A Test of the Congruence Hypothesis in the Think Manager—Think Male Research Paradigm

Matthew W. Lauritsen
University of Minnesota Duluth

Researchers who design studies based on Schein's (1973, 1975) think manager-think male (TMTM) research paradigm ubiquitously conclude that the greater incongruence between beliefs about managers and women, compared to managers and men, is problematic for women in leadership roles. However, this hypothesis is never directly tested. Using polynomial regression analysis, the congruence hypothesis was tested via the evaluations of the respondents' current supervisors as criteria. The hypothesized congruence effect was not found. In other words, there was no observed relationship between the congruence of people's beliefs about men, women, and leaders, and their subsequent evaluations of women leaders. The results of this study raise concerns about the validity of the conclusions of TMTM studies.

Keywords: think manager-think male, agency, communion, polynomial regression, gender stereotypes, leadership perceptions

INTRODUCTION

Research on congruence, fit, or similarity, has been a reoccurring theme in industrial-organizational psychology research (Kristof-Brown, Zimmerman, & Johnson, 2005; Schneider, 2001). Probably the most well-known example of congruence hypotheses comes from person-environment fit theory (see Kristof-Brown & Guay, 2011). Leadership researchers have also hypothesized that congruence matters for the perceptions of leaders (e.g., implicit leadership theory, Lord, Foti & De Vader, 1984). The general form of many of the congruence hypotheses state that individuals who are perceived to be more leader-like will be evaluated more positively.

The think manager-think male (TMTM) research paradigm (Schein, 1973, 1975) posits that the greater congruence between men and managers compared to women and managers leads to prejudice and discrimination against women. However, TMTM studies never empirically test whether this incongruence relates to an outcome (see Koenig, Eagly, Mitchell, & Ristikari, 2011). In fact, almost no empirical evidence has demonstrated that the congruence between categories (e.g., women and managers) relates to leader evaluations (Tavarez, Sobral, Goldszmidt, & Araujo, 2018). In other words, the congruence hypothesis is rarely, if ever, tested. This is concerning when studies assume congruence is operative but never directly test, or even measure, congruence (e.g., Koenig & Eagly, 2014). The question of whether or not congruence matters is a scientific one, but has up to this point been ignored in TMTM studies. This paper aims to empirically test this hypothesis and answer the question: does the similarity in trait ratings of gender and leader stereotypes actually matter for the evaluations of individual managers?

LITERATURE REVIEW

Schein's (1973, 1975) TMTM paradigm examines the similarity between respondent ratings of men, women, and managers. Studies under this paradigm randomly assign participants to rate one of three targets (men, women, or some managerial role) on a number of traits or characteristics and then compare the between-group ratings. TMTM researchers ubiquitously conclude that the greater distance found between managers and women compared to managers and men is problematic for the perception of women and decision-making which affect women in various leadership roles (e.g., Carli, Alawa, Lee, Zhao, & Kim, 2016; Fischbach, Lichtenthaler, & Horstmann, 2015). The alleged hidden barriers posed by the incongruence is assumed to play a large role in prejudicial attitudes towards women leaders (e.g., Eagly & Karau, 2002; Koenig, et al., 2011; Schein, 2001). Since the initial studies (Schein, 1973, 1975), numerous replications have been produced which show that the characteristics that people commonly ascribe to leaders are highly similar to the characteristics that people commonly ascribe to men (e.g., Carli, et al., 2016; Duehr & Bono, 2006; Fischbach et al., 2015; Koenig, et al., 2011; Schein, 1975; Schein, 2001). However, it has not been empirically demonstrated that congruence actually matters for the perception of leadership in real people.

The conclusion sections across TMTM papers are nearly identical. First, they conclude that women are perceived to be less qualified for leadership positions because of trait content incongruence. Second, the incongruence leads to lower expectations for women's performance in leadership positions. Third, women find it more difficult to be leaders than men, and subordinates may respect them less. Thus, researchers conclude that incongruence between gender stereotypes and leader stereotypes create hidden psychological barriers for women. Such studies do not even mention an alternative hypothesis. For example, Fischbach et al. (2015) conducted a TMTM study using emotions as trait content and concluded:

> We propose [gender] stereotypes are a strong barrier for women's career development and a challenge in their daily work because they may result in prejudice and devaluation of their work performance …The findings indicate the well-documented strong-male leader similarity and the weak-woman leader similarity hold ...this is [a] reason why management positions appear less attractive to women and are so difficult for them to achieve (Fischbach, et al., 2015, p 161).

