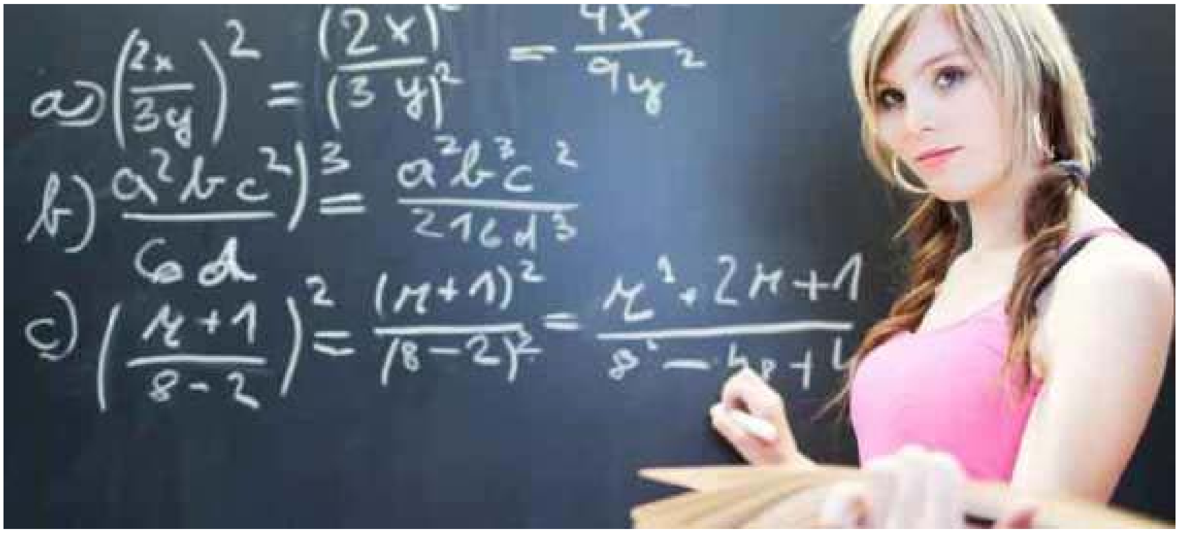
**Replication-Index**

*Improving the replicability of empirical research*

# *Tag Archives: Stereotype Threat and Women’s Math Performance*



# Why are Stereotype-Threat Effects on Women’s Math Performance Difficult to Replicate?

January 6, 2015 r-index, R-Index4Science, Replicability, science Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis, Gender Differences in Math, Girls and Math, Publication Bias, Reporting Bias, Spencer Steele Quinn 1999, Stereotype Threat and Women’s Math Performance, Stereotype-Threat

The R-Index of Studies on Stereotype Threat and Schoolgirls’ Math Performance

**Background**

In a seminal article, Spencer, Steele, and Quinn (1999) proposed the concept of stereotype threat. They argued that women may experience stereotype-threat during math tests and that stereotype threat can interfere with their performance on math tests.

The original study reported three experiments.

**STUDY 1**

Study 1 had 56 participants (28 male and 28 female undergraduate students). The main aim was to demonstrate that stereotype-threat influences performance on difficult, but not on easy math problems.

A 2 x 2 mixed model ANOVA with sex and difficulty produced the following results.

Main effect for sex, *F*(1, 52) = 3.99, p = .051 (reported as p = .05), z = 1.96, observed power = 50%.

Interaction between sex and difficulty, *F*(1, 52) = 5.34 , p = .025, z = 2.24, observed power = 61%.

The low observed power suggests that sampling error contributed to the significant results. Assuming observed power is a reliable estimate of true power, the chance of obtaining significant results in both studies would only be 31%. Moreover, if the true power is in the range between 50% and 80% power, there is only a 32% chance that observed power to fall into this range. The chance that both observed power values fall into this range is only 10%.

Median observed power is 56%. The success rate is 100%. Thus, the success rate is inflated by 44 percentage points (100% – 56%).

**The R-Index for these two results is low, Ř = 12 (56 – 44).**

Empirical evidence shows that studies with low R-Indices often fail to replicate in exact replication studies.

