MASTER'S THESIS

Title

A Multiverse Meta-Analysis of Stereotype Threat for Women: Conceptual Instability and Dissemination Bias

submitted by

Beatrice Schreier BA BSc

in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Degree programme code as it appears on the student record sheet:	UA 066 840
Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Supervisor:	Assoz. Prof. Mag. Dr. Jakob Pietschnig Privatdoz.

Table of Content

Abstract

The present multi-level meta-analysis examined the phenomenon of stereotype threat among women in STEM. It synthesized evidence from 92 studies with a combined sample size of 11,608 female participants, focusing on theoretical boundary conditions and robustness analyses. The overall effect of stereotype threat for women's performance in STEM was small (d = –0.18). However, this effect proved trivial considering indicators of dissemination bias. Moreover, the wide prediction interval and the predominance of within-study compared to between-study heterogeneity indicated that the pooled mean is not an informative estimator for future studies. The theoretical conditions domain identification, task difficulty, and self-threat (C. M. Steele, 1997), did not substantially explain heterogeneity in threat effects. This challenged the robustness and theoretical coherence of the concept and further limited the interpretability of the pooled estimate. Multiverse and specification curve analysis highlighted the sensitivity of observed effects to analytical decisions, with nearly half of all computed specifications yielding non-significant summary effects. Indicators of potential dissemination bias emerged: earlier studies reported larger effects, whereas more recent studies found no or even positive effects (decline effect). Moreover, small-study effects suggested that less precise studies disproportionately contributed to the adverse summary effect. A Three-Parameter Selection Model (3PSM) found no evidence for selective reporting. The findings point to poorly defined boundary conditions and inconsistent operationalizations of central constructs. They also reveal conceptual overlaps between related constructs (jingle-jangle fallacies) within research on stereotype threat among women. Taken together, these findings challenge the robustness and theoretical precision of stereotype threat for women in STEM. To move the field forward, future research should more clearly specify the underlying theoretical framework and differentiate between psychological phenomena and theoretical constructs. Rather than continuing to test the questionable stereotype threat hypothesis, research should adopt a more general, phenomenon-driven approach to examine potential implications and interindividual differences in stereotype activation among women in STEM.

Introduction

Women are a minority in science, technology, engineering and mathematics (STEM) occupations and educational pathways around the world. A recent UNESCO report (2024) shows that women make up for only a quarter of the workforce in science, technology and engineering. The so-called "STEM gender gap" varies substantially between subfields. For example, while women hold 26% of jobs in data and artificial intelligence, they only constitute 12% in cloud computing (UNESCO, 2024). With regard to mathematics, the data are more limited; however, women make up approximately one quarter of researchers in this field (27%; Elsevier, 2024).

The gap is also evident in education. Between 2018 and 2023, only 35% of STEM graduates worldwide were women (UNESCO, 2024). This proportion varies between countries, yet in only 9 out of 122 were women the majority of STEM graduates, predominantly in Arab States (UNESCO, 2024). Accordingly, the report revealed that only 15% of degree programs women choose fall within the STEM domains. But low graduation rates are only part of the picture. While the educational choices influence the gender gap, other factors play a meaningful role as well. For example, in the European Union, only one in four women with a degree in information technology (IT) ends up working in the sector (UNESCO, 2024). In contrast, more than half of men with the same degree pursue a career in this field.

Longstanding research suggests that prejudices, defined as negative attitudes towards a group of people, influence educational and career paths of marginalized groups, such as racialized individuals and women (Allport, 1979; E. Aronson et al., 2021; J. Aronson et al., 1998). Eaton et al. (2020) demonstrated how biases among gatekeepers in physics and biology can relate to the diversity of those fields. In their study, they manipulated gender and race on otherwise identical CVs of post-doctoral candidates. In a sample of 251 professors, the results revealed racial biases, disadvantaging Latinx and Black against Asian and White candidates in both departments. Women were perceived as more likeable but less competent and hireable than men in Physics, however not in Biology. The authors linked this finding to the gender distribution in the departments, (90% male in physics vs. 65% in biology), suggesting that "masculine cultures" may favor male over female applicants (Eaton et al., 2020, p. 26).

In addition to behavioral correlates, prejudices also have a cognitive component: Stereotypes (E. Aronson et al., 2021). A large-scale study involving 350,000 participants across 66 nations found that implicit and explicit gender-science stereotypes relate to the female enrollment in science courses (Miller et al., 2015). Stereotypical traits associated with women are often categorized as communal, such as kindness, emotionality, and devotion, and they fit traditionally feminine gender roles picturing women as primary caregivers (Abele, 2003). This contrasts to agentic traits, like self-confidence, independence and decisiveness, which are stereotypically linked to men (Abele, 2003).

Beyond shaping career opportunities of women in STEM, gender stereotypes also affect their career paths (Thébaud & Charles, 2018). They shape women's perceptions of themselves and their abilities, their persistence and sense of belonging in STEM fields, among other psychological processes (J. Aronson et al., 1998; Eaton et al., 2020; Thébaud & Charles, 2018).

In an effort to explain the underperformance of African Americans, C. M. Steele and Aronson (1995) developed the concept of stereotype threat. Stereotype threat is defined as a self-threat that interferes with cognitive functioning due to a stereotype becoming relevant in an evaluative situation. Shortly after, C. M. Steele (1997) and Spencer et al., (1998) reported analogous effects on women in mathematics. They proposed that the concept could explain short-term underperformance in quantitative fields, as well as, over time, disidentification from stereotyped domains.

Research on stereotype threat has grown continuously over the past three decades, as indicated by increasing numbers of publications. For example, a Scopus search (conducted on August 12, 2025) for the term "stereotype threat" yielded 16 results in 2000, 79 in 2010, and 177 in 2020. Over time, the concept has been applied to diverse contexts, such as social class (Croizet & Claire, 1998), older adults (Armstrong et al., 2017), individuals with learning disabilities (Haft et al., 2019), and women in sports (A. Gentile et al., 2018). As of July 2025, a search for the term "stereotype threat" returns approximately 23,000 results. The concept has also received attention in the public discourse, and is often discussed as a psychological mechanism contributing to the underrepresentation and underperformance of women in STEM (e.g., Delaume & Lupu, 2022; Elsesser, 2019; Herrmann, 2021; Katter, 2018; Vedantam, 2012; Weber & Winkler, 2019). To address these disparities, efforts are made to reduce the impact of stereotype threat through interventions (Liu et al., 2021).

Despite its application and popularity in the discourse on group differences in education and occupations, stereotype threat is a controversial concept (E. Aronson et al., 2021; Eronen & Bringmann, 2021; Warne, 2022). Researchers criticize the reliability of effects, the validity of operationalizations and conclusions, theoretical precision, as well as misrepresentations of the findings and their consequences in many publications (Flore &

Wicherts, 2015; Picho-Kiroga et al., 2021; Sackett & Ryan, 2012; Sackett et al., 2004; Tomeh & Sackett, 2022; Warne, 2022; Whaley, 1998; Whaley, 2018; Zigerell, 2017).

The present study addresses the debate on the validity and relevance of stereotype threat for women in STEM by conducting a multi-level meta-analysis. Specifically, the validity of the theoretical boundary conditions is evaluated by examining whether they account for heterogeneity in effect sizes. Moreover, the generalizability and robustness of stereotype threat effects on women are examined through meta-regressions and subgroup analyses. The study further investigates dissemination bias and evaluates the sensitivity to analytical decisions through a multiverse and specification-curve analytical approach.

The Origins of Stereotype Threat Theory

Stereotype threat as a phenomenon was initially introduced in 1995, when Claude M. Steele and Joshua Aronson published a series of experiments on African American students underperforming on intellectual tests. The central finding was that, when controlling for prior ability (SAT scores), framing a test as diagnostic of ability led to underperformance among Black Americans. This effect arose compared to Black participants in the non-diagnostic group and White participants in both groups. C. M. Steele and Aronson (1995) argued that framing a task as "explicitly scholastic or intellectual" activated the threat of confirming the stereotype that Black Americans have lower intellectual ability than White Americans (p. 797).

To better understand the mechanisms and experiences of stereotype threat, they conducted a series of experiments where they further manipulated study framing, race salience, and let participants fill out questionnaires. The conclusion was that stereotype threat was a situational threat independent of performance expectations (C. M. Steele & Aronson, 1995). The authors theorized that it likely activated a stereotype-related anxiety or fear in Black participants, which interfered with their cognitive capacity and caused underperformance in ability measures.

In the following years, the theory was tested on women in mathematics (Spencer et al., 1999; C. M. Steele, 1997). Spencer et al. (1999) conducted three experiments that investigated whether stereotype threat led to underperformance in mathematics in women with high mathematical ability. No gender differences appeared on easy items, but women underperformed on more complex problems (Spencer et al., 1999). Since this pattern could also stem from other mechanisms, for example, reflect higher discriminative power of difficult tests rather than stereotype threat alone, they further employed different experimental designs.

To test the stereotype threat interpretation of the observed effect, Spencer et al. (1999) manipulated the framing of a test as one that showed or did not show gender differences. A pattern analogous to the first experiment was found: Women in the gender differences condition performed worse than all other groups, whereas women in the no-threat condition performed equally to men.

However, due to a floor effect in some items, the authors excluded easy items, and the results were limited in their generalizability. Thus, they investigated stereotype threat in a third experiment, in which a no-information condition (interpreted as a present baseline stereotype threat) was compared to a no-gender-differences condition. Consistent with their hypothesis, the results showed that women scored lower than men in the control condition and performed equally well as men in the removal condition. They also tested anxiety, self-efficacy, and evaluation apprehension as possible mediator variables of stereotype threat and found a "marginally significant effect" (p = .09; p. 18) on anxiety only.

Spencer et al. (1998) thus reported evidence for the stereotype threat phenomenon on women in mathematics. Based on their third experiment, they concluded a broad generalizability of stereotype threat effects on standardized test performances of stereotyped groups.

In the key theoretical paper, C. M. Steele (1997) differentiates between short-term and long-term effects of stereotype threat. In acutely threatening situations, stereotype threat impairs performance in individuals for whom it is important to perform well. Over time, repeated threats lead to the process of disidentification; individuals remove the stereotyped domain from the self-concept, which affects their motivation and achievement in that domain (C. M. Steele, 1997). The framework implies that stereotype threat effects depend on specific conditions, which are detailed in the following section.

Theoretical Boundary Conditions of Stereotype Threat

The theoretical framework proposed several conditions central to the concept's definition. The conditions were believed necessary for threat-induced underperformance or assumed to moderate the strength of the effects (Spencer et al., 1999; C. M. Steele, 1997; C. M. Steele & Aronson, 1995; J. Steele et al., 2002). However, the distinction between necessary and moderating conditions is theoretically ambiguous and treated inconsistently across early publications. The relations between the conditions and stereotype threat were oftentimes implied rather than systematically outlined (J. Aronson et al., 1998; Spencer et al., 1999; C. M. Steele, 1997; C. M. Steele & Aronson, 1995; J. Steele et al., 2002). As a result, empirical studies vary in how they define, operationalize, and implement them.

The conditions below are derived from the original papers on stereotype threat, in which the first experiments were reported, and the theoretical framework was presented:

1. **Stereotype Relevance.** Stereotype threat arises when a stereotype becomes salient and relevant to the performance in a situation (C. M. Steele & Aronson, 1995).
2. **Group Identification.** The individual must have a group identity to which a negative stereotype exists. For the stereotype to become self-relevant, they must feel like they belong to the stereotyped group (C. M. Steele, 1997; C. M. Steele & Aronson, 1995).
3. **Stereotype Awareness.** The individual must be aware of the stereotype, even though not necessarily believe in it (J. Aronson et al., 1998; C. M. Steele, 1997).
4. **Domain Identification.** The domain in which performance is evaluated must be relevant to the individual's self-concept or self-evaluation (J. Aronson et al., 1998; C. M. Steele, 1998).
5. **Self-Threat.** The stereotype poses a threat to the self, for example, through a fear confirming a negative stereotype of limited ability to themselves or others (J. Aronson et al., 1998; C. M. Steele, 1997; C. M. Steele & Aronson, 1995).
6. **Task Difficulty.** The task has to be sufficiently challenging to allow for a stereotype-related interference with the performance (C. M. Steele & Aronson, 1995). Later, C. M. Steele (1997) argued that the task must be of moderate difficulty. According to the paper, very easy and very hard tasks are nondiagnostic of ability and thus unlikely to produce threat-induced underperformance.

Subsequent research raised questions about the necessity of these theoretical variables, with meta-analyses reporting null or inconsistent effects for key concepts (Flore & Wicherts, 2015; Picho et al., 2013; Shewach et al., 2019; Nguyen & Ryan, 2008; Picho-Kiroga et al., 2021). This meta-analysis investigates whether theoretical boundary conditions can explain variability in stereotype threat effects. Thereby, it evaluates the conceptual coherence of the stereotype threat framework in the context of women in STEM. While meta-regressions cannot evaluate causality, consistent patterns of effect sizes depending on the presence or absence of theoretical conditions would support the construct's validity. On the other hand, absence may suggest conceptual ambiguity and question the relevance of the proposed conditions.

As this meta-analysis only includes stereotype threat studies that manipulate gender/sex salience in a STEM context, the first two conditions are satisfied by design. Furthermore, large-scale evidence indicates that a substantial majority of women worldwide recognize the gender-science stereotype (Miller et al., 2015; Nosek et al., 2009), beginning in early childhood (Cvencek et al., 2011; Shenouda et al., 2024). Because stereotype awareness is near-universal, the third theoretical condition is treated as a presupposition of stereotype threat research rather than a coded variable. This leaves domain identification, self-threat, and task difficulty as primary variables of interest (described in detail in the Methods section).

In the publications on stereotype threat, the authors did not specify a unified mechanism that explained how stereotype threat led to underperformance (Spencer et al., 1999; C. M. Steele, 1997; C. M. Steele & Aronson, 1995). Rather, the authors proposed a range of psychological mechanisms that might mediate the relationship. Since then, researchers have aimed to identify and test these mechanisms. The following section provides a general overview to deepen the understanding of the concept.

Mechanisms of Stereotype Threat

In their seminal paper, C. M. Steele and Aronson (1995) proposed a series of mechanisms that explain how stereotype threat interferes with performance. The proposed mechanisms were largely derived post hoc and remain speculative in nature. They include emotional, cognitive, and motivational processes such as increased anxiety, cognitive interference, divided attention, overeffort, or withdrawal of effort. Underperformance because of lowered expectations was discussed but rejected, since the manipulation of expectations did not predict performance. Finally, disidentification was introduced as a long-term effect influencing academic performance; however, as stated above, it was excluded as a short-term outcome in the original 1995 publication on stereotype threat.

In the subsequent theoretical elaboration, C. M. Steele (1997) refrained from specifying the mechanism but emphasized the role of a stereotype-induced anxiety. He conceptualized this anxiety as a form of “spotlight anxiety”, namely, as a fear of being evaluated or judged based on an existing stereotype (C. M. Steele, 1997).

J. Aronson, together with McGlone (2009), later reviewed psychological processes that could mediate the influence of stereotype threat for performance, reporting ambiguous findings on most of them. Based on their evaluation of available evidence, they identified reduced working memory, i.e., the limited ability to temporarily store and manipulate information, as a confirmed mechanism of stereotype threat. Anxiety was presented as a mediator with ambiguous evidence. Lowered expectations were reported as a mechanism in some studies; however, drawing from previous studies, they questioned the findings. Lowered effort, avoidance of challenge, and disidentification were proposed as long-term consequences of the threat rather than short-term mediators.

C. M. Steele and Aronson (1995) stated that “several such mechanisms may be involved simultaneously, or different mechanisms may be involved under different conditions” (p. 809). By doing so, they left room for divergent patterns and complicated the identification of core mechanisms of stereotype threat. Experimental studies show mixed results across different contexts and groups. A review of mediators of stereotype threat concludes that evidence for individual mechanisms is inconsistent (Pennington et al., 2016). The effects seem to depend on factors such as the stereotyped group, the type of stereotype threat induction, and the mediators and performance measures used. Taken together, the mechanisms through which stereotype threat manipulations lead to underperformance are complex and not well understood to date (Pennington et al., 2016; Shapiro & Neuberg, 2007).

Experimental Paradigms

Stereotype threat for women in STEM has been investigated using two main experimental paradigms: manipulating the diagnosticity of a test and directly manipulating the salience of the group identity or existing stereotype. The diagnosticity paradigm was first introduced by C. M. Steele and Aronson (1995) to investigate academic underperformance in Black Americans. It operates through the logic that the stereotype is implicitly activated through a domain-relevant and evaluative context.

The stereotype salience paradigm serves to investigate whether the observed effects of underperformance are attributable to the stereotype becoming relevant (Spencer et al., 1998). In this paradigm, the threat is induced by telling women that a test does or does not show gender differences (Spencer et al., 1998). Results showed that women performed worse than men in the gender differences condition. However, in the no-gender-differences condition, the performance gap disappeared. These findings were interpreted as evidence that the threat in high-stakes mathematics testing could stem from stereotype threat, rather than alternative explanations, such as biological differences in ability (Spencer et al., 1998; C. M.Steele, 1997).

Over time, the salience paradigm became the dominant method in stereotype threat research on women. Although originally interpreted as analogous, researchers suggest they may reflect distinct types of threats (Shapiro & Neuberg, 2007). In a meta-analytic context, comparing both paradigms risks introducing the “apples and oranges” problem (Sharpe, 1997): the silent control condition of the gender salience paradigm is, in some studies, equivalent to the threat condition in the test diagnosticity paradigm. Analyzing both studies together risks pooling effect sizes that are not conceptually equivalent.

To ensure conceptual clarity, only salience-paradigm studies were included in the present meta-analysis. Nevertheless, even with stereotype threat research limited to this paradigm, threat manipulations differ greatly, ranging from explicitly stating group differences (e.g., Campbell & Collaer, 2009) to identifying one's group membership (Anderson, 2001) to viewing stereotypic images (Good et al., 2010) or sex composition (Inzlicht & Ben-Zeev, 2000). This heterogeneity likely contributes to the ambiguity in stereotype threat effects, or even constitutes distinct forms of threat (Shapiro & Neuberg, 2007), problems related to internal and construct validity.

Conceptual and Methodological Criticism

Although the theory of stereotype threat has been widely cited and for a long time was regarded as a well-established psychological phenomenon (Eronen & Bringmann, 2021), it has faced criticism for its conceptual ambiguity and inconsistent operationalizations. Some authors have questioned the validity of its core assumptions and the robustness of its effects across contexts and populations (e.g., Picho-Kiroga et al., 2021; Sackett et al., 2004; Tomeh & Sackett, 2021; Warne, 2022; Whaley, 2018).

Several methodological issues specific to the original stereotype threat experiments on women by Spencer et al., (1999) were pointed out, questioning the validity of conclusions and reliability of the results (Stoet & Geary, 2012). In the first study of Spencer et al.'s (1999) paper, the observed pattern could represent actual gender ability differences in the data instead of differences based on stereotype threat. The second study excluded half of the data due to floor effects, raising concerns about test reliability. From the third study, some participants were discarded because they "did not make a reasonable effort" (Spencer et al., 1999, p. 15), raising concerns about the validity of the effects (Stoet & Geary, 2012).

Relevance of Stereotype Threat Research on Women in STEM

Stereotype threat is proposed to negatively affect skilled women's performance on difficult quantitative tests and to contribute to women's disidentification from STEM domains (Spencer et al., 1999). Based on the first stereotype threat experiments on women in mathematics, Spencer et al. (1999) proposed to change situational "predicaments" and remove the circumstantial threats to reduce women's underperformance in quantitative fields.

Since then, researchers have construed stereotype threat effects as a well-established phenomenon with robust consequences on women's performance and career aspirations in STEM fields (Stoet & Geary, 2012a). Consequently, the popular media have followed (e.g., Herrmann, 2022; Elsesser, 2019; Katter, 2018; Quirk, 2022; Vedantam, 2012; Weber & Winkler, 2019). However, whether stereotype threat is a real phenomenon and the effects are relevant is subject to an ongoing debate (Picho-Kiroga et al., 2021; Sackett et al., 2004; Stoet & Geary, 2012; Warne, 2022; Whaley, 1998, 1998; Zigerell, 2017).

The persistent gender disparities in STEM underscore the importance of understanding the mechanisms that trigger and perpetuate them and ultimately, how they can be overcome (UNESCO, 2024). C. M. Steele (1997) discussed the practical consequences of his research on students and proposed several strategies to reduce stereotype threat. Since then, researchers have developed interventions to mitigate the threat. These interventions were found to be generally effective in increasing task performance (Liu et al., 2021). Nevertheless, it is of great interest to clarify whether stereotype threat is a real and replicable phenomenon, under which conditions it occurs, and who is most affected. Gaining clarity on the underlying concept is crucial for intervention research, as it helps draw causal inferences, understand the mechanisms at work, and ultimately optimize interventions.

Furthermore, in light of the replication crisis (Open Science Collaboration, 2015), the replicability and stability of stereotype threat have been questioned (e.g., Inzlicht, 2024; Sailer, 2024; Schimmack, 2017). Inzlicht (2024) reflected about stereotype threat research in general and his own studies on the topic and that the effects were "produced by methods that we now consider questionable" (para. 14). The concept has also been discussed in the context of theory crisis, a fundamental problem of psychological research that relates to a lack of clear and falsifiable psychological theories (Eronen & Bringmann, 2021). Psychological theories often do not provide clear predictions, and empirical findings are rarely used for theoretical progress. Psychological scientists often develop new theories, but the evidence does not result in a consensual acceptance or refutation of theories; they rather fade away once the evidence base becomes too inconsistent (Eronen & Bringman, 2021). It is the responsibility of researchers to critically assess (controversial) theories, and if they do not hold, improve or discharge them (Eronen & Bringmann, 2021). Especially meta-analyses investigating stereotype threat for women in STEM have been debated and re-evaluated, reaching contrary conclusions (Nguyen & Ryan, 2008; Picho-Kiroga et al., 2021a; Ryan & Nguyen, 2017; Warne, 2022; Zigerell, 2017). The popularity of stereotype threat research contrasts with the ambiguous evidence. Given the far-reaching implications for education, policy, and social equity, it is crucial to provide a systematic synthesis and critical evaluation of the cumulative evidence.

Previous Meta-Analytic Evidence

A large and growing body of literature has investigated the concept of stereotype threat over the past three decades. In addition to primary studies, it has also been the subject of numerous meta-analyses that summarized evidence on stereotype threat in general, as well as group-specific stereotype threats (Appel et al., 2015; Appel & Weber, 2021; Armstrong et al., 2017; Flore & Wicherts, 2015; Gentile et al., 2018; Haft et al., 2023; Lamont et al., 2015; Nguyen & Ryan, 2008; Picho-Kiroga et al., 2021a; Priest et al., 2024; Shewach et al., 2019; Smittick, 2022). However, meta-analyses and their results have been subject to criticism, and subsequent revisions have led to contradicting conclusions (e.g., Nguyen & Ryan, 2008; Picho-Kiroga et al., 2021; Warne, 2022, Zigerell, 2017). Moreover, the considerable heterogeneity in results has remained largely unexplained, and there is little agreement on the role of theoretical conditions in the size and direction of effect sizes.

The overall effects have varied between small and moderate, from $d = -0.22$ for girls in mathematics (Flore & Wicherts, 2015) to $d = -0.63$ for immigrants (Appel et al., 2015). Synthesized effects of stereotype threat for female participants in STEM were on the lower end of the range (d = –0.22, Flore & Wicherts, 2015; $d = -0.21$, Nguyen & Ryan, 2008; $d = -0.33$, Shewach et al., 2019).

Besides estimating the summary effect of stereotype threat, meta-analyses covered an array of research questions, including numerous predictor variables. The influence of the theoretical conditions produced ambiguous results. Nguyen and Ryan (2008) found that stereotype relevance, operationalized as the strength of the threat removal strategy, differentially affected stereotype threat effects depending on the stereotyped group. Domain identification was found to have a quadratic association with stereotype threat effects, with moderately STEM-identified women being most impaired by the threat, whereas women with low or high identification were less affected; for ethnic minorities, primary data were insufficient (Nguyen & Ryan, 2008). In later meta-analyses, domain identification did not predict stereotype threat effects on women or ethnic minorities (Shewach et al., 2019; Stoevenbelt et al., 2022). Test difficulty also produced ambiguous findings, with more difficult tests yielding stronger stereotype threat effects in ethnic minorities and women (Nguyen & Ryan, 2008) or no effect on threat outcomes (Flore & Wicherts, 2015; Stoevenbelt et al., 2022).

