Journal of Advanced Academics
1–16
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1932202X211061517
journals.sagepub.com/home/joaa
SAGE

No Strong Evidence of Stereotype Threat in Females: A Reassessment of the Picho-Kiroga et al. (2021) Meta-Analysis

Russell T. Warne

Abstract

Recently, Picho-Kiroga (2021) published a meta-analysis on the effect of stereotype threat on females. Their conclusion was that the average effect size for stereotype threat studies was $d = .28$, but that effects are overstated because the majority of studies on stereotype threat in females include methodological characteristics that inflate the apparent effect size. In this response, I show that Picho-Kiroga et al. (2021) committed fundamental errors in their meta-analysis that undermine confidence in the article and warrant major corrections. But even if the data were not flawed, the conclusion that Picho-Kiroga et al. (2021) should have reached is that their results are most consistent with a population effect size of zero. There is no compelling evidence that stereotype threat is a real phenomenon in females.

Keywords

stereotype threat, statistical power, effect sizes, statistical methods, females

Recently, the *Journal of Advanced Academics* published a meta-analysis by Picho-Kiroga et al. (2021) on stereotype threat in females taking tests and assessments that measure performance in mathematics. The authors concluded that methodological shortcomings of primary studies inflated the apparent strength of stereotype threat

Utah Valley University, Orem, Utah, USA.

Corresponding Author:
Russell T. Warne, Department of Behavioral Science, Utah Valley University, 800 W. University Parkway MC 115, Orem, UT 84058.
Email: rwarne@uvu.edu

effects, and that "... the common notion that stereotype threat significantly contributes to gender gaps in STEM is more than likely an overstatement" (Picho-Kiroga et al., 2021, p. 253). Nevertheless, Picho-Kiroga et al. (2021) affirmed the reality of the stereotype threat, while acknowledging that other psychosocial influences might have a more powerful effect on females' academic achievement in mathematics and related areas.

In this response, I show that Picho-Kiroga et al. (2021) committed several methodological errors. I also show that their conclusions do not go far enough: not only are stereotype threat effects in females overstated, the data are most consistent with a hypothesis that stereotype threat effects in females are not real. Any apparent effects can be explained by publication bias, abysmally low statistical power in studies, and questionable research practices, such as *p*-hacking.

Background for the Picho-Kiroga et al. (2021) Meta-Analysis

Stereotype threat was a phenomenon first described by Steele and Aronson (1995) in a well cited article that reported four laboratory studies. In these studies, Steele and Aronson (1995) gave reminders—some subtle, some not—of negative stereotypes about the cognitive prowess of African Americans to a mixed group of African American and White examinees. Across all four studies, Steele and Aronson (1995) reported that exposure to these reminders lowered scores in African Americans, while African American examinees who experienced neutral prompts had unchanged scores. The prompts had no apparent effect on White examinees' scores.

Steele & Aronson's (1995) article garnered widespread attention, and many commentators and social sciences have seen stereotype threat as a cause of lower average academic achievement and/or test scores for Black Americans, Hispanics, and (in mathematics domains) females (e.g., Cross & Cross, 2017; Dweck, 2009), and some researchers have produced studies supporting this conclusion (e.g., Spencer et al., 2016; Walton & Spencer, 2009). However, this interpretation has been contested, with one reason being that Steele & Aronson's (1995) study showed that stereotype threat created *new* score gaps (Sackett et al., 2004), contrary to Steele's (1997) claim that stereotype threat explains existing gaps. Skeptics have also argued that (1) stereotype threat effects explain only a small percentage of the score difference between groups, (2) the methodological variation in studies makes it difficult to draw conclusions, and (3) the causal mechanism of how stereotype threat could lower scores—and for whom and under what conditions—is not clear (Ganley et al., 2013; Warne, 2020; Wax, 2009; Whaley, 1998).

Skeptics of stereotype threat also have data to support their beliefs. Large-scale attempts to observe stereotype threat on actual high-stakes tests have been unsuccessful (e.g., Walker and Bridgeman, 2008), and as testing situations become more realistic, the effect diminishes to the point where the effect size is close to zero (Shewach et al., 2019). The case against stereotype threat has grown in recent years as psychology has grappled with the replication crisis (see Nelson et al., 2018, for a thorough treatment of the history and progression of the replication crisis in psychology).

While exact replications of stereotype threat studies remain rare (see Finnigan & Corker, 2016; Gibson et al., 2014; and Moon & Roeder, 2014, for exceptions), it is apparent that stereotype threat research shares many of characteristics found in research that does not replicate, including small sample sizes, high researcher flexibility in creating studies and analyzing data, and strong social incentives to find statistically significant effects (Warne, 2020).

