**对*Engstrom, H. R., Laurin, K., Zuroff, D. C., & Schmader, T.*(2025)研究结果的计算可复现性检验**

小组成员分工

|  |  |  |  |
| --- | --- | --- | --- |
| 组长 | 毕雅茹 | | |
| 组员 | 高诗雨、毕雅茹、张馨予 | | |
| 分工 | | | |
| 数据分析 | 高诗雨、毕雅茹 | PPT 制作 | 张馨予 |
| 文字报告制作 | 高诗雨、毕雅茹、张馨予 | PPT 展示 | 张馨予 |

\* 同一名同学可负责多个部分；如同一内容由多位同学负责，可按百分比注明贡献占比

**摘要:**本研究通过对Engstrom等人（2025）关于群体性别影响领导行为的五项研究进行计算可复现性检验，重点验证了研究1中“领导者对男性群体的支配性行为显著高于女性群体”这一假设。研究采用两因素被试间设计（N=368），通过在线平台招募美国成年参与者，随机分配至“领导男性群体”或“女性群体”条件。测量工具包括8项支配性行为量表和2项社群性感知量表。结果显示，男性组支配性评分（M=2.83, SD=1.32）显著高于女性组（M=2.34, SD=1.12, t=3.84, p<0.001, d=0.40），且社群性感知的中介效应显著（95% CI=[-0.10, 0.01]）。复现过程中，所有统计指标（均值、效应量、p值）与原文献完全一致（δ=0%），仅在进行中介检验时，发现z值出现微小偏差（0.008%）。研究表明，群体性别对领导行为的塑造具有跨情境稳健性，领导者对员工的性别刻板印象通过“低群体性”认知偏差驱动支配性行为差异。研究证实了原文献方法学的严谨性，同时提示未来需关注计算工具版本控制和开放科学实践的标准化。

**关键词:**计算可复现性；性别刻板印象；领导行为；群体性别效应

**1 引言**

**1.1 所选文献信息**

**表 1 文献信息表**

|  |  |  |  |
| --- | --- | --- | --- |
| **1 文献基本信息** | | | |
| 所选文献 | Engstrom, H. R., Laurin, K., Zuroff, D. C., & Schmader, T. (2025). Do people lead men and women differently? Multimethod evidence that group gender affects leaders’dominance.*Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001735 | | |
| 数据来源 | https://osf.io/d5m4k/ | | |
| **2 文献选取** | | | |
| 文献主题是否包含不止一篇研究？ | * 是，且包含元分析研究   √是，但不包含元分析研究   * 否 | 文献此前被其他研究者重复过？ | * 是(附上原文链接)   √否 |
| **3 研究假设选取** | | | |
| 重复的研究假设 | 研究一：领导者对男性群体的支配性行为显著高于女性群体 | | |
| 重复的研究假设是否在其他研究中经过重复？ | * 是(附上原文链接)   √否 | 文献共几个实验，重复的研究假设是第几个实验中的？ | 6（5项研究和一个预实验），重复的研究假设是研究1中的。 |
| 选择该假设的原因 | (若未按照指南推荐的优先顺序，需额外说明) | | |
| **4 数据集选取** | | | |
| 是否采用原始数据？ | √是   * 否 | 是否对样本量进行修改？ | * 是(说明原因)   √否 |
| 若修改，报告原文样本量大小和修改后的样本量大小 | 无 | 若修改，报告使用 G-power 计算的修改后的样本量对应的效应量 | 无 |

**1.2 文献介绍**

**1.2.1 研究背景**​​

性别刻板印象对领导行为的塑造作用长期受到组织行为学领域的关注。传统研究多聚焦于领导者自身性别差异（如男性更倾向于专制型领导风格），却忽视了群体性别构成如何动态影响领导者的行为选择。社会角色理论指出，性别刻板印象可能导致领导者通过调整行为策略以适应不同性别群体的期待。近年实验研究表明，男性领导者常被赋予更多支配性行为的合法性，而女性领导者则需平衡合作性与权威性。然而，现有研究多基于自然情境下的观察或单一方法设计，缺乏对群体性别效应的系统验证，尤其是对混合性别群体及行为外显化的考察不足。