Picho-Kiroga et al. (2021) investigated how the implementation of the conditions predicts stereotype threat by computing an “essential conditions” variable containing the three concepts stereotype salience, stereotype awareness, and domain identification. They found that only 6.9% adhered to all three conditions, and that the non-implementation of the conditions “artificially inflated the magnitude of stereotype threat effects” (Picho-Kiroga et al., 2021, p. 250). However, their meta-analysis was subject to criticism regarding the methodology and reasoning. Warne (2022) argued that, among other conceptual errors, publication bias results were misinterpreted, falsely reporting no evidence of bias. He considered the claim of "artificial inflation" of effect sizes in studies that did not follow the concept's theoretical boundaries to be an underestimation. Rather, the true effect of stereotype threat for women in mathematics was most consistent with zero.

Meta-analyses that have assessed dissemination bias in stereotype threat literature have, in general, reported evidence for bias. For example, all assessments adopted by Flore and Wicherts (2015) indicated biases, and the trim and fill method reduced the overall estimate from $d = –0.22$ to $d = –0.07$, changing an overall small stereotype threat in girls to essentially no effect. A reassessment of Nguyen and Ryan's (2008) meta-analysis using several dissemination bias methods indicated the presence of bias, reducing the confidence in the found summary effect (Zigerell, 2017). In line with this, Stoevenbelt et al., (2022) found small study effects with more precise studies clustering around the zero effect.

To conclude, numerous meta-analyses have studied stereotype threat effects across different contexts and have consistently found small to moderate effect sizes with substantial, largely unexplained heterogeneity. Support for theoretical moderators remains inconsistent. Overall, the robustness of effects is questionable, as a series of bias assessments have indicated inflated stereotype threat effects. This study, therefore, focuses on evaluating the influence of bias in stereotype threat research on women in STEM, applying a broad analytical framework to assess the stability of results.

Scope of the Present Study

The present meta-analysis aims to clarify the current state of evidence on stereotype threat for women in STEM. Beyond providing an updated summary of stereotype threat effects on women's performance in STEM, this study assesses whether the occurrence of the effects depends on specific theoretical boundary conditions proposed by C. M. Steele (1997). It further addresses questions on the generality of the phenomenon: Is stereotype threat a stable phenomenon that can be observed broadly, or does it depend on specific contextual factors?

The present analysis also evaluates the impact of potential biases. The credibility of the literature and the reliability of the overall effect size depend on the extent to which the estimates for stereotype threat are distorted, as well as the form of bias. By systematically examining evidence for different forms of bias, this study contributes to clarifying why the research on stereotype threat for women in STEM has produced conflicting findings and whether stereotype threat can be considered a robust and trustworthy phenomenon. This is essential given the continuing debate about the stability and replicability of stereotype threat effects.

Research Questions

1. What is the overall effect of stereotype threat for women in STEM?
2. To what extent can the heterogeneity in effect sizes be explained by the presence or absence of the theoretical conditions domain identification, self-threat, and task difficulty?
3. Do stereotype threat effects on women differ across different STEM subdomains?
4. Do stereotype threat effects on women in STEM differ depending on publication status?
5. How stable are stereotype threat effects on women in STEM over time?
6. Can the observed effects be generalized across different study specifications?
7. Is there evidence of dissemination bias?

Method

In the following sections, I describe my sampling plan, data exclusions, coding procedures and measures. This research was preregistered through the Open Science Framework: https://osf.io/mp9hq

Literature Search

To identify relevant studies, I conducted a literature search in August 2024. The databases CORE, ERIC, PsycInfo, Scopus, and Web of Science were searched with the search string (stereotype AND threat) AND (female OR wom*n OR girl* OR children OR sex OR gender), PSYNDEX was searched with (stereotype AND threat) AND (frau* OR weiblich* OR mädchen OR kind* OR geschlecht OR gender). The literature search in CORE was terminated when no relevant paper appeared on three consecutive pages with 30 hits per page. This strategy yielded a total number of 6,233 potentially eligible articles.

After automatically removing duplicates using Zotero (Corporation for Digital Scholarship, 2024), papers were excluded that were published before 1995, the date in which stereotype threat was first introduced as a concept (C. M. Steele & Aronson, 1995). Then, all abstracts were screened to detect papers focusing on stereotype threat. Three studies were written in languages other than English or German and were translated using the free version of the online translation tool DeepL (DeepL SE, 2025). This first screening process resulted in 1,136 studies that were retrieved and assessed for eligibility regarding the inclusion criteria discussed below.

Finally, I examined the studies cited by Picho-Kiroga et al. (2021), a recent meta-analysis of stereotype threat for women in STEM, to find other suitable studies.

Criteria for Inclusion/Exclusion

Studies were included in the analysis according to five criteria listed below. Of the 1,136 potentially eligible studies, five could not be retrieved, and 1,058 were excluded post-reading. All studies included in the meta-analysis are marked with an asterisk (*) in the bibliography.

1. **Experimental Manipulation.** The study had to be an experiment investigating stereotype threat for women in STEM. Specifically, the threat had to be induced by manipulating gender salience in at least one group of female participants, with at least one group of female participants serving as a comparison group. The mere manipulation of test diagnosticity , that is, informing whether a test is diagnostic or not, was not regarded as a sufficient activation of stereotype threat. While test diagnosticity has been discussed as an equal manipulation of stereotype threat, the

comparability of the conditions remains questionable. It was, however, regarded in the analysis as it was part of the operationalization of self-threat.

2. **No intervention study.** Studies aiming to reduce stereotype threat effects (i.e., intervention studies) were excluded.
3. **STEM Outcome.** The dependent variable had to fall within a STEM domain or involve a non-verbal reasoning task closely related to STEM.
4. **Individual-Level Data.** Only studies that investigated stereotype threat effects on the individual level were included.
5. **Statistical Information.** Sufficient statistical information had to be reported to calculate Cohen's *d*s. Specifically, studies had to include the means (*M*) and standard deviations (*SD*s) or standard errors (*SE*s) for the outcome variable for women in both the experimental and comparison group, as well as the sample size.
   Relative scores (e.g., percentage correct and corresponding *SD*s or *SE*s) were accepted only if sufficient information was available to reconstruct the original distribution, i.e., the number of items. Studies were excluded if the outcome measures were transformed in a way that prevented reconstruction of raw values (e.g., point deductions for incorrect answers or *z*-transformations), or if only accuracy scores were reported (i.e., number correct divided by number attempted), as these formats limited the calculation of *d*s and reduced comparability across studies.
   Adjusted outcomes were only included when they were controlled for ability, meaning that performance outcomes were statistically adjusted for participants' prior ability level. This procedure is widespread in stereotype threat research and was a variable of interest in the specification curve analysis. However, I contacted the corresponding authors to request the raw means, standard deviations, and asmple sizes of the outcome variables (unadjusted for prior ability) and used these, where provided (see below for details on the author contact procedure).

Coding Procedure

An elaborate coding form was used to code the included studies. I read and coded all studies that met the inclusion criteria after full-text ($k = 93$). The coding included study characteristics, names of the experimental and comparison groups, sample characteristics, methodological information, effect size-related data, and all other variables of interest described below.

To ensure accuracy, I reviewed the coding sheet by revisiting each study independently in a second round. If the entries were consistent, I color-coded them as confirmed. If I encountered inconsistencies or uncertainty, I corrected the value if necessary, and left the cell unmarked to allow for additional review at a later stage.

Since a multilevel meta-analysis was calculated, multiple effect sizes per study were coded, when they were found relevant (Assink & Wibbelink, 2016). Using this approach, dependency in the data is modeled, and coding several effect sizes per study is regarded as an approach superior to selecting a single effect size or aggregating effects within studies (Rodgers & Pustejovsky, 2021).

Multiple effect sizes from studies were extracted when they reflected different threat manipulations, comparison groups or outcome variables. Additionally, multiple outcomes per study were coded when variables of interest were manipulated or naturally differed among conditions or groups (e.g., education status or domain identification). Finally, multiple effect sizes were allowed where conditions or groups differed in terms of methodological or study-design factors, for example manipulations of the sex of the experimenter (Drążkowski et al., 2017), the sex composition (Moè, 2018) or when participants were split into groups, for example based on testosterone levels (Josephs et al., 2003) or school type (Picho & Stephens, 2012). The reasoning behind this decision was that such cases created multiple dependent effects within studies, which were accounted for in the multi-level structure of meta-analysis.

When studies or conditions were confounded with variables other than listed above, they were not coded because they did not allow for a clear interpretation of the effects.

In cases where studies met the criteria, but statistical information was missing, the corresponding authors were contacted and asked to provide the missing information. When authors did not respond, they were contacted a second time after one month, and a third time again after another month. If they did not provide the required information or did not respond, the study was excluded from further analysis.

Definition of Variables

Each included study was systematically coded on a set of variables. These are described below.

Performance in STEM. Performance scores were the outcome of interest and were coded as raw means and corresponding *SD*s for women in the experimental and comparison groups. Relative data was transformed into raw scores, with the approach delineated below. Ability-adjusted data was coded if the authors failed to send the raw data, as described above in the Criteria for Inclusions section.

For outcome variables where higher scores indicated poorer performance (e.g. angular errors or number of trials to solve a problem), values were reverse coded to ensure that the direction of the effect was consistent across studies.

Comparison groups. Comparison groups were distinguished in control and removal groups. They were coded as control groups if participants received no treatment or if they received an irrelevant treatment. If the participants received information aiming to reduce or remove stereotype threat for women in STEM, such as information that there were no gender differences in the given test (e.g. as in Sunny et al., 2017), the comparison groups were coded as removal groups.

Education Status. The sample was coded as precollege, when it was a school sample, college, when participants were university students, and mixed when no clear participant type was evident.

STEM Domain. The outcome measure was assigned to one of the STEM domains. If the outcome assessed performance in multiple domains or was assessed by a task type that was relevant to multiple domains (e.g., mental rotation), the outcome domain was coded as general.

Ability Control. Whether the reported performance was adjusted by ability was coded in a binary ("yes" vs. "no").

Setting. The setting in which the study was conducted was coded as such: "lab" for all studies conducted in dedicated testing rooms, "field" for studies conducted in classrooms or other everyday settings, and "online" for studies conducted online.

Threat Activation. The manipulation of stereotype threat was classified as blatant when it involved explicit statements, and as subtle when it was induced implicitly, such as through priming procedures.

Study Location. The location was coded based on the geographical region in which the data were collected, categorized by continent.

Publication Status. Studies that were published in a peer-reviewed journal were coded as published, otherwise they were coded as unpublished. In cases where both a published and unpublished record was available, coding was based on the record that contained all relevant information. If this was the case in both records, the study was coded as published to ensure consistency and relevance for subsequent publication bias analyses.

Domain identification. If the sample was identified with the studied STEM domain of interest was coded binarily (yes vs. no).

C. M. Steele (1997) defined domain identification as a combination of ability and self-relevance. The present meta-analysis did not treat domain-specific ability or performance (such as SAT achievement) as a sufficient indicator of domain identification. Likewise, the enrollment in certain school courses was not regarded as domain-identified, as these often reflect mandatory or constrained choices rather than active self-selection.

At the university level, studying a STEM subject was considered domain-identified. Also, if studies measured domain identification, participants were classified as domain-identified if they scored within the upper 50% percentile on a domain identification scale.

Self-Threat. Self-threat was operationalized based on how the study's purpose was framed to participants and coded using a decision tree and was coded as yes or reduced.

First, I assessed whether the participants in the experimental group received feedback on their performance, or whether they were told their individual performance would be ranked (e.g., "we will be comparing your score to other students", Mangels et al., 2012, p. 233).

If this did apply, I marked the condition as self-threatening, if not, I proceeded to a second step, where I looked whether there was a framing that distanced the task from personal evaluation (e.g., "results obtained today will be used only for program assessment", Anderson, 2001, p. 81; "study about cognitive styles", Drążkowski et al., 2017, p. 417).

If a distancing framing was present, I recorded it before proceeding to the final step, in which I examined whether the task was presented as predictive of personal abilities or future success (e.g., "spatial abilities […] are very important in everyday life", Moè, 2012, p. 22).

When step two was noted as distancing, and step three was not noted as predictive framing, I coded the condition as reduced self-threat, and vice-versa. In cases where both a distancing and a predictive framing was present, I decided based on expert judgment which framing was more salient and thus likely to have guided participants' perception of the task.

Task difficulty. Task difficulty was calculated based on the average percentage of items not solved in the pre-test or in the comparison group(s).

Statistical Analyses

All statistical analyses were performed in the open-source software R using the packages metafor (Viechtbauer, 2010) and robumeta (Fisher et al., 2014).

Effect Size Calculation. Prior to all analyses, standardized mean differences (Cohen's *d*s; Cohen, 1988) were calculated for group differences according to the following formula:

$$\text{Cohen's } d = \frac{M_1 - M_2}{pooled\ SD} \text{ with pooled } SD = \sqrt{\frac{((N_1 - 1) * SD_1{}^2 + (N_2 - 1) * SD_2{}^2)}{N_1 + N_2 - 2}}$$

When *SE*s were reported instead of *SD*s I calculated *SD*s using the following formula:

$$SD = SE * \sqrt{N}$$

Group 1 (i.e., M_1, SD_1, N_1) always referred to the experimental group (i.e., under stereotype threat), and Group 2 (i.e., M_2, SD_2, N_2) referred to the comparison group. Thus, negative values of *d* indicate poorer performance in the stereotype threat condition, which in the context of stereotype threat theory is interpreted as a stereotype threat effect on women in STEM subjects.

In several cases, group-level sample sizes were not reported. When the total sample size was available but no further group information was provided, I assumed equal group sizes. In cases where additional information was available , I estimated the group sizes accordingly, based on the reported test statistics (e.g., degrees of freedom in *t*-tests or ANOVAs). In one study (Steinberg et al., 2012) outcome means and standard deviations were reported for participants categorized as low versus high in domain identification, defined as ±1 *SD* from the sample mean. As no group-level sample sizes were provided, I estimated subgroup sizes based on the total sample assuming a normal distribution of domain identification and allocating approximately 16% of participants to each of the high and low domain identification groups, and approximately 68% to the medium domain identification group.

The escalc() function embedded in metafor (Viechtbauer, 2010) was used to calculate the *d*s, which operates using the formula above.

The effect sizes were interpreted according to Cohen (1988): values of $d = 0.2$, 0.5, and 0.8 were considered the lower thresholds of small, medium, and large effects, respectively.

Meta-Analytic Method. To answer research questions 1 through 5, several multilevel models were computed using the rma.mv function in metafor (Viechtbauer, 2010). A random-effects meta-analysis in the tradition of Hedges & Olkin (1985), in which one effect size per study is extracted, allows for random variation of data between studies. Hence, on level 1 the variance component represents variation within studies (sampling error), and level 2 represents variation between studies (Konstantopoulos, 2011; Van Den Noortgate et al., 2013).

Because in the present meta-analysis multiple effect sizes were extracted per study, a third level was modelled to account for dependency in the data; therefore, three variance components were modelled: sampling variation for outcomes (level 1), variation over outcomes within studies (level 2), and variation between studies (level 3; Assink & Wibbelink, 2016; Van Den Noortgate et al., 2013).

Summary Effect. For the estimation of the overall effect of stereotype threat for women in STEM (Research Question 1), an random-effects multilevel null model was calculated, which allows for random effects on the within-study and between-study levels (Assink & Wibbelink, 2016). The parameter was estimated using the Restricted Maximum Likelihood method (REML).

To assess whether the model fitted the data significantly better than a two-level model, a Likelihood-Ratio Test (LRT) was performed by comparing an random-effects multilevel null model with Maximum Likelihood parameter estimation to a second model with the within-study variance fixed to zero (Assink & Wibbelink, 2016).

Heterogeneity was quantified with σ_1^2, σ_2^2, Q and I^2. To interpret heterogeneity, I followed Assink & Wibbelink's (2016) approach by estimating I^2 for the three meta-analytic levels. I^2 was interpreted according to Higgins et al. (2003), with 25%, 50%, and 75% as low, moderate and high levels of inconsistency.

Meta-Regression on Theoretical Conditions. A central objective of the present study was to investigate the amount of variance between effect sizes in the literature on stereotype threat for women in STEM that can be explained by the conditions posed in stereotype threat theory (Research Question 2). To examine this, a mixed-effects multilevel model with the three predictors domain identification, self-threat and task difficulty posed was fitted.

The categorical predictors domain identification and self-threat were dummy coded with “domain-identified” and “self-threat present” as the reference categories, because these represent the theoretically central conditions under which stereotype threat effects are expected to be strongest. The metric predictor task difficulty was mean centered prior to the analysis to ensure that the intercept reflects the estimated effect at the average level of task difficulty.

Since a non-linear relationship between task difficulty and the outcome was theoretically plausible with strongest threat effects at moderate difficulty (C. M. Steele, 1997), a mixed-effects multilevel model with a quadratic term for task difficulty was additionally calculated. The quadratic term allows for testing whether stereotype threat effects increase up to a certain level of task difficulty and then decrease again, rather than changing in a strictly linear fashion. This analysis was not preregistered and is therefore considered exploratory.

Sensitivity Analyses. To examine the generalizability and robustness of stereotype threat effects on women in STEM, a series of analyses were conducted. These analyses are described in detail below.

STEM Domain. First, a subgroup-analysis was conducted to analyze whether the effects differed between differing STEM subdomains (Research Question 3), by including the outcome’s domain as a categorical predictor.

Publication Status. Possible differences in effect sizes depending on publication status were investigated a mixed-effects multilevel model with publication status as a dichotomous predictor (Research Question 4). Additionally, two separate meta-analyses were conducted for the published and unpublished subsets. These subgroup analyses were performed not as formal tests of publication bias. The goal was rather to provide descriptive estimates of effect sizes and heterogeneity within each subset, thereby contributing to a better understanding of the heterogeneity observed across all stereotype threat studies on women in STEM.

Publication Year. Changes in effect sizes over time were examined with mixed-effects multilevel models with publication year as a continuous predictor to assess whether stereotype threat effects have systematically changed over the past decades (Research Question 5). Publication year was centered at the year 2000, corresponding to the earliest publication included in the analysis (Inzlicht & Ben-Zeev, 2000).

To aid interpretation and assess whether this trend may be influenced by dissemination bias, exploratory analyses were conducted separately for published and unpublished studies. These analyses were not preregistered and aimed to clarify whether the decline effects were observable across all studies or limited to the published subset. A decline effect in both subsets would suggest a real change of stereotype threat effects on women in STEM, whereas a decline effect restricted to published studies would be more consistent with publication bias.

Multiverse and Specification Curve Analysis. To examine robustness across a wide variety of analytical decisions (Research Question 6), a multiverse and specification curve analysis was conducted, originally introduced by Simonsohn et al. (2015) and adapted for meta-analyses by Voracek et al. (2019). This multiverse analysis allows for the modeling of effect sizes across multiple researcher degrees of freedom, specifically decisions about which data to include and how to analyze them (Voracek et al., 2019). The specification curve organizes and visualizes the results.

To conduct this analysis, it was necessary to identify all reasonable specifications, including both "which" factors and "how" factors. I identified the included moderators by systematically reviewing previous meta-analyses on stereotype threat (Appel et al., 2015; Appel & Weber, 2021; Armstrong et al., 2017; Flore & Wicherts, 2015; A. Gentile et al., 2018; Lamont et al., 2015; Liu et al., 2021; Nadler & Clark, 2011; Nguyen & Ryan, 2008; Picho et al., 2013; Picho-Kiroga et al., 2021a; Priest et al., 2024; Shewach et al., 2019). Since the number of potential moderators was very large, including all of them in the present multiverse analysis was neither feasible nor theoretically justified. Conceptually overlapping moderators, (e.g., stereotype- vs. fact-based threat activation; Lamont et al., 2015, strongly overlapped with subtle vs. blatant activation) were excluded to avoid inflating the number of possible specifications without providing additional information. Other potential moderators, such as participants' age, sex composition, or type of outcome measure would have required arbitrary categorization, introducing artefactual patterns and reducing interpretability. Finally, some moderators were expected to be too rarely represented to allow for meaningful comparisons (e.g., use of financial incentives as motivational feature; Shewach et al., 2019).

Therefore, I focused on a set of seven distinct methodological and study design "which" factors: (1) comparison group (control vs. removal), (2) sample type (pre-college vs. college vs. mixed), (3) study setting (lab vs. field vs. online), (4) stereotype threat activation (subtle vs. blatant), (5) ability-control (unadjusted vs. adjusted), (6) study location (North America vs. Europe vs. Asia vs. Africa vs. Oceania), (7) publication status (published vs. unpublished). Additionally, supersets were formed that included all factor levels (Voracek et al., 2019).

Each of these subsets was analyzed according to two "how" factors: (1) effect size (Cohen's *d* vs. Hedges' *g*), (2) meta-analytic model (fixed-effect model vs. random-effects model). Originally, the preregistration listed the estimation method (REML vs. ML) as a 'how' specification factor. The intended factor, however, was the meta-analytic model type (fixed-effect vs. random-effects). In the present specification-curve analysis, I therefore compare fixed-effect models with random-effects models; random-effects models are estimated with REML.

Out of the seven "which" factors, all possible combinations of factor levels were formed (Voracek et al., 2019). The combination of these "which" and "how" factors yielded a total of $3^4 * 4^2 * 6 * 2^2 = 31{,}104$ specifications calculated. Only specifications containing ≥ 2 samples were retained for the final specification curve analysis.

For computational feasibility, a two-level meta-analytic model was used in the multiverse analysis, rather than the three-level model approach applied in the main analyses. Effect sizes were aggregated when they originated from the same study, did not differ in any of the "which" factors relevant to the multiverse analysis, and reflected comparable outcome measures (i.e., identical item length). The aggregation strategy used in this analysis represents a middle ground between two common approaches in meta-analytic practice: a conservative approach that aggregates all effect sizes from the same study to preserve statistical independence (e.g., Borenstein et al., 2021), and a more inclusive approach that treats all effect sizes as independent units, regardless of origin (e.g., Voracek et al., 2019).

Outlier Diagnostics and Combinatorial Meta-Analysis (GOSH). To identify potentially influential cases, outlier diagnostics were conducted using metafor's inf() function (Viechtbauer, 2010). The most extreme outlier was highlighted in the GOSH to visualize its influence on the overall distribution of effect sizes and heterogeneity.

To further account for potentially influential systematic factors not explicitly modelled in the specification curve analysis, a combinatorial meta-analysis (Olkin et al., 2012) was conducted. This method performs meta-analyses on all possible subsets of the included studies, and GOSH visualizes the resulting distribution of effect sizes and heterogeneity (I^2). Because all possible (but not necessarily theoretically justified) combinations are considered, the shape of the distribution provides insight into whether the pattern is unimodal or indicates the presence of influencing factors. As the number of potential subsets grows rapidly with the number of included studies, a random subset of 100,000 combinations was calculated, and the resulting pattern was inspected visually.

Dissemination Bias.

To assess potential distortions due to dissemination bias (Research Question 7), I followed the approach by Petkari et al. (2024) for bias assessment in the context of three-level meta-analysis. Only studies published in peer-reviewed journals were included in the following analyses.

Egger Regression. To model the relation of study precision with effect sizes, an Egger regression was conducted with the standard error on effect sizes (Sterne & Egger, 2001).

Contour-enhanced Confetti Funnel Plot. A funnel plot to visualize dependency and significance regions was modelled, and the Egger regression line was added to visually inspect asymmetry.

Egger Multi Level Meta-Analysis (MLMA). Additionally, the Egger MLMA (Rodgers & Pustejovsky, 2021) technique, a regression with standard error on effect sizes with sandwich estimators, to account for dependency within studies was calculated using the robumeta package in R (Fisher et al., 2014). Here, a multi-level meta-regression is calculated by adding a measure of effect size precision as a predictor.