Re-Analysis of Picho-Kiroga et al. (2021)

It is against this backdrop that Picho-Kiroga et al. (2021) conducted their meta-analysis. Taken at face value, their work supports the existence of a small, but detectable, influence from stereotype threat. Picho-Kiroga et al. (2021) also suggested possible improvements for future studies and new avenues of research. However, there are deficiencies in the article that make some of their conclusions untenable. Instead, Picho-Kiroga et al.'s (2021) meta-analysis is consistent with the null hypothesis and also consistent with the view that apparent effects are due to publication bias and questionable research practices. My analysis of their work focuses on methodological errors and logical problems that undermine the article's central thesis; I will not touch upon every methodological decision the authors made.

Positive Aspects of the Meta-Analysis

Before dissecting Picho-Kiroga et al.'s (2021) errors, it is important to give credit where credit is due. I applaud the authors for recognizing the deficiencies in much of the stereotype threat literature. Most adherents to stereotype threat theory ignore the low methodological quality of the research in their surveys of the research on the topic (e.g., Spencer et al., 2016). Picho-Kiroga et al. (2021) are correct that most studies on the topic suffer from low statistical power and theoretically deficient research designs. People who engage in research on stereotype threat would be wise to incorporate Picho-Kiroga et al.'s (2021) suggestions.

Picho-Kiroga et al. (2021) also made some methodological decisions that are commendable. Focusing on heterogeneity in the research on stereotype threat was a valuable contribution because subjective methodological choices often have an influence on a particular study's results. Understanding the causes of this heterogeneity may help future researchers isolate the signal of a true effect from the noise of methodological heterogeneity.

Picho-Kiroga et al. (2021) are also correct in some of their interpretations. In particular, the researchers' view that many individual studies of stereotype threat show inflated effect sizes is correct. Stereotype threat researchers would benefit from tempering their—and others'—expectations about the power for stereotype threat to cause average score gaps between males and females, and likely between other demographic groups (Warne, 2020).

It is also important to applaud Picho-Kiroga and her colleagues for their behavior in the face of critiques of their meta-analysis. I informed the editors of the *Journal of*

Advanced Academics of problems I identified in the article on January 26, 2021—two days after the article was originally published online. The editors informed Picho-Kiroga and her colleagues of the errors, and some of the mistakes were corrected promptly. The current revised version of the article was issued online no later than February 15, 2021, and appeared in print the next month. Picho-Kiroga responded quickly when I requested the data and syntax files for the meta-analysis. I encourage other scientists who find their work under scrutiny to display the same promptness and civility that Picho-Kiroga did. Even though I disagree with their conclusions, I have respect towards Picho-Kiroga and her colleagues for their willingness to have their work subjected to scrutiny.

Methodological Problems

However, the problems in Picho-Kiroga et al.'s (2021) meta-analysis severely undercut their main thesis that stereotype threat effects are a small but consistent influence that could depress some females' scores on mathematics tests. In this subsection, I will focus on the authors' effect size calculations, their erroneous understanding and use of statistical power, problems in compiling and reporting data, and deficient tests for publication bias.

Errors in calculating effect sizes. When I investigated Picho-Kiroga et al.'s (2021) dataset, it was apparent that the effect sizes were not correct. The reason for these errors is not always clear, but I identified a few errors that occurred multiple times. One error in their dataset was to miscode experimental and control groups, resulting in six effect sizes that showed the opposite sign as the correct results.[1] When I compared their data file to the summary statistics reported in the original publications, I identified six effect sizes that were miscalculated.

These errors aside, Picho-Kiroga et al. (2021) reported the wrong effect size, claiming that their effect sizes were Cohen's *d* values, when in reality they were Hedges's *g** values. While Hedges's *g** is a legitimate choice of effect size, reporting the incorrect statistic is a factual inaccuracy that distorts' readers' understanding of the data, especially because Picho-Kiroga et al. (2021, p. 243) explicitly stated that they used the Cohen's *d* formula and made no mention of Hedges's *g** or the fact that they used a completely different formula when calculating their effect sizes.

Errors in calculating statistical power. Statistical power is the capacity for a study to reject a false null hypothesis (Warne, 2021). Studies with low statistical power are uninformative because they are unlikely to reject a false null hypothesis. Therefore, if they retain a null hypothesis, it is not clear whether that is because such an outcome was always likely or because the null hypothesis is really true. When studies with low statistical power reject a null hypothesis, the result is equally uninformative because their effect sizes and test statistics are unstable, making many statistically significant results the product of capitalization on chance and sampling error. This is why conducting one large study with high statistical power is much more informative than conducting multiple smaller studies with low statistical power (Schimmack, 2012). Assuming that practical constraints of time, money, and logistics are not important considerations, higher statistical power is always preferable to low statistical power.

The use of Hedges's g^* instead of Cohen's d introduced inaccuracies into the statistical power calculations in Picho-Kiroga et al.'s (2021) study. The authors used the computer program G-Power to calculate their power statistics, but the program requires the use of Cohen's d to produce correct power estimates. This discrepancy would make all of the statistical power calculations and the analyses that used them, such as the meta-regression or the moderator analysis, incorrect.

