Gender Stereotype Threats and College Performance: A Meta-Analysis

Academic performance is a key predictor of success in many facets of life, such as employment opportunities and social mobility. However, numerous studies have highlighted the adverse impacts of gender stereotypes on academic achievement, especially in the setting of higher education. Many studies have shed light on how female students in STEM (science, technology, engineering, and mathematics) are more likely to drop out, receive lower exam grades, and perform worse than their male counterparts (Çetinkaya et al., 2020). To address this imbalance in academic performance, various interventions have been introduced. Yet, evaluating their effectiveness is crucial, especially given the ongoing underrepresentation of women in STEM.

Gender Stereotypes and Stereotypes Interventions

Negative stereotypes can influence people's attitudes toward target groups. Vulnerable groups may internalize these negative stereotypes, thus becoming susceptible to their influence (B. Zhang et al., 2023). For example, women could face stereotypes suggesting they are less capable in STEM (science, technology, engineering, and mathematics), causing them to internalize beliefs about their inadequacy in these fields. Similarly, men may face stereotypes suggesting emotional stoicism, leading them to internalize these beliefs.

To combat these harmful stereotypes and lessen their effects, stereotype threat interventions are used. Broadly, there are three main types of stereotype threat interventions: belief, resilience, and identity. Belief-based interventions aim to challenge and reshape individuals' perceptions of stereotypes through exposure to counter-stereotypical information (Chung & Huang, 2021). An example of a belief-based intervention is the values affirmation intervention, which involves individuals reflecting on their core values. This practice has been shown to act as a protective shield against psychological threats posed by stereotypes (Miyake et al., 2010). Secondly, resilience-based interventions focus on equipping individuals with the skills and mindset needed to overcome the adverse effects of stereotypes, fostering resilience and perseverance (Helmreich et al., 2017). An example of resilience-based intervention is educating individuals about stereotype threats, empowering them to recognize and mitigate their influence on their psychological well-being (Johns et al., 2005).

Finally, identity-based interventions empower individuals to embrace and express their unique identities beyond narrow stereotypes, fostering a sense of belonging and affirmation within diverse communities (Oyserman & Destin, 2010). One example of an identity-based intervention is the use of role models from underrepresented groups who

have succeeded in STEM fields. By doing so, researchers illustrate that individuals can succeed regardless of societal norms or expectations (Marx & Roman, 2002). Another identity-based intervention involves dissociating the self from performance outcomes. For example, in one study, individuals used fictitious names to separate themselves from their achievements, thus reducing the impact of stereotype threat (S. Zhang et al., 2013).

College Performance and Gender Stereotypes

College performance could be assessed through a variety of metrics. These may include grade point average (GPA), scores on standardized exams, completion rates, and engagement in extracurricular activities. These measures significantly shape students' career prospects and influence their opportunities and success. Thus, it is crucial that people make the most of their time in college; however, when gender stereotypes threaten their performance, it becomes imperative to address these biases and implement interventions that empower students to overcome such challenges and reach their full potential.

The Proposed Moderators

This meta-analysis aimed to examine moderators of the effects of interventions on college performance. We considered the specific intervention a moderator, distinguishing between affirmation-based interventions and those without affirmation components. Additionally, we used the publication years of the studies as a moderator, intending to discern any variations in the effects of intervention over time.

The Current Meta-Analysis

The present meta-analysis synthesized existing literature that examined the effects of gender intervention on women's performance in college, examined publication bias and addressed moderator questions relevant to theory and practice. Specifically, we examined whether effect sizes would vary by methodological characteristics (e.g., type of intervention, the specific intervention, and the performance measurement) and sample characteristics (e.g., sample mean age and gender composition).

Method

Search Strategy and Study Selection

Electronic searches were conducted in February of 2024 on EBSCO Host's APA PsycINFO databases using the search terms stereotype threat, gender, intervention, and college. The publication years were restricted to 1995-2024 to ensure the inclusion of up-to-date studies and to avoid outdated research. To mitigate the impact of potential publication bias, both published and unpublished papers, such as dissertations, were considered. The search yielded 117 articles, which were then evaluated based on the following inclusion

criteria: studies needed to be (a) experimental or quasi-experimental; (b) have college participants only and be relevant to academic performance; (c) employ a gender threat intervention whether that be belief, resilience, or identity. As depicted in Figure 1, 39 of the 117 articles were subjected to a full paper review, and only 13 studies met all criteria.