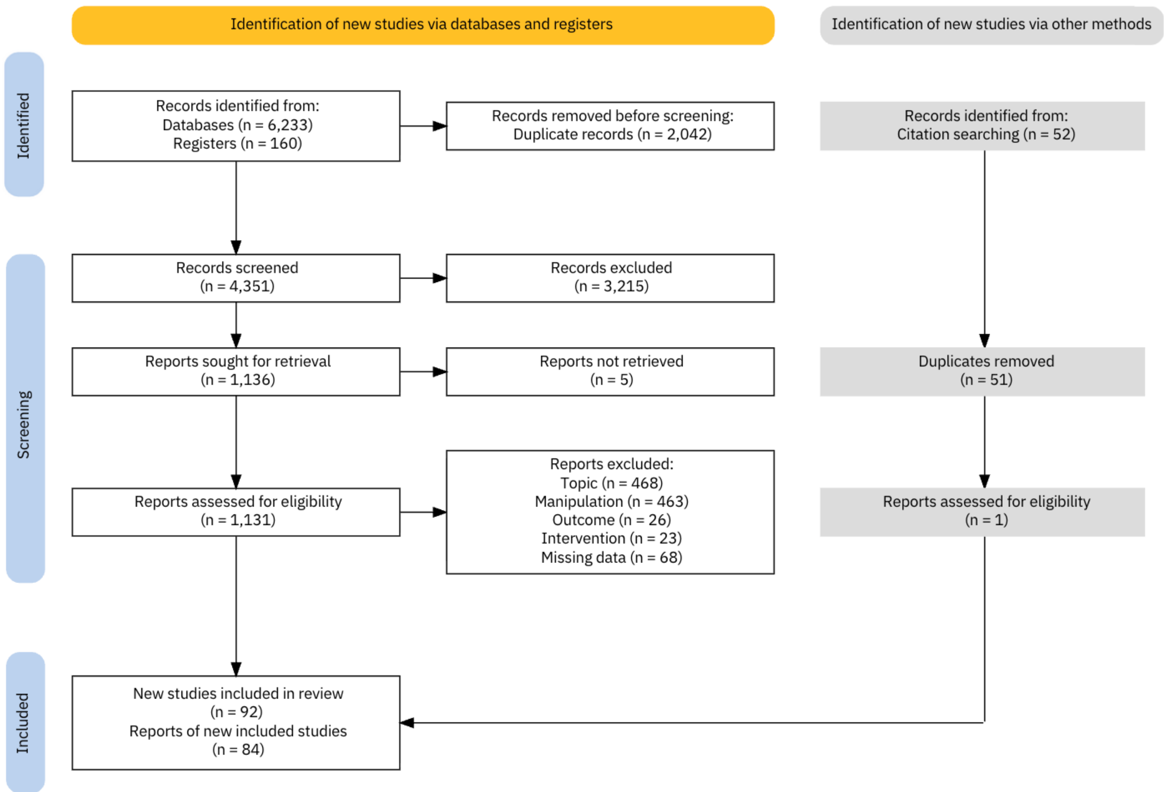
Three-parameter-selection model (3PSM). A 3PSM was conducted to detect a potential bias stemming from selective reporting (Rodgers & Pustejovsky, 2021). The 3PSM estimates the chance of observing a non-significant effect size versus a significant one (ψ). A likelihood-test is used to detect selective reporting with the null hypothesis that there is no selective reporting of significant results (Rodgers & Pustejovsky, 2021).

Results

Literature search yielded 6,233 papers, 4,351 of which were fully screened after duplicates were removed. One additional paper was detected through comparing included studies with the most recent meta-analysis on the topic (Picho-Kiroga et al., 2021). In total, 161 effect sizes from 92 studies, derived from 84 individual reports corresponding to 77 unique citations. The included studies comprised a combined sample size of $N = 11{,}608$ female participants (see Fig. 1).

Figure 1

PRISMA Chart



Note. This PRISMA flow diagram illustrates the process of study selection for the systematic review. Records were identified through database searches, study registers, and citation tracking.

An overlap of $k = 29$ citations (37.67%) was found between the present meta-analysis and Picho-Kiroga et al.,'s (2021). The small overlap was due to differing inclusion criteria (specifically, I applied strict criteria regarding the experimental manipulation and statistical

Table 1

Overview of Included Studies

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Agnoli et al., 2021 (1)	75	-0.27	0.23	Identified	Present	40.1	Subtle	Control	Unadjusted	Math	School	Europe	Published
Agnoli et al., 2021 (2)	80	0.27	0.22	Identified	Present	36.1	Subtle	Control	Unadjusted	Math	School	Europe	Published
Ambady et al., 2004 (1)	40	-0.20	0.32	Not identified	Present	49.5	Subtle	Control	Unadjusted	Math	College	North America	Published
Ambady et al., 2004 (2)	39	-0.84	0.33	Not identified	Present	40.1	Subtle	Control	Unadjusted	Math	College	North America	Published
Anderson, 2001	604	-0.07	0.08	Not identified	Present	46.5	Subtle	Control	Unadjusted	General	College	North America	Unpublished
Appel et al., 2011	40	-0.16	0.32	Identified	Reduced	32.4	Subtle	Removal	Unadjusted	Science	College	Europe	Published
Aramovich, 2014	73	-0.53	0.24	Not identified	Present	78.5	Blatant	Control	Unadjusted	Math	College	North America	Published
Bailey, 2005	41	-0.50	0.32	Identified	Reduced	42.7	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Ben-Zeev & Kirtman, 2012 (1)	45	-0.03	0.30	Not identified	Present	12.4	Subtle	Removal	Adjusted	Math	College	North America	Unpublished
Ben-Zeev & Kirtman, 2012 (2)	43	-1.03	0.32	Not identified	Present	12.4	Subtle	Removal	Adjusted	Math	College	North America	Unpublished
Ben-Zeev et al., 2005	37	-1.00	0.35	Identified	Present	55.9	Subtle	Control	Unadjusted	Math	College	North America	Published
Cadinu et al., 2006 (1)	25	-1.52	0.45	Not identified	Present	34.1	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Cadinu et al., 2006 (2)	34	0.03	0.34	Not identified	Present	34.1	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Campbell & Collaer, 2009 (1)	70	0.00	0.24	Not identified	Present	17.0	Blatant	Control	Unadjusted	General	College	North America	Published
Campbell & Collaer, 2009 (2)	73	-0.61	0.24	Not identified	Present	17.0	Blatant	Removal	Unadjusted	General	College	North America	Published
Casad et al., 2017	498	0.07	0.09	Not identified	Present	38.9	Blatant	Removal	Unadjusted	Math	School	North America	Published
Chalabaev, 2012	21	-0.53	0.44	Not identified	Present	38.0	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Chan & Rosenthal, 2014	37	-1.24	0.36	Not identified	Present	48.0	Blatant	Control	Unadjusted	Math	School	Asia	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Cherney & Campbell, 2011 (1)	204	0.29	0.14	Not identified	Present	55.2	Blatant	Removal	Unadjusted	Math	School	North America	Published
Cherney & Campbell, 2011 (2)	135	0.51	0.17	Not identified	Present	55.2	Blatant	Removal	Unadjusted	Math	School	North America	Published
Conway-Klaassen, 2010 (1)	39	0.06	0.32	Identified	Present	63.5	Blatant	Control	Unadjusted	Science	College	North America	Unpublished
Conway-Klaassen, 2010 (2)	40	-0.03	0.32	Identified	Present	63.5	Blatant	Removal	Unadjusted	Science	College	North America	Unpublished
Danaher, 2011	62	-0.16	0.25	Not identified	Present	43.9	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Davies et al., 2016 (1)	105	-0.52	0.20	Not identified	Present	51.0	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Davies et al., 2016 (2)	105	0.00	0.20	Not identified	Present	44.0	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Davies et al., 2016 (3)	47	-0.32	0.30	Not identified	Present	57.0	Blatant	Removal	Unadjusted	Math	School	Europe	Published
Davies et al., 2016 (4)	47	0.58	0.30	Not identified	Present	58.0	Blatant	Removal	Unadjusted	Math	School	Europe	Published
Delgado & Prieto, 2008 (1)	168	-0.01	0.15	Not identified	Present	25.4	Blatant	Control	Unadjusted	Math	School	Europe	Published
Delgado & Prieto, 2008 (2)	168	-0.28	0.16	Not identified	Present	63.5	Blatant	Control	Unadjusted	General	School	Europe	Published
Deviyanti, 2016 (1)	43	-0.45	0.32	Identified	Reduced	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (2)	21	0.06	0.44	Identified	Reduced	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (3)	90	-0.21	0.21	Identified	Reduced	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (4)	107	0.08	0.19	Not identified	Reduced	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (5)	51	-0.53	0.29	Identified	Present	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (6)	21	0.09	0.45	Identified	Present	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (7)	60	0.19	0.26	Identified	Present	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Deviyanti, 2016 (8)	98	0.05	0.20	Not identified	Present	62.2	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Drazkowski et al., 2017	83	-0.42	0.22	Not identified	Reduced	21.3	Subtle	Control	Unadjusted	General	College	Europe	Published
Dunst et al., 2013	26	-0.46	0.40	Not identified	Present	23.1	Blatant	Removal	Unadjusted	General	School	Europe	Published
Finnigan & Corker, 2016 (1)	89	-0.09	0.21	Not identified	Reduced	53.5	Blatant	Control	Unadjusted	Math	Mixed	North America	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Finnigan & Corker, 2016 (2)	93	-0.03	0.21	Identified	Reduced	53.5	Blatant	Control	Unadjusted	Math	Mixed	North America	Published
Fogliati & Bussey, 2013	54	-0.68	0.28	Not identified	Present	66.7	Blatant	Removal	Unadjusted	Math	College	Oceania	Published
Ford et al., 2004	31	-0.71	0.37	Not identified	Present	36.2	Blatant	Removal	Unadjusted	Math	College	North America	Published
Ganley, 2011 (1)	30	0.16	0.37	Not identified	Present	47.9	Subtle	Control	Unadjusted	Math	School	North America	Unpublished
Ganley, 2011 (2)	67	0.14	0.24	Not identified	Present	47.9	Subtle	Control	Unadjusted	Math	School	North America	Unpublished
Ganley, 2011 (3)	77	-0.27	0.23	Not identified	Present	47.9	Subtle	Control	Unadjusted	Math	School	North America	Unpublished
Ganley, 2011 (4)	65	0.06	0.25	Not identified	Present	63.0	Subtle	Control	Unadjusted	Math	College	North America	Unpublished
Gentile, 2009	63	0.04	0.25	Not identified	Present	45.4	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Good et al., 2010 (1)	34	-0.69	0.35	Not identified	Present	25.9	Subtle	Removal	Unadjusted	Science	School	North America	Published
Good et al., 2010 (2)	33	-0.28	0.35	Not identified	Present	25.9	Subtle	Control	Unadjusted	Science	School	North America	Published
Grand et al., 2011 (1)	113	0.12	0.19	Not identified	Present	43.5	Blatant	Control	Unadjusted	Engineering	College	North America	Published
Grand et al., 2011 (2)	123	0.23	0.18	Not identified	Present	48.9	Blatant	Control	Unadjusted	Engineering	College	North America	Published
Grand et al., 2011 (3)	113	-0.06	0.19	Not identified	Present	50.4	Blatant	Control	Unadjusted	Math	College	North America	Published
Grand et al., 2011 (4)	123	0.24	0.18	Not identified	Present	55.0	Blatant	Control	Unadjusted	Math	College	North America	Published
Gringard, 2017 (1)	61	0.13	0.26	Not identified	Present	55.7	Blatant	Control	Unadjusted	Math	College	Europe	Unpublished
Gringard, 2017 (2)	38	-0.97	0.34	Not identified	Present	55.7	Blatant	Control	Unadjusted	Math	College	Europe	Unpublished
Gringard, 2017 (3)	38	-0.07	0.32	Not identified	Present	55.7	Blatant	Control	Unadjusted	Math	College	Europe	Unpublished
Gringard, 2017 (4)	76	-0.45	0.23	Not identified	Present	55.7	Blatant	Control	Unadjusted	Math	College	Europe	Unpublished
Guizzo et al., 2019	71	0.06	0.24	Not identified	Present	79.7	Blatant	Removal	Unadjusted	General	Mixed	Europe	Published
Hermann & Vollmeyer, 2016	54	-0.58	0.28	Not identified	Present	37.5	Subtle	Control	Unadjusted	Math	School	Europe	Published
Hermann & Vollmeyer, 2022	72	-0.11	0.24	Not identified	Present	55.0	Subtle	Control	Unadjusted	Math	School	Europe	Published
Inglis & O'Hagan	719	0.09	0.07	Not identified	Reduced	61.8	Subtle	Control	Unadjusted	Math	School	Europe	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Inzlicht & Ben-Zeev, 2000 (1)	36	-0.56	0.34	Not identified	Present	72.8	Subtle	Control	Adjusted	Math	College	North America	Published
Inzlicht & Ben-Zeev, 2000 (2)	28	-0.20	0.38	Not identified	Present	60.2	Subtle	Control	Adjusted	Math	College	North America	Published
Jamieson & Harkins, 2009 (1)	32	-1.37	0.39	Not identified	Present	86.3	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2009 (2)	32	0.40	0.36	Not identified	Present	65.4	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2009 (3)	32	-1.83	0.42	Not identified	Present	87.5	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2009 (4)	32	0.70	0.36	Not identified	Present	68.3	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2010 (1)	26	0.09	0.39	Not identified	Reduced	56.9	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2010 (2)	26	-1.30	0.43	Not identified	Present	56.9	Blatant	Removal	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2012 (1)	34	-0.56	0.35	Not identified	Reduced	68.6	Subtle	Control	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2012 (2)	34	-0.99	0.36	Not identified	Reduced	68.6	Subtle	Control	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2012 (3)	50	0.30	0.29	Not identified	Present	68.6	Blatant	Control	Unadjusted	Math	College	North America	Published
Jamieson & Harkins, 2012 (4)	50	-0.93	0.30	Not identified	Present	68.6	Blatant	Control	Unadjusted	Math	College	North America	Published
Johnson, 2012	74	0.18	0.23	Not identified	Present	45.6	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Jones, 2005	63	-0.24	0.25	Not identified	Present	55.3	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Josephs et al., 2003 (1)	37	-0.80	0.34	Not identified	Present	31.9	Subtle	Control	Adjusted	Math	College	North America	Published
Josephs et al., 2003 (2)	39	0.22	0.32	Not identified	Present	31.9	Subtle	Control	Adjusted	Math	College	North America	Published
Kahalon et al., 2018 (1)	41	0.19	0.32	Not identified	Present	27.0	Subtle	Control	Unadjusted	Math	College	Asia	Published
Kahalon et al., 2018 (2)	42	0.50	0.32	Not identified	Present	27.0	Subtle	Control	Unadjusted	Math	College	Asia	Published
Kahalon et al., 2018 (3)	39	0.28	0.32	Identified	Present	27.0	Subtle	Control	Unadjusted	Math	College	Asia	Published
Kahalon et al., 2018 (4)	39	-0.17	0.32	Identified	Present	27.0	Subtle	Control	Unadjusted	Math	College	Asia	Published
Keller, 2002	37	-0.40	0.34	Not identified	Present	48.5	Blatant	Control	Unadjusted	Math	School	Europe	Published
Keller, 2007	55	0.13	0.27	Not identified	Present	55.1	Blatant	Removal	Adjusted	Math	School	Europe	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Lesko, 2006	68	-0.61	0.25	Not identified	Present	54.9	Blatant	Removal	Adjusted	Math	College	North America	Published
Lester, 2007 (1)	66	0.06	0.26	Not identified	Reduced	51.2	Subtle	Control	Unadjusted	Math	College	North America	Unpublished
Lester, 2007 (2)	66	-0.15	0.26	Not identified	Reduced	45.6	Subtle	Control	Unadjusted	Math	College	North America	Unpublished
Macchione & Sacco, 2023	159	0.12	0.16	Not identified	Present	51.7	Blatant	Removal	Unadjusted	Math	College	North America	Published
Mangels et al., 2012 (1)	35	-0.18	0.34	Not identified	Present	40.3	Blatant	Removal	Unadjusted	Math	College	North America	Published
Mangels et al., 2012 (2)	33	-0.21	0.36	Identified	Present	40.3	Blatant	Removal	Unadjusted	Math	College	North America	Published
Marchand & Taasoobshirazi, 2013 (1)	101	0.04	0.20	Not identified	Reduced	76.0	Subtle	Control	Unadjusted	Science	School	North America	Published
Marchand & Taasoobshirazi, 2013 (2)	90	-0.58	0.22	Not identified	Reduced	76.0	Blatant	Removal	Unadjusted	Science	School	North America	Published
McGlone & Aronson, 2006	45	-0.63	0.31	Not identified	Present	49.0	Subtle	Control	Unadjusted	General	College	North America	Published
Mingle, 2014 (1)	110	0.20	0.19	Not identified	Present	39.0	Blatant	Removal	Unadjusted	Math	School	North America	Unpublished
Mingle, 2014 (2)	115	0.28	0.19	Not identified	Present	77.0	Blatant	Control	Unadjusted	Math	School	North America	Unpublished
Mingle, 2014 (3)	99	-0.16	0.20	Not identified	Present	64.0	Blatant	Control	Unadjusted	Math	School	North America	Unpublished
Moè, 2018 (1)	43	-0.54	0.31	Not identified	Present	32.2	Blatant	Removal	Unadjusted	General	College	Europe	Published
Moè, 2018 (2)	35	0.46	0.34	Not identified	Present	32.2	Blatant	Removal	Unadjusted	General	College	Europe	Published
Muzzatti & Agnoli, 2007 (1)	35	0.05	0.36	Not identified	Present	49.1	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (2)	68	0.23	0.24	Not identified	Present	33.7	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (3)	64	0.13	0.25	Not identified	Present	39.0	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (4)	42	-0.42	0.31	Not identified	Present	33.7	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (5)	42	0.03	0.31	Not identified	Present	65.6	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (6)	48	0.15	0.29	Not identified	Present	69.5	Subtle	Control	Unadjusted	Math	School	Europe	Published
Muzzatti & Agnoli, 2007 (7)	30	-1.20	0.40	Not identified	Present	67.5	Subtle	Control	Unadjusted	Math	School	Europe	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Neuburger et al., 2012	72	-0.14	0.24	Not identified	Present	25.9	Blatant	Removal	Unadjusted	General	School	Europe	Published
Pennington et al., 2019	37	-0.64	0.34	Identified	Present	10.0	Blatant	Control	Unadjusted	Math	College	Europe	Published
Picho & Schmader, 2018 (1)	63	-0.01	0.25	Not identified	Present	83.0	Blatant	Removal	Unadjusted	Math	School	Africa	Published
Picho & Schmader, 2018 (2)	62	-0.30	0.26	Not identified	Present	83.0	Blatant	Removal	Unadjusted	Math	School	Africa	Published
Picho & Stephens, 2012 (1)	38	-0.74	0.34	Not identified	Present	63.5	Blatant	Control	Unadjusted	Math	School	Africa	Published
Picho & Stephens, 2012 (2)	51	-0.14	0.28	Not identified	Present	63.5	Blatant	Control	Unadjusted	Math	School	Africa	Published
Rivardo et al., 2011	45	0.04	0.30	Not identified	Present	68.3	Blatant	Control	Unadjusted	Math	College	North America	Published
Rosenthal & Crisp, 2006 (1)	24	0.35	0.41	Not identified	Reduced	52.0	Subtle	Control	Unadjusted	Math	College	Europe	Published
Rosenthal & Crisp, 2006 (2)	31	0.47	0.36	Not identified	Present	47.9	Blatant	Control	Unadjusted	Math	College	Europe	Published
Rucks, 2008 (1)	34	0.66	0.35	Identified	Reduced	44.4	Subtle	Control	Unadjusted	Math	College	North America	Unpublished
Rucks, 2008 (2)	34	-0.27	0.34	Identified	Present	44.4	Subtle	Control	Unadjusted	Math	College	North America	Unpublished
Rydell et al., 2009 (1)	56	-1.36	0.30	Not identified	Present	56.5	Blatant	Control	Unadjusted	Math	College	North America	Published
Rydell et al., 2009 (2)	56	-1.28	0.29	Not identified	Present	60.8	Blatant	Control	Unadjusted	Math	College	North America	Published
Seitchik et al., 2014 (1)	36	-0.54	0.34	Not identified	Reduced	75.6	Subtle	Control	Unadjusted	Math	College	North America	Published
Seitchik et al., 2014 (2)	36	0.10	0.33	Not identified	Reduced	73.0	Subtle	Control	Unadjusted	Math	College	North America	Published
Seitchik et al., 2014 (3)	36	0.23	0.33	Not identified	Present	73.0	Blatant	Control	Unadjusted	Math	College	North America	Published
Seitchik et al., 2014 (4)	36	-0.82	0.35	Not identified	Present	75.6	Blatant	Control	Unadjusted	Math	College	North America	Published
Seitchik, 2014 (1)	43	-1.34	0.34	Not identified	Present	38.0	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Seitchik, 2014 (2)	43	1.82	0.36	Not identified	Present	47.3	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Seitchik, 2014 (3)	30	0.92	0.38	Not identified	Present	47.9	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Seitchik, 2014 (4)	73	-0.88	0.25	Not identified	Present	49.7	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Seitchik, 2014 (5)	73	0.79	0.24	Not identified	Present	54.0	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Simmons, 2010	96	-0.50	0.21	Identified	Present	75.6	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Smetacková & Rubin, 2015	138	-0.15	0.17	Not identified	Present	47.0	Blatant	Control	Unadjusted	Math	School	Europe	Published
Sorenson, 2022	42	-0.18	0.31	Identified	Present	35.6	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Spencer, 2005	40	-0.12	0.32	Not identified	Present	22.3	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Steinberg et al., 2012 (1)	11	-0.64	0.63	Identified	Present	66.0	Blatant	Removal	Unadjusted	Math	College	North America	Published
Steinberg et al., 2012 (2)	47	-0.36	0.30	Identified	Present	66.0	Blatant	Removal	Unadjusted	Math	College	North America	Published
Steinberg et al., 2012 (3)	11	-0.55	0.62	Identified	Present	66.0	Blatant	Removal	Unadjusted	Math	College	North America	Published
Steinberg et al., 2012 (4)	11	-0.43	0.62	Identified	Present	66.0	Blatant	Control	Unadjusted	Math	College	North America	Published
Steinberg et al., 2012 (5)	47	-0.76	0.30	Identified	Present	66.0	Blatant	Control	Unadjusted	Math	College	North America	Published
Steinberg et al., 2012 (6)	11	-1.78	0.72	Identified	Present	66.0	Blatant	Control	Unadjusted	Math	College	North America	Published
Sunny et al., 2017 (1)	37	0.01	0.33	Identified	Present	64.1	Blatant	Control	Unadjusted	Science	College	North America	Published
Sunny et al., 2017 (2)	37	0.37	0.33	Identified	Present	64.1	Blatant	Removal	Unadjusted	Science	College	North America	Published
Tagler, 2003	131	-0.21	0.18	Not identified	Present	72.6	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Taillandier-Schmitt et al., 2012	40	-0.41	0.33	Not identified	Present	36.7	Blatant	Control	Unadjusted	Math	College	Europe	Published
Wang et al., 2022 (1)	54	0.60	0.28	Identified	Present	69.9	Subtle	Control	Unadjusted	Math	College	Asia	Published
Wang et al., 2022 (2)	68	0.55	0.25	Identified	Present	69.9	Subtle	Control	Unadjusted	Math	College	Asia	Published
Wicherts et al., 2005 (1)	94	-0.34	0.21	Not identified	Present	72.6	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (10)	58	-0.43	0.27	Not identified	Reduced	46.7	Blatant	Removal	Unadjusted	Math	College	North America	Published
Wicherts et al., 2005 (11)	58	0.18	0.26	Not identified	Reduced	36.2	Blatant	Removal	Unadjusted	Math	College	North America	Published
Wicherts et al., 2005 (2)	95	-0.05	0.21	Not identified	Present	72.6	Blatant	Control	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (3)	94	-0.59	0.21	Not identified	Present	63.2	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (4)	95	-0.75	0.21	Not identified	Present	63.2	Blatant	Control	Unadjusted	Math	College	Europe	Published

Study	*N*	*d*	*SE*	Domain ID	Self-Threat	Task Difficulty	ST Activation	Comparison	Ability Control	Domain	Sample	Region	Publication
Wicherts et al., 2005 (5)	94	-0.38	0.21	Not identified	Present	71.5	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (6)	95	-0.24	0.21	Not identified	Present	71.5	Blatant	Control	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (7)	94	0.12	0.21	Not identified	Present	81.0	Blatant	Removal	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (8)	95	0.05	0.21	Not identified	Present	81.0	Blatant	Control	Unadjusted	Math	College	Europe	Published
Wicherts et al., 2005 (9)	58	0.45	0.27	Not identified	Reduced	68.2	Blatant	Removal	Unadjusted	Math	College	North America	Published
Wille et al., 2018	163	-0.09	0.16	Not identified	Present	47.2	Subtle	Control	Unadjusted	Math	School	Europe	Published
Wilson, 2009 (1)	30	-0.56	0.37	Not identified	Present	28.8	Blatant	Removal	Unadjusted	Math	College	North America	Unpublished
Wilson, 2009 (2)	27	0.01	0.39	Not identified	Present	28.8	Blatant	Control	Unadjusted	Math	College	North America	Unpublished
Wister et al., 2013	38	-0.19	0.33	Not identified	Present	46.1	Subtle	Control	Unadjusted	Math	College	North America	Published
Wu & Zhao, 2021	112	-0.62	0.19	Not identified	Present	41.1	Blatant	Control	Unadjusted	Math	College	Asia	Published
Yagan & Avci, 2023 (1)	81	-0.13	0.22	Not identified	Present	67.9	Blatant	Control	Unadjusted	Math	School	Europe	Published
Yagan & Avci, 2023 (2)	80	-0.34	0.23	Not identified	Present	67.9	Blatant	Control	Unadjusted	Math	School	Europe	Published

Note. Each row represents one experimental comparison included in the meta-analysis. *d* = standardized mean difference; *SE* = standard error of *d*. Domain ID = domain identification (Identified vs. Not identified); Self-Threat = presence or reduction of stereotype threatening framing; ST Activation = type of stereotype threat activation (Subtle = indirect cues such as gendered images, group composition, or implicit priming; Blatant = explicit statements about gender differences in test performance); Comparison = type of control condition (Control vs. Removal); Ability Control = whether outcome was ability-adjusted; Domain = STEM domain of the outcome; Sample = participant type (School vs. College vs. Mixed); Region = study location; Publication = publication status. Task Difficulty was coded as a percentage (low = easy, high = hard).

information, see methods section), and a thorough search for grey literature, which represented 27.27% (k = 21) of the included citations.