But this discrepancy does not fully account for the statistical power results reported by Picho-Kiroga et al. (2021). The statistical power values they report are often not realistic. The most obvious indication is that the mean observed power was .59, but the weighted mean effect size was .28, and the median sample size was $n = 40$. Using these effect size (d) and n values to calculate power of the typical study produces an estimate of .219. When statistical power (based on a mean effect size of $d = .28$) was calculated for each study, the average power was .189. Both of these numbers are less than half of the mean power that Picho-Kiroga et al. (2021) reported. As a result, the percentage of effect sizes that came from studies with statistical power of .80 or higher is also inflated. Instead of 23 of 102 (or 22.5%) effect sizes having power of .80 or higher, the true number (assuming a population effect size of $d = .28$) is just one. (Tellingly, the effect size for this study is $d = \text{-}.069$.)

The first author stated in an email that she used G-Power to calculate post hoc power, using a study's observed effect size (erroneously g^* instead of d) and the observed group sample sizes (K. Picho-Kiroga, personal communication, June 3, 2021). Using this procedure does not reproduce the statistical power calculations in the authors' data file. Furthermore, examination of the data file confirms the problems with their estimates; some of the power values are mathematically impossible. One common error is for different samples—with different n and effect size values—to have the same power estimate. Calculating incorrect power estimates is another error that compromises the statistical power estimates and all analyses that used these values.

Conceptual errors in using statistical power. At several points in their article, Picho-Kiroga et al. (2021) warned their readers about the problems of low statistical power. This is why it puzzling that they do not fully understand the difference between two types of statistical power: *a priori* power and *post hoc* (or *observed*) power. A priori power estimates the probability that a study could reject a null hypothesis, given an expected effect size (which is a metric of how wrong the null hypothesis is). This expected effect size can be drawn from the results of an earlier study, the average effect size in a meta-analysis, a value based on theoretical considerations (such as the smallest effect size of interest), or an arbitrary benchmark, such as Cohen's (1988) widely used standards of "small," "medium," and "large" effect sizes (though see Warne et al., 2012, for criticisms of these benchmarks). The mathematical process is the same; what differs is the source of the effect size and sample size data.

The apparently subtle difference in the choice of effect size is actually extremely important. Using the observed effect size for calculating statistical power capitalizes on sampling error and results in a power estimate that is nothing more than a mathematical transformation of the p-value (Hoenig & Heisey, 2001). The statistical

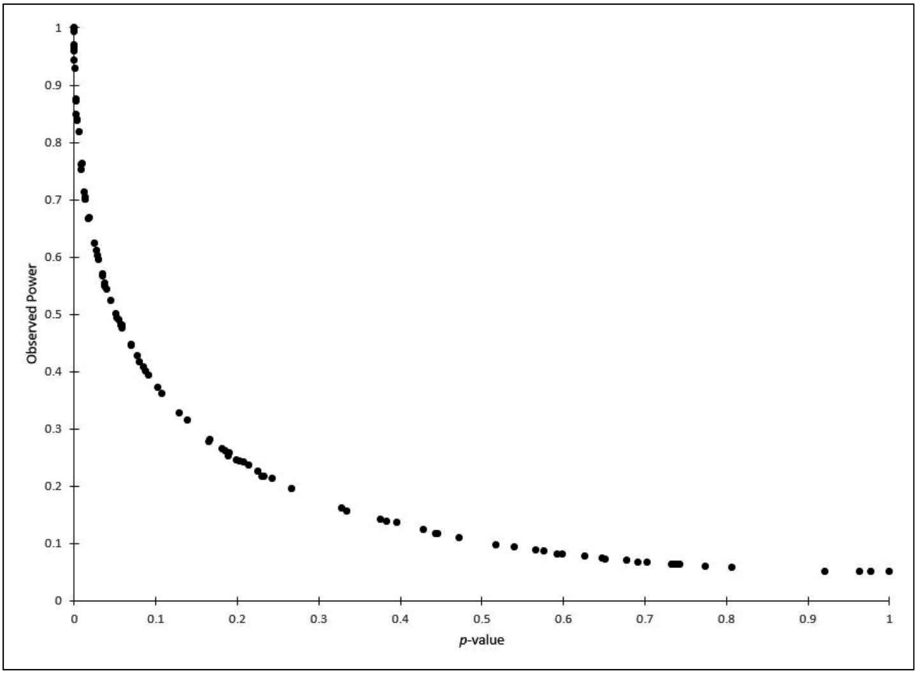

Figure 1. Scatterplot showing the relationship between an effect size's *p*-value and observed power. Note that effect sizes are Cohen's *d* values recalculated from the correct descriptive statistics in the body of studies in Picho-Kiroga et al.'s (2021) meta-analysis.

relationship between observed power and the *p*-value is clear in Figure 1, which is a scatterplot showing the relationship between the two statistics for the effect sizes in the Picho-Kiroga et al. (2021) data. Because of this mathematical relationship, post hoc power values are inflated when the null hypothesis is rejected. Additionally, post hoc power is uninformative in meta-analysis because the average post hoc power in a body of studies is approximately equal to the proportion of rejected null hypotheses (Schimmack, 2012).