Study Characteristics Extraction

Six coders independently extracted study characteristics. Two coders coded each study, and all six coded 50% of the studies to ensure reliability. Inconsistencies were resolved by consensus. We coded methodological characters (type of intervention, the specific intervention, and the performance task) and sample characteristics (mean age and gender composition).

Type of intervention

We coded whether the intervention was belief-based, resilience-based, or identity-based. We delved deeper, specifying interventions such as self-affirmation or educating participants on stereotypes.

Performance Measurement

We coded the measure of performance used in each study, whether it was a post-exam assessment, GPA, or other means of measuring academic performance.

Sample's Mean Age

The sample's mean age was extracted. When the mean age was not reported, we left it as N/A. Since we concluded in the title and abstract screening that the study is about college students, even when the sample mean age was not reported, we kept the study in our analysis.

Sample's Gender Composition

The gender composition of the sample was extracted. When not reported, it was recorded as N/A.

Effect Size Extraction

The targeted effect size was Hedge's g because it reflects the standardized difference between two means and adjusts for small sample sizes. When hedge's g is not reported, we extracted other measures of effect size that could be converted later, such as Cohen's d. Otherwise, we prioritized pre-post difference score means and standard deviation over posttrial means and standard deviation to consider pre-trial differences.

Then, following standard procedures, we used these means and standard deviation to compute Cohen's d, which was then used to calculate Hedge's g.

Analytical Approach

All analyses were conducted in R-Studio 4.1.3 (R Core Team, 2022) using the package metafor (Viechtbauer, 2010). We conducted a random-effects model, estimating 2 using restricted maximum likelihood estimation and utilizing the t-distribution for significance testing. Next, we identified influential cases based on conventional cutoffs for the difference in fits (larger than $3 \times 1/(k - 1)$), Cook's distance (chi-square significance test), and hat values (larger than $3 \times (1/k)$; Viechtbauer, 2010). Any identified influential case(s) were excluded, and the summary effect size was recomputed.

We used a funnel plot, the trim-and-fill method, and Egger's test to evaluate publication bias. The funnel plot displays each effect size against its precision so that asymmetrical distribution suggests potential publication bias. The trim-and-fill method removes effect sizes until symmetry is achieved, recomputing the summary effect size and filling the plot with hypothetical missing studies. Egger's test performs significance testing on the degree of symmetry and examines standard error as a moderator.

We conducted subgroup analyses for descriptive purposes, examining extracted characteristics. Our analysis did not include any continuous methodological or sample characteristics. Categorical variables for which the k of each coded level was at least five were stratified. Initially, our intended moderator was the type of intervention (belief, resilience, and identity). However, resilience and identity did not meet the criteria.

Consequently, we examined the specific interventions used, categorizing them into affirmative interventions (k = 14) and non-affirmative interventions (k = 5). Additionally, studies published from 2010 and after were classified as "Recent Years" (k = 12), while those from 1995 to 2009 were labeled as "Earlier Years" (k = 7), thus forming a distinct subgroup for analysis by cutting the publication year at the median.

Finally, we conducted meta-regressions to examine the moderating effects by studying characteristics. We performed meta-regressions when at least five studies were present per level for categorical variables. Like earlier procedures, the type of intervention was regrouped into affirmative versus non-affirmative intervention, and influential cases (k = 1) were excluded from our analysis.

Results

Descriptive Statistics

The current meta-analysis synthesized data from 13 studies examining the effectiveness of gender threat interventions in diminishing women's relative performance ability, incorporating 20 unique effect sizes. These studies collectively involved 4052 participants. However, in instances where information regarding the cell sizes of intervention and control groups was absent, we assumed equal cell sizes. We rounded up cell sizes if decimal values were obtained.

Among all the studies, 85% implemented belief intervention (k = 17), 10% utilized identity intervention (k = 2), and the remaining 5% (k = 1) employed resilience intervention strategies. We acknowledge the limited number of studies employing identity and resilience intervention strategies, which may need to be revised to generalize our findings.

