounted for less than half of one percent (.005) of the multivariate variance. Therefore, we describe these interactions for illustrative purposes, but are cautious about placing too much significance on them since the omnibus multivariate effects were not significant.

The significant interaction at the scale level in the univariate ANOVAs was between Leader gender and Rater source on ratings of Operational, $F(2, 13) = 3.29$, $p < .05$. This interaction indicated that all rating sources rated women leaders about $d = .12$ standard units higher than they rated men leaders, except subordinates who only rated them $d = .07$ higher.

There were two 2-way and one 3-way significant interactions in the univariate ANOVAs in the subscale analyses. A two-way interaction between Leader gender and Rater gender on the Participative subscale, $F(1, 13) = 5.04$, $p < .05$ indicated that female raters rated women leaders the highest and men leaders the lowest ($d = .11$), whereas male raters rated women and men leaders much more similarly ($d = .04$). The other two-way interaction, between Leader gender and Rater source on the Efficiency subscale, $F(2, 13) = 3.60$, $p < .05$, indicated that all rater groups rated women marginally higher than men leaders (ds = .01 to .03) except peers, who rated women leaders lower than men leaders ($d = -.06$). Finally, the three way interaction was on the Direction subscale, $F(2, 13) = 3.46$, $p < .05$. Male and female raters in most rater groups rated men leaders higher than women (ds = .06 to .11). Male peers showed the same trend, but female peers did not: they rated men leaders lower than they rated women leaders ($d = -.04$).

Taken together, these analyses provide virtually no *prima facie* evidence of bias against women leaders. None of the interactions with leader gender were statistically significant in the omnibus, multivariate analyses. In the univariate analyses, which must be interpreted with caution because of the increased potential for Type I errors (Tabachnick & Fidell, 2007) and the modest size of the effects, only four of the 48 interaction effects indicated the possibility of gender bias. Only one of these interactions could be interpreted as suggesting bias against women (where male peers rated women leaders disproportionately lower on Efficiency). In other words, there was far greater consistency than difference in how male and female raters from the various coworker groups rated women leaders.

Phase 2: Evidence of Differential Prediction

The next set of analyses examined whether the ratings of leadership behavior were differentially related to criterion variables representing leadership effectiveness. In these analyses, bias is indicated by differences in the regression lines relating behavior and effectiveness for women compared to men leaders (Aguinis et al., 2010; Cleary, 1968; SIOP, 2003). The regression lines may differ in two distinct ways: they could have different intercepts and/or different slopes. These differences are illustrated in Figure 2.

Graph A in Figure 2 depicts regression lines with different intercepts. In this example, the ratings of behavior would be biased against women because they underestimate the effectiveness of women leaders compared with men leaders who have the same behavior rating. Another way to see the bias is that women have to be rated higher on the behavior to be equally effective as men, which is a reflection of the double-standard bias (Foschi, 1996; Lyness & Thompson, 2000). Graph B depicts regression lines with different slopes. In this example, the ratings of behavior are valid for men leaders, but not for women leaders; the ratings of the behavior are biased against women because they are unrelated to their effectiveness. Finally, Graph C depicts regression lines with different intercepts and slopes, which combines both forms of differential prediction.

Leader behavior scores. To simplify the presentation of these analyses, we first focus on the overall effect where ratings of the Forceful, Enabling, Strategic, and Operational scales were combined to form a single summary variable called the Versatility score (see Kaiser, Overfield, & Kaplan, 2010, or Kaiser & Overfield, 2010 for the computational details and validation of this procedure). Versatility scores represent how closely leaders were rated to doing "the right amount" of each behavior and range from 0% to 100%, where higher scores indicate more ratings closer to "the right amount."

Because the *prima facie* analyses found little evidence that some classes of coworkers rated the women and men leaders differently than other classes, we felt justified in aggregating the behavior ratings across female and male raters into a composite score that equally represented the superior, peer, and subordinate perspectives. This approach has been used in previous research and has been shown to produce scores that are more reliable and valid than relying on ratings from any one single source (Oh & Berry, 2009). We computed scores by first calculating the mean rating across raters within the superior, peer, and subordinate groups and then calculated the grand mean across all three groups for each target manager.

Statistical justification for the rating aggregation was sought by calculating the degree of rating similarity using the $r_{wg(j)}$ interrater agreement coefficient (James, Demaree, & Wolf, 1984) and the one-way random effects intraclass correlation coefficient to estimate interrater reliability (McGraw & Wong, 1996). We calculated ICC(1) to estimate the reliability of an individual rater and ICC(2) to estimate the reliability of the average rating across k raters (where k equaled the median number of two for superiors, six for peers, five for subordinates, and three for the aggregate rating across all three sources). We computed these statistics within superior, peer, and subordinate groups and for the aggregation across the three groups (LeBreton, Burgess, Kaiser, Atchley, & James, 2003). As the results in Table 6 show, the degree of interrater agreement and reliability both within and across the