Such a conclusion may be warranted, but TMTM studies do not include a criterion. The lack of criterion validation is troubling considering the conclusion sections of the papers discuss their findings as though there was a demonstrated empirical connection to an outcome. Given the weak evidence of the relationship between stereotypes and the behavior towards real people (see Blanton, Jaccard, Klick, Mellers, Mitchell, & Tetlock, 2009; Jussim, Cain, Crawford, Harber, & Cohen, 2009), it would seem the conclusions are much stronger than the available evidence warrants. Thus, the conclusion of the effects of the TMTM phenomenon are suspect.

Limitations of TMTM Interpretations

There are several limitations to common conclusions of TMTM studies. First, these studies do not include a criterion variable. Typically, social science researchers conduct some kind of multivariate analysis in order to make claims about the relationship between variables (e.g., multiple linear regression). In the TMTM paradigm, the incongruence itself, computed as an intraclass correlation, is taken as evidence of current and future prejudicial and discriminatory consequences for women. But, TMTM studies only indicate a difference in similarity between ratings of men, women, and managers. There is no evidence of the consequences.

Second, the relative degree of difference between ratings of categories (e.g., men and leaders) is not automatically meaningful. Categorical boundaries are fuzzy and overlap (Rosch, 1973; Lakoff, 1987). Lakoff (1987) described categorization as containing many distinct properties. According to Lakoff (1987), categories are embodied with regard to their content, which means that they can be understood without

necessarily being defined. For example, we know that a "dog" is something furry and may be friendly, but we would have trouble exhaustively describing its content. We see a dog and we know it is a dog, we understand this without thinking. Abstract categories (e.g., the dog's breed) are applied more slowly and require more attention and effort (Lakoff, 1987).

Within categories, a degree of membership exists (Rosch, 1973). Some members of categories have more prototypical traits than other members, which is to say that a penguin, for example, is a bird, but a hawk is more bird-like because of its ability to fly. To say that separate categories (e.g., men and women) do or do not belong to another category (e.g., leader) is not a straight forward process. It is not necessarily the similarity of the content of the sub-categories to the higher category that makes a "man" a "leader". Is a man a leader? Obviously, to ask the question absent any other information is nonsensical. It is not necessarily the characteristics of categories that make them part of the category per se, rather, it is something about their potential for interaction with us (Lakoff, 1987). For example, the category "bed" contains all objects that serve some function. Is a pile of hay a bed? Yes, if you can sleep on it. Is a woman a leader? Yes, if she can inspire the group towards a common goal.

Combining concepts and deriving the meaning from that combination is no simple matter either. Fodor & Lepore (1996) provide the example of "pet fish". The prototype for "pet" and the prototype for "fish" are different and the associated trait descriptions would be different as well. The prototype for pet fish, on the other hand, conjures to mind a more specific range of possibilities. The distance of an object from the prototypical pet fish is not a function of its distance from the prototypical pet and its distance from the prototypical fish. Knowing that "pet" and "fish" have prototypes does not allow one to predict if the prototypical "pet fish" is more like a "gold fish" or a "cat". This is because categories are not prototypes, instead they define the necessary conditions for categorization and not statistical correlates (Fodor & Lepore, 1996). Thus, it is not the nature of their statistical properties that categories form meaning through their combination.

Furthermore, it is possible to know exactly what is meant by "male managers" but have no idea which male managers are prototypical. This is because the specific combinations of categories are not mere combinations of the basic level category attributes (Fodor & Lepore, 1996). We do not determine the meaning of the combination of categories by combining the prototypes' attributes. Pet fish means something different than the combined similarities of pet and fish (Fodor & Lepore, 1996). Even if we knew what was common between two categories we could not determine what the combination of those categories would look like in the real world. We know what pet fish means because we have knowledge about the real world—we know what kinds of fish people keep as pets. Combining individual prototypes to determine the meaning of the new combination would be an irrational strategy for inferring meaning. Yet, the statistics in TMTM studies are computed and interpreted based off of this exact process.

The lack of criterion-related validity, taken together with the possibility that the similarity between trait ratings of categories is not necessarily meaningful, suggests that the claims of TMTM researchers may be overstated.

The Present Study

According to the TMTM paradigm (Schein, 1973)¬ and similar but more recent theories (e.g. role congruity theory; Eagly & Karau, 2002; lack of fit theory; Hielman, 1983), the following two hypotheses could be generated:

Hypothesis 1a. Greater congruence between respondents' leader prototype traits and male gender stereotypes will be positively related to male supervisor evaluations.

Hypothesis 1b. Greater congruence between respondents' leader prototype traits and female gender stereotypes will be positively related to female supervisor evaluations.