Overall Effect Size

We observed a significant overall effect size, g = .64, SE = .19, CI_{95} [.26,1.02], p < .001, indicating a small to moderate beneficial effect of the interventions on women's performance ability. The 95% prediction interval was -0.99 to 2.27, suggesting that the true effect size could be expected to fall within this range in 95% of future studies. The heterogeneity test suggests excess variability was observed beyond expected if variability was solely due to sampling variance, Q (df = 19) = 155.3667, p < .0001. Furthermore, I^2 statistics suggest that about 96% of the total variance was due to heterogeneity, and about 4% was due to sampling variance. Finally, τ^2 was estimated at 0.66 (τ = 0.81), suggesting that, on average, effect sizes deviated from the overall effect size by about .81 Hedge's g units. The forest plot presents the distribution of the effect sizes and their precision (Figure 2).

We then conducted diagnostic tests based on conventional cutoffs for differences in Fits, Cook's distance, and hat values of the overall effect size model. One influential case was identified. After excluding the effect size, the summary effect size remained significant, g = .47, SE = .11, CI_{95} [.26 .69], PI_{95} [-.36,1.31], such that the beneficial effect of interventions on women's performance ability remained to be small to moderate. To be conservative, we removed the influential case from all subsequent analyses.

Publication Bias

As depicted in Figure 3, the funnel plot displays effect sizes as a function of its standard error. As we hypothesized a positive Hedge's g, missing effect sizes to the left of the summary effect size would suggest potential publication bias—a trim-and-fill test estimated three studies missing on the left side. Following the trim-and-fill adjustment, the new summary effect size decreased but remained significant, g = 0.38, SE = .12, CI_{95}

[.15, .60], $p < .001$. Additionally, Egger's test for funnel plot asymmetry yielded a non-significant p-value of 0.6742, indicating minimal evidence of publication bias.

Subgroup Analyses

We conducted subgroup analyses for all extracted characteristics, comparing levels as is for categorical variables. As summarized in Table 1, the intervention effects remained significant regardless of publication years (recent years: $g = .28$; older years: $g = .77$). However, affirmative interventions had significant effects, while non-affirmative interventions were not (affirmative: $g = .43$; non-affirmative: $g = .61$). It is important to note that subgroup analyses provide descriptive statistics only, and any inferences regarding differences in effect size by methodological or sample characteristics should be drawn from the meta-regression analyses reported below.

Meta-Regression Results

As summarized in Table 2, the type of intervention was not a significant moderator, $b = - .16$, $SE = .26$, $p = .53$. Since our p-value is greater than the alpha criterion of .05, we fail to reject the null hypothesis that the association between publication year and effect size is equal to 0. Thus, there is no association between the type of intervention and effect sizes.

Additionally, the year the paper was published was a significant moderator, $b = -0.49$, $SE = .19$, $p = 0.011$, $R^2 = 34.41\%$. Using the same alpha criterion, we reject the null hypothesis that the association between publication year and effect size is equal to 0 because our p-value is less than the alpha criterion. Thus, there is an association between publication year and effect sizes; more specifically, the effect size tends to decrease as the year increases. In other words, there is a negative relationship between publication year and effect sizes, suggesting that interventions may have been less effective in more recent studies than earlier ones.

Discussion

The present meta-analysis synthesized data from 13 studies examining the effectiveness of gender threat interventions in improving women's performance in college, incorporating 20 unique effect sizes. The findings reveal a significant small-to-moderate beneficial effect of these interventions, indicating that, on average, women who participated in gender threat interventions experienced an improvement in their academic performance. This supports the idea that interventions targeting gender stereotypes can positively impact educational outcomes for women, aligning with previous research highlighting the detrimental effects of gender stereotypes on academic achievement. The

success of these interventions could be instrumental in narrowing the gender gap in STEM fields, as evidenced by numerous studies (Miyake et al., 2010; S. Zhang et al., 2013).