Most of the effect sizes (k = 139, 86.34%) represented stereotype threat effects on women in mathematics. Additional domains included general STEM (k = 11, 6.83%), Science (k = 9, 5.59%), and engineering (k = 2, 1.24%). The reports were conducted in laboratory (k = 49, 63.64%), field (k = 25, 32.47%), and online settings (k = 3, 3.90%).

Task difficulty ranged from 10.00% to 87.50% (M = 53.82%, SD = 16.93%). The first and third quartiles were at 40.30% and 66.00%, respectively. Most effect sizes were derived from participants not preselected or grouped by domain identification (k = 129, 80.12%). Self-threat was present in 85.09% (k = 137) of the conditions and reduced in 14.91% (k = 24).

Summary Effect

Research Question 1 addressed the overall effect of stereotype threat for women in STEM. The overall effect size was significant and small in magnitude, $d = –0.18$, $SE = 0.04$, $z = –4.96$, $p < .001$, 95% CI [–0.25, –0.11]. The model indicated significant heterogeneity on Level 1 ($Q(162) = 457.75$, $p < .001$).

The I^2 statistic indicated that 69.10% of the variance between effect sizes could not be attributed to sampling variance, suggesting a moderate to large amount of unexplained heterogeneity across effect sizes. The sampling variance (Level 1) was between low and moderate (I^2 = 30.9%), variance within studies was moderate to high ($\sigma_2{}^2 = 0.13$, I^2 = 69.1%), and variance between studies negligible ($\sigma_3{}^2 = 0.00$, I^2 < 0.01%). The substantial amount of heterogeneity is emphasized by a wide prediction interval (95% PI [–0.89, 0.53]).

A likelihood ratio test indicated that the full model provided a significantly better fit than the model with fixed within-study variance, χ^2 (2)= 58.48, p <.001, and showed a higher likelihood (log likelihood = –113.39 vs. –142.52).

The complete model did not significantly fit better than the model with fixed between-study variance, χ^2 (2) = 0.00, p = 0.50, and did not show a higher likelihood (log likelihood = –113.28 vs. –113.28). The three-level model primarily captured heterogeneity at the within-study level, and between-study variance was negligible when within-study variance was modelled.

Theoretical Conditions

The impact of theoretical conditions on stereotype threat for women in STEM (Research Question 2) was investigated using a mixed-effects multilevel model, including the three predictors domain identification (identified vs. not identified), self-threat (present vs. reduced) and task-difficulty.

The intercept reflected the estimated stereotype threat effect under the reference conditions of the model, i.e., for domain-identified participants under self-threat and at average task difficulty. According to stereotype threat theory, these conditions should produce the strongest performance decrements. The intercept was not significant, indicating no stereotype threat effect on women in STEM for domain-identified participants under self-threat and average difficulty. The overall test of moderators was not significant, $Q_M(3) = 0.77$, $p = .857$, and none of the individual predictors reached statistical significance (see Table 2). The test for residual heterogeneity was significant, $Q_E(159) = 1.69$, $p < .001$.

Table 2

Meta-Regression on Theoretical Conditions of Stereotype Threat

Predictor	Estimate	*SE*	*z*	*p*	95% CI
Intercept	–0.16	0.09	–1.82	.069	[–0.34, 0.01]
Domain Identification	–0.03	0.10	–0.32	.746	[–0.22, 0.16]
Self-Threat	0.07	0.10	0.70	.487	[–0.13, 0.27]
Task Difficulty	0.00	0.00	–0.41	.681	[–0.00, 0.00]

Note. Reference category for domain identification was "identified" and for self-threat "present". Task difficulty was mean-centered ($M = 52.82$, $SD = 16.93$).

The conditions formulated in stereotype threat theory, domain identification, self-threat, and task difficulty do not explain heterogeneity within the data. The summary effect from studies on women in STEM that implemented these conditions was not significant.

To explore a possible quadratic relationship between task difficulty and stereotype threat, a post-hoc meta-regression was conducted (see Table 3). The omnibus test of moderators was not significant, $Q_M(2) = 5.19$, $p = .075$, but the quadratic term reached significance ($b < –0.00$, SE 0.00, $z = –2.25$, $p = 0.024$, 95% CI [–0.00, 0.00).

Task difficulty reduced 1.93% in the total observed variance, compared to the null model.

Sensitivity Analyses

STEM Domain. A mixed-effects multilevel model tested whether stereotype threat effects on women varied across STEM subjects (Research Question 3), with the outcome's STEM

Table 3

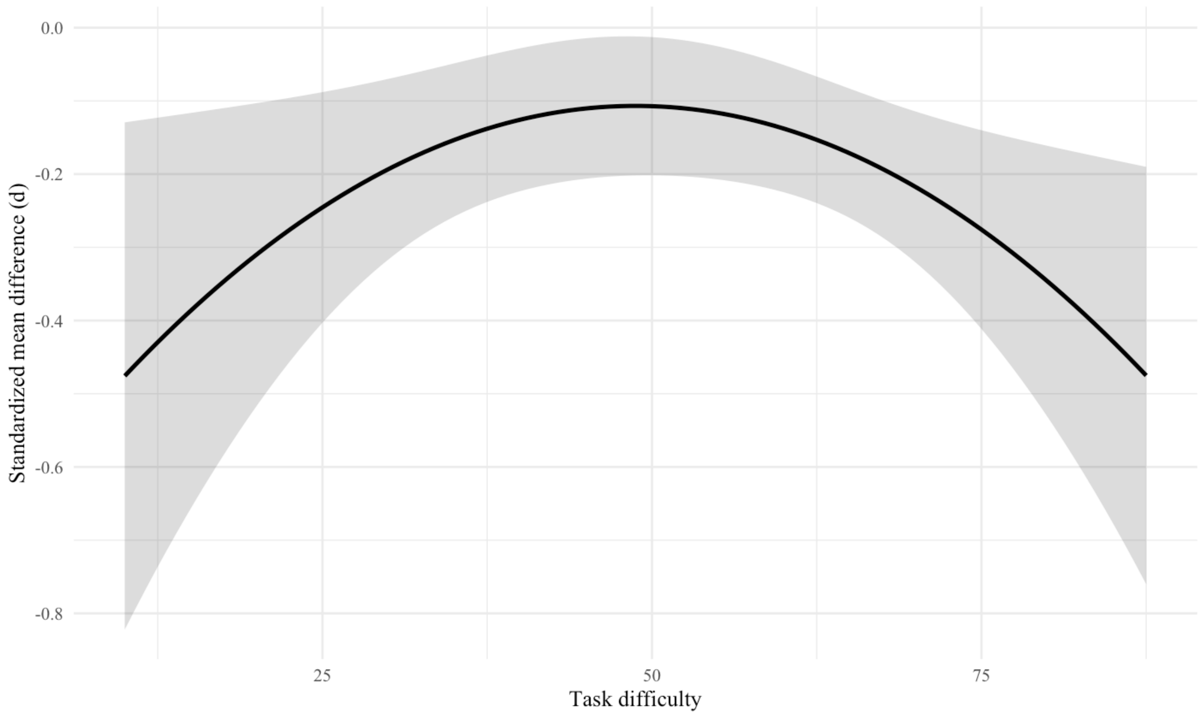
Meta-Regression on Task Difficulty

Predictor	Estimate	*SE*	*z*	*p*	95% CI
Intercept	–0.11	0.05	–2.35	0.019	[–0.20, –0.019]
Task Difficulty	–0.002	0.002	–0.89	0.372	[–0.006, 0.002]
Task Difficulty (sq)	–0.0002	0.0001	–2.25	0.024	[–0.0005, 0.0000]

Note. Task difficulty was mean-centered (M = 52.82, SD = 16.93). The model includes both a linear and a quadratic term for Task Difficulty to examine potential non-linear effects.

Figure 2

Stereotype Threat as a Quadratic Function of Task Difficulty



Note. Predicted effect sizes (*d*) as a function of task difficulty. The shaded ribbon denotes the 95 % confidence intervals.

domain included as a categorical predictor. The reference category was mathematics, as it represented the largest subgroup of studies. The intercept was significant, b = –0.18, SE = 0.04, z = –4.61, p < 0.001, 95% CI[–0.26, 0.10]. The overall test of moderators was not

Table 4 Meta-regression on Outcome Domain

Meta-regression on Outcome Domain

Predictors	Estimate	*SE*	*z*	*p*	95% CI
Intercept (Math)	–0.18	0.04	–4.68	< .001	[–0.26, –0.11]
Science	0.04	0.16	0.25	.803	[–0.28, 0.36]
Engineering	0.36	0.29	1.22	.223	[–0.22, 0.93]
General	–0.05	0.14	–0.35	.730	[–0.32, 0.22]

Note. Math was used as the reference category for outcome domain.

significant $Q_M(3) = 1.69$, $p = .638$, and none of the contrasts reached statistical significance (see Table 4). Test for residual heterogeneity was significant, $Q_E(159) = 451.77$ $p < .001$.

Publication Status. A meta-regression and subgroup analyses assessed whether stereotype threat effects differ based on publication status (Research Question 4). Publication status did not significantly explain heterogeneity, $b = 0.13$, $SE = 0.08$, $z = 1.60$, $p = .110$, 95% CI [–0.03, 0.29].

The meta-analytic summary effect of the published subset ($k = 116$) was significant, $d = –0.21$, $SE = 0.04$, $z = –4.95$, $p < .001$, 95% CI [–0.31, –0.13], with moderate heterogeneity ($I^2 = 68.38\%$).

The effect for the unpublished subset ($k = 45$) was not significant, $d = –0.09$, $SE = 0.07$, $z = –1.26$, $p = .207$, 95% CI [–0.22, 0.05], $I^2 = 69.75\%$.

Publication Year. Research Question 5 addressed the stability of stereotype threat effects over time. Publication year was centered at the year 2000, the publication year of the earliest study included in the analysis. A mixed-effects multilevel model with publication year as a continuous predictor on the complete sample indicated significant changes in effect sizes over time (see Table 5). The test for residual heterogeneity was significant, $Q_E(116) = 440.52$, $p < .001$.

Table 5

Meta-Regression on Publication Year

Predictor	Estimate	*SE*	*z*	*p*	95% CI
Intercept	–0.34	0.09	–3.95	< .001	[–0.52, –0.17]
Publication Year	0.01	0.01	2.07	.038	[0.00, 0.03]

Note. Publication year was centered at the year of the first study included in the analysis (2000).

A second analysis was performed on published data only and explored the relation between publication year and effect sizes. The estimated summary effect at the year 2000 was significant, $b = –0.434$, $p < 0.001$, and publication year significantly predicted effect sizes in published studies, $b = 0.018$, $p = 0.019$ (see Table 6 for full model results; Figure 3 illustrates the trend visually).

Table 6

Meta-Regression on Publication Year (Published Studies Only)

Predictor	Estimate	*SE*	*z*	*p*	95% CI
Intercept	–0.43	0.10	–4.21	< .001	[–0.64, –0.23]
Publication Year	0.02	0.01	2.31	0.021	[0.00, 0.03]

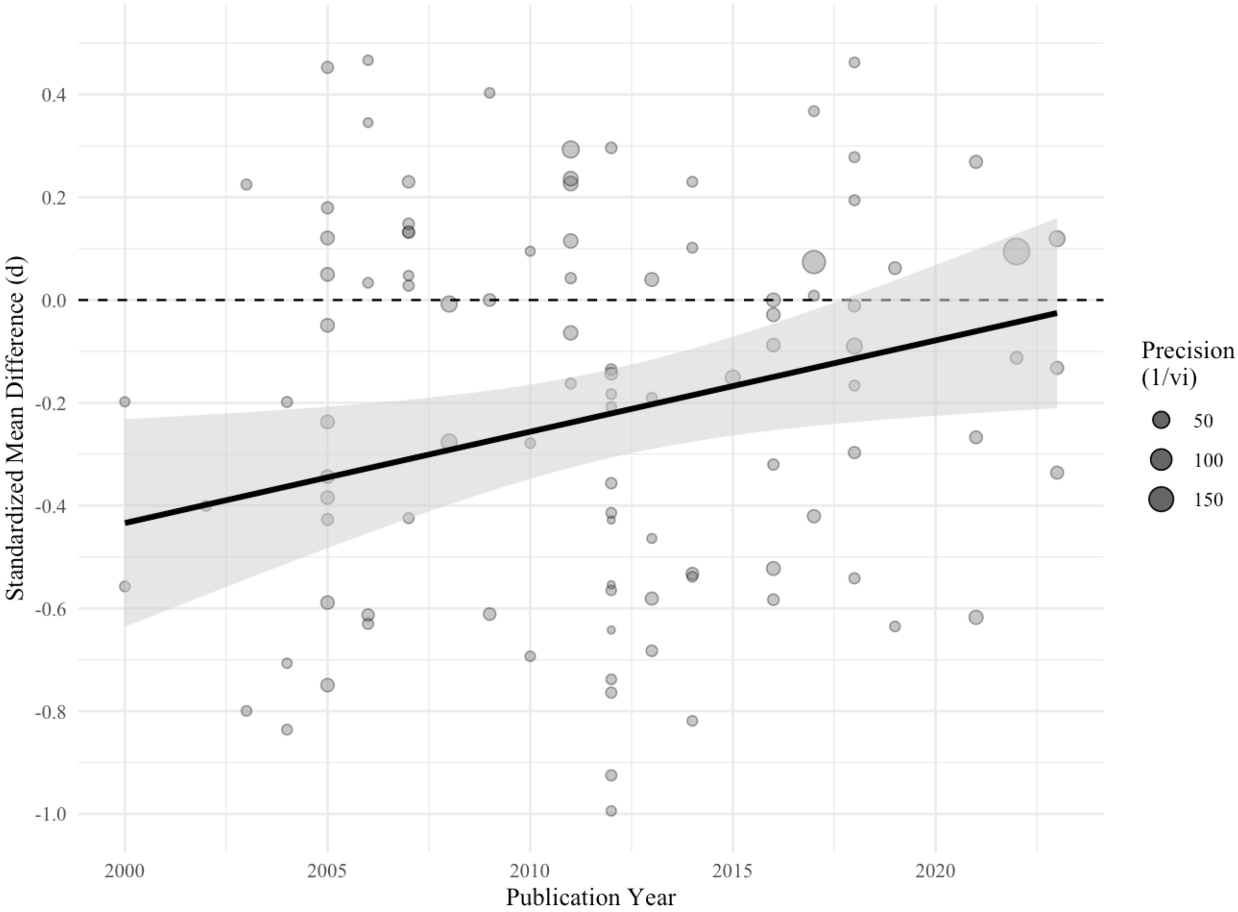
Note. This analysis includes only published studies. Publication year was centered at the year of the first included study (2000).

The publication year model was compared to the null model on published studies only (described above in the section on publication status).

The model including the predictor showed better fit than the baseline model, $\chi^2(1) = 5.28$, $p = .022$, with a higher log likelihood (–77.70 vs. –80.34). The test for residual heterogeneity was significant, $Q_E(116) = 330.22$, $p < .001$. Publication year reduced 5.80% in the total observed variance, compared to the null model.

Figure 3

Stereotype Threat as a Function of Publication Year



Note. Predicted standardized mean differences (*d*) are plotted as a function of publication year in published studies. The solid line reflects the fitted meta-regression model, with the shaded area representing the 95% confidence interval. Point size indicates precision (1/vi). The slope and confidence intervals suggest that stereotype threat effects became weaker and non-significant in more recent publications.

A third model was run on unpublished studies only. Neither the intercept (b = –0.12, SE = 0.20, z = –0.58, p = 0.561, 95% CI [–0.51, 0.28]) nor the predictor were significant (b = 0.00, SE = 0.02, z = 0.15, p = 0.879, 95% CI [–0.03, 0.03]).

Multiverse and Specification Curve Analysis

The multiverse addressed the generalizability of stereotype threat effects across a series of researchers' degrees of freedom, and results were subsequently organized and visualized through the specification curve (Research Question 6). For computational feasibility, a two-level model served as a basis for the analysis. The summary effect was estimated with d = –0.17 (SE = 0.03, z = –5.27, p < .001, 95% CI [–0.23, –0.11]). Heterogeneity was moderate to large ($Q(122)$ = 287.46, p < .001, I^2 = 59.32%).

Across the multiverse, from a combination of 31,104 possible meta-analytic specifications, a total of 1,632 meta-analyses (based on ≥ k = 2 samples) were calculated, yielding standardized mean differences (SMD) from –0.86 to = +0.39. In total, 53.8% (k = 878) of summary effects produced CIs that did not include zero. The specification curve revealed an overall pattern of instability depending on analytical decisions (see Figure 4). Specifications with small sample sizes (warmer colors) were more likely to produce extreme estimates.

Further visual inspection of the curve showed that the meta-analytic specifications exclusively containing unpublished-only studies were concentrated in the right half of the curve. These specifications yielded moderate negative to small positive effects (SMD = –0.29 to +0.12), and only 13.5% (k = 26) of the confidence intervals excluded the zero. In contrast, the published-only specifications spanned the entire curve, from large negative to positive estimates of moderate size (SMD = –0.86 to = +0.39). Half of all published specifications (50.14%, k = 319) yielded significant summary effects.

Another visual pattern arised based on study setting: Field studies were concentrated on the right side of the curve and 79.62% (k = 332) CIs including zero. On the other hand, laboratory-only specifications were clustered on the left of the curve, and most of these specifications (90.00%, k = 378) yielded significant summary effects.

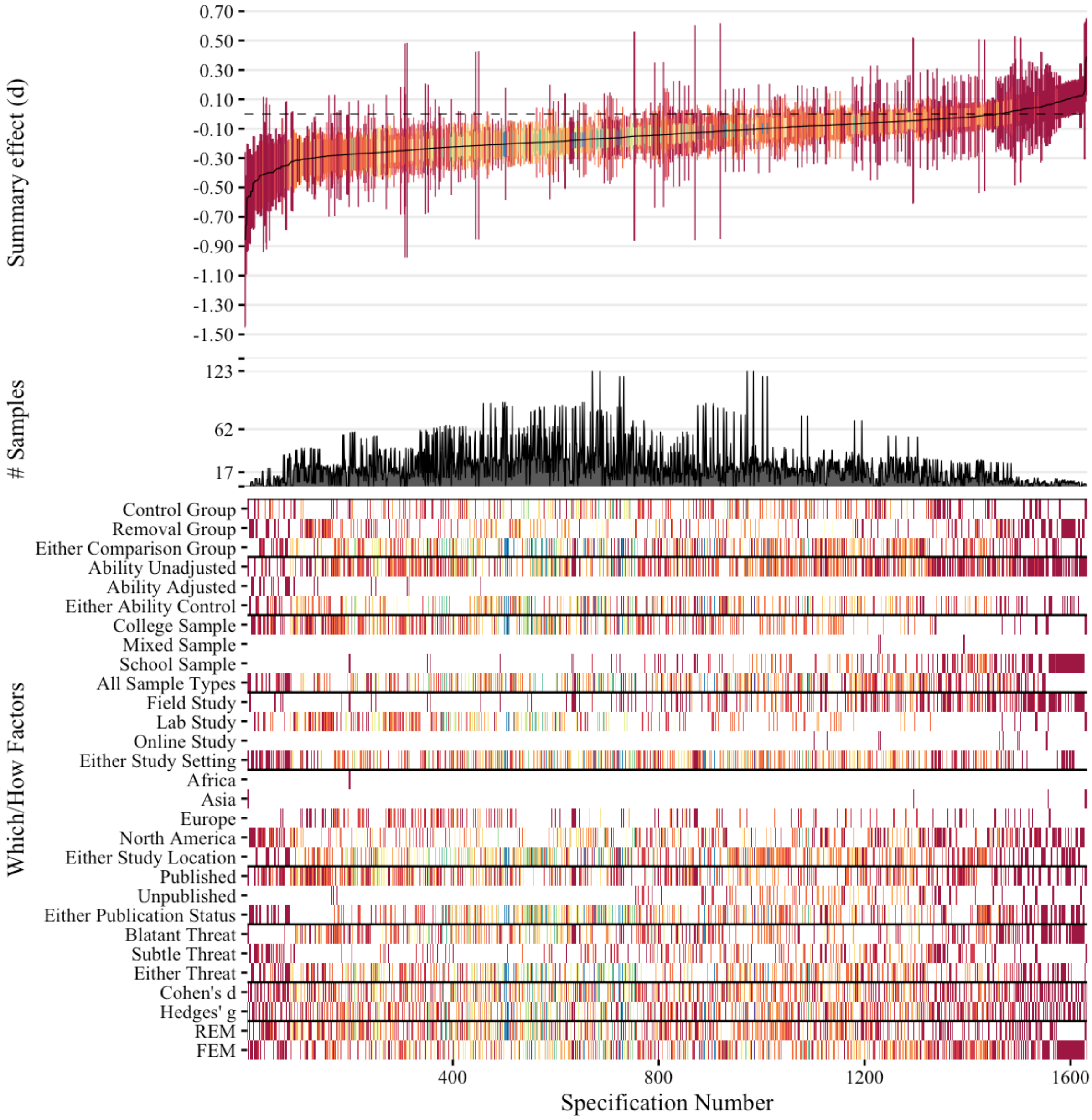
Taken together, the multiverse and specification curve analysis indicated that stereotype threat effects on women in STEM varied considerably across analytic specifications.

Outlier diagnostics and Combinatorial Meta-Analysis (GOSH)

Outlier diagnostics using the influence-function in metafor (Viechtbauer, 2010) revealed both Rydell et al. (2009) studies as outliers (Fig. 5). To quantify the influence, a meta-analysis was calculated after removing the two outliers. The summary effect slightly decreased from d = –0.17 (see above for whole model results) to d = –0.15 (SE = 0.03, z = –4.96, p <.001, CI [–0.21, –0.09]). Heterogeneity decreased by nearly 6% (I^2 = 59.32% vs. 53.14%).

Figure 4

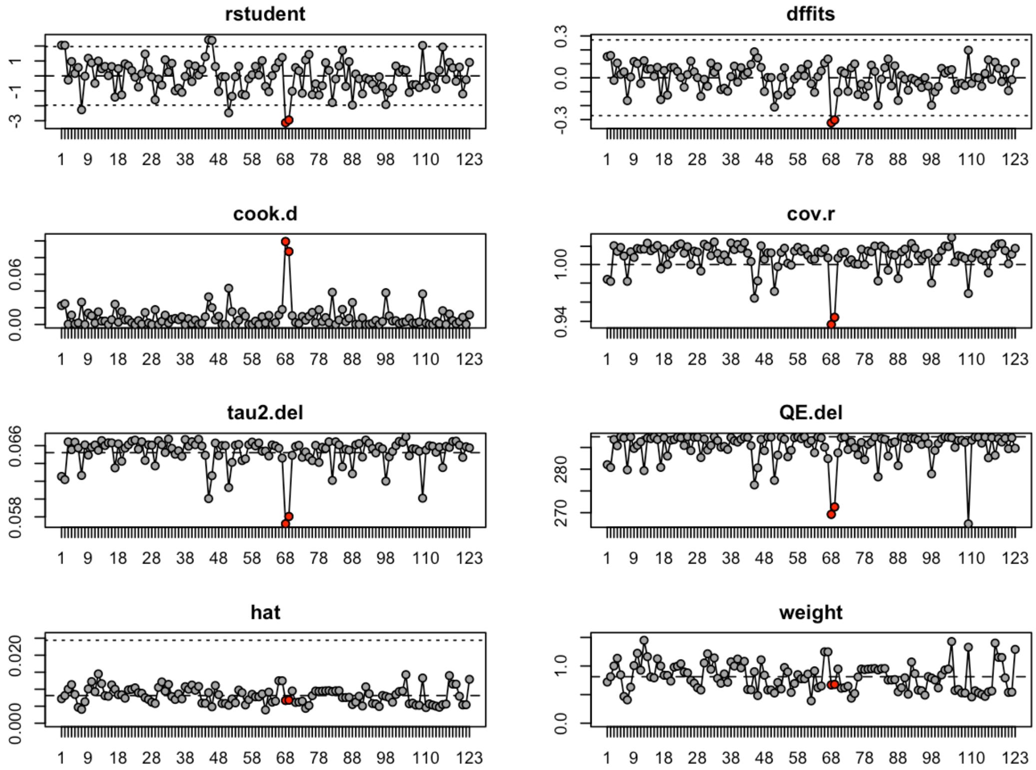
Specification Curve of Stereotype Threat Effects



Note. Specification-curve analysis of stereotype threat for women in STEM. Top panel: Summary effects (Cohen's *d*) with 95% confidence intervals for the 1,632 retained specifications ($k \geq 2$), sorted from smallest to largest and connected by the solid black curve. A dashed horizontal line at $d = 0$ denotes no effect. Middle panel: Number of aggregated samples (k) per specification, with ticks at $k = 17$ (mean), $k = 62$ (midpoint) and $k = 123$ (maximum). Bottom panel: Combination of Which-factors (e.g. comparison group, ability control, setting) and How-factors (effect metric, meta-analytic model) defining each specification. Columns are color-coded on a spectral scale (warm hues = fewer samples; cool hues = more samples). FEM = fixed-effect model; REM = random-effects model.