However, as argued previously, there are reasons to suspect that this logic is flawed. Given that no research to date has directly tested the TMTM hypothesis, the present study aims to empirically investigate

these claims. In order to test the fundamental claim of the TMTM studies, one needs to replicate the general research design in the original Schein (1973) studies and compare these results to some outcome. This immediately creates a problem: the between-group nature of the design does not produce variability in the congruence statistics. For example, the correlation between ratings of men and leaders is one summary number (i.e., the intraclass correlation coefficient). Conducting the study as a repeated measures design, however, does allow for variation. Specifically, one can compute a congruence statistic (e.g., the similarity of ratings between women and managers) for each person. This can then be correlated with some outcome variable. This then allows for a direct test of the TMTM congruence hypothesis.

In order to test the TMTM hypothesis, some trait content needs to be selected for participants to rate. For the present study, agency and communion will be used as rating content. Agency and communion have been found to relate to the judgement of others, self, group, and gender perceptions (Abele & Wojciszke, 2007; Fiske, 2019). The agency dimension relates to dominance, competence, and assertiveness. The communion dimension relates to nurturance, interdependent self-construal, and consideration. Patterns of agency and communion have been identified in perceptions of self (Wojciszke, Baryla, Parzuchowski, Szymkow & Abele, 2011; Gebauer, Wagner, Sedikides, & Neberich, 2013) and others (Leach, Ellemers, & Berreto, 2007; Brambilla, Sacchi, Rusconi, Cherubini, & Yzerbyt, 2012) as well as implicated as the basis for many stereotypes (Fiske, 2019). Agentic behavior of others is indicative of competence, which leads to respect, while communal behaviors are interpreted as warmth, which leads to liking (Wojciszke et al., 2009). According to the stereotype content model (Fiske, 2019), warmth judgements follow from a perception of cooperation which include both friendliness and morality (Kervyn, Fiske, & Yzerbyt, 2015). Communal traits and behaviors are likely to be interpreted as warmth and communal leaders are more likely to be liked than leaders low on communion. Thus, given their role in the perception of leaders and gender stereotypes, the content is deemed appropriate for rating.

The present study seeks to answer the following question: Does the TMTM phenomenon have an observable effect on the judgement of real managers?

METHODOLOGY

Participants and Data Collection Procedure

Participants were recruited through Amazon's Mechanical Turk (MTurk), an online platform which hosts tasks for people to complete Human Intelligence Tasks (HITs) in exchange for monetary compensation. The survey was restricted to those with a HIT approval rate greater than 95%, greater than 5,000 HITs approved, and were located in the United States. MTurk workers were paid $0.50 each.

The initial data collection yield 968 responses. Due to manipulation check failures, inattentive responding, or non-purposeful responding (e.g., excessively endorsing the same anchors) some cases were excluded. At the end of the survey, participants were asked which targets they rated, cases were retained if they correctly selected men, women, leaders, and their current supervisor. Survey responses were examined for visual patterns of inattentive responding (e.g., a diagonal response pattern, selecting "1" for every response, etc.). In addition, participants who entered bogus responses to open-ended questions regarding their position or industry were not included. In total, 288 cases were removed. Thus, the final sample consisted of 680 responses.

Participants of the final sample were 53% female and 40 (SD = 15.29) years old on average. They were mostly employed full time (62%), others part time (20%), and others unemployed (18%). They have been in their current job for an average of 5.7 (SD = 6.63) years. They were more likely to have a male supervisor (55%) than a female supervisor and have worked with him or her for an average of 4.3 (SD = 4.66) years. Participants worked in a variety of self-described industries (over 200 unique responses). The five most common industries were education, retail, healthcare, manufacturing, and information technology.

Study Design

This study adopts Schein's (1973) TMTM research design but instead is a repeated-measures design. Each participant responded to questionnaires about three targets: a leader, men in general, and women in

general. The leader target is used as to be more generalizable with a variety of TMTM studies (e.g., Koenig et al., 2011). Participants were asked to rate how characteristic various traits are of each target. They also filled out a survey evaluating their current supervisor.

One concern that has been raised about within-subject designs is the potential for carry over effects. To control for this, the questionnaire order was randomized. The concern about carry over effects may be overstated as the lack of context in between-group studies may be a bigger problem (Birnbaum, 1999). In addition, meta-analytic research suggests there is no statistically significant difference between results for within-subjects and between-subjects research designs for studies examining gender discrimination (Davison & Burke, 2000).

Measures

Agency

The agency scale (Abele, Uchronski, Suiner, & Wojciszke, 2008) consists of 8 item (e.g., able, independent, rational) and was on a 5-point scale. Respondents were asked to rate how characteristics each item was for each target.