Our analysis extends the current literature by quantifying the overall effect size across various studies, encompassing various types of interventions and performance measures. However, it's essential to acknowledge potential biases in the literature favoring the publication of studies with positive results. To assess the impact of publication bias, we conducted a trim and-fill analysis, which identified and adjusted for missing studies. This process involved trimming one study and inputting three additional studies, resulting in a total of k = 22 studies included in the analysis. Additionally, we conducted Egger's test, a statistical method for detecting funnel plot asymmetry, to further evaluate publication bias. The results of Egger's test indicated no significant asymmetry in the funnel plot ($z = 0.4203$, $p = 0.6742$), suggesting minimal evidence of publication bias. While potential biases exist in the literature, our findings suggest that they did not substantially influence the observed effect size. Thus, we can conclude that the overall effect size of our meta-analysis was not predominantly driven by publication bias. However, further investigation is warranted to assess how much publication bias may have influenced the findings.

Furthermore, we observed substantial between-study heterogeneity in effect size and revealed two moderators that partially explained such heterogeneity. First, the type of intervention was not a significant moderator. However, from subgroup analyses, we observed that affirmative interventions had a significant effect while non-affirmative interventions did not. This disparity may arise from the nature of affirmative interventions, wherein participants engage in repeated affirmations of their core values, a component that non-affirmative interventions do not have (Çetinkaya et al., 2020). The repetition of affirming their core values instills these principles in participants, serving as a protective barrier against psychological threats posed by stereotypes. This process equips them with the ability to discern what is truly important to them, thus mitigating the impact of trivial, irrelevant stereotypes.

However, this meta-analysis lacked statistical power to examine specific non-affirmative interventions. As previously mentioned, our original intention was to separately analyze belief-based, resilience-based, and identity-based interventions. However, due to limited data availability, we regrouped our intervention moderator into affirmative (a type of belief-based intervention) and non-affirmative interventions. It is plausible that other interventions are just as effective as affirmative interventions. Nonetheless, our analysis revealed that all interventions were effective, suggesting that any intervention is preferable to no intervention. Further research is needed to explore the comparative effectiveness of various intervention approaches.

Secondly, the year the paper was published was a significant moderator, implying that there is an association between the publication year and effect sizes; more specifically, the effect size tends to decrease as the publication year increases. In other words, there is a negative relationship between publication year and effect sizes, suggesting that interventions may have been less effective in more recent studies than earlier ones. However, despite this trend, subgroup analyses revealed that intervention effects remained significant across all publication years. Thus, interventions have consistently demonstrated an ability to enhance women's academic performance, regardless of the year of publication.

The smaller effect sizes in recent studies could be attributed to multiple reasons. First, shifts in societal attitudes, changes in educational environments, and the evolving prevalence of gender stereotypes may influence the effectiveness of interventions differently over time. Secondly, recent studies may be incorporating a more diverse pool of participants, potentially resulting in fluctuations in effectiveness. Furthermore, recent interventions may be more intricate or multifaceted compared to earlier ones, presenting challenges in achieving similar levels of effectiveness. For instance, a study conducted in 2019 introduced a novel approach by adapting the values affirmation stereotype threat into a multi-stereotype threat connectionist model, which explores how stereotypes influence individuals' behavior and performance for those who face threats related to multiple stereotypes simultaneously (Çetinkaya et al., 2020).

However, these are all potential reasons why effect sizes have decreased over the years. Further research is necessary to determine the exact cause of these differences. For instance, extracting the racial composition of these studies could help infer whether a diverse participant pool leads to smaller effect sizes. Similarly, we could extract socioeconomic status (SES) or other relevant demographic variables, providing additional insights into the factors influencing intervention effectiveness. Moreover, it would be beneficial to consider other factors, such as intervention duration and delivery methods (e.g., in-person sessions, video modules, text-based interventions) as potential moderators in future analyses. By exploring these additional variables, we can gain a more comprehensive understanding of how various factors interact to impact the effectiveness of interventions targeting gender stereotypes.

This meta-analysis provides empirical evidence that intervention effectively enhances women’s academic performance. Findings highlight that affirmative interventions improve women’s academic performance more than non-affirmative ones. Further research is necessary to strengthen the evidence concluded from this paper. Additional studies could possibly delve deeper into the mechanisms underlying the

effectiveness of affirmative interventions and explore potential moderators or mediators of intervention outcomes. Moreover, longitudinal studies could provide valuable insights into the long-term effects of these interventions on women's academic achievement and overall well-being.

