JOIBS: April 2024. ISSN 2992-9253



# Schein (1973) Revisited: The Think Manager-Think Male Effect is a Statistical Artifact

Matthew W. Lauritsen, Independent Scholar. E-mail: matthewlauritsenphd@gmail.com

Funding: None.

Competing interests: None.

Citation: Lauritsen, M. W. (2024). Schein (1973) revisited: The think manager-think male effect is a statistical artifact. *Journal of Open Inquiry in the Behavioral Sciences.* https:/doi.org/10.58408/issn.2992-9253.2024.02.02.0001

**Abstract**

Schein (1973) is a highly cited article in research on sex and gender biases. The original article concluded that people are biased against women regarding requisite management characteristics. However, the present paper replicates Schein (1973) and demonstrates that the findings were a result of an imbalanced ratio of items which exhibited mean differences between men and women targets. In addition, the use of intraclass correlations creates an illusion of large differences or similarities between targets when the actual mean rating differences are practically trivial and statistically nonsignificant. A bias against women, against men, and no bias are obtained by altering the number of male and female items, or by applying the intraclass correlation to more appropriate data. The implications of the results for the measurement of sex and gender biases are discussed. Broader concerns are raised about ideological biases which allow for conclusions and theories to propagate without empirical support.

*Keywords*: think manager-think male, gender, bias

## Introduction

Schein (1973) reported that "managers are perceived to possess those characteristics, attitudes and temperaments more commonly ascribed to men in general than to women in general" (p. 99). Schein (1973) is the foundational work of the think manager-think male paradigm (Koenig, Eagly, Mitchell, & Ristikari, 2011). The study was published about 50 years ago but is still being cited in leading journals as support for sex and gender bias in organizational settings (e.g., Ma, Rosette & Koval, 2022). The study also laid the groundwork for theories of sex and gender discrimination in leadership (e.g., role congruity theory, Eagly & Karau, 2002; lack of fit model; Heilman, 1983). The broad conclusions of which are that the phenomenon has negative consequences for aspiring women leaders. The original effect has been replicated many times and continues to be found even in relatively recent articles (e.g., Carli, Alawa, Zhao, & Kim, 2016). Researchers claim that sex and gender biases will continue to be a problem for women as long as evidence for the think manager-think male effect continues to be found, (Heilman & Eagly, 2008).

Research studies in this paradigm typically consist of three key features: (1) participants are randomly assigned to rate men, women, or managers on various adjectives in a between-groups experimental design, (2) the adjective ratings are from a 92-item measure known as the Schein Descriptive Index, and (3) intraclass correlations are computed to test the hypothesis that men and managers are more similar to each other than women and managers. The studies tend to find a large intraclass correlation for men and managers and a relatively small intraclass correlation for women and managers.

Conclusion sections across these papers are nearly identical. They commonly claim that women experience barriers to leadership, men discriminate against women, or hiring and promotion decision-makers are biased against women. While these statements may ultimately be true, the claims are unjustified because none of these studies have ever empirically linked their findings to the perceptions of the qualifications of women, selection decisions regarding women, or any outcome variable. Yet, these studies continue to be published using the same methodology and statistical analysis and purport the same types of consequences since that of the early 1970s (e.g., Schein, 1973, 1975; Duehr & Bono, 2006; Koenig et al., 2011; Carli et al., 2016). Continued publications of the effect give credence to the idea that gender-based prejudice and discrimination are as strong as they ever were (Heilman & Eagly, 2008).

This paper will argue that the findings obtained in the think manager-think male paradigm do not provide evidence of anything more than a statistical artifact.

## The Schein (1973) Study

Participants were randomly assigned to rate men, women, or successful middle managers on 92 items on the Schein Descriptive Index in a between-subjects experimental design. The items consisted of mostly trait-like adjectives such as "logical" but also included more ability-like phrases such as "able to separate feelings from ideas." Prior to the main study, in developing the Index, 131 items that differentially described males and females were rated by 24 male and female college students. Half of the students rated these items for how characteristic they were of "men in general" and the other half rated "women in general." Item ratings across the two conditions were then compared. A key decision by the researcher was to retain items only if they exhibited statistically significant mean differences, as the result of a t-test, between the men and

women conditions. This decision was intentional in order to "maximize the differences in the descriptions of Women and Men" (Schein, 1973, p. 96). The final item set consisted of 92 various items which exhibited mean differences between men and women.

The Index contains items which are rated on a 5-point scale, ranging from 1 (not characteristic) to 5 (characteristic) with a neutral rating of 3 (neither characteristic nor uncharacteristic). The instructions on each of the three forms of the Index were as follows:

> On the following pages you will find a series of descriptive terms commonly used to characterize people in general. Some of these terms are positive in connotation, others are negative, and some are neither very positive nor very negative. We would like you to use this list to tell us what you think (women in general, men in general, or successful middle managers) are like. In making your judgments, it may be helpful to imagine that you are about to meet a person for the first time and the only thing you know in advance is that the person is (an adult female, an adult male, or a successful middle manager). Please rate each word or phrase in terms of how characteristic it is of (women in general, men in general, or successful middle managers).

There was no additional information about the target provided. Based on the instructions, participants likely rated typical versions of men, women, and successful middle managers, as opposed to an "ideal" version of these categories. However, the adjective "successful" likely inflated the ratings for the manager category. Given the lack of further instruction, it is not possible to know the extent to which respondents aggregated their real experiences with members of these categories or are relying purely on stereotypes of the categories when making their ratings. The psychometric properties of this instrument were not reported.

To determine the "degree of resemblance" (Schein, 1973, p. 97), intraclass correlation coefficients (ICCs) were computed on the means of item means. The ICC was computed twice, once for the men and managers conditions together, and then again for women and managers. The same manager item means were used in both analyses. A large and statistically significant ICC was found for men and managers. A small and nonsignificant ICC was found for women and managers. No ICC was reported for men and women. No mean difference values were reported. The original analysis was further broken down by the age of the participants. Schein reported slight variation in ICCs across three different age groups. However, substantive interpretation of subgroup variation in ICCs in this research design is ambiguous and should probably be avoided (Lauritsen, 2022). In addition, Schein (1973) also ran an analysis of variance (ANOVA) but did so in a manner identical to a t-test. Typically, an ANOVA is run by including all three of the experimental groups together, followed by a post-hoc analysis, rather than just two conditions at a time. An ANOVA was not run to test the difference of the men and women conditions. Schein (1973) reported a statistically significant difference in means between men and managers and no difference between women and managers.