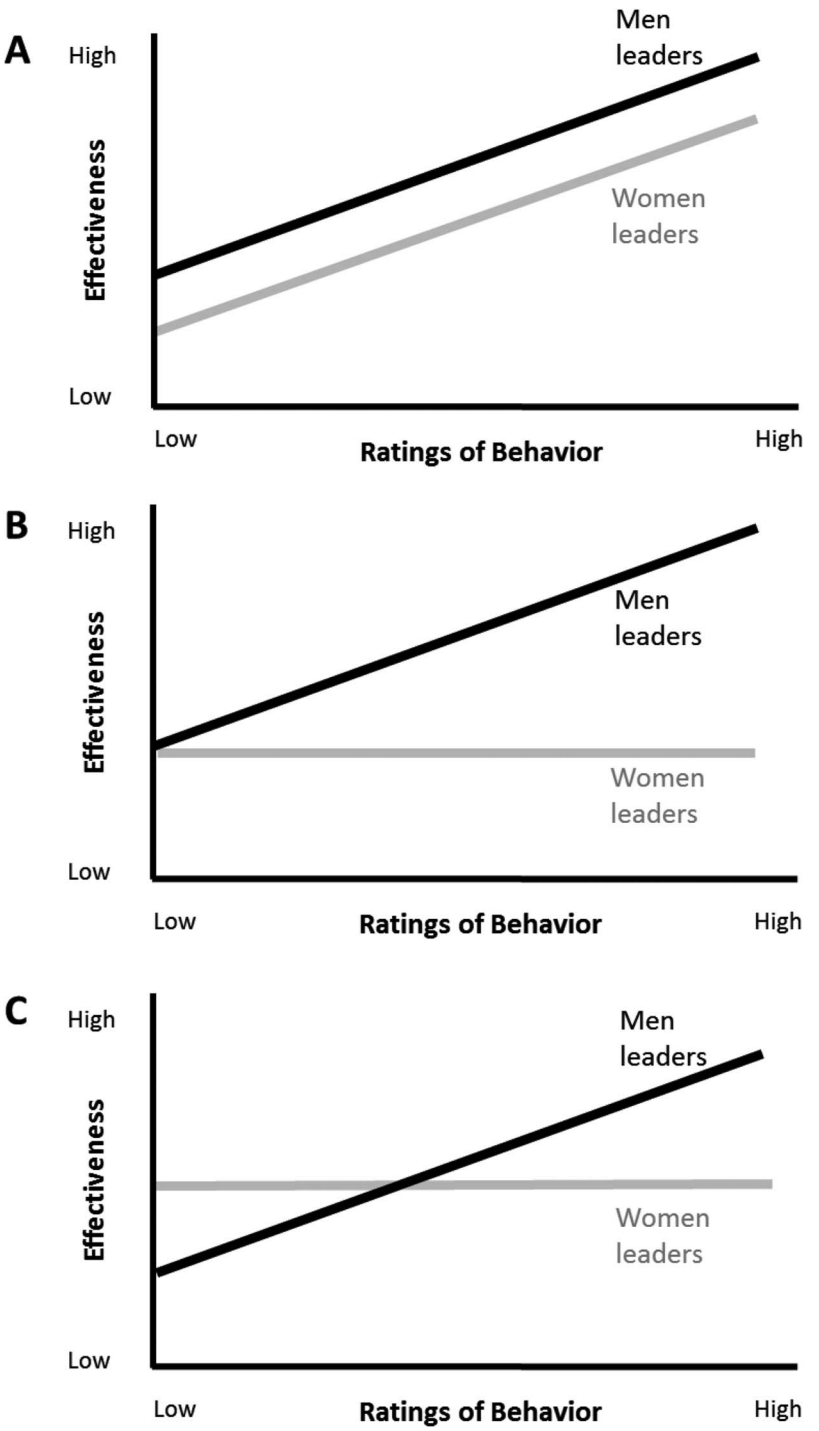

Figure 2. Examples of different ways that regression lines relating behavior and effectiveness variables could differ for women and men leaders.

Table 6
Inter-Rater Reliability and Inter-Rater Agreement for Leader Behavior Variables

	Superiors			Peers			Subordinates			Aggregated across sources		
Scale/subscale	ICC(1)	ICC(2)	$r_{WG(j)}$	ICC(1)	ICC(2)	$r_{WG(j)}$	ICC(1)	ICC(2)	$r_{WG(j)}$	ICC(1)	ICC(2)	$r_{WG(j)}$
Forceful	.34	.61	.96	.26	.72	.94	.27	.74	.94	.53	.67	.99
Take charge	.27	.54	.91	.19	.64	.87	.21	.67	.87	.51	.66	.98
Decisive	.32	.59	.90	.23	.69	.86	.23	.69	.88	.48	.63	.97
Demanding	.24	.48	.91	.18	.62	.87	.22	.68	.87	.40	.54	.97
Enabling	.22	.46	.97	.20	.65	.96	.18	.63	.96	.40	.55	.99
Empowering	.16	.37	.93	.14	.54	.91	.14	.56	.92	.28	.41	.98
Participative	.15	.36	.94	.15	.57	.92	.12	.50	.91	.29	.43	.98
Encouraging	.18	.40	.93	.19	.64	.92	.20	.65	.92	.40	.55	.98
Strategic	.21	.46	.97	.12	.50	.97	.16	.59	.97	.29	.43	.99
Direction	.19	.42	.92	.07	.36	.92	.11	.49	.92	.20	.31	.98
Expansion	.23	.48	.92	.15	.57	.90	.19	.63	.90	.37	.51	.97
Innovation	.10	.25	.95	.08	.39	.93	.10	.47	.94	.25	.38	.99
Operational	.15	.36	.98	.14	.54	.97	.18	.62	.96	.37	.52	1.00
Execution	.14	.34	.94	.13	.52	.91	.18	.62	.90	.33	.47	.98
Efficiency	.13	.31	.94	.10	.46	.92	.10	.46	.91	.23	.35	.99
Order	.10	.25	.95	.11	.48	.92	.17	.60	.92	.34	.49	.99

Note. ICC(2) was based on 2 raters for superior ratings, 6 for peer ratings, 5 for subordinate ratings, and 3 groups for ratings aggregated across sources. $r_{WG(j)}$ values represent the average $r_{WG(j)}$ statistic computed across all focal managers.

superior, peer, and subordinate sources was high and justified aggregation (LeBreton & Senter, 2008).

Leader effectiveness variables. Three measures were used as dependent variables to represent a broad view of leadership effectiveness—perceived effectiveness of the individual leader, team attitudes, and team results (Kaiser, Hogan, & Craig, 2008). *Perceived effectiveness* was measured with ratings to the item, "Please rate this manager's overall effectiveness as a leader on a 10-point scale where 5 is adequate and 10 is outstanding." These single-item ratings were aggregated as the grand mean of the mean superior, mean peer, and mean subordinate rating. We used the same procedure as above to determine whether there was sufficient similarity within and across rater groups to justify aggregation. Support was found with Mean $r_{wg} = .96$ and ICC(2) = .51 for the aggregated, 360-degree rating of perceived effectiveness.

Team attitudes were measured with subordinate ratings on a three-item scale labeled *Team Vitality* which represents the degree of morale, engagement, and cohesion among members of the teams for which the managers in our sample are responsible. Subordinate ratings were used because the construct of interest concerns *their* attitudes. Team results were measured by superior ratings on a three-item scale labeled *Team Productivity*, which represents the quantity, quality, and overall output of the teams or units for which the managers are responsible. Superiors are in perhaps the best position to gauge the productivity of the teams as they use this information to manage and evaluate the performance of the study participants.

The Team Vitality and Productivity items were rated on a traditional 5-point response scale, where higher ratings indicate more of the attribute. Prior research has shown that these scales are factorially distinct and valid measures that correlate as expected with similar measures of team performance (Kaiser et al., 2010). In the present sample, correlations among the three leadership effectiveness variables were statistically significant ($p < .001$) but small to moderate in size, suggesting some degree of convergence but also an adequate degree of discriminate validity: perceived effectiveness-vitality $r = .47$, perceived effectiveness-productivity $r = .51$, and vitality-productivity $r = .18$.

Table 7
Descriptive Statistics and Reliabilities for Variables Used in Tests of Differential Prediction

	Versatility score			Perceived effectiveness		Team vitality			Team productivity		
Subject	*M*	*SD*	α	*M*	*SD*	*M*	*SD*	α	*M*	*SD*	α
Women leaders	.86	.05	.95	7.96	.55	3.81	.52	.88	3.87	.49	.85
Men leaders	.85	.05	.96	7.97	.56	3.81	.53	.89	3.83	.52	.86
Overall	.86	.05	.96	7.97	.56	3.81	.53	.89	3.85	.51	.86

Note. $N = 857$ women, 857 men, and 1,714 overall leaders.

Descriptive statistics and reliabilities for the aggregated behavior variable and the three effectiveness variables are presented in Table 7. There were no significant gender differences in mean scores on these measures.

Analyses. We used the Cleary (1968) model to test for differential prediction of women compared to men leaders. This approach to assessing bias utilizes a moderated multiple regression framework and involves regressing a dependent variable (e.g., effectiveness) on three independent variables: a predictor (in the present case, ratings of behavior summarized in the Versatility score), a dichotomous group membership variable (leader gender, coded 0 for male and 1 for female), and the interaction between the predictor and group membership variables (the cross-product of the Versatility score times the gender variable). In these analyses, a significant effect for leader gender indicates different intercepts for the regression lines for the women compared to men leaders (e.g., graph A in Figure 2) and a significant effect for the Versatility score-leader gender interaction term indicates different slopes (e.g., graph B in Figure 2).

Table 8 presents the results of tests for the differential prediction of each of the effectiveness variables. In all three cases, the Versatility score, leader gender, and Versatility-gender interaction terms were significant, indicating that the behavior ratings were valid predictors of the effectiveness criteria, but that the regression lines had both different intercepts and different slopes for women and men leaders. To interpret these differences, separate regressions were computed for the women and men leaders and their regression lines were plotted and compared (Aguinis et al., 2010; Cleary, 1968). The regression equations are presented in Table 9 and the regression lines are plotted in Figure 3.