It is even more problematic that Study 1 was supposed to demonstrate just the basic phenomenon that women perform worse on math problems than men and that the following studies were designed to move this pre-existing gender difference around with an experimental manipulation. If the actual phenomenon is in doubt, it is unlikely that experimental manipulations of the phenomenon will be successful.

**STUDY 2**

The main purpose of Study 2 was to demonstrate that gender differences in math performance would disappear when the test is described as gender neutral.

Study 2 recruited 54 students (30 women, 24 men). This small sample size is problematic for several reasons. Power analysis of Study 1 suggested that the authors were lucky to obtain significant results. If power is 50%, there is a 50% chance that an exact replication study with the same sample size will produce a non-significant result. Another problem is that sample sizes need to increase to demonstrate that the gender difference in math performance can be influenced experimentally.

The data were not analyzed according to this research plan because the second test was so difficult that nobody was able to solve these math problems. However, rather than repeating the experiment with a better selection of math problems, the results for the first math test were reported.

As there was no repeated performance by the two participants, this is a 2 x 2 between-subject design that crosses sex and treat-manipulation. With a total sample size of 54 students, the n per cell is 13.

The main effect for sex was significant, *F*(1, 50) = 5.66, p = .021, z = 2.30, observed power = 63%.

The interaction was also significant, F(1, 50) = 4.18, p = .046, z = 1.99, observed power = 51%.

Once more, median observed power is just 57%, yet the success rate is 100%. Thus, the success rate is inflated by 43% and the **R-Index is low, Ř = 14%**, suggesting that an exact replication study will not produce significant results.

**STUDY 3**

Studies 1 and 2 used highly selective samples (women in the top 10% in math performance). Study 3 aimed to replicate the results of Study 2 in a less selective sample. One might expect that stereotype-threat has a weaker effect on math performance in this sample because stereotype threat can undermine performance when ability is high, but anxiety is not a factor in performance when ability is low. Thus, Study 3 is expected to yield a weaker effect and a larger sample size would be needed to demonstrate the effect. However, sample size was approximately the same as in Study 2 (36 women, 31 men).

The ANOVA showed a main effect of sex on math performance, F(1, 63) = 6.44, p = .014, z = 2.47, observed power = 69%.

The ANOVA also showed a significant interaction between sex and stereotype-threat-assurance, *F*(1, 63) = 4.78, p = .033, z = 2.14, observed power = 57%.

Once more, the R-Index is low, **Ř = 26** (MOP = 63%, Success Rate = 100%, Inflation Rate = 37%).

**Combined Analysis**

The three studies reported six statistical tests. The R-Index for the combined analysis is low **Ř = 18** (MOP = 61%, Success Rate = 100%, Inflation Rate = 41%).

The probability of this event to occur by chance can be assessed with the Test of Insufficient Variance (TIVA). TIVA tests the hypothesis that the variance in p-values, converted into z-scores, is less than 1. A variance of one is expected in a set of exact replication studies with fixed true power. Less variance suggests that the z-scores are not a representative sample of independent test scores. The variance of the six z-scores is low, Var(z) = .12. A chi-square test shows that the probability of this event is **less than 1 out of 1,000,000,000,000,000**, chi-square (df = 5) = 105.

**Conclusion**

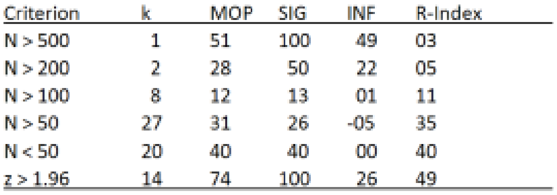
The replicability analysis of Spencer, Steele, and Quinn (1999) suggests that the original data provided inflated estimates of effect sizes and replicability. Thus, the R-Index predicts that exact replication studies would fail to replicate the effect.

**Meta-Analysis**

A forthcoming article in the Journal of School Psychology reports the results of a meta-analysis of stereotype-threat studies in applied school settings (Flore & Wicherts, 2014). The meta-analysis was based on 47 comparisons of girls with stereotype threat versus girls without stereotype threat. The abstract concludes that stereotype threat in this population is a statistically reliable, but small effect (d = .22). However, the authors also noted signs of publication bias. As publication bias inflates effect sizes, the true effect size is likely to be even smaller than the uncorrected estimate of .22.