Figure 5

Influence Diagnostics



Note. Influence diagnostics identified Rydell et al.'s (2009) two studies as outliers.

Combinatorial meta-analysis was calculated by meta-analyzing 100,000 random subsets using the gosh-function in metafor (Viechtbauer, 2010). The GOSH provided a univariate distribution with no indication of distinct clusters (Fig. 6).

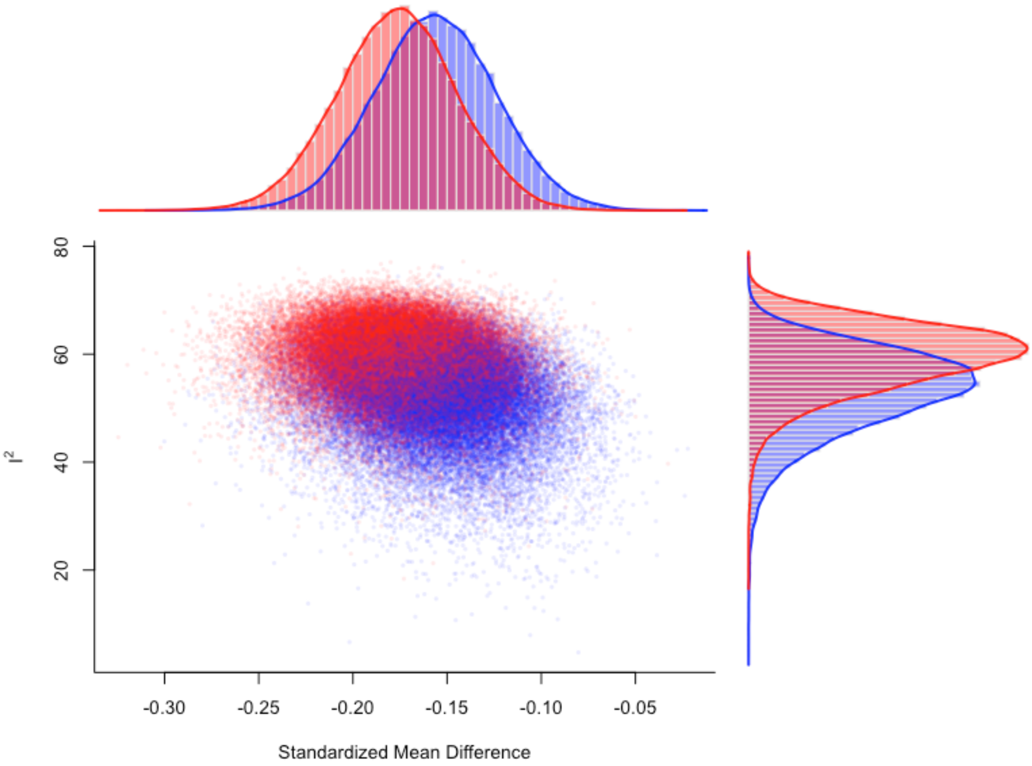
The large spread of summary effects and the I^2-axis emphasized the variability across study subsets and heterogeneity in meta-analytic results, dependent on the included studies. The more extreme Rydell et al. (2009) study ($d = -1.36$ vs. $d = -1.28$) was highlighted in red.

Dissemination Bias

Egger Regression. Egger's regression detected a correlation between effect sizes and standard errors, $b = -1.95$, $SE = 0.47$, $t = -4.11$, $df = 114$, $p < .001$, 95% *CI* [–2.88, –1.01], indicating that when the standard error increased by one unit, the effect size decreased by 1.95 units.

Figure 6

Graphical Display of Study Heterogeneity (GOSH)



Note. GOSH shows the two-level random-effects meta-analytic summary effects (*d*) and the between-study variance (I^2) for a random sample of 100,000 subsets. The distributions of summary effects and I^2 values are depicted on the top and to the right, respectively. Subsets including Rydell et al.'s (2009) first study are highlighted in red.

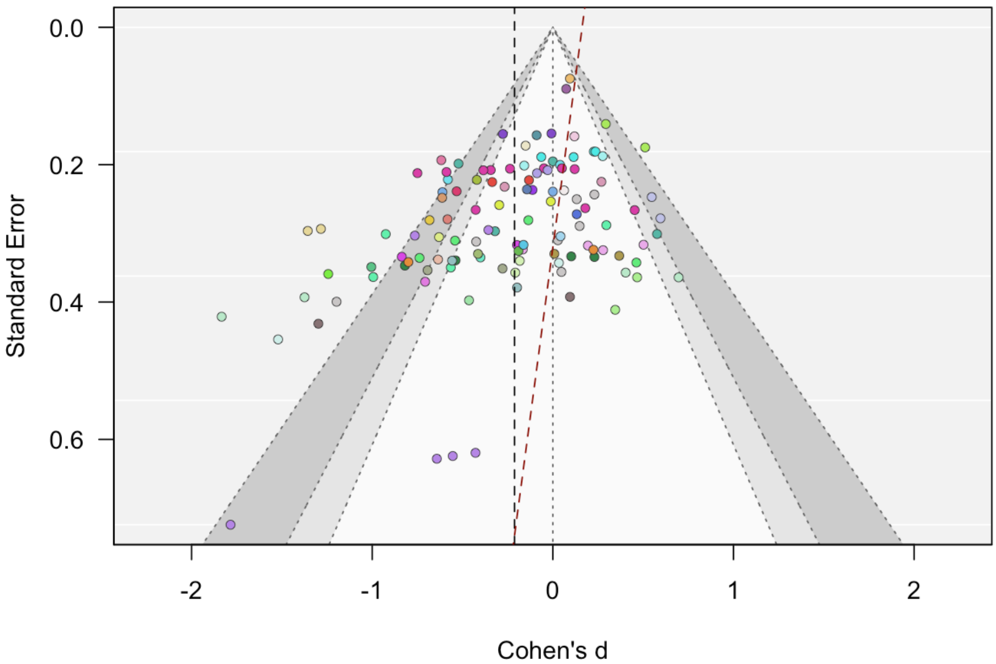
The intercept was also statistically significant, $b = +0.31$, $SE = 0.13$, $t = 2.30$, $p = 0.024$, 95% CI [0.04, 0.58]. This indicates an estimate of $d = +0.31$ with the standard error hypothetically fixed to zero. This stands in contrast to the overall negative meta-analytic effect ($d = –0.18$). In the context of stereotype threat theory, a positive effect speaks for a reversed stereotype threat (i.e., stereotype lift).

Funnel Plot. A funnel plot was created to visually inspect the distribution of published stereotype threat effects on women in STEM (Fig. 7). The basis for the plot was the three-level model on the published subset (reported above in the section on publication status).

Visual inspection of the contour-enhanced confetti funnel plot indicated asymmetry. Studies with high precision ($SE = 0.0–0.2$) appear symmetrically distributed around the zero-effect line and are missing in the predicted negative direction. At moderate precision ($SE = 0.2–0.4$), statistically significant effects cluster on the left side of the plot (consistent

Figure 7

Contour-Enhanced Confetti Funnel Plot



Note. Funnel plot containing published studies. Effect sizes deriving from a single study are color-coded. The white, light-grey, and dark-grey contours indicate significance (p =.10, .05, and .01, respectively). The plot indicates absence of significant studies on the contrary to the predicted effect (right side of the plot). The meta-analytic summary effect is indicated by the black dashed line ($d = –0.21$). The Egger regression is presented as the red dashed line ($b = –1.95$, $SE = 0.47$, $p < .001$).

with a stereotype threat effect), while no significant studies in the opposite side appeared in this region. One study with low precision appeared in the predicted (negative) direction; none existed in the unexpected direction. Funnel plot asymmetry was statistically confirmed by the red dashed Egger regression line.

Egger Sandwich Test. Egger's test with sandwich estimators indicated a correlation between standard errors and effect sizes ($b = –2.11$, $SE = 0.40$, $t = –5.3$, df = 15.8, $p < 0.001$, 95% CI [–2.95, –1.26]).

The intercept was significant ($b = 0.30$, $SE = 0.10$, $t = 3.09$, $df = 12.0$, $p = 0.009$, 95% CI [0.09, 0.2]), indicating that, in hypothetically perfectly precise studies, the estimated effect would be small and positive.

3PSM. A three-parameter selection model was run on published studies only. The chance of observing a non-significant effect size versus a significant one was $\psi = 0.64$. The test was not significant ($\chi^2 = 0.40$, $p < .526$).

Discussion

Stereotype threat has long been considered a well-established psychological mechanism contributing to the underrepresentation of women in STEM (Eronen & Bringmann, 2021). Over the past three decades, an extensive body of research has examined this phenomenon across diverse groups and contexts. However, the theoretical clarity and empirical consistency of stereotype threat effects have been called into question. The present meta-analysis aimed to systematically evaluate the magnitude and variability of stereotype threat effects on women in STEM and their relations to the theoretical conditions.

The finding that stereotype threat has a small negative effect on women's performance in STEM aligns with previous meta-analyses on stereotype threat for female participants in STEM (Flore & Wicherts, 2015; Picho et al., 2013; Picho-Kiroga et al., 2021). Although small in size, the effect is not trivial: for example, a difference of this size roughly equates to one year of learning in 15-year-old students across OECD countries (OECD, 2023). At the same time, the present evidence reveals that this summary effect is not a robust or trustworthy indicator of the general magnitude of stereotype threat effects. This underscores the importance of understanding the conditions under which these effects emerge or not.

Heterogeneity of Stereotype Threat Effects

The summary estimate of stereotype threat for women in STEM was accompanied by a wide confidence interval, indicating an imprecise overall estimate with considerable heterogeneity. An inspection of the variance revealed that more than two thirds of the heterogeneity could not be attributed to sampling error. The wide prediction interval indicates that future studies could produce effect sizes ranging from large negative to moderate positive effects, highlighting the instability of the effects. Overall, the sources of the observed heterogeneity remain largely unexplained.

The three-level model showed that the heterogeneity almost exclusively stemmed from within-study variance, whereas no meaningful variance was found between studies. This pattern suggests that, compared to effect sizes reported within the same studies, effect sizes stemming from different studies were rather similar. Different effect sizes within studies reflected, for example, subtle vs. blatant threat conditions (e.g., Marchand & Taasoobshirazi, 2013); control vs. removal groups (e.g., Campbell & Collaer, 2009); samples from same-sex vs. mixed-sex schools (e.g., Picho & Stephens, 2012), samples of different age groups (e.g., Ganley et al., 2013), domain-identified vs. not-identified participants (e.g., Kahalon et al., 2018); or participants under self-threat vs. reduced self-threat (e.g., Jamieson & Harkins, 2010).

The pattern in which almost all observed variance in a multi-level meta-analysis is located within rather than between studies is an extreme scenario (Harrer et al., 2021). This reflects the complexity and heterogeneity of stereotype threat research on women in STEM. Together, these findings suggest that the summary effect of $d = -0.18$ is not a trustworthy estimate for stereotype threat (Harrer et al., 2021). Subsequently, I discuss the results of various meta-regressions and robustness analyses to address the question of where the heterogeneity stems from, and which factors may explain it.

Theoretical Conditions. To answer research question 2, how much of the heterogeneity in stereotype threat literature could be explained by the presence or absence of C. M. Steele's (1997) theoretical conditions, a meta-regression was fitted with the predictors domain identification, task difficulty, and self-threat. The theoretical predictors self-threat, domain identification and task difficulty did not explain variation of effect sizes. More importantly, no significant stereotype threat effect for studies on domain-identified participants under self-threat doing tasks of moderate difficulty was found. As these are conditions central to the theoretical definition of stereotype threat, the absence of a significant moderation as well as an effect in this subset of studies raises doubts about the robustness of stereotype threat effects in the field of women in STEM and aligns with the possibility of a true effect of zero (Warne, 2022; Zigerell, 2017).

An exploratory analysis revealed a quadratic link between task difficulty and stereotype threat, contrasting with C. M. Steele's (1997) reasoning that the threat should be most pronounced at moderate difficulty. According to the theory, tasks should be diagnostic of ability, hence not too difficult or easy (C. M. Steele, 1997). However, stereotype threat effects appeared strongest at the extremes of task difficulty. This finding challenges the notion that observed stereotype threat effects stem from frustration driven by the stereotype as an explanation for underperformance (C. M. Steele, 1997). Instead, performance reductions on very difficult and very easy tasks could reflect disengagement and withdrawn effort, rather than interrupted cognitive functions due to anxiety or fear. However, these considerations remain speculative and are based on exploratory findings. Moreover, the model accounted for only 1.9% of heterogeneity, indicating that task difficulty explains only a marginal amount of variance.

Picho-Kiroga et al. (2021) also examined the role of theoretical conditions on stereotype threat effects in women, focusing on the variables math identification, stereotype awareness, and stereotype salience. Thus, only domain identification overlapped with the current analysis. Their approach involved creating a composite variable that indicated the number of theoretical conditions present in each experiment (ranging from zero to three), and they found that only 6.9% met all theoretical conditions. Despite differences in the specific theoretical conditions investigated, both analyses had a similar aim, and the results were consistent: the absence of the theoretical conditions inflated stereotype threat effects (Picho-Kiroga et al., 2021).

In a reassessment of the meta-analysis by Picho-Kiroga et al. (2021), Warne (2022) identified methodological shortcomings and concluded that "stereotype threat in females is likely not a real phenomenon" (p. 181). The findings from the present analysis provide support for this conclusion: The theoretical conditions do not account for a meaningful proportion of heterogeneity in the data, and stereotype threat, when operationalized according to its core theoretical assumptions, does not produce underperformance in women in STEM.

Robustness Analyses. Several analyses were performed to examine the robustness of stereotype threat effects on women in STEM. The majority of the $k = 163$ analyzed effect sizes reflected stereotype threat effects in math, and only $k = 22$ covered other subdomains. No differences were found between stereotype threat effects based on the STEM domain. Due to the limited data in science, technology, and engineering, systemic differences may exist, but were not identified due to an underpowered meta-regression.

Heterogeneity could not be significantly explained by publication status. However, subgroup meta-analyses suggested a notable difference: For published studies, the summary effect was at small and significant ($d = –0.21$; 95% CI [–0.31, –0.13]); for unpublished studies, the null hypothesis was not rejected ($d = –0.09$, 95% CI [–0.22, 0.05]). Although the meta-regression was not statistically significant, it conveys a piece of information that contributes to the broader information of the evidence base. Stereotype threat effects are only apparent in published studies, and grey literature does not provide support for the phenomenon's existence in the context of women in STEM. High heterogeneity in both subgroups and the resulting wide confidence intervals likely limited the statistical power to detect a significant difference between the studies.

Year of publication significantly predicted stereotype threat effects on women in STEM. In the year 2000 (the first year covered in this meta-analysis), the estimated effect was $d = –0.34$. The standardized mean difference decreases by $d = 0.014$ per year, resulting in an estimated effect size of $d = +0.01$ by 2025. This could indicate that stereotype threat is losing relevance in the context of women in STEM domains. For instance, UNESCO reported that the performance gap in mathematics is narrowing across the globe (UNESCO, 2022).

However, exploratory subgroup analyses revealed that the published studies primarily drove this trend. In published studies, the summary effect was estimated with $d = –0.44$ at the year 2000 and decreased by $b = 0.018$ per year, resulting in an estimated summary effect of $d = +0.02$ by 2025. In contrast, no temporal trend was observed in unpublished studies. The analysis of grey literature included 45 effect sizes from 21 studies between 2001 and 2022, suggesting that the absence of an effect is unlikely due to limited data. If stereotype threat effects had changed due to cultural or societal shifts, a similar pattern would be expected in unpublished studies as well.

Instead, the stereotype threat effects were most pronounced in early published literature, while more recent publications and unpublished studies more often reported null results. This suggests that the decline of effect sizes over time reflects a dissemination bias, rather than a genuine temporal trend in the phenomenon itself. This pattern is consistent with the notion of a false-positive decline effect, in which early significant findings gradually regress to a summary effect considerably weaker than initially observed effect sizes (Pietschnig et al., 2019). Because the summary effect regresses to zero, the results are more consistent with a false-positive decline effect than with an overestimation of effect sizes (inflated decline effect; Pietschnig et al., 2019). If this interpretation holds, future meta-analytic summary effects of stereotype threat for women in STEM will continue to approach zero.

The lacking robustness of stereotype threat effects was also observed based on the multiverse investigating a total of 1,632 meta-analytic subsets. Specifically, 53.8% of the specifications yielded a significant stereotype threat effect on women in STEM, while nearly half did not. Visual inspection of the specification curve revealed a high variability of summary effects across different specifications and indicating a sensitivity to researcher degrees of freedom. Specifically, the pattern suggested that, in addition to publication status, study setting (lab vs. field) and sample type (college vs. school) may be factors feeding into the heterogeneity. However, the spread of the estimates within each category was substantial, underscoring the high degree of heterogeneity across all analytical decisions.

In sum, multiverse and specification curve analysis illustrated a general instability of the findings. The GOSH of a random subset of 100,000 meta-analyses yielded a unimodal distribution of effect sizes and I^2 values. The visual pattern, therefore, suggested an absence of meaningful subgroups or clustering structures in the data, and thus, does not point to unaccounted sources of heterogeneity. Taken together, the specification curve analysis and GOSH clearly emphasize that the observed stereotype threat effects on women in STEM are highly sensitive to analytical decisions and do not converge on a robust underlying effect. Consequently, stereotype threat effects cannot be assumed to generalize to the population of females in STEM.

Dissemination-bias analyses further underscored the instability of stereotype threat for women in STEM. The results indicated that smaller, less precise studies disproportionately contributed to the negative summary effect, whereas larger and more precise studies tended to yield effect sizes close to zero or even small positive. This pattern indicates that the negative overall estimate is, at least in part, a consequence of small-study effects. The 3PSM did not detect a statistically significant suppression of non-significant results. Given the high within-study heterogeneity observed in this meta-analysis, this lack of detection is not surprising; many reported effect sizes were insignificant.

High heterogeneity has long been recognized in stereotype threat research. Although its sources often remained unidentified, the phenomenon was nonetheless regarded as stable and well-established (Flore & Wicherts, 2015; Nguyen & Ryan, 2008). It was commonly assumed that stereotype threat was a context-dependent phenomenon with many different moderators and mediators, and that overlooked but theoretically relevant factors could be able to explain the a substantial amount of the observed heterogeneity (Flore & Wicherts, 2015; Nguyen & Ryan, 2008; Pennington et al., 2016).

As dissemination bias analyses were applied to the body of research, concerns were raised about the magnitude of the effect (Flore & Wicherts, 2015; Shewach et al., 2019; Zigerell, 2017). In addition, recent meta-analyses emphasized the role of methodological and analytical decisions, as well as the study design to the overall effect (Picho-Kiroga et al., 2021a; Shewach et al., 2019). The present findings suggest that stereotype threat effects on women in STEM may be even less robust and generalizable than previously assumed. Specifically, the multiverse and specification curve analysis reveals a lack of stability in the body of research. While stereotype threat effects seem to be influenced by some methodological decisions, the effects are highly volatile across the board.

Furthermore, the analyzed theoretically grounded predictors do not relate to the effects as postulated by the originators of stereotype threat theory (J. Aronson et al., 1998; Spencer et al., 1999; C. M. Steele, 1997; C. M. Steele & Aronson, 1995; J. Steele et al., 2002); in direct contrast to what C. M. Steele (1998) theorized, stereotype threat effects were least pronounced at moderate task difficult. These findings add to the evidence that the observed effects are not derived from studies designed in consistence with the theory and therefore cannot be seen as confirmations of the construct (Picho-Kiroga et al., 2021; Warne, 2021). Taken together, these findings point to deeper issues in the field, raising questions about the conceptual clarity and theoretical coherence of stereotype threat research.

Conceptual and Methodological Challenges

The variability between effect sizes could not be explained by the included theoretical variables. The process of coding and reading the primary studies revealed deep conceptual issues that provide context for the observed heterogeneity in the data. From its early formulations, stereotype threat theory lacked conceptual precision. C. M. Steele (1997) presented the phenomenon in abstract and metaphorical terms, and every-day situations. Key psychological processes and mechanisms were ambiguously used in different contexts, blurring the meanings. This imprecision may partly explain why primary studies have often operationalized stereotype threat in highly diverse ways and without consistently adhering to the theoretical conditions; It likely contributes to the high observed heterogeneity in effects that could not be attributed to sampling variance.

Although C. M. Steele's (1997) paper is presented as a "general theory of domain identification" (p. 613), the paper lacked a concise definition and operationalization of the concept. While domain identification is a central theoretical condition, only 19.9% of studies in this meta-analysis met the criterion. Here, domain identification was regarded as present if participants were above the mean on an identification scale, or they were enrolled in STEM-degrees. However, operationalizations vary widely and include validated scales (e.g., Picho & Schmader, 2018), self-made questionnaires (e.g., Wille et al., 2018), previous academic performance (e.g., Ganley, 2011), enrollment in certain degree programs, such as psychology (e.g., Macchione & Sacco, 2023) or science (e.g., Kahalon et al., 2018), or a combination of the above (e.g., Mingle, 2014). Language from domain identification measures is derived from other constructs, such as interest in a domain, the importance of a domain, and academic self-concept (Ruff, 2013). This variability reflects that there is no field-wide consensus on how to define and operationalize domain identification, with most studies disregarding this condition entirely.

This lack of a coherent conceptualization helps explain divergent meta-analytic findings on the role of domain identification. The present analysis did not find that a relationship between domain identification and stereotype threat for women in STEM, a finding that was also reported by Shewach et al. (2019). Other researchers found that, contrary to stereotype threat theory, less identified women experienced higher performance decrements in stereotype threat studies (Nguyen & Ryan, 2008; Picho-Kiroga et al., 2021).

Similar conceptual inconsistencies appeared regarding task difficulty. Picho-Kiroga et al. (2021) stated there was no objective way to measure task difficulty and did not investigate its influence on stereotype threat effects. The present analysis operationalized difficulty with the percentage of items that could not be solved and found that most studies employed moderately difficult tasks. However, the contradictions in the foundational papers complicate the internal logic of stereotype threat research. C. M. Steele (1997) claimed that stereotype threat should only emerge under moderate difficulty, where performance is matched to a person's ability and relevant to the self-concept. Yet, the original experiments used difficult Graduate Record Examination (GRE) items and described the task as "very difficult" to the participants (C. M. Steele & Aronson, 1995, p. 799). In the early paper, no U-pattern of task difficulty was proposed or tested, contradicting the later posed condition. Such contradictions are more than interpretive; they reflect a lack of conceptual precision that undermines the internal coherence of stereotype threat theory. The present meta-analysis reflects these conceptual inconsistencies as it did not find a relation between task difficulty and stereotype threat effects in the expected direction. An exploratory analysis even suggested that effects were strongest at the extremes of difficulty, directly contradicting the theoretical framework. Together, this pattern indicates that the lack of conceptual precision directly translates into inconsistent empirical findings and contributes to the unexplained heterogeneity of stereotype threat effects.

Validity Concerns

The present analysis supports broader concerns about the validity of stereotype threat research. Theoretically central conditions did not predict stereotype threat effects in the expected direction. Instead, the effects of stereotype threat for women in STEM were largely driven by small-study effects and showed a systematic decline-effect over time. Precise and recent studies estimated stereotype threat for women in STEM to be non-existent, or even reversed in the form of a small performance lift. These findings resonate with broader arguments that stereotype threat is an example of a psychological phenomenon that once was regarded as well-established but turned out to be a replication failure (Eronen & Bringman, 2021).