**1.2.2 主要研究问题与假设**​​

本研究围绕以下核心问题展开：群体性别是否系统性影响领导者的支配性行为？其作用机制是否源于性别刻板印象（如对男性“低社群性”的认知）？基于此提出以下假设：

假设1：领导者对男性群体的支配性行为显著高于女性群体；

假设2：性别刻板印象（尤其是对男性“低社群性”的感知）是群体性别效应的中介变量；

假设3：混合性别群体的支配性行为强度介于全男性与全女性群体之间，且受刻板印象强度调节。

**1.2.3 研究方法与结果**​​

本研究通过五项研究验证假设，涵盖实验操控、行为编码与跨群体评价。

研究1采用了单因素两水平的被试间实验设计（N=368），参与者被随机分配至“想象领导男性群体”或“女性群体”条件，完成8项主导行为量表（α=0.93）及社群性感知评估（α=0.85/0.88）。 实验结果表明，男性群体条件下的支配性评分显著更高（t(366) =3.84，p < .001，d = 0.40），且领导者对女性群体的社群性评分显著更高（b=0.44, p<.001）。一个2（领导者性别：女性vs男性）×2（追随者性别：女性vs男性）的方差分析重复了追随者性别的效应（F(1, 363) =15.40，p < .001）。该分析还显示，领导者性别的效应仅为前者的四分之一左右（F(1, 363) =1.03，p = .311，d = 0.10）。与此同时，中介模型表明，参与者认为女性群体比男性群体更具共同性（b = 0.44，z = 3.20）。

研究2采用了单因素三水平的被试间实验设计（N=368），参与者随机分配至领导男性/女性/混合性别群体条件，完成主导行为量表及性别刻板印象问卷（7点李克特量表）。 实验结果表明，领导男性群体的支配性评分最高（M=4.35），显著高于女性（M=3.82, p<.001）和混合组（M=3.96, p=.003）。中介分析显示，刻板印象解释39%的效应量（95% CI [0.28, 0.52]）。

研究3采用单因素多水平被试间设计进行实地研究（N=314），研究者让儿童体育教练（78%男性）报告其对全男生/女生队伍的领导行为，采用支配性-声望行为双维度量表。 结果表明，全男生队伍的支配性评分更高（M=4.10 vs. 3.65, p=.002），但声望行为无差异（p=.18）。

研究4采用被试内拉丁方实验设计（N=161），研究人员邀请参与者录制3分钟领导演讲视频，编码员盲评支配性行为（如指令性语言频率）及语言风格（LIWC词典）。研究结果表明，男性群体条件下，支配性行为频次更高（M=12.3 vs. 8.7, t(159)=4.81, p<.001, d=0.38），且社群性词汇使用减少（M=15.2 vs. 18.5, p=.001）。

在研究5中，参与者观看领导视频后评价其效能（1-7分量表）及团队凝聚力。研究结果表明，男性参与者对高支配性领导的评价更高（β=0.17, p=.004），而女性参与者呈负相关（β=-0.12, p=.03）。

**1.2.4 结论**​​

本研究通过多方法设计系统验证了群体性别对领导行为的塑造作用。结果表明，领导者对男性群体的支配性倾向具有跨情境稳健性，其核心机制在于性别刻板印象对“低社群性”的认知偏差。尽管男性领导者的支配性效应略强于女性，但两类群体均表现出显著差异，挑战了传统研究中性别二元对立的简化假设。研究提示，组织需警惕性别刻板印象对领导效能的潜在负面影响，例如过度依赖支配性策略可能削弱团队协作与创新。未来研究可进一步探索文化差异、领导特质（如敌意性别偏见）及群体动态（如权力分配）的调节作用，为性别平等的领导力培养提供理论依据。

**2 方法**

**2.1 样本**​

样本数据集包含368名美国成年参与者（131名男性，232名女性，4名非二元性别者，1名未报告性别），年龄均值为35.83岁（标准差未明确报告）。样本通过在线平台Prolific Academic招募，覆盖多种族背景：57%为白人，22%为黑人，14%为西班牙裔，13%为亚裔，其余5%为其他或混合种族。参与者需满足视力正常或矫正视力正常、英语流利等基本条件，实验过程中无中途退出记录。

样本的关键排除标准包括：27名参与者因未完整观看实验材料（图片未加载或误操作）被剔除，最终有效样本为368人。与此同时，在本研究中，所有参与者均签署知情同意书，实验程序通过Prolific Academic伦理审查委员会审核。

**2.2 原研究方法简介**

原研究采用了被试间设计（Between-Subjects Design），通过随机分配参与者至两种实验条件：

条件1：领导由6名男性照片组成的群体；

条件2：领导由6名女性照片组成的群体。

与此同时，本研究采用使用芝加哥面部数据库（Chicago Face Database）中的63张女性和57张男性照片作为实验材料，实现群体性别的操纵。筛选标准包括：

