Journal of Applied Psychology
2017, Vol. 102, No. 8, 1159–1168

0021-9010/17/$12.00 http://dx.doi.org/10.1037/apl0000188

COMMENT

Potential Publication Bias in the Stereotype Threat Literature: Comment on Nguyen and Ryan (2008)

L. J. Zigerell
Illinois State University

Stereotype threat is a widely cited psychological phenomenon with purported important real-world consequences. Reanalysis of data from the Nguyen and Ryan (2008) stereotype threat meta-analysis indicated the presence of small study effects in which the effect size for less precise studies was larger than the effect size for more precise studies. Four methods to adjust the meta-analysis effect size for potential publication bias produced divergent estimates, from essentially no change, to a 50% decrease, to a reduction of the estimated effect size to near zero. Caution is therefore warranted both for citing Nguyen and Ryan (2008) as evidence of a meaningful stereotype threat effect and for claiming that the stereotype threat effect size is negligible based on these adjustments, given that the detected small study effects might be due to unexplored moderators instead of publication bias.

Keywords: stereotype threat, publication bias, race, sex

Supplemental materials: http://dx.doi.org/10.1037/apl0000188.supp

Stereotype threat refers to the phenomenon in which a negative stereotype of a group causes members of that group to conform to the stereotype (Steele & Aronson, 1995). Research on stereotype threat has important implications; for example, based on evidence that stereotype threat causes certain groups to underperform academically relative to their true ability, a set of experimental psychologists submitted an amicus curiae brief in support of the use of race in university admissions for the U.S. Supreme Court case *Fisher* v. *University of Texas at Austin* (Aronson et al., 2015).

Nguyen and Ryan (2008) is a highly cited meta-analysis that reported a pooled stereotype threat effect size estimate of $d = -0.26$ that was different than zero at conventional levels of statistical significance. NR2008 considered the possibility of publication bias in the stereotype threat literature, by, among other things, reporting statistical tests of failsafe *N* (Rosenthal, 1979) and presenting a funnel plot of sample size against effect size (Light & Pillemer, 1984). NR2008 described the funnel plot as "[resembling] a relatively symmetrical inverted funnel, indicating the absence of publication bias in the data set" (p. 1325) and did not report estimates adjusted for potential publication bias. However, as reported below, multiple statistical tests indicated that the funnel plot was not symmetric.

Dataset Adjustment

A funnel plot based on the sample sizes and effect sizes in NR2008 Table 3 did not match the funnel plot in NR2008 Figure 1, so I contacted Dr. Nguyen, who provided a spreadsheet of data. The spreadsheet contained effect sizes that, when combined with sample sizes in NR2008 Table 3, produced a funnel plot that matched or at least closely matched the funnel plot in NR2008 Figure 1. These data are plotted in the top left panel of Figure 1.

I made three changes to the NR2008 sample. I removed studies from two articles that have been retracted since the publication of NR2008,[1] I changed the coding for Edwards (2004) from a female stereotype threat study to a race/ethnicity stereotype threat study, and I added a female stereotype threat study from Stricker and Ward (2004) Study 2, with a sample size of 694 and an effect size

I thank Hannah Nguyen for making data available and answering questions about the Nguyen and Ryan (2008) article and data. Data and code to reproduce the study's analyses will be available at the author's Dataverse. Before publication, results for some of the analyses reported on in the study were posted on Twitter.

Correspondence concerning this article should be addressed to L. J. Zigerell, Schroeder Hall 401, Department of Politics and Government, Illinois State University, Normal, IL 61790-4540. E-mail: ljzigerell@ilstu.edu

[1] The retracted articles are Marx and Stapel (2005) and Marx and Stapel (2006), which contributed three studies to the 116 studies in the NR2008 sample. Diederik Stapel was a coauthor on another article in the NR2008 sample (Marx et al., 2005), but, by the date of my analysis, that article had not been retracted and was thus retained in the sample.