Communion

The communion scale (Abele et al., 2008) consists of 8 items (e.g., caring, sympathetic, understanding) and was on a 5-point scale. Respondents were asked to rate how characteristics each item was for each target.

Evaluations

Participants evaluated their supervisor on three dimensions: respect, liking, and effectiveness. The respect scale (Wojciszke et al., 2009) consists of 3 items ($\alpha = .94$) and includes items such as "I respect my supervisor"; the liking scale (Wojciszke et al., 2009) consists of 3 items ($\alpha = .93$) and includes items such as "I like my supervisor"; and the effectiveness scale consists of 3 items ($\alpha = .92$) and includes items such as "my supervisor is competent at his/her job". When the participant reported being unemployed, they were prompted to think about their previous supervisor instead.

Data Analysis

Polynomial regression analyses were run and corresponding surface graphs were plotted for the significant regression models in order to test the hypotheses that congruence related to supervisor evaluations (Edwards, 2002). First, predictors were mean-centered in order to reduce their multicollinearity and to ease interpretation of interactions (Aiken & West, 1991). The mean of the predictors was chosen in order to analyze the specific claims in Schein (1973) paradigm studies and are of special interest because the mean of a category represents the perception of the most typical member of that category. Second, higher-order terms were produced from agency and communion within each condition. For example, when comparing women and leaders on agency, women-agency squared, leader-agency squared, and the product of women and leader agency were created. Third, evaluations were regressed on these five predictors. Before recording the results for each polynomial regression, Mahalanobis' distance was computed to detect multivariate outliers as outliers can influence polynomial regression results (Edwards, 2002). Cases which exceeded the critical value of 20.52 (df = 5, p <.001) were excluded from the subsequent analyses.

Testing multiple hypotheses can increase Type 1 error, and thus, Holm's sequential Bonferroni correction was applied to limit this when examining the overall model's R^2 values for polynomial regressions (Holm, 1979). The Holm's correction involves a stepwise increase of the significance level (often based on $p < 0.05$). The method divides .05 by the number of tests and uses that value as the new rejection level for the largest effect size. Then, the rejection level is divided by (n minus 1) to examine the second largest relationship, and so on.

If the overall model (R2) of a given polynomial regression equation ($Z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + e$) was significant, a response surface graph was plotted. The surfaces can then be analyzed to test congruence hypotheses and examine the fit patterns. There are two references for a surface: a congruence

line and an incongruence line. Each line has two parameters: slope (a_1 and a_3) and curvature (a_2 and a_4). Thus, each surface has four parameters which help in the interpretation of the graph. The slope of the congruence line is given by $a_1 = (b_1 + b_2)$, where b_1 is the unstandardized beta coefficient for the X variable and b2 is the unstandardized beta coefficient for the Y variable. Curvature along the line of congruence is assessed by calculating $a_2 = (b_3 + b_4 + b_5)$, where b_3 is the unstandardized beta coefficient for X squared, b_4 is the unstandardized beta coefficient for the cross-product of X and Y, and b_5 is the unstandardized beta coefficient for Y squared. The slope of the line of incongruence is assessed by calculating $a_3 = (b_1 - b_2)$. The curvature of the line of incongruence is assessed by calculating $a_4 = (b_3 - b_4 + b_5)$. Significance tests for these response surface parameters was computed using a bias-corrected percentile method (Efron & Tibshirani, 1993) on 10,000 bootstrapped samples as suggested by Edwards (2002).

To support a congruence effect, several conditions of the response surface must be met (Humberg, Nestler, & Back, 2019). First, the line of congruence must not differ significantly from zero, and the line of incongruence must not be significantly different from one. a_4 must be significantly negative and a_3 must not be significantly different from zero. If any of these conditions are violated, a strict congruence effect is rejected. However, the general fit pattern can still be examined via the response surface visualization as long as the overall polynomial regression model is statistically significant. A strict congruence effect is not a requirement to support the TMTM hypothesis in the current study, but some congruence effect ought to be observable if the hypothesis is to be supported.

RESULTS

Prior to hypothesis testing, an exploratory factor analyses was conducted combining agency and communion ratings across men, women, and leaders, including evaluations of the current supervisor, to assess potential common method variance issues. Using an oblique rotated factor analysis, the first factor explained 27.59% of the variance and all of the factors combined explained 63.28%. In addition, construct validity evidence was examined. The exploratory factor analysis yielded satisfactory results according to the guidelines of simple structure (Kline, 2002). However, the evaluation criteria (i.e., respect, liking, and effectiveness) were not adequately distinct (i.e. all factor loadings were above .8 and only on one factor) and so these were combined into one construct for future analyses ($\alpha = 0.96$). This factor is referred to as Evaluations.