For each of the three dependent variables, the regression lines for women had a higher intercept and a somewhat less steep slope. The less steep slope indicates that the behavior ratings were less highly correlated with the effectiveness criteria for women leaders and the higher intercept indicates

Table 8
Results From Moderated Multiple Regression Analyses Testing for Differential Prediction

Variable	Perceived effectiveness β	Team vitality β	Team productivity β
Leader gender	.76*	.90*	.93*
Versatility score	.72***	.51***	.41***
Gender × Versatility	−.78**	−.92**	−.91*
Model R^2	.45***	.21***	.13***

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 9
Separate Regression Equations for Women and Men Leaders

	Dependent variables					
	Perceived effectiveness		Team vitality		Team productivity	
Equation	Women	Men	Women	Men	Women	Men
Constant (intercept)	2.41	1.36	.38	−.67	1.45	.40
Unstandardized *B* (slope)	6.70	7.74	4.11	5.25	2.95	4.02
Model R^2	.39	.52	.16	.27	.09	.16

Note. All coefficients significant, $p < .001$.

that the behavior ratings *over*predicted their effectiveness. A closer examination of the different regression lines shows that they diverge the most at the low end but converge toward the high end of the continuum of Versatility scores. For instance, for Versatility scores of 60%, women are predicted to have Perceived Effectiveness scores that are $d = .77$ SDs higher, Team Vitality scores that are $d = .69$ SDs higher, and Team Productivity scores that are $d = .80$ SDs higher than are predicted for men. For Versatility scores at the mean, 85%, these values diminish to *d*s = .30, .15, and .28, respectively. This suggests a bias in favor of the women who received lower Versatility scores: their effectiveness is higher than would be expected based on their relatively unfavorable ratings of behavior.

The ratings of behavior also accounted for less variance in the effectiveness criteria for women leaders compared to men leaders (39% for women vs. 52% for men in Perceived Effectiveness;16% for women vs. 27% for men in Team Vitality; and 9% for women vs. 16% for men in Team Productivity). This does not mean ratings of women leaders were not valid; they were significantly correlated with the effectiveness variables. But the correlations were smaller for women leaders compared to men leaders. Evidently something other than these leadership behaviors influenced the effectiveness of women leaders but not men leaders.

To understand how the behavior dimensions contributed to this overall effect, we conducted an analysis of differential prediction for each of the four LVI scales. The results indicated greater differential prediction for the Enabling and Strategic dimensions compared with the Forceful and Operational dimension. As an illustration, Figure 4 presents the results for the Perceived Effectiveness criterion (recall that the LVI behaviors are rated on the "Too little/Too much" scale where the ideal score is "0, the right amount" and both low and high scores are unfavorable). The curvilinear regression lines for women and men leaders were virtually identical for Forceful and Operational. However, for Enabling and Strategic, ratings near "the right amount" were associated with equally high perceived effectiveness, but there was overprediction for women leaders at both the "too little" and the "too much" extremes. The amount of variance explained in the perceived effectiveness of

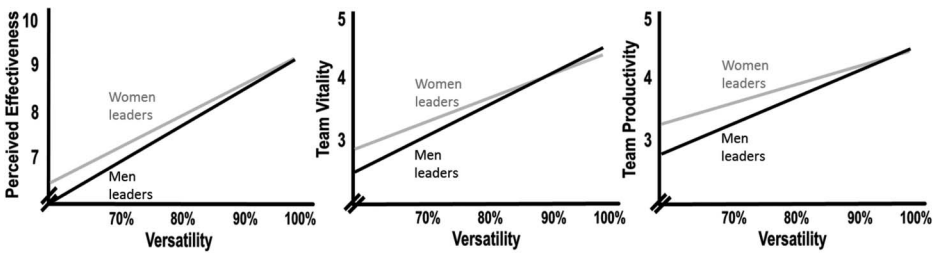

Figure 3. Separate regression lines relating a summary of behavior ratings (Versatility scores) to effectiveness criteria for women and men leaders.

women and men leaders was not significantly different for Forceful (11% for women, 10% for men) and Operational behaviors (4% for both women and men), but was for Enabling (8% for women, 14% for men) and Strategic behaviors (17% for women, 22% for men). Raters of women leaders seem to be taking into account additional factors that compensate for suboptimal levels of Enabling and Strategic behavior in their evaluations of effectiveness.

Phase 3: Substantive Gender Differences in Behavior

The preceding analyses yielded nearly no evidence of systematic bias negatively distorting the ratings of women leaders. Ratings from one class of coworkers were not disproportionately worse than ratings from any other class of coworkers and the behavior ratings did not underpredict the effectiveness of women leaders. We interpreted these results as indicating that the ratings are likely to be valid representations of actual behavior, and thus proceeded to compare the leadership behavior of women and men leaders using the scale and subscale scores computed as the grand mean of the mean superior, mean peer, and mean subordinate ratings.

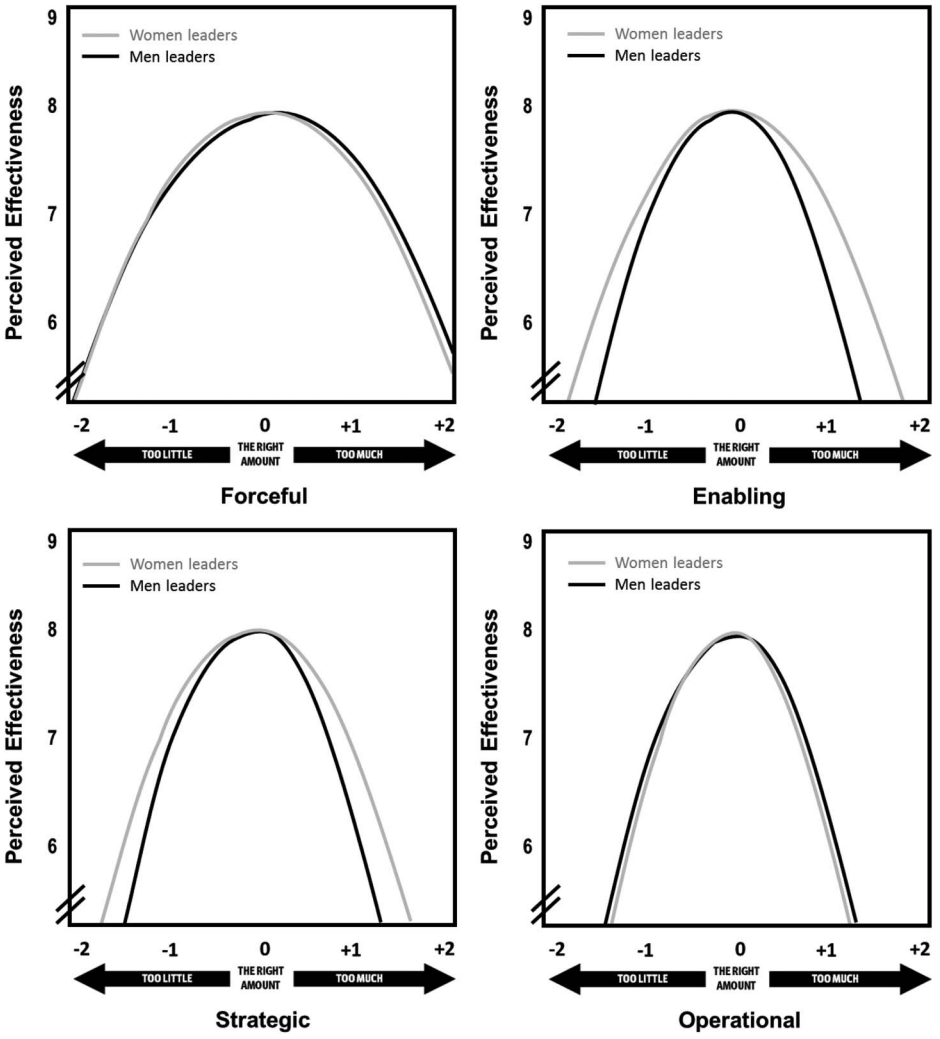

Figure 4. Separate curvilinear regression lines relating ratings of specific leadership behavior scales to perceived effectiveness for women and men leaders.