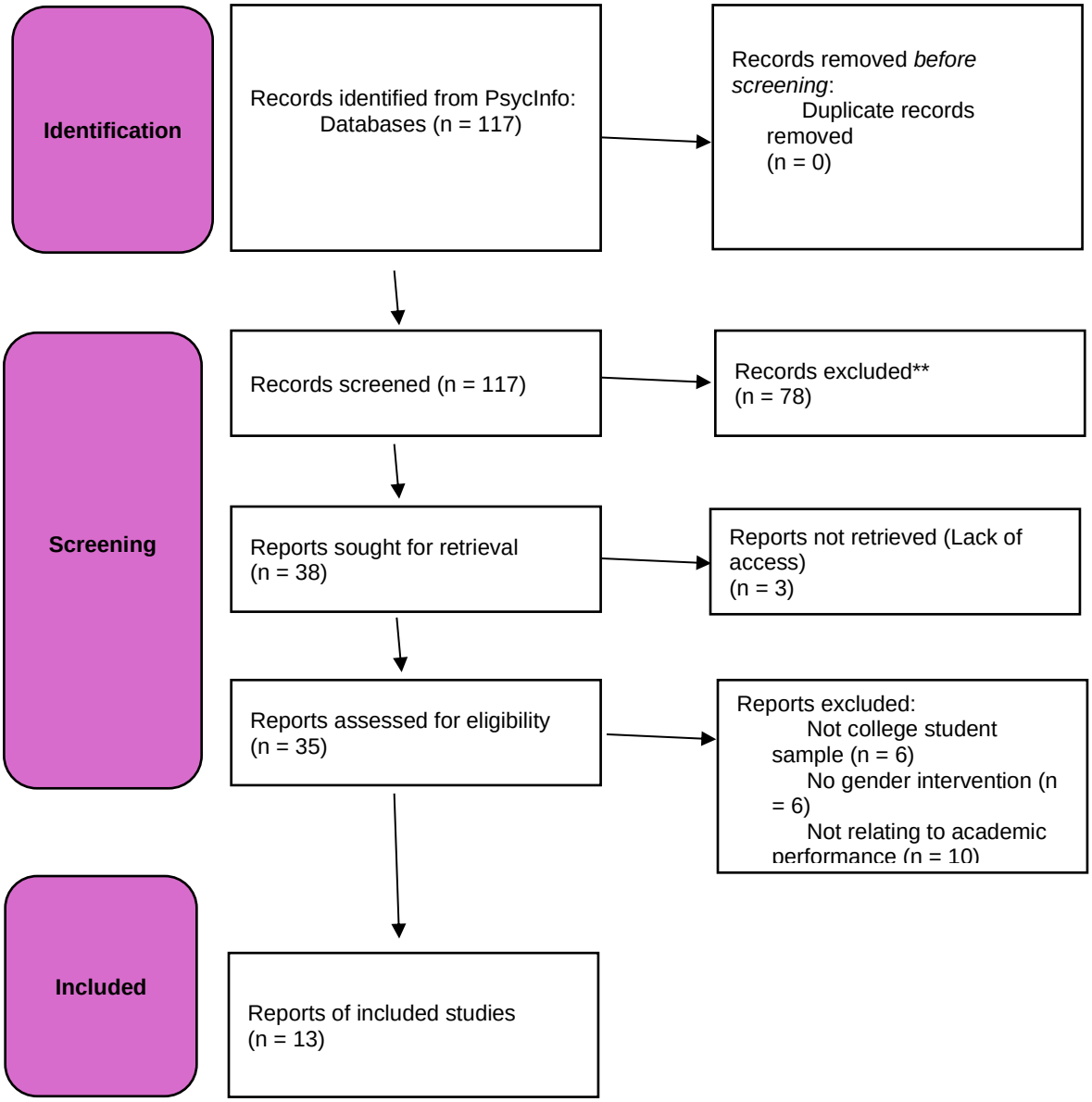

Figure 1. PRISMA flow diagram.