Congruence Hypotheses

Results of polynomial regressions are interpreted based on slopes and curvatures along the line of congruence and incongruence. A significant increase in the slope along the congruence line indicates that outcomes increase when ratings of supervisors and leader increase. Results of each predictor and the surface values are reported in Table 2. However, strict conditions for congruence also require the significance testing of parameters of the first principal axis (Edwards, 2002; Humberg et al., 2019). The first principal axis can be expressed as a linear equation that relates Y to X such that $Y = p_{10} + p_{11}X$. The values of p_{10} and p_{11} are computed from the coefficients in the polynomial regression equation.

To test the TMTM hypotheses that gender-leader congruence relates to supervisor evaluations, polynomial regression analyses were conducted. The different models (see Table 2) incorporate different stereotype content depending upon the gender of the supervisor. For example, if the respondent's supervisor was male, only male stereotype ratings were used to create the congruence variable.

The polynomial regression models were significant for both men supervisors (agency; $R^2 = .054$, $p < .05$, communion; $R^2 = .063$, $p < .05$) and women supervisors (agency; $R^2 = .121$, $p < .05$, communion; $R^2 = .108$, $p < .05$). See Figure 1 for the associated response surfaces. For men-agency, a strict congruence effect cannot be supported ($p_{10} \neq 1$). The men-agency line of incongruence has a significant negative slope ($a_3 = -.450$, $p < .05$) suggesting evaluations are higher when leader agency ratings exceed agency ratings for men. The men-communion surface shows a general congruence effect with an additive main effect ($p_{10} = 0$, $p_{10} = 1$, $a_1 = .420$, $p < .05$, $a_4 = -.470$, $p < .05$). The men-communion line of incongruence has a negative curvature ($a_4 = -.470$, $p < .05$) suggesting that as the discrepancy between men and leader communion

ratings increases, evaluations declined more sharply. For both women-agency and women-communion a strict congruence effect cannot be supported ($p_{10} \neq 0$, $p_{10} \neq 1$). Both women conditions show an additive main effect.

TABLE 1
MEANS, STANDARD DEVIATIONS, AND CORRELATIONS

Variable	M	SD	1	2	3	4	5	6	7
1. Men-A	3.99	0.57	(.85)						
2. Men-C	3.21	0.68	.55	(.89)					
3. Women-A	3.76	0.67	.51	.45	(.89)				
4. Women-C	4.18	0.66	.59	.32	.71	(.91)			
5. Leader-A	4.37	0.56	.51	.27	.37	.43	(.84)		
6. Leader-C	3.80	0.75	.32	.42	.39	.32	.53	(.91)	
7. Evaluations	3.78	1.03	.18	.21	.19	.20	.21	.27	(.96)

Note. $n = 680$. A = Agency, C = Communion.
Variables ranged from 1 to 5.
All reported correlations are statistically significant $p < .05$.
Reliabilities are given in parentheses.

TABLE 2
RESULTS FROM POLYNOMIAL REGRESSION ANALYSES

Surface	b					R^2	Position of first principal axis		Shape of surface long lines LOC		LOIC	
	X	Y	X^2	XY	Y^2		p_{10}	p_{11}	a_1	a_2	a_3	a_4
Men Agency	-0.01	**0.44**	-0.31	0.06	0.14	0.054*	0.947	**15.066**	**0.430**	-0.110	**-0.450**	-0.230
Men Communion	**0.20**	**0.22**	-0.10	0.21	-0.16	0.063*	0.145	0.754	**0.420**	-0.050	-0.020	**-0.470**
Women Agency	**0.36**	**0.31**	**-0.56**	-0.03	0.30	0.121*	**21.113**	**-64.537**	**0.670**	-0.283	0.050	-0.230
Women Communion	**0.24**	**0.32**	**-0.39**	0.05	0.01	0.108*	**-4.271**	**15.232**	**0.557**	-0.329	-0.081	-0.432

Note. N = 680 but cases were only used if the supervisor's sex matched the gender category, thus, n's ranged from 365 for men and 293 for women supervisors.
b unstandardized regression coefficient, X values are the relevant gender category, Y values are the leader category.
Bold indicates the 95% bias corrected percentile method confidence intervals from 10,000 bootstrapped samples do not include zero.
The position of the first principal axis in the X-Y-plane is given by Y = p_{10} + p_{11}X. Strict congruence effects must show that the 95% confidence interval includes 0 for p_{10} and includes 1 for p_{11}.
LOC line of congruence, the slope and curvature are given by a_1 and a_2 respectively, LOIC line of incongruence, the slope and curvature are given by a_3 and a_4, respectively.