Table 10
Scale Reliabilities and Correlations Between Scales and Subscales

Scale	Forceful	Enabling	Strategic	Operational	F1	F2	F3	E1	E2	E3	S1	S2	S3	O1	O2	O3
Forceful	(.93)															
Enabling	−.71	(.91)														
Strategic	.41	−.07	(.90)													
Operational	.03	−.03	−.31	(.78)												
F1. Take charge	.91	−.64	.36	.13	(.86)											
F2. Decisive	.89	−.61	.39	−.11	.74	(.87)										
F3. Demanding	.86	−.66	.33	.07	.70	.63	(.89)									
E1. Empowering	−.56	.80	.00	−.19	−.61	−.40	−.51	(.89)								
E2. Participative	−.61	.86	−.08	.07	−.51	−.61	−.51	.52	(.82)							
E3. Encouraging	−.63	.87	−.09	.05	−.50	−.54	−.67	.51	.69	(.85)						
S1. Direction	.23	−.01	.82	−.12	.21	.22	.20	.05	−.04	−.04	(.89)					
S2. Expansion	.54	−.23	.89	−.34	.50	.51	.44	−.15	−.21	−.22	.55	(.86)				
S3. Innovation	.20	.14	.82	−.34	.17	.21	.14	.16	.12	.07	.51	.67	(.77)			
O1. Execution	.37	−.25	.00	.68	.48	.19	.33	−.43	−.07	−.12	−.07	.07	−.02	(.78)		
O2. Efficiency	−.33	.22	−.46	.69	−.27	−.35	−.24	.15	.19	.21	−.20	−.55	−.43	.10	(.69)	
O3. Order	−.04	.01	−.26	.81	.02	−.14	.02	−.08	.06	.06	−.02	−.32	−.35	.31	.51	(.74)

Note. Results based on sample of $N = 1{,}714$ combined women and men leaders. Values along the diagonal are internal consistency reliability estimates (coefficient α). Correlations greater than $|r| = .06$ significant at $p < .01$.

Two one-way MANOVAs were conducted with Leader gender as the independent variable, one using the four primary scales—Forceful, Enabling, Strategic, and Operational—as dependent variables, and the other using the 12 subscales as dependent variables. Reliabilities and intercorrelations among the aggregated scale and subscale scores are presented in Table 10. The multivariate effect for leader gender was significant in both analyses and accounted for 5.3% of the multivariate variance in the primary scales (Wilks' $\lambda = .947$, $F(4, 1709) = 23.83$, $p < .001$) and 8.7% in the subscales (Wilks' $\lambda = .913$, $F(12, 1701) = 13.58$, $p < .001$).

The MANOVA models tested for the significance of gender differences in mean scores. To further aid in interpretation of differences on the behavior scores, we also considered the proportion of women and men who were rated "too little," "the right amount," and "too much" on each scale and subscale. According to Cohen (1988), scores that differ by .6 standard units represent a "moderate effect." Therefore we classified scale and subscale scores as "the right amount" if they fell within ±.6 SDs of "0, the right amount." We classified scores exceeding this range as "too much" and those below this range as "too little." The significance of mean differences based on the univariate ANOVAs as well as the descriptive statistics for women and men leaders and the proportions classified "too little," "the right amount," and "too much" are presented in Table 11.

The results indicate that women scored significantly higher on the Forceful ($d = .28$ standard units) and Operational ($d = .27$) scales, but lower on the Strategic ($d = -.32$) and Enabling ($d = -.12$) scales. The subscale results further clarified that women scored higher on all three Forceful subscales—Take charge ($d = .25$), Decisive ($d = .14$), and Demanding ($d = .30$)—and two of the Operational subscales, Execution ($d = .31$) and Order ($d = .17$). Further, more women were rated "too much" and more men "too little" on the three Forceful subscales and one Operational subscale, Execution, whereas the primary difference on the other Operational subscale, Order, was that more women were rated "the right amount" and more men "too little." The major differences where

Table 11
Descriptive Statistics and Mean Differences for Leader Behavior Scores

	Women leaders					Men leaders						
Scale/subscale	*M*	*SD*	Too little	Right amt.	Too much	*M*	*SD*	Too little	Right amt.	Too much	*d*	*F*(1, 1712)
Scales												
Forceful	−.01	.38	28%	49%	23%	−.11	.35	37%	48%	15%	.28	28.58***
Enabling	−.20	.26	51%	45%	4%	−.17	.26	47%	46%	7%	−.12	6.38*
Strategic	−.32	.25	74%	25%	1%	−.24	.25	70%	28%	3%	−.32	13.13***
Operational	−.05	.18	31%	55%	14%	−.10	.19	42%	47%	12%	.27	24.42***
Subscales												
Take charge	+.06	.42	20%	53%	27%	−.04	.39	27%	52%	21%	.25	26.65***
Decisive	+.02	.43	24%	51%	26%	−.04	.41	28%	51%	21%	.14	8.94**
Demanding	−.14	.41	41%	44%	14%	−.26	.41	53%	38%	9%	.30	39.33***
Empowering	−.23	.32	47%	48%	5%	−.14	.32	35%	55%	10%	−.28	33.92***
Participative	−.23	.29	55%	40%	5%	−.24	.28	55%	41%	4%	.00	.01
Encouraging	−.14	.31	35%	55%	9%	−.13	.33	35%	52%	12%	−.02	.15
Direction	−.40	.31	74%	25%	1%	−.32	.30	65%	33%	2%	−.25	27.87***
Expansion	−.35	.36	62%	34%	4%	−.32	.36	60%	34%	5%	−.08	2.59
Innovation	−.22	.22	61%	36%	2%	−.20	.23	58%	37%	5%	−.11	5.48*
Execution	−.02	.28	26%	53%	21%	−.10	.27	38%	48%	15%	.31	41.41***
Efficiency	−.06	.22	32%	53%	15%	−.06	.25	35%	48%	17%	.02	.11
Order	−.08	.24	31%	56%	12%	−.12	.27	41%	47%	12%	.17	12.79***

Note. $N = 857$ women leaders and 857 men leaders. $d = (M_{women} - M_{men})$/pooled *SD* *F* values based on univariate ANOVAs.
* $p < .05$. ** $p < .01$. *** $p < .001$.

women scored lower on the subscales indicated more of them were rated "too little" on Empowering ($d = -.28$), Direction ($d = -.25$), and Innovation ($d = -.11$).

Discussion

The preceding analyses were designed to first determine whether ratings of women leaders were biased and second to identify real gender differences in leadership behavior. A notable strength of the study was the use of a large, matched sample of women and men from six different organizations. The evaluation of bias included two complementary analytic strategies, which are commonly used together in the study of bias in cognitive tests. Neither analysis provided evidence of bias against women leaders. However, there was some evidence of bias *in favor* of women leaders. Finally, the analysis of substantive gender differences found significant differences in the behavior of women and men leaders. Women were described with a more forceful-operational style and men were described with a more strategic-enabling style. Next we interpret the gender bias results and the gender differences results separately, and then proceed to a more general discussion of the implications for practice and research.