A priori power, in contrast, tells researchers the actual probability that a study would reject a null hypothesis, given the effect size and other methodological characteristics. Because that effect size is not derived from the observed effect size in the study, it cannot merely be a transformation of *p*-values (except in the rare case where the selected effect size is precisely equal to the study's observed effect size).

Picho-Kiroga et al. (2021) admitted to using post hoc power (p. 243) in all their calculations, which was confirmed in an email (K. Picho-Kiroga, personal communication, June 4, 2021). In January 2021, when the editors informed the authors of the problem of using post hoc power, the authors' only change was to alter the phrase "post hoc power" to "observed power." Needless to say, this does not fix the problem of using the wrong type of power because the problem is one of statistics, not semantics.

This error has downstream consequences when the statistical power calculations are used in later analyses. For example, Picho-Kiroga et al. (2021) used post hoc power calculations as a source of data to measure of methodological quality of studies. This had the effect of making studies that rejected the null hypothesis have higher methodological quality scores because of the relationship between observed value of the p-value. Coding studies with higher post hoc power as having higher methodological quality is functionally equivalent to an editor or peer reviewer who incorrectly believes that studies that reject the null hypothesis are better than those that do not. Treating the methodological quality and p-values as being related is a major contributor to publication bias (Dickersin & Min, 1993; Franco et al., 2014).

In reality, methodological quality is independent of the results of a study. A well-designed study does not suddenly become a poorly designed study if the null hypothesis is retained. Conversely, rejecting the null hypothesis does not miraculously change a study with methodological flaws into a high-quality study. But using post hoc power to be a factor in evaluating methodological quality will result in these types of judgments for some studies.

Problems in reporting and compiling data. Comparing the raw data with the summary statistics and data reported in the current version of Picho-Kiroga et al.'s (2021) article reveals many inconsistencies. For example, Rows 9–12 on p. 245 in Table 1 are not in the data file and appear to be duplicate data of the following four rows. Four other rows in the raw data do not correspond to rows in Table 1 and seem to be missing. Sample sizes in Table 1 also often do not match the sample sizes reported in Picho-Kiroga et al.'s dataset. Figure 1 has an inconsistency in reporting 192 full-text articles assessed for eligibility, of which 138 were excluded and 53 were included, but these latter two numbers do not sum to 192. These types of errors caused me to contact the editors of the *Journal of Advanced Academics* and inform them of the problems in the Picho-Kiroga et al. (2021) article. Many of the errors I reported were corrected, though the ones mentioned above remain in the current version. These errors undercut a reader's confidence in the authors' work.

Picho-Kiroga et al. (2021) also had the problem of non-redundant sample members in their data. Some effect sizes share a control group, while others report dependent variables from the same sample members as if they were separate samples. How these non-redundant sample members handled is not explained; they were apparently treated as independent, non-overlapping groups.

Evaluation of publication bias. Another major deficiency in the Picho-Kiroga et al. (2021) meta-analysis is the evaluation of publication bias. To their credit, Picho-Kiroga and her colleagues did use three methods to investigate the possibility of publication: the fail-safe N value, Begg's test, and a funnel plot. Their interpretation of all three analyses was that there was no evidence of publication bias. However, two of these procedures are deficient for examining publication bias, and Picho-Kiroga et al. (2021) misinterpreted the results of two of the three methods.

The fail-safe N procedure attempts to estimate the number of null findings that would have to be suppressed in order for a meta-analytic effect size to move from statistically significant to statistically non-significant. Contrary to Picho-Kiroga et al.'s

(2021, pp. 247–248) claim, a high fail-safe *N* value does *not* indicate a lack of publication bias because the statistic assumes that all missing studies would perfectly fit the null hypothesis—an unlikely proposition because all effect sizes are sample statistics that are subject to variability due to sampling error.[2] Moreover, the fail-safe *N* procedure also assumes that an entire study would be suppressed, when in the real world, studies that produce unfavorable findings can still be published through selective reporting of results and by massaging statistics through questionable research practices (Lakens et al., 2016).

Begg's test is better for detecting publication bias, though mostly because the fail-safe *N* procedure is so bad at it. Begg's test is known to have low statistical power for detecting typical levels of publication bias (Sterne et al., 2000), and it is completely unsurprising that the test would not detect publication bias ($p = .481$) in the meta-analysis. Picho-Kiroga et al. (2021, p. 248) interpreted this result as indicating that there was "no publication bias." However, this interpretation is incorrect; it would be correct to say that Begg's test *failed to detect* any publication bias. Absence of evidence is not evidence of absence.