*$p < .05$, for R^2 values using the Holm's sequential Bonferroni correction.

FIGURE 1
THE EFFECTS OF LEADER-GENDER FIT ON SUPERVISOR EVALUATIONS
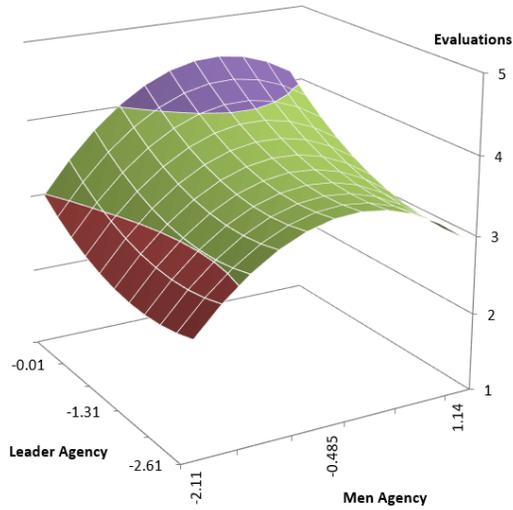
Figure 1a
Leader-Men Agency

Figure 1b
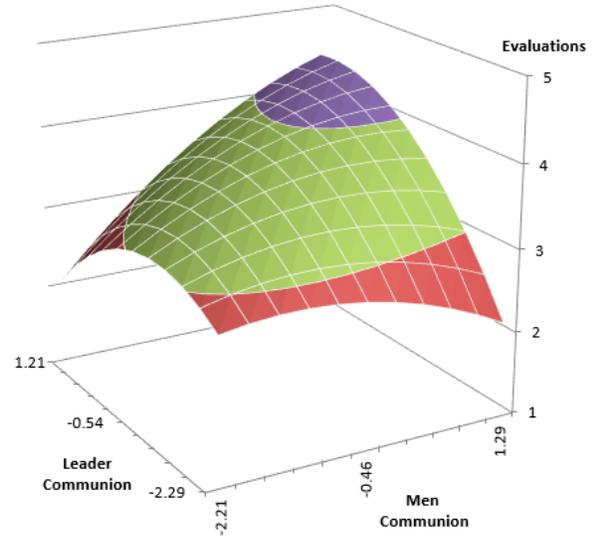*Leader-Men Communion*

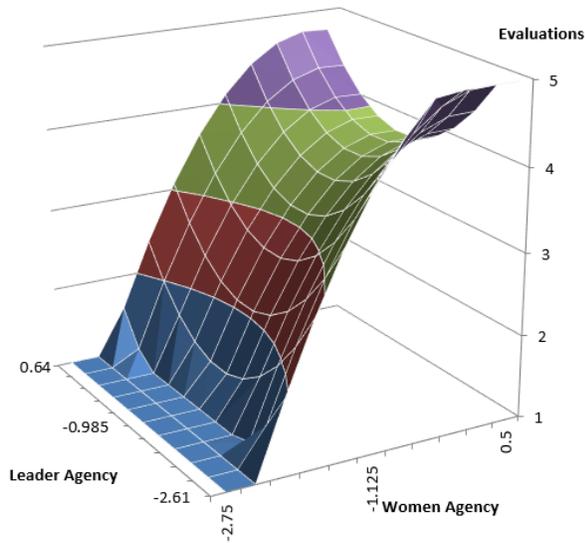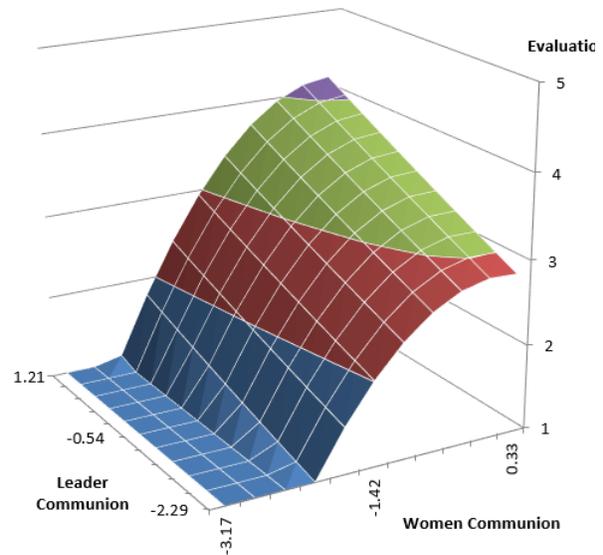**Figure 1c**
Leader-Women Agency

Figure 1d
Leader-Women Communion

CONCLUSION

The purpose of the present study was to formally test the claim made by Schein's (1973) TMTM paradigm. Specifically, the purpose was to examine the congruence, or fit pattern, between peoples' general views of men, women, and leaders, and then determine the potential consequences of (in)congruence on the evaluations of their current supervisors. None of the models produced the congruence effect implied by

the TMTM research paradigm. In other words, the disagreement between the stereotypes of women and beliefs about leadership do not lead to lower evaluations for women in managerial roles.