These concerns are reflected on the experimental level, where the internal validity of stereotype threat manipulations has been repeatedly questioned (e.g., Picho-Kiroga et al., 2021; Whaley, 1998, 2018, 2021). Scholars have argued that stereotype salience manipulations produced an awareness for stereotypes rather than a general threat (Marsden, 2014). This indicates that stereotype threat manipulations are “fat-handed”: they manipulate several psychological variables at once (Eronen, 2020). Fat-handedness explains why stereotype activation can elicit negative, positive, or null effects on performance and subsequently where heterogeneity in effect sizes observed in the present analysis may come from.

The lack of precision in manipulations not only undermines internal validity but also fuels conceptual vagueness. To date, many different psychological variables were proposed as mediators of stereotype threat effects depending on different contexts (Pennington, 2016). The notion that ”the type and degree of this threat vary from group to group and, for any group, across settings” (C. M. Steele, 1997, p. 618) implies that every psychological phenomenon observed in an experiment could be integrated in the stereotype threat framework, either as a mechanism or a proxy of threat. This stands for a theoretical vagueness that, essentially, makes the theory hard to test and falsify (Eronen & Bringmann, 2021; Meehl, 1978).

This lack of theoretical specificity, combined with the broad range of possible threat manipulations can lead to the jingle fallacy: the problem that different constructs are researched under the same label (Hanfstingl et al., 2024). Moreover, following inconsistencies in stereotype threat research, scholars introduced related concepts like a more general social identity threat (J. Aronson & McGlone, 2009; Inzlicht et al., 2011; Thoman et al., 2014) or stereotype susceptibility (Shih et al., 1999). These concepts highlight different nuances; however, they overlap and oftentimes remain underdefined. Frequently, researchers use them interchangeably, a validity problem called the jangle fallacy (Hanfstingl et al., 2024). The broadening of the theory, as well as the lack of discriminatory power between different concepts, makes it challenging to understand which phenomenon is being studied, and, consequently, which conditions and assumptions apply. These conceptual ambiguities introduce problems to cumulative research, exacerbating problems of generalizability and comparability. Specifically, the jingle/jangle fallacies add to the plausible explanations of the heterogeneity observed in the present meta-analysis and help to account for the ambiguous findings regarding predictors of stereotype threat effects on women in STEM.

A widespread approach in stereotype threat research is adjusting experimental data using previous ability scores, which was met with methodological criticism. Adjusting stereotype threat data with prior ability violates the core assumptions of ANCOVA (Stoet & Geary, 2012; Wicherts, 2005). Within C. M. Steele’s (1997) reasoning, academic ability should be highly related to domain identification; this implies an interaction between the covariate (ability) and the experimental manipulation (Wicherts, 2005). Crucially, the covariate itself should be affected by stereotype threat: stereotyped students, by definition, experience stereotype threat in high-stakes testing situations such as the SAT. This poses a serious theoretical and statistical problem for using it as a covariate in a group comparison (Wicherts, 2005). As the present meta-analysis approached study authors for unadjusted scores where they were not reported, only eight ability-adjusted scores were included, which likely resulted in a smaller summary effect that more precisely reflects the true underlying effect (Shewach et al., 2019).

Taken together, these conceptual and methodological issues contextualize the high heterogeneity found in this meta-analysis. As discussed above, stereotype threat effects on women in STEM are heavily influenced by small-study effects as well as dissemination biases reflected in the decline effect. Additionally, the volatility of the phenomenon may also reflect a lack of theoretical coherence and construct validity.

C. M. Steele (1997) originally introduced the concept as an explanation of situational underperformance. As the gender-based performance gap in mathematics (which constitutes most stereotype threat studies on females) has been largely mitigated (UNESCO, 2022), it becomes increasingly unclear what stereotype threat research aims to explain. Without a coherent theoretical framework, consistent operationalizations, and a clearly defined scope, stereotype threat effects cannot be meaningfully and validly interpreted. Its contribution to explaining or mitigating the STEM gender gap remains questionable.

Limitations

While the present thesis analyzed a total of $k = 161$ effect sizes, some subgroup analyses were underpowered because only a few studies fell into categories of interest (e.g., STEM domains other than mathematics). This also limited the information available for the multiverse and specification curve analysis, as not all possible combinations could be computed.

As discussed in the previous chapters, operationalizations of core concepts in stereotype threat research vary widely. This concerns both the concept of stereotype threat itself and predictors of stereotype threat effects. Consequently, the quality of primary studies sets a limit to the quality of cumulative research.

The present research focused specifically on the concept of stereotype threat for women in STEM and was restricted to the stereotype salience paradigm to ensure comparability of effect sizes. Naturally, this limits generalizability to this population and this particular experimental mechanism. The high heterogeneity suggests that variability in effect sizes may not only stem from characteristics of women in STEM, but also from broader methodological and conceptual issues. For this reason, it is plausible that the discussed problems and conclusions extend to stereotype threat research more generally. Nevertheless, this generalization should be viewed with caution and requires further empirical investigation.

Future Research

Instead of continuing to test the stereotype threat hypothesis on women in STEM, research might benefit from taking a step back and investigating the effects of common stereotype threat activations. It is crucial to examine whether specific manipulations activate a stereotype, if they are absent or less salient in silent control groups, and if removal techniques nullify them. The validity of commonly used stereotype threat inductions can be evaluated through approaches that assess the internal validity of stereotype threat effects to examine whether they trigger equivalent responses.

Research on stereotype susceptibility, vulnerability, and stigma consciousness examines individual differences in how people respond to stereotype activation. As evidence speaks against stereotype threat as a universal phenomenon for women in STEM, putting more emphasis on inter-individual differences could offer valuable insights into the observed heterogeneous effects. The gathered evidence should then be synthesized to develop a new, evidence-based, testable framework of stereotype-related effects.

The stereotype threat theory does not contribute to explaining and mitigating the gender gap in STEM. Future research may benefit from a phenomenon-driven approach, rather than applying and re-interpreting theoretical frameworks that lack coherence and empirical support. A research approach driven by robust empirical findings and real-world relevance may offer insights into the causal links between stereotypes and women's success in the STEM domains.

Altogether, research under the name of stereotype threat would benefit from a clear recommitment to the original theoretical premises, or, alternatively, a clear redefinition of the concept. In general, authors should explicitly outline the theoretical framework guiding their study and discuss the findings within the specific theoretical context. It is necessary to strictly differentiate between psychological phenomena (e.g., underperformance in stereotype-related domains) and specific theoretical constructs (like stereotype threat). Only by addressing the conceptual and methodological inconsistencies can the field advance toward a coherent understanding of how, when, and for whom threats emerge from stereotypes.

Conclusion

This study provides a critical meta-analytic update on the evidence base of stereotype threat for women in STEM, revealing a small significant overall effect that is marked by high heterogeneity and low theoretical consistency. The findings challenge core assumptions of stereotype threat theory, especially regarding the roles of domain identification, self-threat, and task difficulty. Given the sensitivity of results to analytical decisions and the signs of dissemination bias, the meta-analytic summary effect of $d = -0.18$ is not a robust and generalizable estimate of stereotype threat for women in STEM. These results highlight the need for precise operationalizations, theory-driven designs, and a strict and concise distinction between investigated concepts. As the relevance of stereotype threat to women in STEM is questionable, researchers should adopt a phenomenon-driven approach to investigate the gender gap in STEM.

References

Abele, A. E. (2003). The dynamics of masculine-agentic and feminine-communal traits: Findings from a prospective study. *Journal of Personality and Social Psychology*, *85*(4), 768–776. https://doi.org/10.1037/0022-3514.85.4.768

Agnoli, F., Melchiorre, F., Zandonella Callegher, C., & Altoè, G. (2021). Stereotype threat effects on italian girls' mathematics performance: A failure to replicate. *Developmental Psychology*, *57*(6), 940–950. https://doi.org/10.1037/dev0001186

Allport, G. W. (1979). *The nature of prejudice* (Unabridged, 25th anniversary ed). Addison-Wesley Pub. Co.

Ambady, N., Paik, S. K., Steele, J., Owen-Smith, A., & Mitchell, J. P. (2004). Deflecting negative self-relevant stereotype activation: The effects of individuation. *Journal of Experimental Social Psychology*, *40*(3), 401–408. https://doi.org/10.1016/j.jesp.2003.08.003

*Anderson, R. D. (2001). *Stereotype threat: The effects of gender identification on standardized test performance* (Doctoral dissertation). ProQuest Dissertations & Theses Global. https://www.proquest.com/docview/304768413/fulltextPDF/3E4910B9F39A40BCPQ/1?accountid=14682&sourcetype=Dissertations%20&%20Theses

Appel, M., Kronberger, N., & Aronson, J. (2011). Stereotype threat impairs ability building: Effects on test preparation among women in science and technology. *European Journal of Social Psychology*, *41*(7), 904–913. https://doi.org/10.1002/ejsp.835

Appel, M., & Weber, S. (2021). Do mass mediated stereotypes harm members of negatively stereotyped groups? A meta-analytical review on media-generated stereotype threat and stereotype lift. *Communication Research*, *48*(2), 151–179. https://doi.org/10.1177/0093650217715543

Appel, M., Weber, S., & Kronberger, N. (2015). The influence of stereotype threat on immigrants: Review and meta-analysis. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00900

Aramovich, N. P. (2014). The effect of stereotype threat on group versus individual performance. *Small Group Research*, *45*(2), 176–197. https://doi.org/10.1177/1046496414523508

Armstrong, B., Gallant, S. N., Li, L., Patel, K., & Wong, B. I. (2017). Stereotype threat effects on older adults' episodic and working memory: A meta-analysis. *Gerontologist*, *57*, S193–S205. Scopus. https://doi.org/10.1093/geront/gnx056

Aronson, E., Wilson, T. D., & Sommers, S. R. (2021). *Social psychology* (Tenth edition, global edition). Pearson.

Aronson, J., & McGlone, M. S. (2009). Stereotype and social identity threat. In T. D. Nelson (Ed.), *Handbook of prejudice, stereotyping, and discrimination*. (pp. 153–178). Psychology Press.

Aronson, J., Quinn, D. M., & Spencer, S. J. (1998). Stereotype threat and the academic underperformance of minorities and women. In J. K. Swim & C. Stangor (Eds.), *Prejudice: The target's perspective.* (pp. 83–103). Academic Press; https://doi.org/10.1016/B978-012679130-3/50039-9

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology*, *12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Bailey, A. A. (2005). *Effects of stereotype threat on females in Math and Science fields: An investigation of possible mediators and moderators of the threat-performance relationship* (Issues 11-B) [ProQuest Information & Learning]. https://uaccess.univie.ac.at/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2005-99010-068&site=ehost-live

Ben-Zeev, A., & Kirtman, N. (2012). Stereotype threat beyond the laboratory: Do single-gender colleges signal a safety in the air? In *Psychology of Threat* (pp. 75–93). Nova Science Publishers, Inc.; Scopus. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84892236953&partnerID=40&md5=7aa8826e77cf4585bf4f62d4cb26ccf2

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*(2), 174–181. https://doi.org/10.1016/j.jesp.2003.11.007

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothenstein, H. R. (Eds.). (2021). *Introduction to meta-analysis* (Reprint ed.). Wiley.

Cadinu, M., Maass, A., Lombardo, M., & Frigerio, S. (2006). Stereotype threat: The moderating role of locus of control beliefs. *European Journal of Social Psychology*, *36*(2), 183–197. https://doi.org/10.1002/ejsp.303

*Campbell, S. M., & Collaer, M. L. (2009). Stereotype Threat and Gender Differences in Performance on a Novel Visuospatial Task. *Psychology of Women Quarterly*, *33*(4), 437–444. https://doi.org/10.1111/j.1471-6402.2009.01521.x

Casad, B. J., Hale, P., & Wachs, F. L. (2017). Stereotype threat among girls: differences by gender identity and math education context. *Psychology of Women Quarterly*, *41*(4), 513–529. https://doi.org/10.1177/0361684317711412

Chalabaev, A., Major, B., Sarrazin, P., & Cury, F. (2012). When avoiding failure improves performance: Stereotype threat and the impact of performance goals. *Motivation and Emotion*, *36*(2), 130–142. Scopus. https://doi.org/10.1007/s11031-011-9241-x

Chan, N. H. W., & Rosenthal, H. E. S. (2014). Working memory moderates stereotype threat effects for adolescents in Hong Kong. *Revue Internationale de Psychologie Sociale*, *27*(3–4), 103–118.

Cherney, I. D., & Campbell, K. L. (2011). A league of their own: do single-sex schools increase girls' participation in the physical sciences? *Sex Roles: A Journal of Research*, *65*(9–10), 712–724. https://doi.org/10.1007/s11199-011-0013-6

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Conway-Klaassen, J. M. (2010). *Stereotype Threat's Effect on Women's Achievement in Chemistry: The Interaction of Achievement Goal Orientation for Women in Science Majors*. ProQuest LLC. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2011-99090-259&site=ehost-live

Corporation for Digital Scholarship. (2024). *Zotero* (Version 6.0.37) [Computer software]. https://www.zotero.org/

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype threat to social class: the intellectual underperformance of students from low socioeconomic backgrounds. *Personality and Social Psychology Bulletin*, *24*(6), 588–594. https://doi.org/10.1177/0146167298246003

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Danaher, K. (2011). *Reminders of a positively stereotyped identity when facing stereotype threat: Identity consistency and identity accessibility as mediating mechanisms* (Issues 4-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2012-99200-

540&site=ehost-live

Davies, L. C., Conner, M., Sedikides, C., & Hutter, R. R. C. (2016). Math question type and stereotype threat: Evidence from educational settings. *Social Cognition*, *34*(3), 196–216. https://doi.org/10.1521/soco.2016.34.3.196

DeepL SE. (2025). *DeepL Translator* [Online translation tool]. https://www.deepl.com/

Delaume, R., & Lupu, I. (2022, June 2). *When Career Development Programs Trigger Gender Stereotype Threat*.

Delgado, A. R., & Prieto, G. (2008). Stereotype threat as validity threat: the anxiety-sex-threat interaction. *Intelligence*, *36*(6), 635–640. https://doi.org/10.1016/j.intell.2008.01.008

Deviyanti, D. (2016). *The role of women's identification with math and academic major in women's susceptibility to stereotype threat and stereotype lift* (Issues 10-B(E)) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2016-16233-152&site=ehost-live

*Drążkowski, D., Szwedo, J., Krajczewska, A., Adamczuk, A., Piątkowski, K., Jadwiżyc, M., & Rakowski, A. (2017). Women are not less field independent than men—The role of stereotype threat. *International Journal of Psychology*, *52*(5), 415–419. https://doi.org/10.1002/ijop.12238

Dunst, B., Benedek, M., Bergner, S., Athenstaedt, U., & Neubauer, A. C. (2013). Sex differences in neural efficiency: Are they due to the stereotype threat effect? *Personality and Individual Differences*, *55*(7), 744–749. https://doi.org/10.1016/j.paid.2013.06.007

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in stem: professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*(3–4), 127–141. https://doi.org/10.1007/s11199-019-01052-w

Elsesser, K. (2019, December 11). Lawsuit claims sat and act are biased—here's what research says. *Forbes*. https://www.forbes.com/sites/kimelsesser/2019/12/11/lawsuit-claims-sat-and-act-are-biased-heres-what-research-says/

Elsevier. (2024). *Progress Toward Gender Equality in Research & Innovation: An in-depth analysis of research participation, career progression and research contributions across the globe*. Elsevier.

https://assets.ctfassets.net/o78em1y1w4i4/3W1JKBd7ez41lAJKFWkB4l/d903f7578230d316bf18e702e870a137/Progress-Toward-Gender-Equality-in-Research-Innovation_2024-Review.pdf?utm_source=newsletter&utm_medium=email&utm_campaign=newsletter_axiosscience&stream=science

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality*, *63*, 36–43. https://doi.org/10.1016/j.jrp.2016.05.009

Fisher, Z., Tipton, E., & Zhipeng, H. (2014). *robumeta: Robust Variance Meta-Regression* (p. 2.1) [Dataset]. https://doi.org/10.32614/CRAN.package.robumeta

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25–44. https://doi.org/10.1016/j.jsp.2014.10.002

Fogliati, V. J., & Bussey, K. (2013). Stereotype threat reduces motivation to improve: effects of stereotype threat and feedback on women's intentions to improve mathematical ability. *Psychology of Women Quarterly*, *37*(3), 310–324. https://doi.org/10.1177/0361684313480045

Ford, T. E., Ferguson, M. A., Brooks, J. L., & Hagadone, K. M. (2004). Coping sense of humor reduces effects of stereotype threat on women's math performance. *Personality and Social Psychology Bulletin*, *30*(5), 643–653. https://doi.org/10.1177/0146167203262851

*Ganley, C. M. (2011). *Gender differences in math performance across development: Exploring the roles of anxiety, working memory, and stereotype threat* (Doctoral dissertation). ProQuest Dissertations & Theses Global. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2012-99030-219&site=ehost-live

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, *49*(10), 1886–1897. https://doi.org/10.1037/a0031412

Gentile, A., Boca, S., & Giammusso, I. (2018). 'You play like a woman!' effects of gender stereotype threat on women's performance in physical and sport activities: A meta-analysis. *Psychology of Sport and Exercise*, *39*, 95–103. https://doi.org/10.1016/j.psychsport.2018.07.013

Gentile, L. (2009). *The relationship between gender identity and the effects of stereotype threat on women's math performance* (Issues 3-A) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-99170-290&site=ehost-live

*Good, J. J., Woodzicka, J. A., & Wingfield, L. C. (2010). The effects of gender stereotypic and counter-stereotypic textbook images on science performance. *The Journal of Social Psychology*, *150*(2), 132–147. https://doi.org/10.1080/00224540903366552

Grand, J. A., Ryan, A. M., Schmitt, N., & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance*, *24*(1), 1–28. https://doi.org/10.1080/08959285.2010.518184

Gringard, H. M. (2017). *She saved the world a lot: On the effect and effectivity of general counterstereotypes in stereotype threat situations*. pdx. https://search.ebscohost.com/login.aspx?direct=true&db=pdx&AN=0326187&site=ehost-live

Guizzo, F., Moè, A., Cadinu, M., & Bertolli, C. (2019). The role of implicit gender spatial stereotyping in mental rotation performance. *Acta Psychologica*, *194*, 63–68. https://doi.org/10.1016/j.actpsy.2019.01.013

Haft, S., de Magalhaes, C., & Hoeft, F. (2023). A Systematic Review of the Consequences of Stigma and Stereotype Threat for Individuals With Specific Learning Disabilities. *Journal of Learning Disabilities*, *56*(3), 193–209. https://doi.org/10.1177/00222194221087383

Haft, S. L., Duong, P. H., Ho, T. C., Hendren, R. L., & Hoeft, F. (2019). Anxiety and attentional bias in children with specific learning disorders. *Journal of Abnormal Child Psychology*, *47*(3), 487–497. https://doi.org/10.1007/s10802-018-0458-y

Hanfstingl, B., Oberleiter, S., Pietschnig, J., Tran, U. S., & Voracek, M. (2024). Detecting jingle and jangle fallacies by identifying consistencies and variabilities in study specifications – a call for research. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1404060

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis* (1st Edition). Elsevier Science.

Hermann, J. M., & Vollmeyer, R. (2016). Stereotype Threat in der Grundschule = Stereotype threat in primary school. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*, *48*(1), 42–49. https://doi.org/10.1026/0049-8637/a000143

Hermann, J. M., & Vollmeyer, R. (2022). Gender stereotypes: Implicit threat to performance or boost for motivational aspects in primary school? *Social Psychology of Education*, *25*(2–3), 349–369. https://doi.org/10.1007/s11218-022-09693-8

Herrmann, L. (2021, March 12). Faktencheck: Stereotype über Frauen im MINT-Bereich. *Global Digital Women*. https://global-digital-women.com/faktencheck-stereotype-ueber-frauen-im-mint-bereich/

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Inglis, M., & O'Hagan, S. (2022). Stereotype threat, gender and mathematics attainment: A conceptual replication of Stricker & Ward. *PLoS ONE*, *17*(5 May). Scopus. https://doi.org/10.1371/journal.pone.0267699

*Inzlicht, M., & Ben-Zeev, T. (2000). A threatening intellectual environment: why females are susceptible to experiencing problem-solving deficits in the presence of males. *Psychological Science*, *11*(5), 365–371. https://doi.org/10.1111/1467-9280.00272

Inzlicht, M., Tullett, A. M., Legault, L., & Kang, S. K. (2011). Lingering effects: stereotype threat hurts more than you think. *Social Issues and Policy Review*, *5*(1), 227–256. https://doi.org/10.1111/j.1751-2409.2011.01031.x

Jamieson, J. P., & Harkins, S. G. (2009). The effect of stereotype threat on the solving of quantitative GRE problems: A mere effort interpretation. *Personality and Social Psychology Bulletin*, *35*(10), 1301–1314. https://doi.org/10.1177/0146167209335165

Jamieson, J. P., & Harkins, S. G. (2010). Evaluation is necessary to produce stereotype threat performance effects. *Social Influence*, *5*(2), 75–86. https://doi.org/10.1080/15534510903512409

Jamieson, J. P., & Harkins, S. G. (2012). Distinguishing between the effects of stereotype priming and stereotype threat on math performance. *Group Processes & Intergroup Relations*, *15*(3), 291–304. https://doi.org/10.1177/1368430211417833

Johnson, J. L. (2012). *A mediated-moderation model of stereotype threat and performance pressure effects* [Auburn University]. ProQuest LLC.

Jones, P. R. (2005). *The effects of stereotype threat on the test performance and task choices of women* (2005-99024-121; Issues 6-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2005-99024-121&site=ehost-live

*Josephs, R. A., Newman, M. L., Brown, R. P., & Beer, J. M. (2003). Status, testosterone, and human intellectual performance: Stereotype threat as status concern. *Psychological Science*, *14*(2), 158–163. https://doi.org/10.1111/1467-9280.t01-1-01435

*Kahalon, R., Shnabel, N., & Becker, J. C. (2018). Positive stereotypes, negative outcomes: Reminders of the positive components of complementary gender stereotypes impair performance in counter-stereotypical tasks. *British Journal of Social Psychology*, *57*(2), 482–502. https://doi.org/10.1111/bjso.12240

Katter, J. (2018, February 24). Opinion: Steminism. *Los Angeles Times High School Insider*. https://highschool.latimes.com/opinion/opinion-steminism/

Keller, J. (2002). Blatant stereotype threat and women's math performance: self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles: A Journal of Research*, *47*, 193–198. https://doi.org/10.1023/A:1021003307511

Keller, J. (2007). Stereotype threat in classroom settings: the interactive effect of domain identification, task difficulty and stereotype threat on female students' maths performance. *British Journal of Educational Psychology*, *77*(2), 323–338. https://doi.org/10.1348/000709906X113662

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis: Three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Lamont, R. A., Swift, H. J., & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat: Negative stereotypes, not facts, do the damage. *Psychology and Aging*, *30*(1), 180–193. https://doi.org/10.1037/a0038586

Lesko, A. C., & Corpus, J. H. (2006). Discounting the difficult: how high math-identified women respond to stereotype threat. *Sex Roles: A Journal of Research*, *54*(1–2), 113–125. https://doi.org/10.1007/s11199-005-8873-2

Lester, C. C. (2007). *Stereotype threat and math test performance of undergraduate females* (2007-99010-103; Issues 11-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2007-99010-103&site=ehost-live

Liu, S., Liu, P., Wang, M., & Zhang, B. (2021). Effectiveness of stereotype threat interventions: a meta-analytic review. *Journal of Applied Psychology*, *106*(6), 921–949. https://doi.org/10.1037/apl0000770

Macchione, A. L., & Sacco, D. (2023). Comparing the efficacy of two forms of self-affirmation to reduce stereotype threat effects on women's math performance. *Journal for STEM Education Research*, *6*(1), 181–198. https://doi.org/10.1007/s41979-023-00085-x

*Mangels, J. A., Good, C., Whiteman, R. C., Maniscalco, B., & Dweck, C. S. (2012). Emotion blocks the path to learning under stereotype threat. *Social Cognitive and Affective Neuroscience*, *7*(2), 230–241. https://doi.org/10.1093/scan/nsq100

*Marchand, G. C., & Taasoobshirazi, G. (2013). Stereotype threat and women's performance in physics. *International Journal of Science Education*, *35*(18), 3050–3061. https://doi.org/10.1080/09500693.2012.683461

Marsden, A. J. (2014). *On the construct validity of stereotype threat measures: Some possible boundary conditions*. University of Missouri-Saint Louis.