To clarify the process for analyzing ICCs, Figure 1 contains a hypothetical data structure necessary for calculating an ICC for seven items and three measurements or observations. ICCs are calculated by aggregating the participant ratings and then rotating the data so that the rows are the items and the columns are the conditions. The Men-Managers ICC is calculated by excluding the Women column. Then, the remaining columns are used as the data for the ICC calculation. The Women-Managers ICC is calculated the same way except that the Men column is excluded.

This analytical procedure produces two ICCs—one for Men-Managers and one for Women-Managers. The difference between these values is usually not tested for statistical significance in the think manager-think male paradigm because the two ICCs share the Managers column and are thus dependent.

Figure 1. Example Dataset for ICC Calculations.

| Item | Men | Women | Managers |
|---|---|---|---|
| 1 | 3 | 2 | 4 |
| 2 | 4 | 1 | 5 |
| 3 | 2 | 3 | 3 |
| 4 | 5 | 4 | 6 |
| 5 | 1 | 5 | 2 |
| 6 | 3 | 2 | 4 |
| 7 | 2 | 4 | 3 |

Note. Values under the conditions represent the average ratings across participants.

In Figure 1, the means for Men are closer to the means for Managers (1-unit apart), while Women are further away (2-units apart on average). This will typically result in a higher ICC for the Men-Managers comparison than it will for the Women-Managers comparison. It should be noted that item ratings are not always more similar for Men-Managers as is the case for Item 3 in Figure 1. Had the item list consisted of a higher proportion of items similar to Item 3, one could imagine a situation where the Women-Managers means would be closer together than the Men-Managers means.

Schein's (1973) analytical procedure produces a problem of statistical inference that may not be immediately obvious. Comparing two ICCs as calculated in the above manner is analogous to calculating the difference of the differences between the column means. The difference of a difference using a common comparison results in the common comparison being immaterial to the result. In other words, because Schein's (1973) paradigm compares Men and Women against the same column (i.e., Managers), the difference between the ICCs is a function of the difference between the Men and Women columns. This can be demonstrated algebraically:

Step 1 $(Mgr - M) - (Mgr - W) = X$

Step 2 $Mgr - M - Mgr + W = X$

Step 3 $W - M = X$

Where X is the result of the difference of the differences, Mgr is the average Managers ratings, M is the average Men ratings, and W is the average Women ratings. Simplifying the equation shows that the result of the difference of the differences is the same as the difference between Men and Women.

To be fair, Schein (1973) does not simply take the difference of the average values. However, ICCs are calculated using the variance between columns. In most cases, greater mean differences in average ratings between columns increase the variance between them. Schein (1973) attempted to maximize the difference between ratings of Men and Women and thus the focal ICCs are inevitably different. If there were no mean differences between Men and Women in the initial

survey, the ICCs for Men-Managers and Women-Managers would be nearly identical.

The key understanding from the above exercise is that the relative distance between ICCs is not substantially meaningful as it relates to the perceptual similarity of men, women, and managers. No matter what the ICC point estimate is for one comparison, the other will be higher or lower by an amount directly proportional to the difference between Men and Women. In other words, mean differences embedded in the Schein Descriptive Index make it nearly impossible to accept the null hypothesis that there are no differences between the ICCs. It is from this problem of statistical inference that a deeper investigation is warranted.

**Background**

Schein (1973) borrowed methodology from Broverman, Broverman, Clarkson, Rosenkrantz, and Vogel (1970). This study concluded that clinicians are biased against women when making judgements about mental health. The methodology in Schein (1973) is nearly identical to Broverman et al. (1970) except that Schein (1973) replaced "normal adult" with "managers in general." Several methodological and conceptual criticisms have been brought against the Broverman et al. study (Stricker, 1977; Widiger & Settle, 1987; Philips & Gilroy, 1985). The most damaging criticism being the empirical demonstration that the results were the consequence of a bias in the measurement, not in the participants (Widiger & Settle, 1987). Widiger and Settle (1987) demonstrated that the higher frequency of male-valued items increases agreement among those items for the male target. This creates the illusion of agreement with the adult target and disagreement with the women target. However, the Broverman et al. (1970) article continued to be accepted even after critical analyses were published in well-known journals (Kelly & Blashfield, 2009). Researchers who later studied sex and gender bias in clinical settings seemed to be unaware that the conclusions of Broverman et al. (1970) were erroneous. This case has been identified as an example of psychological science's failure to self-correct (Kelly & Blashfield, 2009).

The methodological similarities between Broverman et al. (1970) and Schein (1973) have gone seemingly unnoticed in the industrial-organizational psychology literature. There exists only one documented mention of this similarity which appears as a footnote in Koenig et al. (2011):

> The design of the studies in this paradigm appears similar to that of a study by Broverman et al. (1970), whose participants rate a "mature, healthy, socially competent" man, woman, or adult person. Broverman et al. found greater similarity between an adult person and men than between an adult person and women, but this finding has been criticized as artifactual (see Kelly & Blashfield, 2008; Widiger & Settle, 1987, for details). However, the similarity of the Broverman et al. study to the think manager-think male studies is only superficial because Schein's (1973) method offers superior item selection and statistical analysis. Specifically, (a) the typical think manager-think male items (in the Schein Descriptive Index) are moderately balanced between agentic and communal qualities (see Duehr & Bono, 2006), and (b) the statistical analysis with an intraclass correlation is more appropriate to the data (p. 620).

It is the central argument of the present study that neither the item selection nor the statistical analysis are superior to Broverman et al. (1970). Schein (1973) developed the Schein Descriptive Index by selecting items which maximized differences between the man and woman conditions

in her study. Items were not chosen based on their relevance to management nor were they intended to be indicators of an underlying construct. The number of items which were more characteristic of men likely outnumbered those more characteristic of women. The consequence of an unbalanced number of male versus female items is that the Index may be biased towards obtaining a finding consistent with the think manager-think male effect. A comparable hypothesis was tested and supported analyzing the Stereotype Questionnaire from the Broverman et al. (1970) study (see Widiger & Settle, 1987). In Widiger and Settle (1987) it was shown that an imbalance in the number of male and female items determined Broverman et al.'s (1970) findings. In other words, there was no evidence of sex bias in Broverman et al. (1970). Instead, the results were an artifactual measurement bias. The statistical analysis in Schein (1973) differs from Broverman et al., (1970) in that Schein calculated ICCs on item means across conditions. What has not been realized in this paradigm is that even trivial differences in item ratings can cause very large differences in ICC values (Lauritsen, 2022). Given the high degree of similarity between Broverman et al. (1970) and Schein (1973), potential measurement bias in the think manager-think male paradigm warrants investigation.