Analysis of Gender Bias

We recognize that reporting evidence of bias in favor of women leaders may be controversial; it certainly runs counter to popular opinion. Nonetheless, *prima facie* results indicated that no particular class of coworkers systematically provided disproportionately less favorable ratings for women leaders. For instance, we found that men were no more likely to rate women leaders as "too Enabling" (too "soft" or too feminine) or "too Forceful" (too "hard" or too masculine). Rather, raters of both sexes and from superior, peer, and subordinate perspectives provided similar ratings of women leaders.

Furthermore, the analyses of differential prediction were consistent across three distinct measures of effectiveness. In all three cases, unfavorable ratings of behavior were associated with higher levels of effectiveness for women leaders than for men leaders. In particular, suboptimal levels of Strategic and Enabling behavior were not as detrimental to the effectiveness of women as for men leaders. It seems that raters were inclined to elevate their ratings of the effectiveness of less-skilled women leaders, in effect compensating for their insufficient use of Strategic and Enabling behaviors. Overprediction of the effectiveness of women leaders is a matter of fact in this database; what is unclear is the reason for the overprediction.

We propose that the overprediction may represent a tendency for coworkers to give women leaders the benefit of the doubt, especially women who demonstrate fewer Strategic and Enabling behaviors. There has been a tremendous push for diversity and inclusion over the last three decades (Anand & Winters, 2008). Corporate citizens have been sensitized to both the ills of sexism and the goal of increasing the number of women in leadership roles. Thus, coworkers may—through some combination of not wanting to seem sexist and truly desiring to see women succeed—enhance their judgments of the effectiveness of women beyond what would be suggested by their lack of Strategic and Enabling behavior.

This line of theorizing is consistent with other trends in the research literature. For instance, a recent meta-analysis examined changes in the perceived effectiveness of women and men leaders over time (Paustian-Underdahl, Walker, & Woehr, 2014). Before 1982 women were, on average, rated as less effective than men. However, since that time, women have been, on average, rated about the same as, or slightly more effective than, men. On the other hand, we are aware of no studies that report the leadership skills and behavior of women to be similarly improving over that time span. Rather, the increase in the rated effectiveness of women leaders is often attributed to changing social norms and greater acceptance and receptivity to women in leadership roles (Eagly & Carli, 2003; Paustian-Underdahl et al., 2014).

We want to be cautious in not overgeneralizing the absence of bias against women leaders in our results. The ratings we analyzed were based on performance in a current role, collected under

conditions of confidentiality, and to be used for strictly development purposes—and explicitly not for selection or succession decisions. Thus, it is in this context that we did not find bias against women leaders. It would be inappropriate to conclude that there is no unconscious bias at all against women in leadership positions. For instance, studies in the "paper people" paradigm—where applications or written descriptions of leaders are presented—often find harsher evaluative judgments when the leader is presented as a female versus an otherwise identical male (Heilman, 1984). It is possible that implicit judgments that are not formally expressed (such as a private opinion) or even explicit judgments made for different purposes, such as performance appraisal or a promotion or succession decision, may show evidence of bias.

Analysis of Gender Differences

Because the ratings of behavior showed little evidence of bias, a high degree of interrater reliability and agreement, and significant correlations with three different outcome variables, we assume that they are valid representations of how the women and men in our sample lead. The analysis of gender differences found that women leaders demonstrated more Forceful and Operational behavior and less Strategic and Enabling behavior than men.

In particular, the women were rated as demonstrating more Take charge, Decisive, and Demanding behavior on the Forceful side, and more Execution and Order behavior on the Operational side. Generally, more women were rated "too much" on the Take charge, Decisive, Demanding, and Execution behaviors and "the right amount" on the Order behaviors whereas more men were rated "too little" on all of them. This pattern suggests more women were acting as assertive executors with a hands-on focus on short-term objectives whereas men were more *laissez fare* and less involved in the details of execution, a pattern found in a meta-analysis of transactional leadership (Eagly et al., 2003). On the other hand, women were rated as using less Empowering behavior on the Enabling side and less Direction behavior on the Strategic side. More women leaders were rated "too little" on these behaviors, whereas more men were rated "the right amount" on them (and, to a lesser extent, rated "too much" on the Empowering behaviors).

These gender differences are consistent with findings from other research on the behavior of upper-level managers. For instance, as noted earlier, one meta-analysis (Eagly & Johnson, 1990) did not find that women use a more stereotypically feminine, people-oriented style (Enabling in our study) or that men use a more masculine, task-oriented style (Forceful and Operational in our study). Real corporate managers do not fit common popular stereotypes. In fact, studies of individuals going through executive development processes like coaching reveal that many women feel pressure to emulate an assertive and results-oriented masculine style and find the transition to a more people-oriented style difficult (Lyons, 2002). Further, a recent study of executive education participants at the French business school, INSEAD, concluded that the current generation of women leaders "owe their success to a strong command of the technical elements of their jobs and a nose-to-the-grindstone focus on accomplishing quantifiable objectives" (Ibarra & Obodaru, 2009, p. 64). The focus of that article, titled "Women and the Vision Thing," was how many women leaders are not seen as possessing sufficient strategic thinking skills. The present results converge with this research in showing how women leaders are more often seen as forceful operators than strategic enablers.

However, popular discussions of the underrepresentation of women in senior leadership frequently suggest that these sorts of gender differences are artifacts of bias (instead of real differences in behavior). A recent article in the *Wall Street Journal*, for instance, reported findings from an as-yet unpublished study comparing the language used in performance reviews of women and men. The central finding was that women were more likely than men to be critiqued for coming on too strong, a trend observed in evaluations from both male and female bosses (Silverman, 2015). The article went on to explain this finding as evidence of unconscious bias and reported on efforts to curb these unfair evaluations.

However, according to the logic behind the standard methodology for identifying bias, because these evaluations were made by both male and female bosses an alternative explanation may be that the women really were more likely to come on stronger than the men. It would not be implausible if one assumes that women in management are driven and ambitious and feel pressure to emulate

a stereotypically masculine ideal. In the remainder of this article we explore the possibility that findings like these, and from our study, do reflect real gender differences in behavior and then follow the implications to new directions for helping more qualified women reach senior leadership roles.

Toward a New Narrative

Like many gender and leadership scholars (e.g., Eagly & Carli, 2003; Vecchio, 2002), we believe that the popular media tends to oversimplify the issues in discussions of the gender imbalance in leadership. This can lead to misguided advice. For instance, the theme of the best-seller, *Lean In* (Sandberg, 2013), is that women should assert themselves more like their male colleagues. This advice may be helpful for the minority of the corporate women in our sample who were rated "too little" Forceful, but surely is less helpful for the majority who were rated "the right amount" or "too much" Forceful. In this spirit, we propose strategies for achieving gender parity that follow from our findings. These suggestions are not part of the popular narrative, but offer some new directions for practice as well as research.

Implications for practice. One criticism of the popular narrative is that it is disempowering to women. On the other hand, our results suggest some things aspiring women managers can do to further their leadership careers. Specifically, they need to be wary about carving out a niche as a results-oriented implementer and proactively seek out broadening experiences to develop skills and a reputation for being a strategic leader.

The forceful-operational style that was more common among women is associated with the tactical management of implementation core to lower-level supervisory, middle-management, or functional roles, whereas the strategic-enabling style more common among men is associated with higher-level executive leadership roles (Charan, Drotter, & Noel, 2001; Kaiser, 2011). This is consistent with how women are rated as more effective leaders than men in middle management, but not executive, positions (Paustian-Underdahl et al., 2014) and how they are more likely to be hired for staff roles with support responsibilities than for line roles responsible for running a business and profits and losses (Lyness & Heilman, 2006).