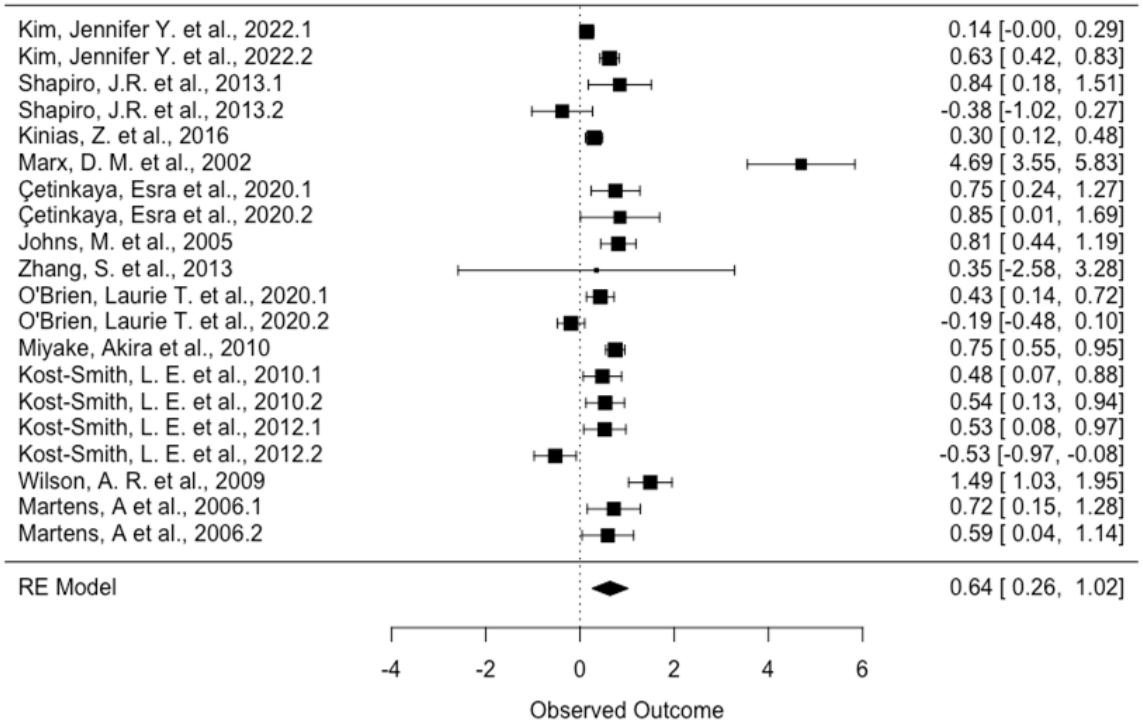

Figure 2: A forest plot of all included effect sizes (depicted as squares), their 95% confidence intervals (depicted as bars), as well as the summary effect size (depicted as a diamond with its width being the 95% confidence interval).