If the measurement and statistical analysis within Schein (1973) are flawed, then so are extant replications (e.g., Koenig et al., 2011; Carli et al., 2016) and any future replications. The potential flaws would also discredit adjacent studies in the paradigm (e.g., Fischbach, Lichtenthaler, & Horstmann, 2015). In addition, some theories of gender discrimination in leadership rely on the findings of the think manager-think male paradigm as an integral component to their theories (e.g., role congruity theory, Eagly & Karau, 2002; lack of fit theory; Heilman, 1983). The central tenet to these theories is a lack-of-fit hypothesis which states that the mismatch between stereotypes of women and desirable work roles causes biased workplace evaluations (Heilman & Eagly, 2008). Such negative consequences include refusal to hire, promote, or provide other advancement opportunities to women. If studies such as Schein (1973) which claim to demonstrate the existence of a lack-of-fit mechanism are flawed, then the lack-of-fit hypothesis loses some support. Think manager-think male studies continue to be published and thus the bibliographies supporting the lack-of-fit hypothesis continues to grow. However, it will be demonstrated in the present study that such findings ought not to be considered evidence of anything more than a statistical artifact. Thus, even though Schein (1973) is about 50 years old, the scientific integrity of the methods and measurement are relevant for contemporary issues in the study of gender and leadership.

**Overview of the Critiques of Schein (1973)**

The methodology, statistical analysis, and interpretation of the results in the think manager-think male paradigm are argued to be biased. Bias in scientific research refers to a systematic error in the design, method, analysis, or reporting of research that results in a distortion of the results or conclusions (Shadish, Cook, & Campbell, 2002). Bias can also be introduced during the interpretation of study results. For example, confirmation bias can occur if researchers only seek out evidence that supports their hypothesis, and ignore evidence that contradicts it (Nickerson, 1998). Bias in research can have serious consequences, as it can lead to incorrect conclusions and flawed policies or interventions.

Schein's (1973) think manager-think male research design contains several biases which severely degrade the paradigm's validity. First, bias is argued to exist in the measurement instrument itself. Items which are stereotypically male likely outnumber those which are stereotypically female.

Given that the manager category is labeled "successful middle manager," items with a positive connotation (e.g., leadership ability) will tend to be rated higher for this category. As a result, participants might rate both men and managers high on the rating scale but for different reasons, not necessarily because they view the two categories as similar.

Second, bias is argued to occur when think manager-think male researchers use statistics which are inappropriate, or in such a way that will increase the chances of obtaining a result in favor of their hypothesis. The ICC is not suitable to test between-group hypotheses in an experimental research design (Gwet, 2014; McGraw & Wong, 1996). ICCs are intended for use in within-group designs. Further, the confidence intervals of the ICC are typically much wider on their lower bound which increases the chances of accepting the null hypothesis for the women-managers comparison. Even average rating differences in the second decimal place on a 5-point scale can severely attenuate the ICC. This is due to the absolute agreement computation in the types of ICCs used in the paradigm. As the ICC is attenuated due to the built-in mean differences, the lower bound confidence interval widens and Type II error increases. A null finding (i.e., a confidence interval which contains zero) for the women-manager ICC is considered to be strong evidence for the think manager-think male hypothesis. Thus, both the measurement and statistical analysis are biased in favor of the paradigm's hypotheses.

Lastly, bias is argued to exist when researchers jump to conclusions or make assumptions about their findings which cannot be supported by available evidence. In this paradigm, rating dissimilarities between men, women, and managers are assumed to be automatically meaningful such that women face bias, prejudice, and discrimination as a result. However, studies in this paradigm do not contain a criterion, yet the content of the discussion sections appear as though they had. The claim that rating differences in this paradigm affect some outcome is a hypothesis that has never been tested. Typically, if a study wants to show evidence that X relates to Y, it ought to include a measure of Y. For example, many studies in this paradigm claim the results are evidence for the discrimination of women by men in hiring decisions. This is an unjustified claim as the congruence between the categories was never linked to an external criterion such as behavior. Such claims may be true, but the findings obtained in Schein's (1973) paradigm cannot be taken as evidence. A variety of different consequential claims have been made since the paradigm's inception and yet the methodology has not changed (Koenig et al., 2011). Thus, researchers are biased in making assumptions about the implications of their findings.

## The Methodological Problems

### *Item Selection*

Schein (1973) did not carefully consider the consequences of retaining items which exhibited mean differences between men and women for use in the Schein Descriptive Index. This may very well create the same artifactual bias as in Broverman et al. (1970). If men are rated higher than women on most of the items, and most of the items have positive connotations, it is no surprise that mean ratings are closer for men and managers than to women and managers. Two categories can be rated relatively similarly, but for different reasons, and this will be captured as agreement, or consistency, when computing reliability statistics such as the ICC. Likewise, two categories can be more discrepant in their ratings but not for the same reason. This problem has been shown to exist in Broverman et al. (1970) and may very well exist here (Widiger & Settle, 1987).

A related issue is the misinterpretation of the rating scale. Reyna (2018) suggests that ideological biases can affect the interpretation of measures in such a way that favors certain viewpoints. She argues that there is a tendency to focus on relative differences and ignore the absolute values associated with scales when the results would favor a certain viewpoint. This is something she calls the high-low fallacy. For example, people may agree that men possess some attribute and their mean on a 5-point disagree-agree scale is 3.7 while the mean for women is 3.4. If this difference is statistically significant, it is tempting to say that people disagree that women possess this attribute. However, both these values are on the same side of the scale in the agree category. In studies such as Schein (1973) ratings tend to differ in degree more often than in kind. Yet, these differences are treated as though they were on opposite sides of the rating continuum.