We often see a related dynamic in our executive coaching work. Many professional women establish themselves as someone their bosses can count on. They are hardworking, detail-oriented, and good at multitasking. They are driven and willing to do whatever it takes to deliver. They have the technical know-how to understand their functional area inside and out. They are also often very capable and find it easier to do things themselves rather than delegate. As one meta-analysis revealed, they are more hands-on and involved (Eagly et al., 2003). And like the research showing women more often have to influence without formal authority (Lyness & Thompson, 2000; Ohlott, Ruderman, & McCauley, 1994), our clients frequently feel that they have to push their agenda to get things done.

However, distinguishing oneself in this way creates a new kind of double-bind that can be career limiting. Upper management sees these women as reliable executors, and they often get placed in these familiar assignments (Ruderman & Ohlott, 2002). But they get passed over for promotions to larger, more visible roles because they are not seen as strategic enough to lead across the enterprise. Further compounding matters is that the pushing required to overcome the lack of formal authority often strains peer relations, making it politically more difficult to promote these women to be the boss of those peers.

It is noteworthy that women were rated lower on the strategic and enabling behaviors and that the bias observed in favor of women concerned how worse ratings on these behaviors were not reflected in lower effectiveness. One possibility is that coworkers didn't expect these women to demonstrate strategic and enabling behaviors—as if coworkers recognize the strategic and enabling deficits, but do not hold it against them in the evaluation of their current performance. But it may be these same deficits that later prevent them from being promoted into higher-level, senior leadership positions.

There are two complementary strategies upwardly mobile women can use to avoid the double-bind of being rewarded as a tactical doer at the expense of becoming recognized as a strategic

thinker: proactively seeking out the broadening experiences needed to develop strategic leadership skills and signaling these skills appropriately.

Strategic leadership skills are developed through a diversity of job experiences—working in multiple functional areas and in different businesses or industries, holding a line management position responsible for business results, working across and in different cultures, responsibility for large-scope assignments, and leading in different business cycles and through different business challenges (Dragoni, Oh, Vankatwyk, & Tesluk, 2011; Hurley & Sonnenfeld, 1998; McCall, Lombardo, & Morrison, 1988; McCauley, Ruderman, Ohlott, & Morrow, 1994). However, women are less likely to have these sorts of broadening job experiences. Various studies have shown that they are more often in staff roles, have less functional diversity, face a smaller range of business challenges, and have fewer international assignments (Lyness & Heilman, 2006; Lyness & Thompson, 2000; Nicholson & West, 1988; Ohlott et al., 1994; Olson, Frieze, & Good, 1987; Van Velsor & Hughes, 1990).

Further, studies of the few women executives who managed to "break through the glass ceiling" show that they were proactive in setting career goals and career planning, asking for increased responsibilities, and more willing to take on risky assignments to expand their experience (Lyness & Thompson, 2000; Morrison, White, & Van Velsor, 1987). Thus, it behooves women to take the initiative to ask for guidance from more senior leaders on career progression and to seek out the broadening experiences that build strategic leadership skills (Wallace, 2008). Another opportunity for development in place is to make one's current role as strategic as possible. For instance, instead of viewing a Human Resources job as a support staff role fulfilling requests, work more proactively and closely with line leaders to anticipate and provide people solutions for their business problems.

However, according to signaling theory, it is not enough to develop strategic skills; aspiring women leaders also need to be savvy in communicating to key organizational decision-makers that they have these skills (Weiss, 1995). We have observed that many women managers focus their attention and comments on tactical issues. They zero in on details, constraints, resource requirements, and the nuts and bolts of execution. This reinforces their reputation for being a tactical doer instead of a strategic thinker. Many of these women are capable of thinking strategically, but they do not use strategic language or elevate their commentary to a broad enough level to be recognized for it. We coach them to ensure that they also express ideas about commercial opportunities, market dynamics, competitor moves, innovation, and connecting current activities with the long-term direction of the company to establish strategic priorities.

There are also things organizations can do to reinforce these suggestions for how women managers can cultivate stronger strategic leadership skills. First would be to clearly define what it means to be strategic. Many of our women clients have reported struggling to understand feedback that they aren't strategic enough. "OK, but concretely, what should I be doing differently and focusing on to be 'more strategic'?," one recently asked. Second, organizations should consider whether they are providing equally the developmental opportunities associated with strategy skills. We suspect that bias may be more impactful early on in the career trajectory, where men are more often given the kinds of assignments that develop strategic leadership skills. We recommend that HR and talent management professionals build in an equality check to the planning and allocation of development assignments to ensure adequate access for upwardly mobile female talent. One strategy for achieving gender parity in the executive suite may be to intentionally provide a larger number of younger, high-potential women with these career guidance and planning tools and broadening experiences.

Implications for research. We encourage more research on bias in ratings, evaluations, and judgments made of women and men leaders. That includes quantitative studies that use *prima facie* and differential prediction methodologies as well as policy-capturing studies (Karren & Barringer, 2002) that use qualitative methods to understand the factors decision makers use and how they weight them. High-stakes decisions such as performance appraisals, judgments of potential, succession evaluations, and compensation decisions can be studied to determine whether our findings generalize. Is there evidence of bias in favor women in these contexts, or when consequences beyond developmental feedback are on the line is there bias against women?

More research is also needed to better understand gender differences in the career development opportunities associated with rising into senior leadership roles. For instance, men are more likely to be hired for line jobs and women for staff jobs (Hurley & Sonnenfeld, 1998; Lyness & Heilman, 2006)—but why? What factors lead to these disparate outcomes, and how can the playing field be leveled? Biases that disadvantage women managers from getting these and other broadening experiences earlier in their career may compound over time to limit the number of later career-stage women seen as qualified for senior strategic leadership roles.

Also, is our interpretation of the literature on the broadening experiences needed to cultivate strategic leadership skills appropriately applied to aspiring women managers? On the surface, it would seem to. For instance, two highly visible female CEOs, Indra Nooyi at PepsiCo and Meg Whitman formerly CEO of eBay and now at Hewlett-Packard, both fit the pattern of having a diverse array of job experiences, including profit and loss responsibilities in different businesses. However, much of the research on the development of strategic skills and their necessity for advancing to a senior executive role was based on men (e.g., McCall et al., 1988).

Finally, to what extent do our findings generalize across different types of organizations? We created a database representing six companies in five industries on three continents to expand the representativeness of the findings. But gender differences in leadership sometimes depend on contextual factors (Eagly & Carli, 2003; Paustian-Underdahl et al., 2014; Vecchio, 2002)—for instance, women are viewed more favorably in service and education industries and when there are higher proportions of women in the organization. Future research is needed to determine whether, for example, differences in the leadership styles of women who reach the top vary by industry sector or in organizations comprised with more or fewer other women. Or do the few women who become CEOs have a more strategic-enabling style, regardless of the context?

Limitations

Although we took great care in constructing our database and conducting the statistical analyses, there are still limitations to this study. First, as we explained earlier, this study was not carried out in the hypothetico-deductive tradition, the preferred model of the scientific method. Thus, we do not regard our findings as supporting any particular hypothesis, but simply as providing empirical observations that can form the basis of a theory to be tested.