The article also reports that the after a correction for bias, using the trim-and-fill method, the estimated effect size is d = .07 and not significantly different from zero. Thus, the meta-analysis reveals that there is no replicable evidence for stereotype-threat effects on schoolgirls' math performance. The meta-analysis also implies that any true effect of stereotype threat is likely to be small (d < .2). With a true effect size of d = .2, the original studies by Steel et al. (1999) and most replication studies had insufficient power to demonstrate stereotype threat effects, even if the effect exists. A priori power analysis with d = .2 would suggest that 788 participants are needed to have an 80% chance to obtain a significant result if the true effect is d = .2. Thus, future research on this topic is futile unless statistical power is increased by increasing sample sizes or by using more powerful designs that can demonstrate small effects in smaller samples.

One possibility is that the existing studies vary in quality and that good studies showed the effect reliably, whereas bad studies failed to show the effect. To test this hypothesis, it is possible to select studies from a meta-analysis with the goal to maximize the R-Index. The best chance to obtain a high R-Index is to focus on studies with large sample sizes because statistical power increases with sample size. However, the table below shows that there are only 8 studies with more than 100 participants and the success rate in these studies is 13% (1 out of 8), which is consistent with the median observed power in these studies 12%.



| Criterion | k | MOP | SIG | INF | R-Index |
|---|---|---|---|---|---|
| N > 500 | 1 | 51 | 100 | 49 | 03 |
| N > 200 | 2 | 28 | 50 | 22 | 05 |
| N > 100 | 8 | 12 | 13 | 01 | 11 |
| N > 50 | 27 | 31 | 26 | -05 | 35 |
| N < 50 | 20 | 40 | 40 | 00 | 40 |
| z > 1.96 | 14 | 74 | 100 | 26 | 49 |

It is also possible to select studies that produced significant results (z > 1.96). Of course, this set of studies is biased, but the R-Index corrects for bias. If these studies were successful because they had sufficient power to demonstrate effects, the R-Index would be greater than 50%. However, the R-Index is only 49%.

**CONCLUSION**

In conclusion, a replicability analysis with the R-Index shows that stereotype-threat is an elusive phenomenon. Even large replication studies with hundreds of participants were unable to provide evidence for an effect that appeared to be a robust effect in the original article. The R-Index of the meta-analysis by Flore and Wicherts corroborates concerns that the importance of stereotype-threat as an explanation for gender differences in math performance has been exaggerated. Similarly, Ganley, Mingle, Ryan, Ryan, and Vasilyeva (2013) found no evidence for stereotype threat effects in studies with 931 students and suggested that "these results raise the possibility that stereotype threat may not be the cause of gender differences in mathematics performance prior to college." (p 1995).

The main novel contribution of this post is to reveal that this disappointing outcome was predicted on the basis of the empirical results reported in the original article by Spencer et al. (1999). The article suggested that stereotype threat is a pervasive phenomenon that explains gender differences in math performance. However, The R-Index and the insufficient variance in statistical results suggest that the reported results were biased and, overestimated the effect size of stereotype threat. The R-Index corrects for this bias and correctly predicts that replication studies will often result in non-significant results. The meta-analysis confirms this prediction.

In sum, the main conclusions that one can draw from 15 years of stereotype-threat research is that (a) the real reasons for gender differences in math performance are still unknown, (b) resources have been wasted in the pursuit of a negligible factor that may contribute to gender differences in math performance under very specific circumstances, and (c) that the R-Index could have prevented the irrational exuberance about stereotype-threat as a simple solution to an important social issue.

In a personal communication Dr. Spencer suggested that studies not included in the meta-analysis might produce different results. I suggested that Dr. Spencer provides a list of studies that provide empirical support for the hypothesis. At present, the existing evidence suggests that published studies provide inflated estimates of the replicability and importance of the effect.

This blog also provides further evidence that male and female psychologists could benefit from a better education in statistics and research methods to avoid wasting resources in the pursuit of false-positive results.