*Statistical Analysis*

Lauritsen (2022) argued that the use of ICCs within this paradigm is severely problematic. A summary of his main points follows:

1. ICCs are based on the analysis of variance (ANOVA; McGraw & Wong, 1996). The ICC is a measure of the proportion of a variance that is attributable to objects of measurement, often called targets (e.g., Shrout & Fliess, 1979). In the think manager-think male paradigm, the objects of measurements are the conditions (e.g., men, women, managers).
2. ICCs in this paradigm are one-way, single-rater, random effects models which use the absolute agreement definition (McGraw & Wong, 1996). The absolute agreement computation in these kinds of ICCs can severely attenuate the ICC point estimates. As a result, it is not uncommon to obtain negative ICCs in this paradigm.
3. Subject and rater error are confounded due to random assignment of individuals to only one condition. In addition, ICCs across men-managers and women-managers are dependent. Consequently, interpreting ICC variation across condition pairs or participant subgroups (e.g., participant gender) is ambiguous.
4. Schein's (1973) study design does not fit into any of the traditional reliability study design frameworks (Gwet, 2014; McGraw & Wong, 1996).
5. ICCs are meaningless when the measurement conditions are not of the same class (McGraw & Wong, 1996). In other words, because men, women, and managers are not of the same class, it would not be appropriate to compute an ICC across categories.
6. ICC confidence intervals are based on the F statistic and are asymmetrical such that the lower bound is often much wider than the upper bound. Combined with the above issues, this can result in distressingly wide lower bound confidence intervals.

Thus, the solution to this problem is not to use some “more appropriate” ICC as recommended by Koenig et al. (2011; p. 624). The central argument here is that there is no appropriate ICC for this kind of data. The between-group nature of the experiment precludes this possibility. This argument is supported by McGraw and Wong (1996) who state the ICC is only appropriate when there is a common population variance for all measurement conditions:

> Where this assumption is not met, it would be meaningless to calculate [the ICC] or any other ICC. This fact harks back to the original justification for the term intraclass, which is that measurements must be of a single class (p. 37).

To understand the conceptual problem of using reliability statistics in this manner, consider the meaning that one can derive from asking two different people to rate two different targets. Even if the ratings were the same, what exactly do they agree on? Interrater agreement is intended to reflect the agreement of different raters rating the same target. If one combines ratings from two different people on two different targets and computes an ICC between them, one is simply computing the variance between the ratings and nothing more—not unlike a spurious correlation.

Consider a comparison which will hopefully not generate controversy involving an eagle, a penguin, and the bird category. To know if eagles or penguins are more or less bird-like, or to know if they can be categorized as birds at all, it is necessary to know the key characteristics which define birds. Birds are defined by features such as having feathers, beaks, internal fertilization, and endothermic metabolism (Lovette & Fitzpatrick, 2016). Both eagles and penguins meet this requirement and can be categorized as birds. We determine that crocodiles are not birds because they lack several of the key characteristics (e.g., feathers). Discussion of the presence or absence of a trait outside of these characteristics (e.g., height) is irrelevant in terms of categorization. In the think manager-think male paradigm, there is no explanation of which combinations of traits are distinguishing qualities of the manager category. It would be helpful to know what defines a manager when attempting to determine the manager-like qualities of potential members. Given that the think manager-think male researchers do not define any of their study categories, it is difficult to interpret the findings in a meaningful way. Thus, rating differences in the paradigm are not necessarily informative, even if those rating differences are large.

Lauritsen's (2022) study revealed some other problematic ICC behavior. First, the study demonstrated that ICC differences do not necessarily correspond to meaningful differences in actual ratings. It is possible for the ratings of two comparisons to have identical mean differences but for the ICCs to be radically different due to differences in variance. For example, Lauritsen (2022) included two additional conditions, doctors and police chiefs, in addition to the traditional manager condition. The man- and woman- ICCs for -doctor was .62 and .60, respectively. The man- and woman- comparison ICCs for -police chief was .93 and .02, respectively. These comparisons exhibit close to the smallest and largest obtainable difference between ICCs (.02 for doctors and .91 for police chiefs). Yet, the mean differences only ranged from .10 to .35 across the comparisons and the difference of the mean differences is *exactly the same* (.18 on a 5-point scale). Lauritsen (2022) also demonstrated that if you compute the ICC with man, woman, and manager conditions together, as you normally would in an ANOVA, the ICCs are very high (.87 for men, women, and managers). Even readers who disagree with the arguments against ICCs herein must conclude that this would imply that there is a large resemblance between men, women, and managers.

**Politically Biased Interpretation of the Results**

Researchers in this paradigm tend to make assumptions about the meaning of their findings or jump to conclusions which are unjustified. Researchers who adhere to a gender feminist worldview make strong assumptions about gender differences which affect the interpretation of their findings. Namely, gender feminism claims that all differences between men and women are socially constructed, power is the singular social motive, and that people are better understood as members of groups, rather than as individuals (Pinker, 2003; Sommers, 1994).

Rating differences between men and women are not automatically meaningful, nor are they automatically negative. Two randomly selected individuals are highly unlikely to be the same in their knowledge, skills, abilities, or other characteristics relevant to management. The chances that the aggregate of men and women you have met in your lifetime have identical distributions of these characteristics is almost certainly zero—even if we lived in a world free from prejudice and discrimination (Pinker, 2003). The burden of proof that rating differences have negative consequences for women remains with the think manager-think male researchers. However, no direct evidence that rating dissimilarities have negative consequences has ever been presented. Thus, the conclusions stem not from evidence, but from some ideological worldview which sees all differences between men and women as a problem and purports to know the consequence of those differences without evidence.

Lauritsen (2020) treated these types of conclusions as testable hypotheses. He proposed that the think manger-think male phenomenon can be conceptualized as an individual difference. He investigated to what extent individual variation in the tendency to "think manager-think male" on ratings of agency and communion related to variation in the evaluations of peoples' actual male or female supervisors. Using polynomial regression, the results showed that the congruence between men, women, and leaders did not matter for evaluating real people. Interestingly, people who held the belief that men were more like leaders than women rated women more favorably than those who believed the opposite. In other words, "thinking male" may not lead to negative evaluations of women. One possible explanation for this is that people are generally accurate in their ratings of groups (Jussim, Cain, Crawford, Harber, & Cohen, 2009). Because peoples' beliefs about groups are generally accurate, they are probably not systematically biased by stereotypes when rating a real person. Another possibility is that the think manager-think male researchers are simply wrong—rating differences are not automatically consequential.