Also, in our interpretation of the results of the *prima facie* analyses, we are assuming that bias is apparent when one class of raters rated women disproportionately worse than other classes. However, it is possible that *all* classes of raters are equally biased. For instance, if one assumed that gender role expectations are a strong aspect of culture, then women may be just as biased against other women in leadership roles as men are. If this were the case, then our interpretation that more women were rated "too much" Forceful could be recast to simply indicate they were being penalized by both women and men raters for displaying stereotypically masculine behavior. In our view, this interpretation is nihilistic, and is at odds with the position that social reality is culturally constructed and defined by shared perceptions and beliefs (Berger & Luckmann, 1966).

We also do not wish to exaggerate the extent of gender differences in our sample. The mean differences in behavior were statistically significant, but there was also substantial overlap in the distributions. As documented in two separate reviews that each included results from dozens of meta-analyses of gender differences across a broad range of behavioral and psychological variables (Hyde, 2005; Zell, Krizan, & Teeter, 2015), there is more similarity than difference between men and women and leadership behavior appears to be no exception. Across the significant findings in our sample, the difference in how many women versus men were rated as too little, the right amount, or too much ranged from a high of 24% (for Demanding and Execution) to a low of 8% (for Decisive). Nonetheless, statistical differences of this magnitude are sufficient to make a noticeable difference in practice (Abelson, 1985).

Finally, although our sample included organizations based on three continents, they all represent Western, postindustrial societies. It is unclear how our results apply to other cultures, especially those where gender inequality is greater (e.g., the Middle East, Africa, Asia, or Latin America;

House, Hanges, Javidan, Dorfman, & Gupta, 2004). It is best to not generalize our findings to these cultures until additional research examines the appropriateness of doing so.

Conclusion

This project began as a search for evidence of bias against women in ratings of their leadership behavior. As the results began to add up across organizations, we were surprised by the consistency in the lack of bias that is so central to popular discussions about the underrepresentation of women in leadership. It may be that these results reflect some degree of social progress and that, as other researchers have recently noted (Paustian-Underdahl et al., 2014), attitudes about women in leadership are improving in modern Western societies. However, many concerned with gender equality are unsatisfied with the degree of progress and the last 30 years of efforts guided by the popular narrative seem to have reached a limit in getting more women to the top. On the other hand, our findings suggest a more empowering message for ambitious women who desire a senior leadership position: to be more proactive in managing their long-term career goals by focusing on the development of strategic organizational leadership skills and the importance of communicating those skills to key decision-makers. Our findings also suggest new avenues for organizations to pursue in achieving gender parity by ensuring female talent have the opportunities and broadening experiences needed to develop strategic organizational leadership skills starting early in their careers.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129–133. http://dx.doi.org/10.1037/0033-2909.97.1.129

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95,* 648–680. http://dx.doi.org/10.1037/a0018714

Anand, R., & Winters, M. (2008). A retrospective view of corporate diversity training from 1964 to the present. *Academy of Management Learning & Education, 7,* 356–372. http://dx.doi.org/10.5465/AMLE.2008.34251673

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Garden City, NY: Anchor Books.

Catalyst. (2015). *Women CEOs of the S&P 500*. New York, NY: Catalyst.

Charan, R., Drotter, S., & Noel, J. (2001). *The leadership pipeline: How to build the leadership-powered company*. San Francisco, CA: Jossey-Bass.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated universities. *Journal of Educational Measurement, 5,* 115–124. http://dx.doi.org/10.1111/j.1745-3984.1968.tb00613.x

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

Dailey, S. L., & Browning, L. (2014). Retelling stories in organizations: Understanding the functions of narrative repetition. *The Academy of Management Review, 39,* 22–43. http://dx.doi.org/10.5465/amr.2011.0329

Dragoni, L., OH, I.-S., Vankatwyk, P., & Tesluk, P. E. (2011). Developing executive leaders: The relative contribution of cognitive ability, personality, and the accumulation of work experience in predicting strategic thinking competency. *Personnel Psychology, 64,* 829–864. http://dx.doi.org/10.1111/j.1744-6570.2011.01229.x

Eagly, A. H., & Carli, L. L. (2003). The female leadership advantage: An evaluation of the evidence. *The Leadership Quarterly, 14,* 807–834. http://dx.doi.org/10.1016/j.leaqua.2003.09.004

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M. L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129,* 569–591. http://dx.doi.org/10.1037/0033-2909.129.4.569

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin, 108,* 233–256. http://dx.doi.org/10.1037/0033-2909.108.2.233

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109,* 573–598. http://dx.doi.org/10.1037/0033-295X.109.3.573

Exposing hidden bias at Google. (2014, September 25). *New York Times,* p. B1.

Foschi, M. (1996). Double standards in the evaluation of men and women. *Social Psychology Quarterly, 59,* 237–254. http://dx.doi.org/10.2307/2787021

Frenkiel, N. (1984, March). The up-and-comers; Bryant takes aim at the settlers-in. *Adweek*. Special Report: Magazine World.

Heilman, M. E. (1984). Information as a deterrent against sex discrimination: The effects of applicant sex and information type on preliminary employment decisions. *Organizational Behavior & Human Performance, 33,* 174–186. http://dx.doi.org/10.1016/0030-5073(84)90019-9

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues, 57,* 657–674. http://dx.doi.org/10.1111/0022-4537.00234

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology, 89,* 416–427. http://dx.doi.org/10.1037/0021-9010.89.3.416

Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Press.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations. The GLOBE study of 62 societies*. Thousand Oaks, CA: Sage.

Hurley, A. E., & Sonnenfeld, J. A. (1998). The effect of organizational experience on managerial career attainment in an internal labor market. *Journal of Vocational Behavior, 52,* 172–190. http://dx.doi.org/10.1006/jvbe.1997.1603

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581–592. http://dx.doi.org/10.1037/0003-066X.60.6.581

Ibarra, H., Ely, R., & Kolb, D. (2013). Women rising: The unseen barriers. *Harvard Business Review, 91,* 61–66.

Ibarra, H., & Obodaru, O. (2009). Women and the vision thing. *Harvard Business Review, 87,* 62–70.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69,* 85–98. http://dx.doi.org/10.1037/0021-9010.69.1.85

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

Kaiser, R. B. (2011). The leadership pipeline: Fad, fashion, or empirical fact? An introduction to the special issue. *The Psychologist-Manager Journal, 14,* 71–75. http://dx.doi.org/10.1080/10887156.2011.570126

Kaiser, R. B., & Hogan, J. (2011). Personality, leader behavior, and overdoing it. *Consulting Psychology Journal: Practice and Research, 63,* 219–242. http://dx.doi.org/10.1037/a0026795

Kaiser, R. B., Hogan, R., & Craig, S. B. (2008). Leadership and the fate of organizations. *American Psychologist, 63,* 96–110. http://dx.doi.org/10.1037/0003-066X.63.2.96

Kaiser, R. B., & Kaplan, R. E. (2005). Overlooking overkill? Beyond the 1-to-5 rating scale. *Human Resources Planning, 28,* 7–11.

Kaiser, R. B., McGinnis, J. L., & Overfield, D. V. (2012). The how and the what of leadership. *Consulting Psychology Journal: Practice and Research, 64,* 119–135. http://dx.doi.org/10.1037/a0029331

Kaiser, R. B., & Overfield, D. V. (2010). Assessing flexible leadership as a mastery of opposites. *Consulting Psychology Journal: Practice and Research, 62,* 105–118. http://dx.doi.org/10.1037/a0019987

Kaiser, R. B., & Overfield, D. V. (in press). *Leadership Versatility Index: A user guide for version 5.0*. Greensboro, NC: Kaiser Leadership Solutions, LLC.

Kaiser, R. B., Overfield, D. V., & Kaplan, R. E. (2010). *Leadership Versatility Index version 3.0 facilitator's guide*. Greensboro, NC: Kaplan DeVries Inc.