1.性别一致性（1,087名评分者一致判定为男性或女性）；

2.中性表情、典型性别特征（如男性化/女性化面部结构）；

3.种族多样性（4名白人、1名黑人、1名亚洲人）。

**2.2.1 测量工具**​​

1.支配性行为量表：8项Likert量表（1=完全不符合，7=完全符合），示例条目包括“我会试图控制群体成员而非允许他们控制我”（α=0.93）；

2.社群性感知量表：2项单极量表（1=完全不温暖，7=非常温暖），用于评估群体成员的共情性和道德感（r=0.71）。

**2.2.2 数据分析方法**

本研究采用了描述性统计计算两组在支配性评分上的均值（男性组M=2.83，SD=1.32；女性组M=2.34，SD=1.12）；用独立样本t检验比较两组支配性评分差异（t(366)=3.84, p<0.001, d=0.40）；采用方差分析（ANOVA）检验领导者性别与群体性别的交互作用（F(1,363)=15.40, p<0.001, η²=0.04）；使用lavaan包（Rosseel, 2012）进行Bootstrap法中介检验（10,000次抽样），验证社群性感知的中介效应（间接效应95% CI=[-0.10, 0.01]）；

**2.2.3 软件与包**

本研究采用R作为数据分析软件；

使用关键软件包包括：lavaan（结构方程模型与中介分析）；ggplot2（可视化效应量）；tidyverse（数据清洗与预处理）；psych（一般的数据分析、描述性统计）；effecsize（计算和解释效应量）。

**2.3 重复思路说明**

在进行可复现性检验时，重复者经过了“确定文献——下载原始数据和代码——数据预处理——进行描述性统计的可复现性检验——采用原始方法进行推断性统计的可复现性检验——记录信息并撰写报告”的复现过程，对所选文献进行可复现性检验。

重复者选取了文章的研究1进行复现，原因有以下几点：一是其主要使用的分析方法与重复者在R语言课程中学习到的分析方法及相关代码较为一致；二是复现研究1可以帮助重复者经历真实数据分析的全流程、尝试解决真实数据分析中遇到的问题，在实践中加深对统计方法的理解与运用，包括数据预处理、描述性统计、推断统计等，从而更好地理解研究结果的含义；三是复现研究1的结果可以与原始研究结果进行比较，从而发现潜在的差异，并探究可能的原因，以更全面、更批判性的视角看待前人研究并进行验证，为心理学的发展做出自己的贡献。

在原始代码的基础上，重复者另外添加了对被试描述性统计信息的可视化代码，绘制了有关被试种族信息的饼图，以便更加清晰地展示原文中所汇报的被试种族占比的信息。

**3 结果**

**3.1 描述性统计**

对原文献描述性统计进行重复的结果，并汇总表格：

**表 2 描述性统计结果的比较**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 男性员工群体 | | | 女性员工群体 | | |
|  | *N* | Mean | *SD* | *N* | Mean | *SD* |
| 原研究  报告结果 | 183 | 2.83 | 1.32 | 185 | 2.34 | 1.12 |
| 本研究 | 183 | 2.83 | 1.32 | 185 | 2.34 | 1.12 |
| *δ* | 0% | 0% | 0% | 0% | 0% | 0% |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |

**3.2 推断性统计**

**3.2.1 使用与原文献相同方法的推断性统计**

报告对原文献推断性统计进行重复的结果，并汇总表格：

**表 3 推断性统计结果的比较(原文献方法)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 假设1：领导者对男性员工群体的支配性行为显著高于女性员工群体（t检验） | | | | |
| 样本量(*N)\** | 统计量  (如 *t* 或 *F*) | 效应量  (如 Cohen’s *d*) | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | *t* = 3.84 | Cohen’s *d* = 0.40 | *p <* 0.001 |  |
| 本研究 | 368 | *t* = 3.84 | Cohen’s *d* = 0.40 | *p <* 0.001 |  |
| *δ* | 0% | 0% | 0% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 员工性别对领导者支配性的主效应（方差分析） | | | | |
| 样本量(*N)\** | 统计量  (如 *t* 或 *F*) | 效应量  (如 Cohen’s *d*) | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | *F* = 15.40 | *η*²p = 0.04 | *p <* 0.001 |  |
| 本研究 | 368 | *F* = 15.40 | *η*²p = 0.04 | *p <* 0.001 |  |
| *δ* | 0% | 0% | 0% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 领导者性别的主效应（方差分析） | | | | |
| 样本量(*N)\** | 统计量  (如 *t* 或 *F*) | 效应量  (如 Cohen’s *d*) | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | *F* = 1.03 | Cohen’s *d* = 0.10 | *p =* 0.311 |  |
| 本研究 | 368 | *F* = 1.03 | Cohen’s *d* = 0.10 | *p =* 0.311 |  |
| *δ* | 0% | 0% | 0% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 领导者性别与员工性别的交互作用（方差分析） | | | | |
| 样本量(*N)\** | 统计量  (如 *t* 或 *F*) | 效应量  (如 Cohen’s *d*) | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | *F* = 0.46 | *η*²p < 0.01 | *p =* 0.500 |  |
| 本研究 | 368 | *F* = 0.46 | *η*²p < 0.01 | *p =* 0.500 |  |
| *δ* | 0% | 0% | 0% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 领导者对不同性别员工的群体性的认知差异（中介分析前置条件检验） | | | | |
| 样本量(*N)\** | *b* | *z* | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | 0.44 | 3.20 | *p <* 0.001 |  |
| 本研究 | 368 | 0.44 | 3.20 | *p <* 0.001 |  |
| *δ* | 0% | 0% | 0% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 完全一致 | 完全一致 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 领导者对不同性别员工的群体性认知在员工性别与领导者支配性之间的  中介效应（间接效应） | | | |
| 样本量(*N)\** | 置信区间 |  |  |
| 原文献  报告结果 | 368 | 95% CI [−0.10, 0.01] |  |  |
| 本研究 | 368 | 95% CI [−0.10, 0.01] |  |  |
| *δ* | 0% | 0% |  |  |
| 评级 | 完全一致 | 完全一致 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 员工性别对领导者的支配性存在直接影响（直接效应） | | | | |
| 样本量(*N)\** | *b* | *z* | 显著性指标(*p*) |  |
| 原文献  报告结果 | 368 | 0.45 | 3.53 | *p <* 0.001 |  |
| 本研究 | 368 | 0.45 | 3.56 | *p <* 0.001 |  |
| *δ* | 0% | 0% | 0.008% | 0% |  |
| 评级 | 完全一致 | 完全一致 | 偏差较小 | 完全一致 |  |

\* 此前描述性统计已报告，因此此处仅需进行报告，不需作为一个额外的结果

**3.3 对原文计算可复现性进行评估**

**3.3.1 使用与原文献相同方法**

报告原文献的值的评级分布、推论的一致情况，整理成表格，如下表所示：

**表 5 结果可复现性的评估表**

|  |  |  |
| --- | --- | --- |
| 可复现性情况 | 数量及占比 | |
| *N* | *%* |
| 完全一致(*δ* = 0%) | 31 | 96.875% |
| 偏差较小(0% < *δ* < 10%) | 1 | 3.125% |
| 偏差较小(*δ* ＞ 10%) | 0 | 0% |
| 因舍入导致的偏差 | 0 | 0% |

\* 结果数量*N*指在重复分析中，对重复分析结果与原结果进行配对比较的次数。对于每个目标效应，结果包括一组数值，如汇总效应估计(summary estimate，如*t*值/ F值)、置信区间界限(confidence interval bound)、效应量(effect size)样本大小(size effect)等，应将原文中报告的每个数值与重复结果进行比较。例如，在一个*t*检验中，原文献报告了*t*值、95%置信区间、cohen’s d和样本大小，则这个效应中*N*＝4。将各效应的*N*求和即为全体数量。

**表 6 推论的一致性的评估表(原分析方法)**

|  |  |  |
| --- | --- | --- |
| 推论的一致性 | 数量及占比 | |
| *N \** | *%* |
| 一致 | 7 | 100% |
| 不一致 | 0 | 0% |

\* 推论数量*N*指在重复分析中，对效应做出统计推断的次数。例如，仅进行了一个*t*检验，则*N*＝1；如果进行了一个2\*2的方差分析，并进行简单效应分析，则有可能有7个统计推断：两个主效应的推论，一个交互作用的推论，四个可能的简单效应分析的推论，因此*N*＝7。如果报告的*p*值相对于重复的*p*值落在显著性水平边界的另一侧，则被归类为推论不一致；反之为推论一致。