To demonstrate the biased interpretations within the paradigm, consider the kinds of conclusions made by researchers when the results come out differently than they had anticipated. The Schein Descriptive Index was used in 83% of the effect sizes in Koenig et al.'s (2011) meta-analysis. Of those, three studies found the effect in the opposite direction—meaning that the respondents rated women (as compared to men) more like managers. Rather than declare "think manager-think female," one study proposed that male participants were hiding their true feelings by being politically correct (Duehr & Bono, 2006) and another maintained that, despite contradictory findings, psychological barriers are caused by this phenomenon and still exist for women but not for men (Byler, 2000). The third study was an unpublished raw dataset, and no written conclusions could be found (Karau & Hansen, 2005). Serendipitously, Sczesny (2003) found identical similarities between men, women, and managers. Instead of celebrating the egalitarian views of the respondents, the author disparaged the male participants for purposefully skewing their responses. Despite this small sample of studies, the pattern of conclusions when the results do not go as expected is deeply troubling.

The present study seeks to demonstrate that the findings of the studies in the think manager-think male paradigm are erroneous. The present study was inspired by Widiger & Settle (1987) who critiqued Broverman et al. (1970) by demonstrating the problems associated with unbalanced item selection. In the same spirit, this study attempts to demonstrates that similar problems are occurring in the think manager-think male paradigm.

## Methodology

## Procedure

300 Participants were recruited through Amazon's Mechanical Turk, a crowdsourcing platform commonly used to recruit people to participate in academic research. After removing cases which failed the manipulation checks, attention checks, or were identified as bots by failing to correctly respond to the CAPTCHA (a visual recognition test used to determine whether the user is human to reduce spam), 146 participants remained. The participants were randomly assigned to rate one of three conditions (Men, Women, or Managers) on the Schein Descriptive Index and were presented instructions identical to that of Schein (1973). The final sample size for the Men condition was 48, Women was 45, and Managers was 53. The sample was 45% men, and the average age was 37.84 (SD = 11). Participants were required to be at least 18 years old and from the United States.

*Schein Descriptive Index.* The overall ICC for men-managers was .81 [.71, .87], and the ICC for women-managers was .39 [.08, .60]. This overall effect provides evidence of a successful replication of the findings in typical think manager-think male studies. Means for individual items were also computed (see Table 1). Significance tests were performed on the means between men and women conditions. Items which exhibited significant mean differences by way of a *t*-test were labeled as male or female depending on for which sex the item was more characteristic. This process resulted in 37 male items, 33 female items, and 22 neutral items.

Table 1. Schein Descriptive Index Items and Means

| | Condition | | | |
|---|---|---|---|---|
| Trait | Men | Women | Managers | Type |
| Able to separate feelings from ideas | 3.35 | 3.16 | 3.94 | 0 |
| Adventurous | 3.81 | 3.00 | 2.89 | 1 |
| Aggressive | 3.67 | 2.24 | 2.85 | 1 |
| Ambitious | 3.98 | 3.29 | 4.19 | 1 |
| Analytical ability | 3.48 | 3.22 | 3.96 | 0 |
| Assertive | 3.83 | 2.82 | 4.19 | 1 |
| Authoritative | 4.06 | 2.58 | 4.28 | 1 |
| Aware of feelings of others | 2.77 | 3.96 | 3.40 | 2 |
| Bitter | 2.85 | 2.40 | 2.45 | 1 |
| Cheerful | 3.21 | 3.78 | 3.21 | 2 |
| Competent | 3.52 | 3.89 | 4.13 | 2 |
| Competitive | 4.17 | 3.13 | 4.26 | 1 |
| Consistent | 3.38 | 3.49 | 3.94 | 0 |
| Courteous | 2.79 | 3.98 | 3.58 | 2 |
| Creative | 3.13 | 3.98 | 3.38 | 2 |
| Curious | 3.29 | 3.60 | 3.21 | 0 |
| Dawdler and procrastinator | 3.00 | 2.82 | 1.98 | 0 |
| Deceitful | 3.15 | 2.53 | 2.47 | 1 |
| Decisive | 3.75 | 3.00 | 4.21 | 1 |
| Demure | 2.58 | 2.89 | 2.55 | 0 |
| Desire for friendship | 3.44 | 4.11 | 2.85 | 2 |
| Desire to avoid controversy | 3.65 | 3.22 | 4.42 | 1 |
| Desires responsibility | 2.94 | 3.31 | 3.75 | 0 |

| | | | | |
|---|---|---|---|---|
| Devious | 3.04 | 2.40 | 2.42 | 1 |
| Direct | 3.83 | 2.96 | 4.15 | 1 |
| Dominant | 3.73 | 2.38 | 4.09 | 1 |
| Easily influenced | 2.85 | 3.36 | 2.68 | 2 |
| Emotionally stable | 3.21 | 3.33 | 3.96 | 0 |
| Exhibitionist | 3.08 | 2.56 | 2.72 | 1 |
| Fearful | 2.46 | 2.91 | 2.13 | 2 |
| Feelings not easily hurt | 3.19 | 2.49 | 3.66 | 1 |
| Firm | 3.77 | 2.98 | 4.11 | 1 |
| Forceful | 3.71 | 2.44 | 3.28 | 1 |
| Frank | 3.46 | 2.60 | 3.81 | 1 |
| Frivolous | 2.58 | 2.78 | 2.13 | 0 |
| Generous | 2.94 | 3.64 | 3.17 | 2 |
| Grateful | 3.02 | 3.62 | 3.19 | 2 |
| Hasty | 3.08 | 2.67 | 2.43 | 1 |
| Helpful | 3.40 | 4.00 | 3.72 | 2 |
| Hides emotion | 3.71 | 2.73 | 3.55 | 1 |
| High need for autonomy | 3.73 | 3.02 | 3.70 | 1 |
| High need for power | 3.88 | 2.47 | 4.04 | 1 |
| High self-regard | 3.88 | 3.44 | 4.08 | 1 |
| Humanitarian values | 3.00 | 3.96 | 3.23 | 2 |
| Independent | 4.00 | 3.42 | 3.94 | 1 |
| Industrious | 3.67 | 2.93 | 4.02 | 1 |
| Intelligent | 3.48 | 3.82 | 3.92 | 2 |
| Interested in own appearance | 3.35 | 4.07 | 3.85 | 2 |
| Intuitive | 3.08 | 3.76 | 3.55 | 2 |
| Kind | 3.08 | 3.69 | 3.15 | 2 |
| Knows the way of the world | 3.27 | 3.38 | 3.70 | 0 |
| Leadership ability | 3.65 | 3.24 | 4.28 | 1 |
| Logical | 3.71 | 3.11 | 4.02 | 1 |
| Modest | 2.67 | 3.42 | 3.15 | 2 |
| Neat | 2.58 | 3.96 | 4.00 | 2 |
| Nervous | 2.56 | 2.96 | 1.92 | 0 |
| Not comfortable about being aggressive | 2.29 | 3.67 | 2.02 | 2 |
| Not conceited about appearance | 2.83 | 2.49 | 2.91 | 0 |
| Obedient | 2.71 | 3.11 | 3.32 | 2 |
| Objective | 3.15 | 3.24 | 3.72 | 0 |
| Passive | 2.42 | 3.18 | 2.53 | 2 |
| Persistent | 3.94 | 3.42 | 4.11 | 1 |
| Prompt | 3.35 | 3.07 | 3.98 | 0 |
| Quarrelsome | 3.06 | 2.60 | 2.32 | 1 |
| Reserved | 2.75 | 3.20 | 2.51 | 2 |
| Self-confident | 4.00 | 3.47 | 4.17 | 1 |
| Self-controlled | 3.23 | 3.49 | 4.08 | 0 |
| Selfish | 3.23 | 2.69 | 2.64 | 1 |