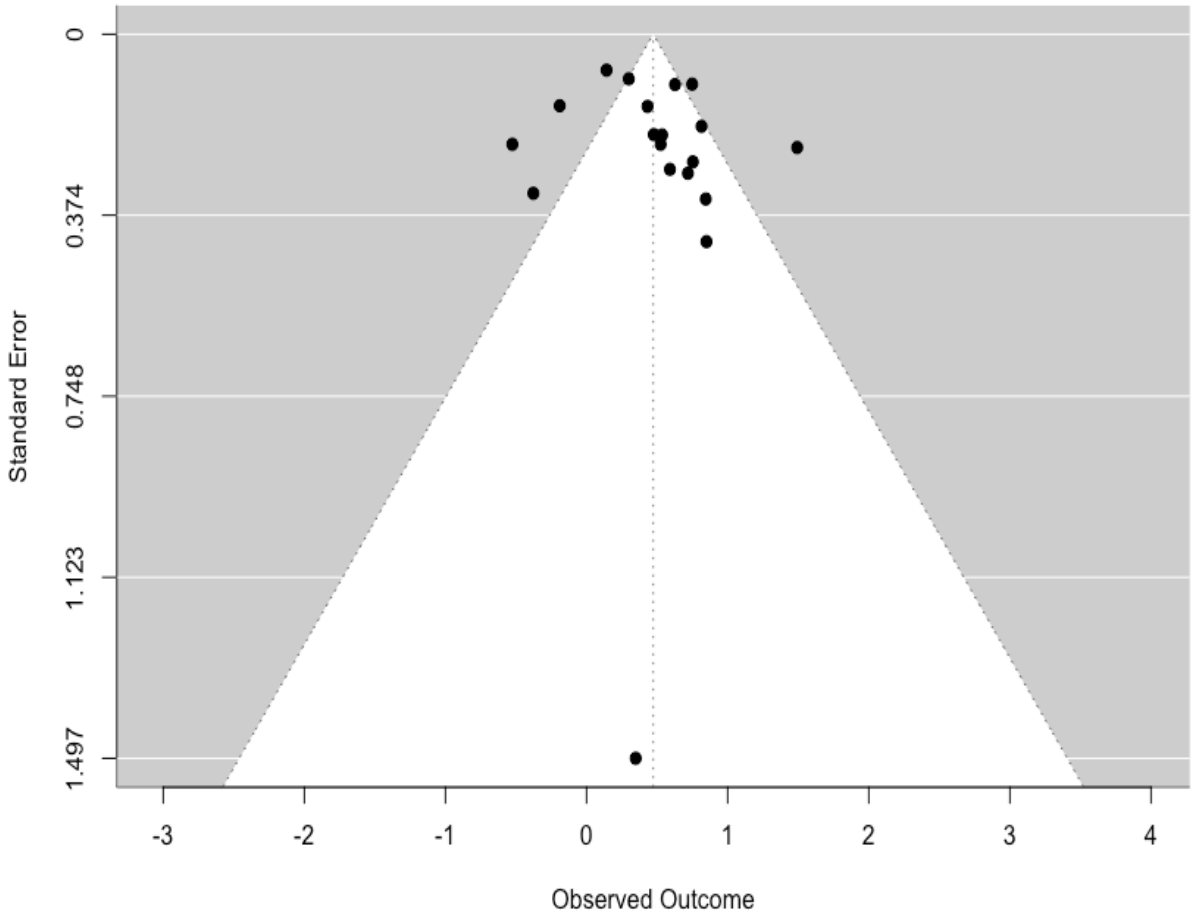

Figure 3: A funnel plot of effect sizes plotted against its standard error. Note that one effect size was identified as an influential case and was excluded from this plot.

Table 1. Subgroup analyses by methodological and sample characteristics.

	k	summary g, 95% CI	I^2	τ^2	Q
Type of Intervention					
Affirmative	14	.43 [.22, .63]	81%	.10	58.59
Non-Affirmative	5	.61 [-.04, 1.26]	91%	.42	42.21
Publication Years					
1995-2009	7	.77 [.53, 1.00]*	59%	.06	13.30
2010-Present	12	.28 [.03, .54]*	86%	.15	51.04

Table 2. Moderation analyses by methodological and sample characteristics.

	k	b (SE), p-value	R^2	I^2	τ^2
Type of Intervention: Affirmative	14	-.164 (0.26), *p* = .53	0%	87%	.18
Publication: 2010-Present	12	-0.486 (.19), *p* = .011	34%	80%	.11

Note. **p* < .05. All binary variables were dummy-coded. Moderators were entered in separate models.

References

Çetinkaya, E., Herrmann, S. D., & Kisbu-Sakarya, Y. (2020). Adapting the values affirmation intervention to a multi-stereotype threat framework for female students in STEM. *Social Psychology of Education*, *23*(6), 1587–1607. https://doi.org/10.1007/s11218-020-09594-8

Chung, Y., & Huang, H.-H. (2021). Cognitive-Based Interventions Break Gender Stereotypes in Kindergarten Children. *International Journal of Environmental Research and Public Health*, *18*(24), 13052. https://doi.org/10.3390/ijerph182413052

Helmreich, I., Kunzler, A., Chmitorz, A., König, J., Binder, H., Wessa, M., & Lieb, K. (2017). Psychological interventions for resilience enhancement in adults. *The Cochrane Database of Systematic Reviews*, *2017*(2), CD012527. https://doi.org/10.1002/14651858.CD012527