**4 讨论**

**4.1 计算可复现性检验结果分析**

结合下表，对原文献进行分析，推测可能导致可复现性检验结果差异的原因。对于重要的原因，逐段进行展开说明。

**表 8 计算上（不）可重复的原因分析表**

|  |  |  |  |
| --- | --- | --- | --- |
| **可能原因** | | | **研究一** |
| **原文献**  **开放性**  **问题** | **一般性开放**  **获取问题** | 几个结果的微小差异，可能是由于分析中使用了没有设置固定种子的随机数； | 在使用R语言进行中介分析时，使用原文献提供的原始数据和代码，进行复现，但出现了文章中汇报的z值和重复者复现的z值不一致的情况，但偏差十分微小，推测可能原因有：复制粘贴错误、软件及软件包版本不一致等问题。 |
| 个别结果的微小差异，可能是由于印刷或复制粘贴错误； |
| 文章文本中程序报告不明确，包括纳入亚组的标准、缺乏或不正确报告用于回归模型的变量、以及未报告的单侧分析； |
| 在文章的开放实践声明中对研究的模糊标记。 |
| **OSF 开放**  **获取特定问题** | OSF 中缺乏对数据和/或代码内容进行说明的文档(readme文档)； | 是 |
| OSF 上的数据与代码文件不一致，如代码中对部分数据进行了操作，但这部分数据在数据文件中无对应； | 否 |
| OSF上的数据存储问题，包括文件损坏或无法下载。 | 否 |
| **数据开放获取特定问题** | 没有提供原始数据； | 已提供 |
| 没有提供处理后的数据； | 没有提供处理后的数据 |
| 没有提供数据处理过程的描述或代码。 | 已提供 |
| **代码开放获取特定问题** | 缺乏共享的分析代码或建模代码； | 原文献只提供了软件包和代码，但并没有提供关于软件版本的信息。 |
| 软件包或软件版本的问题。 |
| **重复**  **过程的原因** | **重复研究与**  **原研究的区别** | 是否使用同样的数据集； | 是 |
| 是否使用同样的数据分析软件及软件包； | 是，但未提供版本信息 |
| 是否使用同样的数据分析方法。 | 是 |
| **重复者**  **相关因素** | 重复者此前是否有过 R 使用经验； | 重复者缺乏使用R语言进行中介分析的知识和操作经验。  原研究的代码只分析了自变量对中介变量的回归、自变量对因变量的直接效应以及中介变量的间接效应，但没有分析中介变量对因变量的影响。重复者对这一分析存在疑惑。 |
| 重复者对关于 R 的知识或操作上存在漏洞，较难理解原文章中的部分操作(可做简单说明)。 |
| **其他**  **影响**  **因素** | **文献年份** | 文献发表年份是否较为久远，是否在开放科学运动之前； | 否 |
| **文献质量** | 文献引用量大小； | 文献引用量：101 |
| 是否有其他研究支持本文献结果； | 否 |
| 是否有其他研究对本文献结果进行了重复，重复结果如何(可做简单说明)。 | 否 |

**4.2 其他思考**

(1) 课程学习方面：通过本学期R语言课程的学习，我们对R的使用有了一些基本的了解 。在本次数据复现过程中，通过实际操作，对R的基本语法、数据结构和常用包有了更深入的了解；

(2) 科学研究方面：通过本学期的R语言课程，我们对科学研究的方法有了新的认识。在数据的收集和处理方面，必须保证数据的准确性和完整性；选择合适的数据分析方法也是至关重要，不同的方法可能会得出不同的结论； R能够使数据的处理和分析更加科学、高效，因此，在心理学科研中，掌握一门编程语言能够使研究者更加灵活高效的进行数据处理和统计分析。

(3) 小组合作方面：在团队合作的过程中，我们感受到了分工明确的重要性，每个成员都贡献了自己的力量，提高了工作效率。同时，成员间的有效沟通也是很重要的，各位成员各抒己见、寻求共识，合力完成了本次的复现工作；

(4) 课程作业方面：本课程一共设计了两次小作业与一次大作业。在homework1的完成过程中，通过实际操作了解了如何实现GitHub仓库与本地文件的连接，提高了学习与工作的效率；在homework2的完成过程中，对课程中学习到的数据分析代码进行了实操练习，进一步巩固了所学知识，进行查缺补漏；在期末作业完成的过程中，对本课程所学的内容进行了综合运用，大大提高了对R语言使用的熟悉度，同时也提高了团队协作的能力；总体来说，我们认为课程作业设计合理、有效。

**参考文献(APA格式)**

Engstrom, H. R., Laurin, K., Zuroff, D. C., & Schmader, T. (2025). Do people lead men and women differently? Multimethod evidence that group gender affects leaders’dominance. *Journal of Experimental Psychology: General.* Advance online publication. https://doi.org/10.1037/xge0001735