These results are in direct conflict with what Schein and others have argued (e.g., Carli et al., 2016; Fischbach et al., 2015; Eagly & Karau, 2002; Schein, 2001). The highest ratings for women supervisors occur when people hold typical (i.e., average) views of women, not when they were incongruent with the leader prototype. TMTM research, together with related ideas such as role congruity theory (Eagly & Karau, 2002), predict lower evaluations for women due to the incongruence hypothesis. The results of this study were near opposite of the TMTM predicted direction.

Strengths and Limitations

There are three specific strengths of this design worth noting: the use of polynomial regression to test congruence, or fit hypotheses; the design similarity to Schein (1973) studies; and the use of agency and communion as rating content.

The present study has limitations in terms of inferring causality. The study did not use an experimental manipulation with random assignment to conditions, and therefore, one cannot make a strong causal inference of the results. Future research could benefit by employing an experiment with manipulation of predictor variables.

Another potential limitation is common method bias. It is unlikely that common method variance biased the results of the study in a major way. One of the possible effects of common method bias is the inflation of observed correlations. However, Simesen, Roth and Oliveira (2010) argued that the interaction terms created in a polynomial regression are not severely susceptible to common method effects. In addition, Holm's sequential Bonferroni was applied to decrease the probability of committing Type 1 errors (Holm, 1979). Although, most of the response surface parameters were not statistically significant. The mean differences of the agency and communion variables across the main conditions were in the same rank order as other between-group designs (e.g., Carli et al., 2016). This suggests that the study was not excessively affected by altering the data collection strategy.

Theoretical Implications

The present study is the first to test the implications of congruence as suggested by Schein's (1973, 1975) TMTM paradigm studies. Empirical findings in the present study did not support the claims of past TMTM studies. This is a significant new contribution, as the studies published using Schein's paradigm have never attempted to link their findings to any criterion. An assumption is made in these studies that incongruence between female gender stereotypes and leader stereotypes is problematic for women leaders. The current study applied the current best practice to test congruence hypotheses (i.e., polynomial regression analysis; Edwards, 2002) and found no support for this common claim.

Practical Implications

The present research demonstrated that the congruence between men, women, and leader stereotypes may not necessarily pose a problem for the evaluations of actual supervisors. Diversity training initiatives should be mindful about overgeneralizing and exaggerating the effects of stereotypes on person perception. A problem in any field is how the layperson interprets and uses scientific evidence. Advocates who run away with scientific findings may be misinterpreting the magnitude of the consequences. In this way, I agree with Eagly (2016)—diversity advocates who misunderstand or ignore scientific research are making matters worse for both society and science.

REFERENCES

Abele, A.E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards and operationalization of fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, *38*(7), 1202–1217. https://doi.org/10.1002/ejsp.575

Abele, A.E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Aiken, L.S., & West, S.G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: SAGE Publications, Inc.

Birnbaum, M. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*, 243–249. 10.1037/1082-989X.4.3.243

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, G., & Tetlock, P.E. (2009, May). Strong claims and weak evidence: Reassessing the predictive validity of the IAT. *Journal of Applied Psychology*, *94*(3), 567–82; discussion 583–603. doi: 10.1037/a0014665. PMID: 19449998

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V.Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology*, *51*(1), 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Carli, L., Alawa, L., Lee, Y., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women ≠ scientists. *Psychology of Women Quarterly*, *40*(2), 244–260. doi:10.1177/0361684315622645

Davison, H.K., & Burke, M.J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*(2), 225–248. https://doi.org/10.1006/jvbe.1999.1711

Duehr, E.E., & Bono, J.E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology*, *59*, 815–846.

Eagly, A.H. (2016). When passionate advocates meet research on diversity, does the honest broker stand a chance? *Social Issues*, *72*(1), 199–222.

Eagly, A.H., & Karau, S.J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598.