| | | | | |
|---|---|---|---|---|
| Self-reliant | 3.96 | 3.31 | 4.04 | 1 |
| Sentimental | 2.65 | 3.98 | 2.51 | 2 |
| Shy | 2.29 | 2.93 | 1.68 | 2 |
| Skilled in business matters | 3.52 | 3.13 | 4.23 | 0 |
| Sociable | 3.40 | 4.18 | 3.77 | 2 |
| Sophisticated | 2.90 | 3.69 | 3.17 | 2 |
| Speedy recovery from emotional disturbance | 3.33 | 2.73 | 3.77 | 1 |
| Steady | 3.63 | 3.38 | 4.04 | 0 |
| Strong need for achievement | 4.02 | 3.20 | 4.25 | 1 |
| Strong need for monetary rewards | 3.81 | 3.09 | 4.06 | 1 |
| Strong need for security | 3.27 | 3.82 | 3.72 | 2 |
| Strong need for social acceptance | 3.54 | 3.56 | 3.40 | 0 |
| Submissive | 2.02 | 2.98 | 2.23 | 2 |
| Sympathetic | 2.65 | 4.02 | 2.96 | 2 |
| Tactful | 3.29 | 3.44 | 3.75 | 0 |
| Talkative | 3.25 | 3.87 | 3.79 | 2 |
| Timid | 2.29 | 2.96 | 2.09 | 2 |
| Uncertain | 2.75 | 2.73 | 1.92 | 0 |
| Understanding | 3.08 | 4.02 | 3.49 | 2 |
| Values pleasant surroundings | 3.29 | 4.31 | 3.66 | 2 |
| Vigorous | 3.50 | 2.96 | 3.49 | 1 |
| Vulgar | 3.19 | 2.02 | 1.89 | 1 |
| Wavering in decision | 2.75 | 3.09 | 2.25 | 0 |
| Well informed | 3.52 | 3.51 | 4.06 | 0 |

Note. Ratings are on a 5-point scale (1=Not characteristic, 5=Characteristic). Type indicates the results of a *t*-test between Men and Women and the codes indicate the sex which scores higher, 1 = male, 2 = female, 0 = no difference.

Given the baseline item frequency (37 male, 33 female, 22 neutral), additional item sets were generated to test the hypothesis that male and female item type frequency affects the overall ICCs. When items needed to be removed from the baseline set to create the new item sets, they were chosen randomly. Random item selection was chosen to reduce the researcher's degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). The following sets were generated: one where the frequency of male items was greater than the number of female items (M > F), one where female items were more frequent than male items (F > M), one where the frequency was equal (M = F), and one where only neutral items were included (Neutral Only). In the item sets, the same 22 neutral items were always retained. This decision was made to keep the total number of items relatively high. The M > F item set contains 33 male, 22 female, and 22 neutral items. The F > M item set contains 22 male, 33 female, and 22 neutral items. The M = F item set contains 22 male, 22 female, and 22 neutral items. The Neutral Only item set contains 0 male, 0 female, and 22 neutral items.

In addition, to demonstrate the deleterious effect of the absolute agreement function within the ICC, mean item ratings were dichotomized. Means were categorized as 1 if the mean was above 3 (characteristic) and 0 if the mean was below 3 (uncharacteristic). The ICCs for the

aforementioned item sets were then recalculated using the dichotomized data. The items in the dichotomous sets are the same as in the continuous sets. 21 items were found to be in the same direction (both either characteristic or uncharacteristic) for men and managers but not for women, while 18 items were found to be in the same direction for women and managers but not for men.

**Results**

The Schein Descriptive Index averages are reported in Table 2. To determine if there were rating differences between the targets, ANOVAs were performed on the means for each item set for the continuous means. No significant differences were found (p-values > .05).

Table 2. Item Set Means and Intraclass Correlations

| Item Set | Means | | | ICCs | | |
|---|---|---|---|---|---|---|
| Continuous | Men | Women | Managers | Men-Managers | Women-Managers | Men-Women |
| Baseline | 3.25 | 3.23 | 3.37 | 0.81 | 0.39 | -0.19 |
| Neutral Only | 3.15 | 3.18 | 3.34 | 0.73 | 0.61 | 0.83 |
| M > F | 3.27 | 3.19 | 3.36 | 0.75 | 0.13 | -0.07 |
| M = F | 3.18 | 3.21 | 3.31 | 0.63 | 0.32 | -0.07 |
| M < F | 3.15 | 3.26 | 3.26 | 0.58 | 0.43 | 0.00 |
| Schein (1973) | - | - | - | 0.62 | 0.06 | - |
| Dichotomous | Men | Women | Managers | Men-Managers | Women-Managers | Men-Women |
| Baseline | 0.71 | 0.62 | 0.64 | 0.51 | 0.47 | 0.05 |
| Neutral Only | 0.63 | 0.73 | 0.68 | 0.90 | 0.90 | 0.80 |
| M > F | 0.71 | 0.61 | 0.69 | 0.63 | 0.37 | 0.07 |
| M = F | 0.68 | 0.59 | 0.65 | 0.46 | 0.62 | 0.09 |
| M < F | 0.65 | 0.62 | 0.64 | 0.47 | 0.53 | 0.05 |

Note. M = male, F = female.