Johns, M., Schmader, T., & Martens, A. (2005). Knowing Is Half the Battle: Teaching Stereotype Threat as a Means of Improving Women's Math Performance. *Psychological Science*, *16*(3), 175–179. https://doi.org/10.1111/j.0956-7976.2005.00799.x

Marx, D. M., & Roman, J. S. (2002). Female Role Models: Protecting Women's Math Test Performance. *Personality and Social Psychology Bulletin*, *28*(9), 1183–1193. https://doi.org/10.1177/01461672022812004

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the Gender Achievement Gap in College Science: A Classroom Study of Values Affirmation. *Science*, *330*(6008), 1234–1237. https://doi.org/10.1126/science.1195996

Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *The Counseling Psychologist*, *38*(7), 1001–1043. https://doi.org/10.1177/0011000010374775

Zhang, B., Hu, Y., Zhao, F., Wen, F., Dang, J., & Zawisza, M. (2023). Editorial: The psychological process of stereotyping: Content, forming, internalizing, mechanisms, effects, and interventions. *Frontiers in Psychology*, *13*, 1117901. https://doi.org/10.3389/fpsyg.2022.1117901

Zhang, S., Schmader, T., & Hall, W. M. (2013). L'eggo My Ego: Reducing the Gender Gap in Math by Unlinking the Self from Performance. *Self and Identity*, *12*(4), 400–412. https://doi.org/10.1080/15298868.2012.687012

Appendix A.

Supplementary Table S1. Effect sizes, methodological characteristics, and sample characteristics for all included studies.

First author, year	Hedge's g	Sample Size	Intervention Type	Specific Intervention	Performance Task	Sample's Mean Age	Female Percentage
Shapiro, J.R. et al., 2013	0.14	1277	belief	individual self-affirmation,	Grade Point Average (GPA)	27.9	41.28
Shapiro, J.R. et al., 2013	0.63	386	belief	individual self-affirmation,	Grade Point Average (GPA)	27.9	41.28
Kinias, Z. et al., 2016	0.84	36	belief	self-affirmation (self as a target for threat	math test (a test similar to the GRE-quant test)	NA	100.00
Marx, D. M. et al., 2002	-0.38	36	belief	self-affirmation (group as a target fo threat)	math test (a test similar to the GRE-quant test)	NA	100.00
Çetinkaya, Esra et al., 202(	0.30	474	belief	standard affirmatior values	core course mea score	28.93	34.99
Çetinkaya, Esra et al., 202(	4.69	44	identity	role model	math test	NA	51.16
Johns, M. et al., 2005	0.75	90	belief	self-affirmation	mental rotation task (number correct)	21.88	100.00
Zhang, S. et al., 2013	0.85	66	belief	group affirmation (under high gender identification)	mental rotation task (number correct)	21.88	100.00
O'Brien, Laurie T. et al., 2020	0.81	117	resilience	teaching-interventio	math test	NA	64.10
O'Brien, Laurie T. et al., 2020	0.35	199	identity	identity-mask; reduce distinctiveness	math test	NA	60.30
Miyake, Akira et al., 2010	0.43	182	belief	educational intervention	STEM GPA (WR women)	18.1	100.00
Kost-Smith, L. E. et al., 201	-0.19	182	belief	educational intervention	STEM GPA (URM women)	18.1	100.00

Kost-Smith, L. E. et al., 201	0.75	399	belief	values affirmation	scores on in-clas exams (three midterms and one final)	NA	29.07
Kost-Smith, L. E. et al., 201	0.48	96	belief	self-affirmation	exam score (2 midterm + 1 fina	NA	31.17
Kost-Smith, L. E. et al., 201	0.54	96	belief	self-affirmation	Force and Motio Concept Evaluation (FMCE)	NA	31.17
Wilson, A. R. et al., 2009	0.53	89	belief	self-affirmation (Study 2 only)	exam score (2 midterm + 1 fina	NA	31.45
Martens, A et al., 2006	-0.53	89	belief	self-affirmation (Study 2 only)	Force and Motio Concept Evaluation (FMCE)	NA	31.45
Martens, A et al., 2006	1.49	92	belief	malleable view	math tasks	21.18	73.60

Note. NA indicates that the study did not report the characteristics.