The best procedure that Picho-Kiroga et al. (2021) used to test for publication bias is the funnel plot. The funnel plot for their study is displayed in Figure 2. Their interpretation was:

> From the funnel plot, it is evident that there are just as many null stereotype threat effects as there are significant effects. That there are studies to the left of the funnel also indicates reverse stereotype threat effects. That is, for some studies, women performed better when exposed to stereotype threat than their control counterparts. (Picho-Kiroga et al., 2021, p. 248)

This is an incorrect interpretation of a funnel plot. Funnel plots are not interpreted by tallying the number of statistically significant and non-significant results and comparing the two. This procedure is called *vote counting* and was discredited long ago (Glass, 1976, 1977; Meehl, 1990). The existence of negative stereotype threat effect sizes is also irrelevant. As a sample statistic with a distribution around a mean and variability measured by a standard deviation, it is natural for there to be some heterogeneity in effect sizes. As a result, it is completely expected for any body of research where the mean effect size is positive and small-to-moderate (which covers many topics in the social sciences and certainly includes stereotype threat) to have some negative effect sizes. Publication bias would have to be 100% efficient to eliminate all of these negative effect sizes; yet, even modest publication bias can inflate effect sizes (Nuijten et al., 2015). Therefore, it is possible for publication bias to coincide with the presence of negative effect sizes.

How should Picho-Kiroga and her colleagues have interpreted their funnel plot? By dividing the plot at the mean effect size (.28) and determining whether there was an asymmetry in the effect sizes on one side compared to the other. To do this, a vertical dashed line representing this mean effect size has been added to Figure 2. A cursory examination of the funnel plot shows an obvious excess of studies (especially with

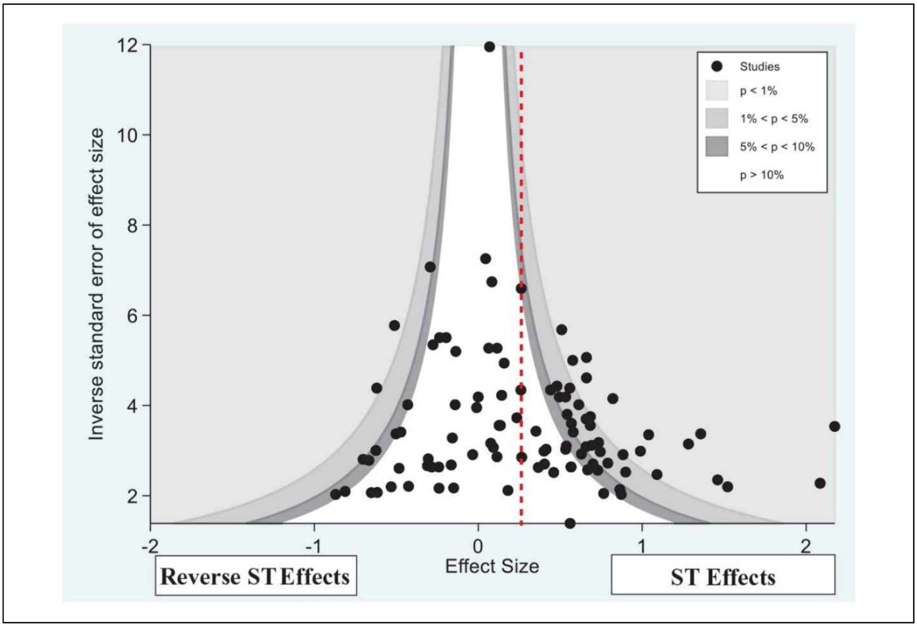

Figure 2. The funnel plot from Picho-Kiroga et al. (2021, p. 249) meta-analysis, modified with a vertical dashed line representing the average effect size of .28. Note that because of undisclosed reverse coding that studies supporting the existence of stereotype threat are shown as having positive effect sizes in this image, whereas in Picho-Kiroga et al.'s (2021) Table 1, these studies have negative effect sizes. This image was modified from Picho-Kiroga et al., 2021, Figure 2.

effect sizes > 0.50). More statistically minded readers will prefer to find the correlation between the absolute value of the effect size and the inverse standard error of the effect size. In the absence of publication bias, this value should be zero; in Picho-Kiroga et al.'s (2021) data file, it is r = -.361 ($p < .001$). Thus, rather than a lack of publication bias, Picho-Kiroga et al.'s (2021) funnel plot indicates the presence of publication bias in the stereotype threat literature. Additionally, the dense population of dots in the funnel plot precisely where p is just below .05 is circumstantial evidence for p-hacking by some stereotype threat researchers.