Edwards, J.R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Dragsow & N. Schmitt (Eds.), *The Jossey-Bass business & management series. Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 350–400). Jossey-Bass.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fiske, S.T. (2019). Stereotype content model. In A. Abele & B. Wojciszke (Eds.), *Agency and Communion in Social Psychology*. London: Routledge. https://doi.org/10.4324/9780203703663

Fischbach, A., Lichtenthaler, P.W., & Horstmann, N. (2015). Leadership and gender stereotyping of emotions: Think manager – think male*? Journal of Personnel Psychology*, *14*(3), 153–162. https://doi.org/10.1027/1866-5888/a000136

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: Why concepts still can't be prototypes. *Cognition*, *58*(2), 253–270. https://doi.org/10.1016/0010-0277(95)00694-X

Gebauer, J.E., Wagner, J., Sedikides, C., & Neberich, W. (2013). Agency-communion and self-esteem relations are moderated by culture, religiosity, age, and sex: Evidence for the "self-centrality breeds self-enhancement" principle. *Journal of Personality*, *81*(3), 261–275. https://doi.org/10.1111/j.1467-6494.2012.00807.x

Heilman, M.E. (1983). Sex bias in work settings: The lack of fit model. *Research in Organizational Behavior*, *5*, 269–298.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pp. 65–70.

Humberg, S., Nestler, S., & Back, M.D. (2019). Response Surface Analysis in Personality and Social Psychology: Checklist and Clarifications for the Case of Congruence Hypotheses. *Social Psychological and Personality Science*, *10*(3), 409–419. doi:10.1177/1948550618757600

Jussim, L., Cain, T., Crawford, J., Harber, K.., & Cohen, F. (2009). The unbearable accuracy of stereotypes. In T. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination* (pp. 199–227). Hillsdale, NJ: Erlbaum.

Kervyn, N., Fiske, S., & Yzerbyt, V. (2015). Forecasting the primary dimension of social perception: Symbolic and realistic threats together predict warmth in the Stereotype Content Model. *Social Psychology*, *46*(1), 36–45. https://doi.org/10.1027/1864-9335/a000219

Kline, P. (2002). *An easy guide to factor analysis*. London: Routledge.

Koenig, A.M., & Eagly, A.H. (2014). Evidence for the social role theory of stereotype content: Observations of groups' roles shape stereotypes. *Journal of Personality and Social Psychology*, *107*(3), 371–392. https://doi.org/10.1037/a0037215

Koenig, A.M., Eagly, A.H., Mitchell, A.A., & Ristikari, T. (2011). Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological Bulletin*, *137*(4), 616–642. https://doi-org.ezproxy.gl.iit.edu/10.1037/a0023557

Kristof-Brown, A., & Guay, R.P. (2011). Person–environment fit. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol. 3. Maintaining, expanding, and contracting the organization* (pp. 3–50). American Psychological Association. https://doi.org/10.1037/12171-001

Kristof-Brown, A.L., Zimmerman, R.D., & Johnson, E.C. (2005). Consequences of individual's fit at work: A meta-analysis of person-job, person-organization, person-group, and person-supervisor fit. *Personnel Psychology*, *58*(2), 281–342. https://doi.org/10.1111/j.1744-6570.2005.00672.x

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Leach, C.W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology*, *93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Lord, R.G., Foti, R.J., & de Vader, C.L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior & Human Performance*, *34*(3), 343–378. https://doi.org/10.1016/0030-5073(84)90043-6

Rosch, E.H. (1973). Natural categories. *Cognitive Psychology*, *4*(3), 328–350. doi:10.1016/0010-0285(73)90017-0. ISSN 0010-0285

Schein, V.E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology*, *57*, 95–100.

Schein, V.E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology*, *60*, 340–344.

Schein, V.E. (2001). A global look at psychological barrier to women's progress in management. *Journal of Social Issue*, *57*, 675–688.

Schneider, B. (2001). Fits About Fit. *Applied Psychology*, *50*, 141–152. doi:10.1111/1464-0597.00051

Siemsen, E., Roth, A., & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, *13*, 456–476. https://doi.org/ 10.1177/1094428109351241

Tavarez, G.M., Sobral, F., Goldszmidt, R., & Araújo, F. (2018). Opening the Implicit Leadership Theories' Black Box: An Experimental Approach with Conjoint Analysis. *Frontiers of Psychology*, *9*, 100–112. doi: 10.3389/fpsyg.2018.00100

Wojciszke, B., Abele, A.E., & Baryla, W. (2009). Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency. *European Journal of Social Psychology*, *39*(6), 973–990. https://doi.org/10.1002/ejsp.595

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A.E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology*, *41*, 617–627.