ICCs were calculated and the original ICCs from Schein (1973) are also reported for reference. The continuous item sets demonstrate the think manager-think male effect. The largest difference between men-managers and women-managers is found in the M > F set (ICCs = .75 and .13, respectively). The lowest is the M < F set (ICCs = .58 for men and .43 for women). The results reveal that the Neutral Only item set largely nullifies the effect. For the men-women comparison, results show that the ICCs are often near zero, except for the Neutral Only set where the ICC (.83) is greater than men-managers (.73) and women-managers (.61). Statistical difference tests were not performed on the ICCs because they are dependent. Across the item sets created in this study, means for men and managers decrease as the number of male items decreases and the means for women increase as the number of female items increases.

Data were dichotomized and the data were then reanalyzed. Results are displayed in Table 2 under the Dichotomous heading. The baseline ICCs now show a near equal effect for men-

managers (.51) and women-managers (.47). The ICCs for M = F and M < F now show an opposite effect than the continuous data. Namely, dichotomizing the data for these item sets reveals a think manager-think female effect. Removing items which exhibit mean differences and dichotomizing them completely removes any evidence of sex or gender bias. The Neutral Only ICCs for men-managers and women-managers are the same (ICCs = .90).

**Discussion**

The results demonstrated that the findings in Schein (1973) were the consequence of a biased measurement process and an inappropriate statistical analysis. By rearranging the item frequency, the ICCs can lead to the conclusions of think manager-think male, -think female, and even that men and women resemble managers more than each other. The misleading nature of Schein's (1973) finding is not surprising when one understands how the ICCs are affected by small differences in mean ratings. None of the obtained Index means are statistically significantly different from each other, yet the ICCs appear to tell a different story. Even if one wanted to interpret the ICCs substantively, one would have to make some bizarre conclusions. When items that exhibit mean differences are removed from the original measure, the highest ICC is among the men-women comparison. Rather than declare *think man-think woman*, it is important to remember that the ICCs are meaningless when applied to ratings of two different categories (McGraw & Wong, 1996). Substantive interpretations aside, the results clearly demonstrate that choices relating to item selection and data analysis can cause one to reach different conclusions.

The measurement problem is particularly important to recognize because it continues to occur decades later (Koenig et al., 2011; Fischbach et al., 2015; Carli et al., 2016). Researchers need to take great care to avoid the use of proxy measures (Reyna, 2018). These are measures which purport to measure phenomenon A while really measuring phenomenon B. In the think manager-think male paradigm, this means purporting to measure sex bias but in actuality measuring the independent ratings of two different categories. Researchers should avoid treating small differences in means on the same side of the measurement continuum as though they belong to different measurement categories (Reyna, 2018). As demonstrated in this study, the Schein Descriptive Index means for men, women, and managers differed by .02 on a 5-point scale on average, yet the ICCs made it appear as though the ratings were wildly different. When researchers find a statistically significant effect, it is tempting to jump to the conclusion that however the alternative hypothesis was framed is therefore correct (Reyna, 2018). Participant ratings more often differed in degree, not in kind. Ignoring the actual values associated with the findings and their meaning is an enormous oversight. Likewise, given that the aggregate ratings in this study hovered around the midpoint of the scale (3, neither characteristic or uncharacteristic) it is possible that people are somewhat ambivalent about their opinions of the traits possessed by men, women, and managers.

The Schein (1973) study was one of the first research studies on sex and gender biases in management and should be recognized as such. As with Broverman et al., (1970), the findings in Schein (1973) tell a convincing story and it is understandable that the flaws have gone unrecognized for this long (Widiger & Settle, 1987). However, it is time that the errors are recognized and that a moratorium of Schein (1973) replications be placed. Allowing the paradigm to continue is not without risks. Permitting unjustified claims to be made based on flawed results may erode public trust in psychological science. The opponents of diversity may examine the literature and publicly identify unjustified claims or dismiss the entire literature as activist

propaganda (Eagly, 2016). If the findings of stereotype research can help reduce bias, prejudice, and discrimination, then the potential public backlash caused by the promotion of false claims presents a barrier to that goal. More generally, the oversimplification of the interpretation of sex and gender differences is detrimental and will cause harm to the reputation of feminism, psychology, and the social sciences if the understanding of these differences is not grounded in science (Stern, 2018). Thus, attending to the issues present in Schein (1973) is important as it has implications beyond what takes place in scientific publications.

A counter argument to the criticisms in the present study is that subsequent theorizing retroactively supports the paradigm's claims (e.g., role congruity theory, Eagly & Karau, 2002). However, the way in which these so-called theories are generated is not scientific. These theories are generated in a way Richard Feynman referred to as cargo cult science (Feynman, 1974). Cargo cult science is a term to describe practices that superficially resemble scientific research but lack the rigor and discipline necessary for true scientific inquiry. This can include practices such as cherry-picking data to support preconceived ideas, using insufficient or inappropriate methods to measure or analyze data, and making sweeping claims based on insufficient evidence.

Cargo cult scientists investigate phenomena in the following manner: (1) an effect is observed, (2) an explanation is invented for the underlying cause of the effect, and (3) every future observation of the effect is taken as evidence of the explanation. For example, any disparity in the representation of women in managerial occupations is taken as evidence of some stereotype-driven lack-of-fit causal mechanism (e.g., Eagly & Karau, 2002). Yet, no one engaged in the scientific process to isolate such a causal mechanism. Managerial occupations which are majority female are ignored or some post hoc rationalization is manufactured instead of reconsidering the validity of the congruence hypothesis. Alternative hypotheses (e.g., men and women choose occupations based on their work interests) are rarely even mentioned (Pinker, 2003). Thus, "theories" that might retroactively validate the paradigm are themselves flawed because they are not based in science, and the tenets are maintained despite being either demonstrably false or unfalsifiable (Abbot et al., 2023).