Another test of publication bias is the test of excessive significance (Ioannidis & Trikalinos, 2007), which compares the expected number of p-values below .05 with the actual number of low p-values in a body of studies. This can be estimated by finding the proportion of results that reject the null hypothesis ($p < .05$) and comparing it to the mean a priori statistical power. In the absence of publication bias, these two values should be equal because the mean a priori power is also the expected proportion of rejected null hypotheses in the population of effect sizes (Schimmack, 2012). In the Picho-Kiroga et al. (2021) data, 31 of 101 effect sizes that I could calculate a p-value for were supportive of stereotype theory and had a $p < .05$. The mean a priori power was .189, indicating that there should be .189(101) = 19.0192 effect sizes with a $p <$ .05 that support stereotype threat theory. Thus, there were about 50% more studies supporting stereotype threat theory than would be expected without publication bias.[3] This result is statistically significant (one variable $\chi^2 = 8.656$, df $= 1, p = .003$), indicating evidence for publication bias in the Picho-Kiroga et al. (2021) meta-analysis.

With this evidence of publication bias, the question then becomes the degree to which publication bias inflates the average effect size. Simulation studies show that in the presence of moderate heterogeneity, publication bias can inflate an observed effect size from zero to $d \approx .30$ or higher (Carter et al., 2019; Renkewitz & Keiner, 2019). Thus, an observed mean effect size of .28, as occurred in Picho-Kiroga et al.'s (2021) meta-analysis is consistent with a true effect size of zero when publication bias is present—as it likely is in the stereotype threat literature.

Logical Deficiencies in Picho-Kiroga et al.'s (2021) Meta-Analysis

In addition to the grave methodological errors, there are also logical deficiencies in the study that make the authors' conclusions erroneous. I will focus on their interpretation of essential components of stereotype threat and how Picho-Kiroga and her colleagues did not recognize that their own data provides evidence against the existence of stereotype threat.

Citing Steele (1997), identified three essential components for stereotype threat to occur in a person:

> Stereotype threat operates under a specific set of conditions. It is contingent upon (a) one's identification with the stereotyped domain, (b) one's awareness of (but not necessarily a belief in) the negative stereotype regarding the stereotyped domain (e.g., women aren't good with numbers), and (c) task difficulty—the task being performed must be relevant to the stereotyped domain and challenging enough to make the treat of stereotype possible. . . (Picho-Kiroga et al., 2021, p. 236)

Picho-Kiroga and her colleagues were savvy in coding the number of these essential conditions that each study in the meta-analysis had. The results, displayed in their Table 2, show that an increasing number of essential conditions in a study led to a weaker stereotype threat effect, until studies with all three essential components had a mean effect size that was statically indistinguishable from zero. The authors interpreted this result to indicate that stereotype threat effects are often inflated and that the true effect may be much weaker than is often seen in studies. While this interpretation is technically correct, it does not go far enough. Instead, the results indicate that stereotype threat effects are so inflated in the literature that the true effect is likely zero.

The finding that stereotype threat effect sizes decrease as essential components of the theory are included in a research design also shows that the "essential" conditions necessary to trigger stereotype threat are not essential at all. If these conditions really were essential for stereotype threat to appear, then the effect sizes would *strengthen* as additional components were added to a study. Instead, adding these "essential" conditions progressively *weakens* the phenomenon until it is indistinguishable from the null hypothesis. This pattern of results is exactly the opposite of what would be predicted by Steele's (1997) theory. This was a strong falsifiability test of stereotype threat theory—and the theory failed it spectacularly.

In the face of this finding, Picho-Kiroga et al. (2021, p. 250, 251) engaged in post hoc theorizing to salvage stereotype threat theory. They claimed that the results they observed could come from the inclusion of sample members who were not susceptible to stereotype threat. However, this is an illogical explanation of their findings. If people not susceptible to stereotype threat are included in a sample, it should attenuate the effect size because their weak or null response to stereotype threat stimuli will dilute the average effect in the study.[4] But, including these people in studies *strengthened* effect sizes in the stereotype threat literature. Even if this were an explanation for the inflated effect sizes in the stereotype threat literature, it only addresses one of the three "essential" conditions that Steele (1997) proposed. It would not explain why studies in which individuals lack an awareness of the stereotype and/or perform a task that is not sufficiently difficult also display inflated effect sizes.

Conclusion

My re-analysis shows that Picho-Kiroga et al.'s (2021) meta-analysis contains severe methodological and logical errors. When the data were re-examined, it is apparent that the results are consistent with a population effect size of zero. The apparent evidence size supporting the existence of stereotype threat is likely the product of publication bias and *p*-hacking in a body of research studies with extremely low statistical power.

Even if Picho-Kiroga et al.'s (2021) meta-analysis is generally correct, the study contradicts the central tenets of stereotype threat theory (Steele, 1997). Adding "essential" components of the theory to a study reduces its effect until it is statistically equal to zero. The authors of the meta-analysis never presented a coherent explanation for why the theory could fail this falsifiability test. Taken together, the evidence indicates that stereotype threat in females is likely not a real phenomenon.