McGlone, M. S., & Aronson, J. (2006). Stereotype threat, identity salience, and spatial reasoning. *Journal of Applied Developmental Psychology*, *27*(5), 486–493. https://doi.org/10.1016/j.appdev.2006.06.003

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006x.46.4.806

Miller, D., Eagly, A., & Linn, M. (2015). Women's representation in science predicts national gender-science stereotypes: evidence from 66 nations. *Journal of Educational Psychology*, *107*(3), 631–644. https://doi.org/10.1037/edu0000005

*Mingle, L. A. (2014). *Threats to success in mathematics: Examining the combined effects of choking under pressure and stereotype threat* (Doctoral dissertation). ProQuest Dissertations & Theses Global. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2014-99170-329&site=ehost-live

Moè, A. (2012). Gender difference does not mean genetic difference: externalizing improves performance in mental rotation. *Learning and Individual Differences*, *22*(1), 20–24. https://doi.org/10.1016/j.lindif.2011.11.001

*Moè, A. (2018). Effects of group gender composition on mental rotation test performance in women. *Archives of Sexual Behavior*, *47*(8), 2299–2305. https://doi.org/10.1007/s10508-018-1245-0

Muzzatti, B., & Agnoli, F. (2007). Gender and mathematics: attitudes and stereotype threat susceptibility in italian children. *Developmental Psychology*, *43*(3), 747–759. https://doi.org/10.1037/0012-1649.43.3.747

Nadler, J. T., & Clark, M. H. (2011). Stereotype threat: a meta-analysis comparing african americans to hispanic americans: stereotype threat: a meta-analysis. *Journal of Applied Social Psychology*, *41*(4), 872–890. https://doi.org/10.1111/j.1559-1816.2011.00739.x

Neuburger, S., Jansen, P., Heil, M., & Quaiser-Pohl, C. (2012). A threat in the classroom gender stereotype activation and mental-rotation performance in elementary-school children. *Zeitschrift Für Psychologie*, *220*(2), 61–69. https://doi.org/10.1027/2151-2604/a000097

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314–1334. https://doi.org/10.1037/a0012702

Nosek, B., Smyth, F., Sriram, N., Lindner, N., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B., Wiers, R., … Greenwald, A. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

OECD (Ed.). (2023). *PISA 2022 Ergebnisse: Lernstände und Bildungsgerechtigkeit* (1st ed.). wbv Media GmbH & Co. KG. https://doi.org/10.3278/6004956w

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH – a graphical display of study heterogeneity. *Research Synthesis Methods*, *3*(3), 214–223. https://doi.org/10.1002/jrsm.1053

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pennington, C., Heim, D., Levy, A., & Larkin, D. (2016). Twenty Years of Stereotype Threat Research: A Review of Psychological Mediators. *PLOS One*, *11*(1). https://doi.org/10.1371/journal.pone.0146487

Pennington, C. R., Litchfield, D., McLatchie, N., & Heim, D. (2019). Stereotype threat may not impact women's inhibitory control or mathematical performance: Providing support for the null hypothesis. *European Journal of Social Psychology*, *49*(4), 717–734. https://doi.org/10.1002/ejsp.2540

Petkari, E., Nikolaou, E., Oberleiter, S., Priebe, S., & Pietschnig, J. (2024). Which psychological interventions improve quality of life in patients with schizophrenia-spectrum disorders? A meta-analysis of randomized controlled trials. *Psychological Medicine*, *54*(2), 221–244. https://doi.org/10.1017/S0033291723003070

Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: a meta-analysis. *The Journal of Social Psychology*, *153*(3), 299–333. https://doi.org/10.1080/00224545.2012.737380

*Picho, K., & Schmader, T. (2018). When do gender stereotypes impair math performance? A study of stereotype threat among ugandan adolescents. *Sex Roles: A Journal of Research*, *78*(3–4), 295–306. https://doi.org/10.1007/s11199-017-0780-9

Picho, K., & Stephens, J. M. (2012). Culture, context and stereotype threat: a comparative analysis of young ugandan women in coed and single-sex schools. *Journal of Educational Research*, *105*(1), 52–63. https://doi.org/10.1080/00220671.2010.517576

Picho-Kiroga, K., Turnbull, A., & Rodriguez-Leahy, A. (2021a). Stereotype threat and its problems: theory misspecification in research, consequences, and remedies. *Journal of Advanced Academics*, *32*(2), 231–264. https://doi.org/10.1177/1932202X20986161

Pietschnig, J., Siegel, M., Eder, J. S. N., & Gittler, G. (2019). Effect declines are systematic, strong, and ubiquitous: a meta-meta-analysis of the decline effect in intelligence research. *Frontiers in Psychology*, *10*, 2874. https://doi.org/10.3389/fpsyg.2019.02874

Priest, R., Griebie, A., Zhou, Y., Tomeh, D., & Sackett, P. R. (2024). Stereotype lift and stereotype threat effects on subgroup mean differences for cognitive tests: a meta-analysis of adult samples. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001185

Rivardo, M. G., Rhodes, M. E., Camaione, T. C., & Legg, J. M. (2011). Stereotype threat leads to reduction in number of math problems qomen attempt. *North American Journal of Psychology*, *13*(1), 5–16. Scopus.

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160. https://doi.org/10.1037/met0000300

Rosenthal, H. E. S., & Crisp, R. J. (2006). Reducing stereotype threat by blurring intergroup boundaries. *Personality and Social Psychology Bulletin*, *32*(4), 501–511. https://doi.org/10.1177/0146167205281009

Rucks, L. J. (2008). *Me, women, and math: The role of personal and collective threats in the experience of stereotype threat* (Issues 11-B(E)) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2017-36666-284&site=ehost-live

Ruff, C. (2013). *Examining and supporting domain identification and student interest in first year college students*.

Ryan, A. M., & Nguyen, H.-H. D. (2017). Publication bias and stereotype threat research: A reply to Zigerell. *Journal of Applied Psychology*, *102*(8), 1169–1177. https://doi.org/10.1037/apl0000242

Rydell, R. J., McConnell, A. R., & Beilock, S. L. (2009). Multiple social identities and stereotype threat: imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology*, *96*(5), 949–966. https://doi.org/10.1037/a0014846

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for african american-white differences on cognitive tests. *American Psychologist*, *59*(1), 7–13. https://doi.org/10.1037/0003-066x.59.1.7

Sackett, P. R., & Ryan, A. M. (2012). Concerns about generalizing stereotype threat research findings to operational high-stakes testing. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application.* (pp. 249–263). Oxford University Press; https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2011-27639-017&site=ehost-live

Seitchik, A. E. (2014). *Stereotype threat, mental arithmetic, and the mere effort account* (2014-99120-212; Issues 12-B(E)) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2014-99120-212&site=ehost-live

Seitchik, A., Jamieson, J., & Harkins, S. (2014). Reading between the lines: Subtle stereotype threat cues can motivate performance. *SOCIAL INFLUENCE*, *9*(1), 52–68. https://doi.org/10.1080/15534510.2012.746206

Shapiro, J. R., & Neuberg, S. L. (2007). From Stereotype Threat to Stereotype Threats: Implications of a Multi-Threat Framework for Causes, Moderators, Mediators, Consequences, and Interventions. *Personality and Social Psychology Review*, *11*(2), 107–130. https://doi.org/10.1177/1088868306294790

Shenouda, C. K., Patel, K. S., & Danovitch, J. H. (2024). Who can do stem?: children's gendered beliefs about stem and non-stem competence and learning. *Sex Roles*, *90*(8), 1063–1074. https://doi.org/10.1007/s11199-024-01493-y

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: a meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514–1534. https://doi.org/10.1037/apl0000420

Shih, M., Pittinsky, T., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. Psychological *Science*, *10*(1), 80–83. https://doi.org/10.1111/1467-9280.00111

Simmons, C. M. (2010). *Does Manipulating Stereotype Threat Condition Change Performance Goal State* [The University]. ProQuest LLC. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2010-99230-211&site=ehost-live

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Smetackova, I. (2015). *Gender stereotypes, performance and identification with math* (T.

Bozkurt & M. Demirok, Eds.; Vol. 190, pp. 211–219). https://doi.org/10.1016/j.sbspro.2015.04.937

Smittick, A. L. (2022). *The in-between: A meta-analytic investigation of stereotype threat effects on mediators of the stereotype threat-performance relationship* (Doctoral dissertation). ProQuest Dissertations & Theses Global. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2022-81481-236&site=ehost-live

Sorenson, S. M. (2022). *Mindfulness as a potential buffer of stereotype threat for underrepresented minority females in STEM* (Issues 11-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2022-70651-129&site=ehost-live

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1998). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 54–68.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Spencer, S. L. (2005). *Stereotype threat and women's math performance: The possible mediating factors of test anxiety, test motivation and self-efficacy* (Issues 4-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2005-99020-067&site=ehost-live

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Steele, C. M. (1998). Stereotyping and its threat are real. *American Psychologist*, *53*(6), 680–681. https://doi.org/10.1037/0003-066X.53.6.680

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Steele, J., James, J. B., & Barnett, R. C. (2002). Learning in a man's world: examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly*, *26*(1), 46–50. https://doi.org/10.1111/1471-6402.00042

*Steinberg, J. R., Okun, M. A., & Aiken, L. S. (2012). Calculus gpa and math identification as moderators of stereotype threat in highly persistent women. *Basic and Applied Social Psychology*, *34*(6), 534–543. https://doi.org/10.1080/01973533.2012.727319

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

Stoet, G., & Geary, D. C. (2012a). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102. https://doi.org/10.1037/a0026617

Stoet, G., & Geary, D. C. (2012b). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology*, *16*(1), 93–102. Scopus. https://doi.org/10.1037/a0026617

Stoevenbelt, A. H., Flore, P. C., Schwabe, I., & Wicherts, J. M. (2022). The uniformity of stereotype threat: Analyzing the moderating effects of premeasured performance. *Intelligence*, *93*. Scopus. https://doi.org/10.1016/j.intell.2022.101655

*Sunny, C. E., Taasoobshirazi, G., Clark, L., & Marchand, G. (2017). Stereotype threat and gender differences in chemistry. *Instructional Science: An International Journal of the Learning Sciences*, *45*(2), 157–175. https://doi.org/10.1007/s11251-016-9395-8

Tagler, M. J. (2004). *Stereotype threat: Prevalence and individual differences* (Issues 12-B) [ProQuest Information & Learning]. https://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2004-99012-013&site=ehost-live

Taillandier-Schmitt, A., Esnard, C., & Mokounkolo, R. (2012). Self-affirmation in occupational training: Effects on the math performance of French women nurses under stereotype threat. *Sex Roles: A Journal of Research*, *67*(1–2), 43–57. https://doi.org/10.1007/s11199-012-0157-z

Thébaud, S., & Charles, M. (2018). Segregation, stereotypes, and STEM. *Social Sciences*, *7*(7), 111. https://doi.org/10.3390/socsci7070111

Thoman, D., Arizaga, J., Smith, J., Story, T., & Soncuya, G. (2014). The grass is greener in non-science, technology, engineering, and math classes: examining the role of competing belonging to undergraduate women's vulnerability to being pulled away from science. *Psychology of Women Quarterly*, *38*(2), 246–258.

https://doi.org/10.1177/0361684313499899

Tomeh, D., & Sackett, P. (2022). On the continued misinterpretation of stereotype threat as accounting for black-white differences on cognitive tests. *Personnel Assessment and Decisions*, *8*(1). https://doi.org/10.25035/pad.2022.01.001

UNESCO. (2022). *Global education monitoring report – gender report: deepening the debate on those still left behind.* UNESCO. https://doi.org/10.54676/rczb6329

UNESCO. (2024). *Support girls and women to pursue STEM subjects and careers: Advocacy brief*. UNESCO. https://doi.org/10.54676/bpsl3344

Van Den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Vedantam, S. (2012, July 12). How stereotypes can drive women to quit science. *NPR*. https://www.npr.org/2012/07/12/156664337/stereotype-threat-why-women-quit-science-jobs

Viechtbauer, W. (2010). Conducting meta-analyses in *r* with the metafor package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how?: a specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift Für Psychologie*, *227*(1), 64–82. https://doi.org/10.1027/2151-2604/a000357

Wang, Z., Guan, J., & Zhao, L. (2022). Priming science identity elicits stereotype boost for chinese science female students: evidence from math performance. *Basic and Applied Social Psychology*, *44*(3–6), 95–108. https://doi.org/10.1080/01973533.2022.2144315

Warne, R. T. (2022). No strong evidence of stereotype threat in females: a reassessment of the meta-analysis. *Journal of Advanced Academics*, *33*(2), 171–186. https://doi.org/10.1177/1932202X211061517

Weber, S., & Winkler, J. (2019). Starke Frauen, schöne Männer? Inhalte und Konsequenzen von Geschlechterstereotypen in den Medien. *In-Mind*, *02/2019*. https://de.in-mind.org/article/starke-frauen-schoene-maenner-inhalte-und-konsequenzen-von-geschlechterstereotypen-in-den

Whaley, A. (2018). Advances in stereotype threat research on African Americans: Continuing challenges to the validity of its role in the achievement gap. *Social Psycholofy of Education*, *21*(1), 111–137. https://doi.org/10.1007/s11218-017-

9415-9

Whaley, A. L. (1998). Issues of validity in empirical tests of stereotype threat theory. *American Psychologist*, *53*(6), 679–680. https://doi.org/10.1037/0003-066X.53.6.679

Whaley, A. L. (2021). Stereotype threat and neuropsychological test performance in the US African American population. *Archives of Clinical Neuropsychology*, *36*(7), 1361–1367. https://doi.org/10.1093/arclin/acab020

Wicherts, J. M. (2005). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist*, *60*(3), 267–269. Scopus. https://doi.org/10.1037/0003-066X.60.3.267

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: a question of measurement invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. https://doi.org/10.1037/0022-3514.89.5.696

*Wille, E., Gaspard, H., Trautwein, U., Oschatz, K., Scheiter, K., & Nagengast, B. (2018). Gender stereotypes in a children's television program: Effects on girls' and boys' stereotype endorsement, math performance, motivational dispositions, and attitudes. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02435

Wilson, A. R. (2009). *Malleable view of intelligence as intervention for stereotype threat: overcoming math underperformance in women*. East Carolina University.

Wister, J. A., Stubbs, M. L., & Shipman, C. (2013). Mentioning menstruation: a stereotype threat that diminishes cognition? *Sex Roles: A Journal of Research*, *68*(1–2), 19–31. https://doi.org/10.1007/s11199-012-0156-0

Woodcock, A., Hernandez, P. R., Estrada, M., & Schultz, P. W. (2012). The consequences of chronic stereotype threat: domain disidentification and abandonment. *Journal of Personality and Social Psychology*, *103*(4), 635–646. https://doi.org/10.1037/a0029120

Yagan, E., & Avci, S. (2023). The effect of stereotype threat for academic success of female students. *International Journal of Psychology and Educational Studies*, *10*(4), 1102–1113.

Zigerell, L. (2017). Potential publication bias in the stereotype threat literature: Comment on Nguyen and Ryan (2008). *Journal of Applied Psychology*, *102*(8), 1159–1168. https://doi.org/10.1037/apl0000188

Appendix

Appendix A: List of Figures

Appendix B: List of Tables

Appendix C: Abstract (Deutsch)

Die vorliegende Multi-Level Metaanalyse untersuchte Stereotype Threat bei Frauen in Wissenschaften, Technik, Ingenieurswesen, Mathematik (engl. STEM). Dazu wurde Evidenz aus 92 Studien mit einer kombinierten Stichprobengröße von 11.608 Studienteilnehmerinnen zusammengeführt. Die Meta-Analyse konzentrierte sich dabei auf theoretische Rahmenbedingungen und Robustheitsanalysen. Der Gesamteffekt von Stereotype Threat auf die Leistung von Frauen in STEM war klein ($d = -0.18$). Dieser Effekt erwies sich angesichts der Hinweise für einen Disseminationsbias als trivial. Darüber hinaus deuteten das breite Vorhersageintervall und die beträchtliche Heterogenität, die vorwiegend innerhalb von Studien und nicht zwischen Studien festgestellt wurde, darauf hin, dass der gepoolte Gesamteffekt kein informativer Schätzwert für zukünftige Studien ist. Die im Rahmen der Stereotype Threat Theorie (C. M. Steele, 1997) aufgestellten wesentlichen theoretischen Bedingungen Domänenidentifizierung, Aufgabenschwierigkeit und Selbstwertbedrohung trugen nicht signifikant zur Varianzaufklärung bei. Dieser Befund stellte die Robustheit und theoretische Kohärenz des Konzeptes infrage und schränkte die Interpretierbarkeit der gepoolten Schätzung weiter ein. Die Multiversums- und Spezifikationskurvenanalyse hob die Sensitivität der beobachteten Effekte gegenüber analytischen Entscheidungen hervor, mit nicht signifikanten Gesamteffekten bei annähernd der Hälfte aller Spezifikationen. Hinweise für einen Disseminationsbias traten auf: Frühere Studien berichteten über große Effekte, während neuere Studien keine oder sogar positive Effekte feststellten (Decline Effect). Darüber trugen kleine, unpräzise Studien überproportional zu dem negativen Gesamtschätzer bei (small study effect). Ein Drei-Parameter-Selektionsmodell fand keine Hinweise auf selektive Berichterstattung. Die Ergebnisse deuten auf unzureichend definierte konzeptuelle Rahmendingungen und inkonsistente Operationalisierungen zentraler Konstrukte hin. Zudem weist dies auf konzeptionelle Überschneidungen zwischen verwandten Konstrukten hin (jingle-jangle fallacies). Zusammengenommen stellen diese Ergebnisse die Robustheit und theoretische Präzision von Stereotype Threat bei Frauen in Frage. Experimentelle Designs sollten künftig klarer ihre theoretischen Rahmenbedingungen darlegen und strikter zwischen psychologischen Phänomenen und theoretischen Konstrukten unterscheiden. Anstatt weiterhin die fragwürdige Stereotype Threat Hypothese zu testen, sollten Studien zukünftig einen allgemeineren, phänomengetriebenen Ansatz verfolgen, um mögliche Implikationen sowie interindividuelle Unterschiede der Stereotypenaktivierung bei Frauen in STEM untersuchen.