Karren, R. J., & Barringer, M. W. (2002). A review and analysis of the policy-capturing methodology in organizational research: Guidelines for research and practice. *Organizational Research Methods, 5,* 337–361. http://dx.doi.org/10.1177/109442802237115

Kay, K., & Shipman, C. (2014). *The confidence code: The science and art of self-assurance– what women should know*. New York, NY: HarperBusiness.

Lawrence, P. R. (1992). The challenge of problem-oriented research. *Journal of Management Inquiry, 1,* 139–142. http://dx.doi.org/10.1177/105649269212007

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6,* 80–128. http://dx.doi.org/10.1177/1094428102239427

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Leonard, H. S. (2003). Leadership development for the postindustrial, postmodern information age. *Consulting Psychology Journal: Practice and Research, 55,* 3–14.

Lyness, K. S., & Heilman, M. E. (2006). When fit is fundamental: Performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology, 91,* 777–785. http://dx.doi.org/10.1037/0021-9010.91.4.777

Lyness, K. S., & Thompson, D. E. (1997). Above the glass ceiling? A comparison of matched samples of female and male executives. *Journal of Applied Psychology, 82,* 359–375. http://dx.doi.org/10.1037/0021-9010.82.3.359

Lyness, K. S., & Thompson, D. E. (2000). Climbing the corporate ladder: Do female and male executives follow the same route? *Journal of Applied Psychology, 85,* 86–101. http://dx.doi.org/10.1037/0021-9010.85.1.86

Lyons, D. (2002). Freer to be me: The development of executives at mid-life. *Consulting Psychology Journal: Practice and Research, 54,* 15–27. http://dx.doi.org/10.1037/1061-4087.54.1.15

Maccoby, M. (2007). *The leaders we need: And what makes us follow*. Cambridge, MA: Harvard Business Review Press.

McCall, M. W., Jr., Lombardo, M. M., & Morrison, A. M. (1988). *Lessons of experience: How successful executives develop on the job*. New York, NY: Free Press.

McCauley, C. D., Ruderman, M. N., Ohlott, P. J., & Morrow, J. E. (1994). Assessing the developmental components of managerial jobs. *Journal of Applied Psychology, 79,* 544–560. http://dx.doi.org/10.1037/0021-9010.79.4.544

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

Mohrman, S. A., & Cohen, S. G. (1995). When people get out of the box: New relationships, new systems. In A. Howard (Ed.), *The changing nature of work* (pp. 365–410). San Francisco, CA: Jossey-Bass.

Morrison, A. M., White, R. P., & Van Velsor, E. (1987). *Breaking the glass ceiling: Can women reach the top of America's largest corporations?* Reading, VA: Addison Wesley.

National Research Council, Committee on Techniques for Enhancement of Human Performance. (2001). *The changing nature of work: Implications for Occupational Analysis*. Washington, DC: National Academy Press.

Nicholson, N., & West, M. (1988). *Managerial job change: Men and women in transition*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511522116

Oh, I. S., & Berry, C. M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology, 94,* 1498–1513. http://dx.doi.org/10.1037/a0017221

Ohlott, P. J., Ruderman, M. N., & McCauley, C. D. (1994). Gender differences in managers' developmental job experiences. *Academy of Management Journal, 37,* 46–67. http://dx.doi.org/10.2307/256769

Olson, J. E., Frieze, J. H., & Good, D. C. (1987). The effects of job type and industry on the income of male and female MBAs. *The Journal of Human Resources, 22,* 532–541. http://dx.doi.org/10.2307/145696

Paustian-Underdahl, S. C., Walker, L. S., & Woehr, D. J. (2014). Gender and perceptions of leadership effectiveness: A meta-analysis of contextual moderators. *Journal of Applied Psychology, 99,* 1129–1145. http://dx.doi.org/10.1037/a0036751

Privitera, G. J. (2015). *Statistics for the behavioral sciences* (2nd ed.). Thousand Oaks, CA: Sage.

Ruderman, M. N., & Ohlott, P. J. (2002). *Standing at the crossroads. Next steps for high-achieving women*. San Francisco: Jossey-Bass.

Sandberg, S. (2013). *Lean in: Women, work, and the will to lead*. New York, NY: Knopf.

Schein, V. E. (2001). A global look at psychological barriers to women's progress in management. *Journal of Social Issues, 57,* 675–688. http://dx.doi.org/10.1111/0022-4537.00235

Silverman, R. E. (2015, September 30). Gender bias at work turns up in feedback. *The Wall Street Journal*. Retrieved from http://www.wsj.com

Smith, C. (2009, July 26). No doubts: Women are better managers. *The New York Times*. Retrieved from http://www.nytimes.com

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Staal, M. A. (2008). Test review of the Leadership Versatility Index. In K. F. Geisinger, R. A. Spies, & J. F. Carlson (Eds.), *The eighteenth mental measurements yearbook* [Electronic version]. Lincoln, NE: Buros Institute of Mental Measurements.

Stogdill, R. M., & Coons, A. E. (1957). *Leader behavior: Its description and measurement*. Columbus, OH: Bureau of Business Research, Ohio State University.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.

United States Department of Labor. (2014). Facts over time: Women in the labor force. Retrieved on September 2, 2014, from http://www.dol.gov/wb/stats/facts_over_time.htm#wilf

Van Velsor, E., & Hughes, M. W. (1990). *Gender differences in the development of managers: How women managers learn from experience* (Tech. Rep. No. 145). Greensboro, NC: Center for Creative Leadership.

Vassar, M. (2008). Test review of the Leadership Versatility Index. In K. F. Geisinger, R. A. Spies, & J. F. Carlson (Eds.), *The eighteenth mental measurements yearbook* [Electronic version]. Lincoln, NE: Buros Institute of Mental Measurements.

Vecchio, R. P. (2002). Leadership and gender advantage. *The Leadership Quarterly, 13,* 835–850.
Wallace, W. (2008). *Reaching the top: Factors that impact the careers and retention of senior women leaders*. Raleigh, NC: Lulu Publishing.
Waller, N., & Lublin, J. S. (2015, September 30). What's holding women back in the workplace? *The Wall Street Journal*. Retrieved from http://www.wsj.com
Weiss, A. (1995). Human capital vs. signaling explanations of wages. *The Journal of Economic Perspectives, 9,* 133–154. http://dx.doi.org/10.1257/jep.9.4.133
Wiggins, J. S., & Trapnell, P. D. (1996). A dyadic-interactional perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality* (pp. 88–162). New York, NY: Guilford Press.
Williams, R. (2012, December 15). Why women may be better leaders than men. *Psychology Today*. Retrieved from http://www.psychologytoday.com/blog/wired-success/201212/why-women-may-be-better-leadersmen
Women in top jobs are viewed as "better leaders" than men. (2010, May 14). *Daily Mail*. Retrieved from http://www.dailymail.co.uk/sciencetech/article-1278009/Women-jobs-viewed-better-leaders-men.html#ixzz1quPyGOPa
Yost, P. R., Yoder, M. P., Chung, H. H., & Voetmann, K. R. (2015). Narratives at work: Story arcs, themes, voice, and lessons that shape organizational life. *Consulting Psychology Journal: Practice and Research, 67,* 163–188. http://dx.doi.org/10.1037/cpb0000043
Yukl, G. A. (2010). *Leadership in organizations* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.
Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist, 70,* 10–20. http://dx.doi.org/10.1037/a0038208

Received October 20, 2015
Latest revision received February 3, 2016
Accepted February 3, 2016 ■