Beyond this specific meta-analysis, there are implications for psychologists and researchers in stereotype threat and related topics. First, stereotype threat researchers should be concerned about the low statistical power in their studies. The median sample size of 40 (i.e., 20 individuals per group) has statistical power of just .1386 to detect an effect size of $d = .28$—indicating that, even if there were a true effect of $d = .28$, a typical study would be almost six times more likely to retain the null hypothesis than reject it. Designing studies that have such a low probability of detecting an effect is methodological negligence; believing that such studies provide evidentiary value borders on incompetence.

Thus, the first takeaway from this re-analysis should be for stereotype threat researchers to increase their statistical power. If the mean effect size of $d = .28$ from Picho-Kiroga et al.'s (2021) meta-analysis is taken as a realistic estimate of the true effect, then the minimum required sample size to achieve a priori statistical power of .80 is 202 subjects per group. To achieve .90 a priori power, a researcher needs 270 subjects per group in a between-subjects design. If the mean effect size is inflated—as even Picho-Kiroga et al. (2021) concluded—then the number of sample members per group increases even further. For a population effect size of $d = .20$, the minimum sample size is 394 people per group to achieve .80 a priori power. If the true population effect size is $d = .10$, then 1,571 people per group are needed to achieve a priori power of .80.

Along with improving statistical power, other aspects of the gender stereotype threat research need to be improved. Including more "essential" components needed to trigger stereotype threat would strengthen the theoretical basis for these studies. Pre-registering studies—especially through the registered report process—is essential for building confidence in the research on a topic that is currently characterized by shoddy research. However, stereotype threat adherents should be warned: so far, pre-registered studies have not supported the theory. Finnigan & Corker's (2016) pre-registered replication of an earlier study showed no evidence for stereotype in females, as did Flore et al.'s (2018) pre-registered study with the largest sample size ever for an experimental study on the topic (n = 1,036 female teenagers) and two replications of a study of stereotype effects in Asian females (Gibson et al., 2014; Moon & Roeder, 2014). The null effects shown in pre-registered studies, including replications, indicate that there is no strong evidence for the existence of gender stereotype threat effects.[5]

Second, the shaky evidence foundation for gender stereotype threat should make scientists cautious about applying stereotype threat theory to real-world situations. In communication to the public and media, researchers should not confidently attribute even a portion of disparate academic outcomes between groups to stereotype threat. Likewise, granting agencies should be more hesitant to fund research on the topic. Spending money to chase after statistical phantoms is not a wise use of financial resources.

The final takeaway from this re-analysis concerns the nature of error detection and correction in science. Picho-Kiroga et al. (2021) touched upon this issue in their article, stating that "self-correction is considered a hallmark of science . . ." and that methodologically poor studies provide "little hope for self-correction . . ." (p. 252). However, the process of correcting Picho-Kiroga et al.'s errors shows that science is *not* self-correcting. "Science" did not inform the editors that there were errors in the original version (or the revised version) of Picho-Kiroga et al.'s (2021) meta-analysis. "Science" did not spend dozens of hours comparing the authors' data to the information reported in different versions of the article, checking data against the original studies, re-calculating statistics, or conducting tests of publication bias. "Science" did not write commentaries explaining the problems of the meta-analysis. And "science" will not issue an expression of concern or corrections to the meta-analysis article. Instead, *people* do these things, often at great inconvenience and opportunity cost. Error correction in science is a time consuming process that only occurs when *people*—authors, readers, editors, and independent researchers—care enough to pursue correction. Even then, correction only occurs if gatekeepers allow corrections.

The scientific community should stop affirming that "science is self-correcting" and instead acknowledge, appreciate, and reward the people who detect and correct errors in scientific research.

To close, I want to reiterate that there are aspects of Picho-Kiroga et al. (2021) meta-analysis that I like, and they *were* correct to conclude that the effect sizes in the stereotype threat research are inflated. Moreover, it was because Picho-Kiroga et al., were good at suggesting moderators, thorough in describing aspects of their study, and willing to share their data that I could conduct a detailed re-analysis and write this critique. I hope is that this response will prompt readers to engage further in the correction process. I believe that after the errors are corrected that this meta-analysis could be reissued in a form that inspires confidence in their work and serves as a strong foundation for progress on understanding stereotype threat.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Russell T. Warne https://orcid.org/0000-0003-4763-3625

Notes

[1] Picho-Kiroga et al. (2021) reverse coded all of the effect sizes but did not disclose this to their readers. It is apparent this happened in the data and in the article, where the majority of effect sizes in Table 1 are negative, but the majority of effect sizes in Figure 2 are positive. This reverse coding does not impact the message of the article, though it should have been disclosed to readers.

[2] Variability introduced by methodological heterogeneity would also make this assumption extremely unrealistic.