Appendix D: R Code

```
library(weightr)
library(metafor)
library(dplyr)
library(readxl)
library(ggplot2)
library(robumeta)
library(extrafont)


#############################
####### PREPARE DATA #######
#############################

# Numeric Columns
numeric_vars <- c(
  "year", "n1", "n2", "age_mean",
  "sexcomposition", "m1", "sd1", "m2", "sd2",
  "taskdifficulty"
)

# Recode Selfthreat
data$selfthreat[data$selfthreat == "possibly"] <- "yes"

# Recode Outcome Domains
data$outcome_domain[data$outcome_domain == "maths"] <- "math"
data$outcome_domain[data$outcome_domain == "chemistry"] <- "science"
data$outcome_domain[data$outcome_domain == "physics"] <- "science"
data$outcome_domain[data$outcome_domain == "spatial ability"] <- "general"
data$outcome_domain[data$outcome_domain == "math and science"] <- "general"

# Convert Columns
data <- data %>%
  mutate(across(all_of(numeric_vars), as.numeric)) %>%
  mutate(across(-all_of(numeric_vars), as.factor))

# Check missing values
data %>%
  select(all_of(numeric_vars)) %>%
  summarise(across(everything(), ~ sum(is.na(.))))

# Set reference categories for categorical variables
data$selfthreat <- relevel(data$selfthreat, ref = "yes")
data$identification <- relevel(data$identification, ref = "identified")
data$outcome_domain <- relevel(data$outcome_domain, ref = "math")
data$publication_status <- relevel(data$publication_status, ref = "published")

############################################
######## GENERAL MULTI-LEVEL-MODEL ########
############################################
```

```
# Compute Cohen's ds and Sampling Variances
es <- escalc(
  measure = "SMD",
  m1i = m1, sd1i = sd1, n1i = n1,
  m2i = m2, sd2i = sd2, n2i = n2,
  data = data
)

# General Model Without Predictors
gmlm <- rma.mv(yi, vi, data = es,
               random = list(~ 1 | effect_id, ~ 1 | study_id),
               method = "REML")
## Model Results
summary(gmlm)

predict(gmlm, digits=3)

gmlm2 <- rma.mv(yi, vi, data = es,
               random = list(~ 1 | effect_id, ~ 1 | study_id),
               sigma2=c(0,NA))

gmlm3 <- rma.mv(yi, vi, data = es,
                random = list(~ 1 | effect_id, ~ 1 | study_id),
                sigma2=c(NA, 0))

anova(gmlm,gmlm2)

anova(gmlm,gmlm3)

##I-squared for multilevel model based on https://www.metafor-
project.org/doku.php/tips:i2_multilevel_multivariate
W <- diag(1/gmlm$vi)
X <- model.matrix(gmlm)
P <- W - W %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% W
100 * sum(gmlm$sigma2) / (sum(gmlm$sigma2) + (gmlm$k-gmlm$p)/sum(diag(P)))

round(gmlm$sigma2 / sum(gmlm$sigma2), 3)

# Determining how the total variance is distributed over the
# three levels of the meta-analytic model;
# Print the results in percentages on screen.
# Approach based on Assink & Wibbelink (2016)
n <- length(es$vi)
list.inverse.variances <- 1 / (es$vi)
sum.inverse.variances <- sum(list.inverse.variances)
squared.sum.inverse.variances <- (sum.inverse.variances) ^ 2
list.inverse.variances.square <- 1 / (es$vi^2)
sum.inverse.variances.square <-
  sum(list.inverse.variances.square)
```

```
numerator <- (n - 1) * sum.inverse.variances
denominator <- squared.sum.inverse.variances -
  sum.inverse.variances.square
estimated.sampling.variance <- numerator / denominator
I2_1 <- (estimated.sampling.variance) / (gmlm$sigma2[1]
                                    + gmlm$sigma2[2] + estimated.sampling.variance)
I2_2 <- (gmlm$sigma2[1]) / (gmlm$sigma2[1]
                           + gmlm$sigma2[2] + estimated.sampling.variance)
I2_3 <- (gmlm$sigma2[2]) / (gmlm$sigma2[1]
                           + gmlm$sigma2[2] + estimated.sampling.variance)
amountvariancelevel1 <- I2_1 * 100
amountvariancelevel2 <- I2_2 * 100
amountvariancelevel3 <- I2_3 * 100
amountvariancelevel1
amountvariancelevel2
amountvariancelevel3

predict(gmlm, digits=3)

##############################################
####### THEORETICAL CONDITIONS MODEL #######
##############################################
# Transform Task Difficulty
## Mean Centering
es$taskdifficulty_c <- scale(es$taskdifficulty, scale = FALSE)
es$tdsq_c <- (es$taskdifficulty_c)^2

# Conditions Model
cmlm <- rma.mv(
  yi, vi,
  mods = ~ identification + selfthreat + taskdifficulty_c,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es,
  method = "REML")

## Model Results
summary(cmlm)

# Task Difficulty Model
tdmlm <- rma.mv(
  yi, vi,
  mods = ~ taskdifficulty_c + tdsq_c,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es,
  method = "REML")

# Plot
df <- data.frame(
  taskdifficulty= seq(
    from = min(es$taskdifficulty, na.rm = TRUE),
```

```
    to   = max(es$taskdifficulty, na.rm = TRUE),
    length.out = 100
  )
)
es$tdsq <- es$taskdifficulty^2

preds <- predict(
  tdmlm,
  newmods = as.matrix(df[, c("taskdifficulty", "tdsq")]),
  digits = 4
)

df$pred      <- preds$pred
df$ci.lb     <- preds$ci.lb
df$ci.ub     <- preds$ci.ub

ggplot(df, aes(x = taskdifficulty, y = pred)) +
  geom_ribbon(aes(ymin = ci.lb, ymax = ci.ub), alpha = 0.2) +
  geom_line(size = 1) +
  labs(
    x     = "Task difficulty",
    y     = "Standardized mean difference (d)",
  ) +
  theme_minimal(base_family = "Times New Roman")  # or another serif font

 1 - (sum(tdmlm$sigma2) / sum(gmlm$sigma2))


################################
######## STEM-DOMAINS ########
################################
stemmlm <- rma.mv(
  yi, vi,
  mods = ~ outcome_domain,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es,
  method = "REML"
)

summary(stemmlm)

######################################
######## PUBLICATION STATUS ########
######################################

# Publication Status Model
ps_mlm <- rma.mv(
  yi, vi,
  mods = ~ publication_status,
  random = ~ 1 | study_id/effect_id,
```

```
  data = es,
  method = "REML"
)

# Model Summary
summary(ps_mlm)

## Published Only Subset
es_pubonly <- subset(es, publication_status == "published")

## Published Only Model
pspub_mlm <- rma.mv(
  yi, vi,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_pubonly,
  method = "REML"
)

## Published Only Summary
summary(pspub_mlm)

## I^2
W <- diag(1/pspub_mlm$vi)
X <- model.matrix(pspub_mlm)
P <- W - W %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% W
100 * sum(pspub_mlm$sigma2) / (sum(pspub_mlm$sigma2) + (pspub_mlm$k-
pspub_mlm$p)/sum(diag(P)))

### Unpublished Only Subset
es_unpub <- subset(es, publication_status == "unpublished")

### Unpublished Only Model
unpub_mlm <- rma.mv(
  yi, vi,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_unpub,
  method = "REML"
)

### Published Only Summary
summary(unpub_mlm)

### I^2
W <- diag(1/unpub_mlm$vi)
X <- model.matrix(unpub_mlm)
P <- W - W %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% W
100 * sum(unpub_mlm$sigma2) / (sum(unpub_mlm$sigma2) + (unpub_mlm$k-
unpub_mlm$p)/sum(diag(P)))
```

```
#####################################
######### PUBLICATION YEAR #########
#####################################
# Centering at Year 2000
es$year_centered <- es$year - 2000

# Publication Year Model
pymlm <- rma.mv(
  yi, vi,
  mods = ~ year_centered,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es,
  method = "REML"
)

# Model Results
summary(pymlm)

## Published Studies Only
es_pubonly <- subset(es, publication_status == "published")

## Model
py_pubmlm <- rma.mv(
  yi, vi,
  mods = ~ year_centered,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_pubonly,
  method = "REML"
)

## Model Results
summary(py_pubmlm)
##I^2
W <- diag(1/py_pubmlm$vi)
X <- model.matrix(py_pubmlm)
P <- W - W %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% W
100 * sum(py_pubmlm$sigma2) / (sum(py_pubmlm$sigma2) + (py_pubmlm$k-
py_pubmlm$p)/sum(diag(P)))

## Plot
py_pubmlm <- rma.mv(
  yi, vi,
  mods   = ~ year,
  random = list(~1|effect_id, ~1|study_id),
  data   = es_pubonly,
  method = "REML"
)

df <- data.frame(
  year = seq(
```

```
    from       = min(es_pubonly$year, na.rm=TRUE),
    to         = max(es_pubonly$year, na.rm=TRUE),
    length.out = 100                    # you only need ~100 points for a smooth line
  )
)

pr <- predict(py_pubmlm, newmods = df$year)
df$pred  <- pr$pred
df$ci.lb <- pr$ci.lb
df$ci.ub <- pr$ci.ub

es_pubonly <- es_pubonly %>%
  mutate(weight = 1/vi)

ggplot() +
  geom_point(
    data  = es_pubonly,
    aes(x = year, y = yi, size = weight),
    shape = 21, fill = "grey40", color = "black", alpha = 0.4
  ) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_ribbon(
    data = df,
    aes(x = year, ymin = ci.lb, ymax = ci.ub),
    fill  = "grey80",
    alpha = 0.5
  ) +
  geom_line(
    data = df,
    aes(x = year, y = pred),
    size  = 1.2,
    color = "black"
  ) +
  scale_size_continuous(
    range = c(1, 5),
    name  = "Precision\n(1/vi)",
    guide = guide_legend(override.aes = list(alpha = 1))
  ) +
  scale_x_continuous(
    breaks = seq(2000, 2025, by = 5),
    limits = range(es_pubonly$year, na.rm = TRUE)
  ) +
  scale_y_continuous(
    breaks = seq(-1.0, 0.5, by = 0.2),
    limits = c(-1.0, 0.5)
  ) +
  labs(
    x     = "Publication Year",
    y     = "Standardized Mean Difference (d)"
  ) +
```

```
  theme_minimal(base_family = "Times New Roman")

### Unpublished Studies
es_unpubonly <- subset(es, publication_status == "unpublished")

### Model
py_unpubmlm <- rma.mv(
  yi, vi,
  mods = ~ year_centered,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_unpubonly,
  method = "REML"
)

### Model Results
summary(py_unpubmlm)

#### LRT
#### Published Only Model, ML
pub_mlm2 <- rma.mv(
  yi, vi,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_pubonly,
  method = "ML"
)

#### Publication Year on Published Only, ML
py_pubmlm2 <- rma.mv(
  yi, vi,
  mods = ~ year_centered,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_pubonly,
  method = "ML"
)

#### LRT
anova(pub_mlm2, py_pubmlm2)

1 - (sum(py_pubmlm$sigma2) / sum(pub_mlm$sigma2))

###### Published Only, REML
pub_mlm <- rma.mv(
  yi, vi,
  random = list(~ 1 | effect_id, ~ 1 | study_id),
  data = es_pubonly,
  method = "REML"
)

##### Model Summary
summary(pub_mlm)
```

```
#####I^2
W <- diag(1/pub_mlm$vi)
X <- model.matrix(pub_mlm)
P <- W - W %*% X %*% solve(t(X) %*% W %*% X) %*% t(X) %*% W
100 * sum(pub_mlm$sigma2) / (sum(pub_mlm$sigma2) + (pub_mlm$k-
pub_mlm$p)/sum(diag(P)))


#########################################
######## SPECIFICATION ANALYSIS ########
#########################################
data <-
read_excel("/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/Coding_Multiverse.xl
sx")

# Numeric Columns
numeric_vars <- c(
  "n1", "n2", "m1", "sd1", "m2", "sd2"
)

# Convert Columns
data <- data %>%
  mutate(across(all_of(numeric_vars), as.numeric)) %>%
  mutate(across(-all_of(numeric_vars), as.factor))

##############################################################################
#######################
# Code for a meta-analytic specification curve analysis in Voracek, Kossmeier, & Tran (2018)
#
# Which data to meta-analyze, and how? A specification-curve and multiverse-analysis
approach    #
# to meta-analysis                                                #
# Contact: michael.kossmeier@univie.ac.at                                    #
# Code last edited: July 31, 2018                                        #
##############################################################################
#######################

# Calculate d and variance
es_mvd <- escalc(
  measure = "SMD",   # Cohen's d
  m1i = m1, sd1i = sd1, n1i = n1,
  m2i = m2, sd2i = sd2, n2i = n2,
  data = data
)

# Calculate g and variance
es_mvg <- escalc(
  measure = "SMDH",  # Hedges g
```

```
  m1i = m1, sd1i = sd1, n1i = n1,
  m2i = m2, sd2i = sd2, n2i = n2,
  data = data
)

es_mvd <- es_mvd %>% rename(yi_d = yi, vi_d = vi)
es_mvg <- es_mvg %>% rename(yi_g = yi, vi_g = vi)

x <- cbind(data, es_mvd[, c("yi_d", "vi_d")], es_mvg[, c("yi_g", "vi_g")])

head(x)


### Baseline Model
tlm <-  rma(yi_d, vi_d,
          data = es_mvd)

tlm

# Construct grouping factors (internal, which factors)
comparison_group <- c("control", "removal", "cg:either")
ability_controlled <- c("no", "yes", "ac:either")
educationstatus <- c("college", "mixed", "precollege", "es:either")
setting <- c("field", "lab", "online", "set:either")
study_location <- c("africa", "asia", "europe", "northern america", "oceania", "sl:either")
publication_status <- c("published", "unpublished", "ps:either")
ST_activation <- c("blatant", "subtle", "sta:either")

# different methods (external, how factors)
effect <- c("Cohen's d", "Hedges g")
ma_method <- c("REM", "FEM")

# Construct all possible combinations of internal and external factors
specifications <- expand.grid(comparison_group = comparison_group, ability_controlled =
ability_controlled,
                         educationstatus = educationstatus, setting = setting, study_location =
study_location,
                         publication_status = publication_status, ST_activation = ST_activation,
                         effect = effect, ma_method = ma_method)
specifications <- data.frame(specifications,
                         mean = rep(NA, nrow(specifications)),
                         set = rep(NA, nrow(specifications)),
                         lb = rep(NA, nrow(specifications)),
                         ub = rep(NA, nrow(specifications)),
                         p = rep(NA, nrow(specifications)),
                         k = rep(NA, nrow(specifications)))


# Conduct specification analyses
for(i in 1:nrow(specifications)) {
```

```
dat <- x

######################################################################
#### Determine specification subset by using "Which" factors #######
######################################################################
# For comparison group
if(specifications$comparison_group[i] == "control") {
  dat <- dat[dat$comparison_group == "control", ]
} else if(specifications$comparison_group[i] == "removal") {
  dat <- dat[dat$comparison_group == "removal", ]
}

# For ability controlled
if(specifications$ability_controlled[i] == "no") {
  dat <- dat[dat$ability_controlled == "no", ]
} else if(specifications$ability_controlled[i] == "yes") {
  dat <- dat[dat$ability_controlled == "yes", ]
}

# For education status
if(specifications$educationstatus[i] == "precollege") {
  dat <- dat[dat$educationstatus == "precollege", ]
} else if(specifications$educationstatus[i] == "college") {
  dat <- dat[dat$educationstatus == "college", ]
} else if(specifications$educationstatus[i] == "mixed") {
  dat <- dat[dat$educationstatus == "mixed", ]
}

# For setting
if(specifications$setting[i] == "lab") {
  dat <- dat[dat$setting == "lab", ]
} else if(specifications$setting[i] == "field") {
  dat <- dat[dat$setting == "field", ]
} else if(specifications$setting[i] == "online") {
  dat <- dat[dat$setting == "online", ]
}

# For study location
if(specifications$study_location[i] == "africa") {
  dat <- dat[dat$study_location == "africa", ]
} else if(specifications$study_location[i] == "asia") {
  dat <- dat[dat$study_location == "asia", ]
} else if(specifications$study_location[i] == "europe") {
  dat <- dat[dat$study_location == "europe", ]
} else if(specifications$study_location[i] == "northern america") {
  dat <- dat[dat$study_location == "northern america", ]
} else if(specifications$study_location[i] == "oceania") {
  dat <- dat[dat$study_location == "oceania", ]
}
```

```
  # For publication status
  if(specifications$publication_status[i] == "published") {
    dat <- dat[dat$publication_status == "published", ]
  } else if(specifications$publication_status[i] == "unpublished") {
    dat <- dat[dat$publication_status == "unpublished", ]
  }

  # For threat activation (ST_activation)
  if(specifications$ST_activation[i] == "subtle") {
    dat <- dat[dat$ST_activation == "subtle", ]
  } else if(specifications$ST_activation[i] == "blatant") {
    dat <- dat[dat$ST_activation == "blatant", ]
  }

  # only compute meta-analytic summary effects for specification subsets with at least two
studies/samples.
  if(nrow(dat) < 2) next

  # Save which study/sample IDs were selected by the "Which" factors for a given
specification.
  specifications$set[i] <- paste(rownames(dat), collapse = ",")

  ##########################################################################
  #### Determine specification analyses by using "How" factors ##########
  ##########################################################################
  # --- "How"-Faktoren korrekt auswerten ---
  if (specifications$effect[i] == "Cohen's d") {
    if (specifications$ma_method[i] == "FEM") {
      mod <- rma(
        yi     = dat$yi_d,
        vi     = dat$vi_d,
        method = "FE",
        control = list(stepadj = 0.5, maxiter = 2000)
      )
    } else if (specifications$ma_method[i] == "REM") {
      mod <- rma(
        yi     = dat$yi_d,
        vi     = dat$vi_d,
        method = "REML",
        control = list(stepadj = 0.5, maxiter = 2000)
      )
    }

    specifications$mean[i] <- mod$b[[1]]
    specifications$lb[i]   <- mod$ci.lb[[1]]
    specifications$ub[i]   <- mod$ci.ub[[1]]
    specifications$p[i]    <- mod$pval[[1]]
    specifications$k[i]    <- nrow(dat)

  } else if (specifications$effect[i] == "Hedges g") {
```

```
    if (specifications$ma_method[i] == "FEM") {
      mod <- rma(
        yi     = dat$yi_g,
        vi     = dat$vi_g,
        method = "FE",
        control = list(stepadj = 0.5, maxiter = 2000)
      )
    } else if (specifications$ma_method[i] == "REM") {
      mod <- rma(
        yi     = dat$yi_g,
        vi     = dat$vi_g,
        method = "REML",
        control = list(stepadj = 0.5, maxiter = 2000)
      )
    }

    specifications$mean[i] <- mod$b[[1]]
    specifications$lb[i]   <- mod$ci.lb[[1]]
    specifications$ub[i]   <- mod$ci.ub[[1]]
    specifications$p[i]    <- mod$pval[[1]]
    specifications$k[i]    <- nrow(dat)
  }
}

# Only keep specifications with at least 2 studies/samples
specifications_full <- specifications[complete.cases(specifications),]

# Only keep unique study/sample subsets resulting from "Which" factor combinations.
specifications_full <- specifications_full[!duplicated(specifications_full[, c("mean", "set",
"effect", "ma_method")]), ]
# Indicator if all studies are included in the set
specifications_full$full_set <- as.numeric(specifications_full$set == paste(1:nrow(x), collapse
=",", sep = ""))

# Save specification data
write.csv2(specifications_full,
"/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/R/specifications_ST.csv")



# plot specification curve (Voracek et al., 2019) ----
specifications_full <-
read.csv2("/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/R/specifications_ST.cs
v")
specifications_full <- specifications_full[, -1]




# grouping factors
comparison_group <- c("control", "removal", "cg:either")
ability_controlled <- c("no", "yes", "ac:either")
```

```
educationstatus <- c("college", "mixed", "precollege", "es:either")
setting <- c("field", "lab", "online", "set:either")
study_location <- c("africa", "asia", "europe", "northern america", "sl:either")
publication_status <- c("published", "unpublished", "ps:either")
ST_activation <- c("blatant", "subtle", "sta:either")


unique(specifications_full$effect)
unique(specifications_full$ma_method)


# methods
effect <- c("Cohen's d", "Hedges g")
ma_method <- c("REM", "FEM")

x_rank <- rank(specifications_full$mean, ties.method = "random")
yvar <- rep(factor(rev(c(comparison_group, ability_controlled, educationstatus, setting,
                    study_location, publication_status, ST_activation, effect, ma_method)),
                levels = rev(c(comparison_group, ability_controlled, educationstatus, setting,
                          study_location, publication_status, ST_activation, effect, ma_method))),
          times = nrow(specifications_full))
xvar <- rep(x_rank, each = length(levels(yvar)))
spec <- NULL

# determine which specifications are observed and which are not
for(i in 1:nrow(specifications_full))
{id <- as.numeric(levels(yvar) %in% as.character(unlist(specifications_full[i, 1:9])))
spec <- c(spec, id)}

plotdata <- data.frame(xvar, yvar, spec)

ylabels <- rev(c("Control Group", "Removal Group", "Either Comparison Group", "Ability
Unadjusted",
            "Ability Adjusted", "Either Ability Control", "College Sample",
            "Mixed Sample", "School Sample", "All Sample Types", "Field Study", "Lab
Study",
            "Online Study", "Either Study Setting", "Africa", "Asia", "Europe",
            "North America", "Either Study Location", "Published", "Unpublished",
            "Either Publication Status", "Blatant Threat", "Subtle Threat",
            "Either Threat", "Cohen's d", "Hedges' g", "REM", "FEM"
))

plotdata$k <- rep(specifications_full$k, each = length(levels(yvar)))
plotdata$fill <- as.factor(plotdata$k * plotdata$spec)
cols <- RColorBrewer::brewer.pal(min(11, length(levels(plotdata$fill)) - 1), "Spectral")

# create specification tile plot
p_spec <- ggplot(data = plotdata, aes(x = xvar, y = as.factor(yvar), fill = fill)) +
  geom_raster() +
  geom_hline(yintercept = c(2, 4, 7, 10, 15, 19, 23, 26, 29, 32) + 0.5) +
```

```
  scale_x_continuous(position = "bottom") +
  scale_y_discrete(labels = ylabels) +
  scale_fill_manual(values = c("white", c(cols[floor(seq(from = 1, to = length(cols), length.out
= length(levels(plotdata$fill)) - 1))]))) +
  labs(x = "Specification Number", y = "Which/How Factors") +
  coord_cartesian(expand = F, xlim = c(0.5, nrow(specifications_full) + 0.5)) +
  theme_bw() +
  theme_minimal(base_family = "Times New Roman")+
  theme(legend.position = "none",
        axis.text = element_text(colour = "black"),
        axis.ticks = element_line(colour = "black"),
        plot.margin = margin(t = 5.5, r = 5.5, b = 5.5, l = 5.5, unit = "pt"))

# create summary forest plot ----------------------------------------------------
specifications_full$xvar <- x_rank
yrng <- range(c(0, specifications_full$lb, specifications_full$ub))
ylimit <- c(yrng[1] - diff(yrng)*0.1, yrng[2] + diff(yrng)*0.1)

placeholder <- paste(rep(" ", 41), collapse = "")

y_breaks_forest <- round(seq(from = round(ylimit[1], 1), to = round(ylimit[2], 1), by = 0.2),
2)
y_labels_forest <- format(y_breaks_forest, nsmall = 2)
y_breaks_forest <- c(ylimit[1], y_breaks_forest)
y_labels_forest <- c(placeholder, y_labels_forest)


p_forest <-
  ggplot(data = specifications_full, aes(x = xvar, y = mean)) +
  geom_errorbar(aes(ymin = lb, ymax = ub, col = as.factor(k)), width = 0, linewidth = 0.25) +
  geom_line(col = "black", linewidth = 0.25) +
  geom_hline(yintercept = 0, linetype = 2, linewidth = 0.25) +
  scale_x_continuous(name = "") +
  scale_y_continuous(name = "Summary effect (d)", breaks = y_breaks_forest, labels =
y_labels_forest) +
  scale_color_manual(values = c(cols[floor(seq(from = 1, to = length(cols),
                                length.out = length(levels(as.factor(specifications_full$k)))))]))
+
  coord_cartesian(ylim = ylimit, xlim = c(0.5, nrow(specifications_full) + 0.5), expand =
FALSE) +
  theme_bw() +
  theme_minimal(base_family = "Times New Roman")+
  theme(legend.position = "none",
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(),
```

```
        panel.grid.minor.y = element_blank(),
        plot.margin = margin(t = 5.5, r = 5.5, b = -15, l = 5.5, unit = "pt"))

# Display plot
p_forest

yrng <- range(c(2, max(specifications_full$k)))
ylimit <- c(2, max(specifications_full$k))

y_breaks_size <- c(17, 62, 123)
y_labels_size <- format(y_breaks_size, nsmall = 0)
y_breaks_size <- c(ylimit[1], y_breaks_size)
y_labels_size <- c(placeholder, y_labels_size)

p_size <-
  ggplot(data = specifications_full, aes(x = xvar, y = k)) +
  geom_area(fill = "gray35", color = "black", size = 0.25) +
  scale_x_continuous(name = "") +
  scale_y_continuous(name = "# Samples", breaks = y_breaks_size, labels = y_labels_size) +
  coord_cartesian(ylim = ylimit, xlim = c(0.5, nrow(specifications_full) + 0.5), expand =
FALSE) +
  theme_bw() +
  theme_minimal(base_family = "Times New Roman")+
  theme(legend.position = "none",
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_text(colour = "black"),
        axis.ticks.y = element_line(colour = "black"),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(),
        panel.grid.minor.y = element_blank(),
        plot.margin = margin(t = 5.5, r = 5.5, b = -15, l = 5.5, unit = "pt"))

# Combine specfication tile plot, subset size indicator and forest plot
p <- gridExtra::arrangeGrob(p_spec, p_size, p_forest,
                            layout_matrix = matrix(c(3, 3, 3, 2, 1, 1, 1, 1, 1), ncol = 1))

head(specifications_full)
specifications_full$mean
specifications_full$lb
specifications_full$ub

p <- ggpubr::as_ggplot(p)
p


ggsave("/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/R/ST_specforest.png", p,
width = 16*1.2, height = 12*1.2, dpi = 600, units = "cm")
```

```
########################
####### GOSH PLOT #######
########################

# Load Two-Level Data
data <-
read_excel("/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/Coding_Multiverse.xl
sx")

# Numeric Variables
numeric_vars <- c(
  "n1", "n2", "m1", "sd1", "m2", "sd2"
)

data <- data %>%
  mutate(across(all_of(numeric_vars), as.numeric)) %>%
  mutate(across(-all_of(numeric_vars), as.factor))

# Compute Effect SIzes
es_mvd <- escalc(
  measure = "SMD",
  m1i = m1, sd1i = sd1, n1i = n1,
  m2i = m2, sd2i = sd2, n2i = n2,
  data = data
)

## Two-Level Model
tlm <- rma(yi = yi, vi = vi,
           measure="SMD",
           data = es_mvd)

## Model Summary
summary(tlm)

### Outliers
inf <- influence(tlm)

### Plot
plot(inf)

### Summary
inf

#### GOSH
png("/Users/beatriceschreier/Documents/Uni/SS23/Masterarbeit/R/GOSH-Rydell.png", width
= 800, height = 600, res = 100)
gosh <- gosh(tlm, subset = 100000)
plot(gosh)
dev.off()
```

```
# New DF Without Outlier (Rydell et al.. 2009)
df_rydell <- es_mvd[-c(68, 69), ]

# New Two-Level Model
tlm_r <- rma(yi = yi, vi = vi, data = df_rydell, method = "REML")

# Model Summary
summary(tlm_r)

##################################################################
####### BIAS ANALYSES BASED ON PETKARI ET AL. (2023) #######
##################################################################
# Numeric Columns
numeric_vars <- c(
  "year", "n1", "n2", "age_mean",
  "sexcomposition", "m1", "sd1", "m2", "sd2",
  "taskdifficulty"
)

data <- data %>%
  mutate(across(all_of(numeric_vars), as.numeric)) %>%
  mutate(across(-all_of(numeric_vars), as.factor))

## Compute Cohen's ds and Sampling Variances
es <- escalc(
  measure = "SMD",
  m1i = m1, sd1i = sd1, n1i = n1,
  m2i = m2, sd2i = sd2, n2i = n2,
  data = data
)

#### Published-Only Subset
es_pub <- subset(es, publication_status == "published")

### General Model Without Predictors
gmlm_pub <- rma.mv(yi, vi, data = es_pub,
               random = list(~ 1 | effect_id, ~ 1 | study_id),
               method = "REML")


summary(gmlm_pub)

##### Multilevel Egger's test according to Rodgers and Pustejovsky (2021)
es_pub$sei <- sqrt(es_pub$vi)

mean(es_pub$sei)
sd(es_pub$sei)

egger.gmlm <- rma.mv(yi ~ 1 + sei,
                   V = vi,
```

```
                    data = es_pub,
                    random = list(~ 1 | reference_id/effect_id),
                    test = "t")
egger.gmlm

###### Funnel Plot
library(randomcoloR)

n_studies <- length(unique(es_pub$reference_id))
cols      <- distinctColorPalette(n_studies)
cols_each <- cols[ as.numeric(factor(es_pub$reference_id)) ]


metafor::funnel(
  gmlm,
  pch     = NA,                        # keine Standard-Punkte zeichnen
  level   = c(0.90, 0.95, 0.99),
  shade   = c("gray98","gray90","gray80"),
  back = "gray95",
  refline = 0,
  yaxt    = "n",
  xlab    = "Cohen's d",
  xlim    = c(-2.25,2.25),
  ylab = ""
)

box()

max_se   <- max(sqrt(es$vi), na.rm=TRUE)
se_ticks <- seq(0, ceiling(max_se*5)/5, by=0.2)
axis(2, at = se_ticks, labels = sprintf("%.1f", se_ticks), las = 1)
mtext("Standard Error", side = 2, line = 2.8)

points(
  unpub$yi,
  sqrt(es_pub$vi),
  pch  = 21,            # gefüllte Kreise
  bg = adjustcolor(cols_each, alpha.f = 0.99),
  col  = "gray20",       # Randfarbe der Punkte
  cex  = 0.8,
  lwd = 0.5
)

abline(v=gmlm_pub$b,lty=2, lwd=1)
abline(egger.gmlm,lty=2,col="darkred",lwd=1)

subset(es_pub, sei > 0.6)


### Egger's sandwich test according to Rodgers and Pustejovsky (2021)
```

```
egger.sand <- robu(formula = yi ~  1 + sei,
                  data = es_pub,
                  studynum = reference_id,
                  var.eff.size = vi)

egger.sand

citation(package='robumeta')

# 3PSM based on sampled data

# get row of es
rows   <- 1:nrow(es_pub)
### get study IDs
sid <- unique(es_pub$study_id)
sn <- c()
for(i in 1:length(sid)) {
  set.seed(i)
  sn <- c(sn, sample(rows[es_pub$study_id == sid[i]], 1))
}

### get sampled data
es_samp <- es_pub[sn, ]

### get 3PSM
psm_fit <- with(es_samp, weightfunct(yi, vi, steps = c(.025, 1)))
lrchisq <- 2 * (abs(psm_fit[[1]]$value - psm_fit[[2]]$value))
pval <- 1 - pchisq(lrchisq, 1L)

### results
psm_fit[[2]]$par[3]
lrchisq
pval
```