There are some illuminating questions to ask about this paradigm which raise concerns about whether the scientific method is being utilized. Why has there not been any incremental progress in the understanding of this phenomenon after 50 years? Why has no one attempted to find under what conditions people think manager-think female or eliminate the effect completely? Why has no one investigated the cause of subgroup variation in this phenomenon? Why has no one asked whether variation in the phenomenon relates to variation in important outcomes? One possibility is that the community of researcher-activists do not really want to understand the world—they want to change it. Their conclusions are designed to support an ideologically left sociopolitical narrative rather than advance science (Crawford & Jussim, 2018; Honeycutt & Jussim, 2022). It is sobering to think of how much scientific progress has been impeded, and how much time has been wasted, by a lack of skepticism in this area.

Researchers involved in this paradigm, and gender studies more broadly, care about gender equality and sincerely want to reduce prejudice and discrimination against women. Over the years, Schein (1973) replications have become a litmus test of society's progress (e.g., Duehr & Bono, 2006). However, the biased nature of the methodology prevents an accurate diagnosis. Koenig et al. (2011) estimated that the relationship between the think manager-think male effect size and time, as measured by year of publication, is zero. Perceptions of, and behavior towards,

women in the workplace have changed considerably since the early 1970s. The statistical artifacts in the paradigm have not.

## Conclusion

Schein's (1970) think manager-think male research paradigm contains severe limitations. Demonstrated herein was the biased measurement and statistical analysis procedure that will virtually guarantee the think manager-think male effect is observed in every future replication or adaptation. This is an important opportunity for psychological science to intervene and self-correct. As pointed out by Kelly and Blashfield (2009), science can fail to self-correct as it did with Broverman et al. (1970) and false conclusions can endure. A healthy scientific discipline ought to at least entertain the possibly that an idea, however old or cherished, might be wrong. Treating ideas and theories as sacrosanct is not in alignment with the spirit of science.

## References

Abbot, D., Bikfalvi, A., Bleske-Rechek, A. L., Bodmer, W., Boghossian, P., Carvalho, C. M., ... & West, J. D. (2023). In defense of merit in science. *Journal of Controversial Ideas, 3*(1), 1. https://doi.org/10.35995/jci03010001.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosenkrantz, P. S., & Vogel, S. R. (1970). Sex role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology, 34*, 1–7.

Byler, V. L. J. (2000). Perceptions of effective athletic administrators and gender schemas. *Dissertation Abstracts International, 61*(12), 4688.

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. https://doi.org/10.1177/0361684315622645.

Crawford, J. T., & Jussim, L. (Eds.). (2018). *Politics of social psychology*. Psychology Press. https://doi.org/10.4324/9781315112619.

Duehr, E. E., & Bono, J. E. (2006). Men, women, and managers: Are stereotypes finally changing? *Personnel Psychology, 59*, 815–846. https://doi.org/10.1111/j.1744-6570.2006.00055.x.

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. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes, 64*, 307-324.

Feynman, R. P. (1974). Cargo cult science: Some remarks on science, pseudoscience, and learning how to not fool yourself. Caltech's 1974 Commencement Address. https://calteches.library.caltech.edu/51/2/CargoCult.htm.

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.

Gwet, K. L. (2014). *Handbook of inter-rater reliability*. Advanced Analytics, LLC.

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

Heilman, M., & Eagly, A. (2008). Gender stereotypes are alive, well, and busy producing workplace discrimination. *Industrial and Organizational Psychology, 1*(4), 393-398. doi:10.1111/j.1754-9434.2008.00072.x

Honeycutt, N., & Jussim, L. (2022). Political Bias in the Social Sciences: A Critical, Theoretical,

and Empirical Review. https://doi.org/10.31234/osf.io/qpn57.

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

Karau, S. J., & Hansen, E. (2005). Cultural influences on perceptions of female managers: A comparison of Sweden and the United States [Unpublished raw data].

Kelly, L. P. & Blashfield, R. K. (2009). An example of psychological science's failure to selfcorrect. *Review of General Psychology, 13*(2), 122-129.

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 /10.1037/a0023557.

Lauritsen, M. W. (2020). Men, women, and leaders: The effect of gender-leader category congruence on supervisor evaluations. ProQuest Dissertations Publishing.

Lauritsen, M. W. (2022). An alternative procedure for subgroup analyses in the think managerthink male paradigm. *Journal of Personnel Psychology, 21*(1), 33-42. https://doi.org/10.1027/1866-5888/a000312.

Lovette, I. J., & Fitzpatrick, J. W. (Eds.). (2016). *Handbook of bird biology* (3rd ed.). John Wiley & Sons.

Ma, A., Rosette, A. S., & Koval, C. Z. (2022). Reconciling female agentic advantage and disadvantage with the CADDIS measure of agency. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0000550.

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

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*, 175–220.

Pinker, S. (2002). *The blank slate: The modern denial of human nature*. Viking.

Phillips, R. D., & Gilroy, F. D. (1985). Sex-role stereotypes and clinical judgments of mental health: The Brovermans' findings reexamined. *Sex Roles, 12*, 179–193.

Reyna, C. (2018). Scale creation, use, and misuse: How politics undermines measurement. In J. T. Crawford & L. Jussim (Eds.), *Politics of Social Psychology* (pp. 81–98). New York, NY: Psychology Press.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331–363.

Rosenkrantz, P., Vogel, S., Bee, H., Broverman, I., & Broverman, D. M. (1968). Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology, 32*, 287–295.

Schein, V. E. (1973). The relationship between sex role stereotypes and requisite management characteristics. *Journal of Applied Psychology, 57*(2), 95–100. https://doi.org/10.1037/h0037128.

Schein, V. E. (1975). Relationships between sex role stereotypes and requisite management characteristics among female managers. *Journal of Applied Psychology, 60*(3), 340–344. https://doi.org/10.1037/h0076637.

Sczesny, S. (2003). Leadership competence: Self- and other-perceptions of male and female managers. *Zeitschrift für Sozialpsychologie, 34*, 133–145. doi:10.1024//00443514.34.3.133.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. *Psychological Bulletin, 86*, 420-428.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Sommers, C. H. (1994). *Who stole feminism?* New York: Simon & Schuster.

Stern, C. (2018). Does political ideology hinder insights on gender and labor markets? In L. Jussim & J. Crawford (Eds.), *The Politics of Social Psychology* (pp. 43-57). Routledge.

Stricker, G. (1977). Implications of research for psychotherapeutic treatment of women. *American Psychologist, 32*, 14–22.

Widiger, T. A., & Settle, S. A. (1987). Broverman et al. revisited: An artifactual sex bias. *Journal of Personality and Social Psychology, 53*, 463–469.