[3] Based on this result, it would take 63.021 missing effect sizes statistically equal to zero or that contradict stereotype threat theory to remove this excessive number of statistically significant results, which would indicate that 38.4% of all effect sizes are unpublished. However, this is an overestimate because this procedure assumes that there is no *p*-hacking in the published studies, which is highly unlikely given the density of effect sizes with *p*-values between .01 and .05 in the funnel plot (see Figure 2). The test of excessive significance also overestimates the number of missing studies when effect size heterogeneity is present—as is the case in Picho-Kiroga et al.'s (2021) data—because it is based on a fixed-effects model (Renkewitz & Keiner, 2019).

[4] This is true, regardless of whether the study uses a within-subjects or between-subjects design. In a within-subjects design, the non-susceptible individuals' scores should not change, and when their zero change scores are included in average change score

calculations, the result will be attenuated. The same attenuation phenomenon will occur in a between-subjects design because the experimental and control groups will both include non-susceptible people, and the mean score difference between groups in this subset of people will be zero. Including their scores in the calculations of each group's averages will reduce the difference in the two groups' means—and therefore drive the effect size closer to zero. For both designs, the severity of this attenuation has a monotonic relationship with the proportion of non-susceptible sample members in the study.

5. It is not clear why Picho-Kiroga et al. (2021) did not include the Flore et al. (2018) study in their meta-analysis, even though it meets the inclusion requirements. My examination of the meta-analysis did not include an investigation of the authors' search procedures.

References

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Cross, T. L., & Cross, J. R. (2017). Challenging an idea whose time has gone. *Roeper Review*, *39*(3), 191–194. https://doi.org/10.1080/02783193.2017.1319000

Dickersin, K., & Min, Y.-I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, *703*(1), 135–148. https://doi.org/10.1111/j.1749-6632.1993.tb26343.x

Dweck, C. S. (2009). Can we make our students smarter? *Education Canada*, *49*(4), 56–61.

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

Flore, P. C., Mulder, J., & Wicherts, J. M (2018). The influence of gender stereotype threat on mathematics test scores of Dutch high school students: A registered report. *Comprehensive Results in Social Psychology*, *3*(2), 140–174. https://doi.org/10.1080/23743603.2018.1559647

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science (new York, N Y)*, *345*(6203), 1502. https://doi.org/10.1126/science.1255484

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

Gibson, C. E., Losee, J., & Vitiello, C. (2014). A replication attempt of stereotype susceptibility. *Social Psychology*, *45*(3), 194–198. https://doi.org/10.1027/1864-9335/a000184

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education*, *5*(1), 351–379. https://doi.org/10.3102/0091732X005001351

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. https://doi.org/10.1177/1740774507079441

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), Article 24. https://doi.org/10.1186/s40359-016-0126-3.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Moon, A., & Roeder, S. S. (2014). A secondary replication attempt of stereotype susceptibility. *Social Psychology*, *45*(3), 199–201. https://doi.org/10.1027/1864-9335/a000193

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172–182. https://doi.org/10.1037/gpr0000034

Picho-Kiroga, K., Turnbull, A., & Rodriguez-Leahy, A. (2021). Stereotype threat and its problems: Theory misspecification in research, consequences, and remedies. *Journal of Advanced Academics*, *32*(2), 231–264. https://doi.org/10.1177/1932202 (20986161

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research: A comparative evaluation of six statistical methods. *Zeitschrift für Psychologie*, *227*(4), 261–279. https://doi.org/10.1027/2151-2604/a000386

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

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487

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

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

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., & 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

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129. https://doi.org/10.1016/S0895-4356(00)00242-0

Walker, M. E., & Bridgeman, B. (2008). *Stereotype threat spillover and SAT scores* (Research Report No. 2008-2). College Board. https://www.ets.org/research/policy_research_reports/publications/report/2008/hspm.

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*(9), 1132–1139. https://doi.org/10.1111/j.1467-9280.2009.02417.x

Warne, R. T. (2020). *In the know: Debunking 35 myths about human intelligence*. Cambridge University Press. https://doi.org/10.1017/9781108593298.

Warne, R. T. (2021). *Statistics for the social sciences: A general linear model approach (2nd ed.)*. Cambridge University Press.

Warne, R. T., Lazo, M., Ramos, T., & Ritter, N. (2012). Statistical methods used in gifted education journals, 2006-2010. *Gifted Child Quarterly*, *56*(3), 134–149. https://doi.org/10.1177/0016986212444122

Wax, A. L. (2009). Stereotype threat: A case of overclaim syndrome? In C. H. Sommers (Ed.), *The science on women and science* (pp. 132–169). AIE Press.

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

About the Author

Dr. Russell T. Warne is an associate professor of psychology at Utah Valley University. He has published over 60 scholarly articles, which have appeared in the prestigious journals Educational Researcher, Learning and Instruction, Intelligence, The American Journal of Psychology, Psychological Bulletin, the Journal of School Psychology, and more. He is also the author of two books published by Cambridge University Press: Statistics for the Social Sciences: A General Linear Model Approach and In the Know: Debunking 35 Myths About Human Intelligence.