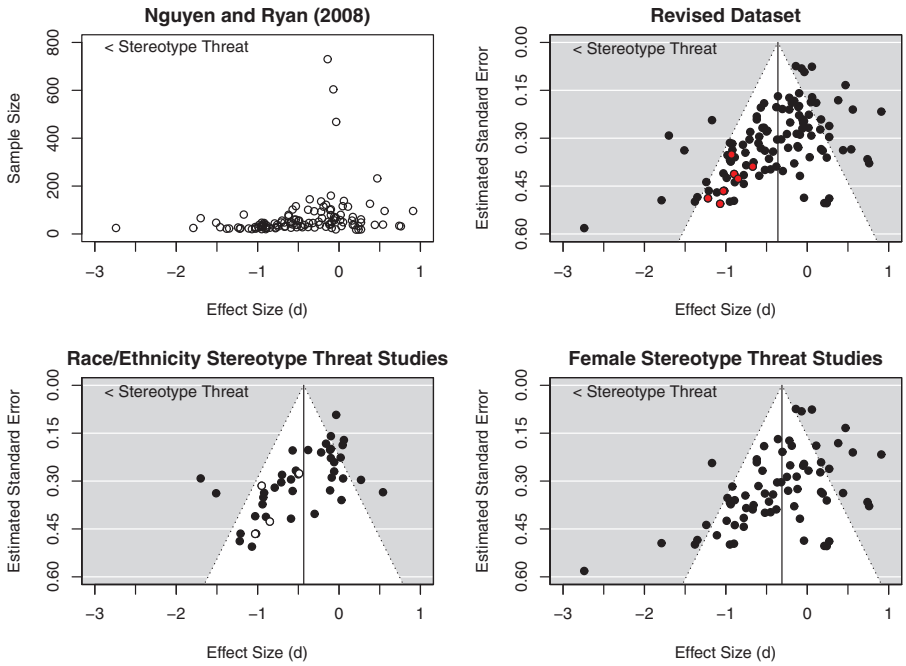

Figure 1. Funnel plots. The top left panel includes all 116 studies from the Nguyen and Ryan (2008) dataset. The remaining panels use only the revised set of 114 studies. The non-black dots in the top right funnel plot indicate sample studies with stereotype threat originators Steele or Aronson as a coauthor. The white dots in the bottom left funnel plot indicate sample stereotype threat studies in which the sample was white participants. Two points from Aronson et al. (1999) largely overlap: both have sample sizes of 23 and nearly the same estimated effect size ($d = -1.02$ and $d = -1.03$). See the online article for the color version of this figure.

d of 0.0575.[2] Data for this revised NR2008 sample are presented in Table 1 and the top right panel of Figure 1, with the bottom panels of Figure 1 presenting the studies disaggregated by test taker type.[3] Standard errors were estimated with formula 7.30 of Hunter and Schmidt (2004; p. 286). The main statistical analyses reported in this comment were conducted in R (R Core Team, 2014) using the metafor package (Viechtbauer, 2010). The Illinois State University Institutional Review Board (IRB) has determined that the data analysis reported on in this study does not require IRB review because the data analysis does not constitute human subjects research.

Evidence of Small Study Effects

The vertical line at −0.36 in the top right panel in Figure 1 indicates the mean effect size for the full 114-study sample based on a random effects meta-analysis. The relatively sparse area in the bottom right quadrant suggests the presence of small study effects in which less precise studies have a different—and, in this case, a larger negative—estimated effect than more precise studies. This visual observation was corroborated with statistical tests, which detected evidence of funnel plot asymmetry at $p < .001$ in Begg's test (Begg & Mazumdar, 1994) and at $p < .001$ in Egger's test (Egger et al., 1997).

The cumulative meta-analysis plot in Figure 2 presents more evidence of small study effects: the top dot and top line represent the effect size estimate and 95% confidence interval for a random-effects meta-analysis based on only the most precise study, as measured by estimated standard error, and each successive dot and line below indicate the effect size estimate and 95% confidence

[2] NR2008 (pp. 1318, 1322) indicated that: "Schmader and Johns (2003, Study 2) and Stricker and Ward (2004, Study 2) conducted studies where test takers' gender was nested within race/ethnicity subgroups (i.e., White men/women vs. Latinos/Latinas vs. African American men/women). Because these studies mainly aimed at examining race-based stereotype threat effects, only the effect sizes as a function of race/ethnicity and stereotype threat activation contributed data points to the overall meta-analytic data set." NR2008 then indicated in footnote 3 (p. 1322) that "Stricker and Ward (2004, Study 1) was an exception to this rule, as the stereotype threat cues were both race-based and gender-based (i.e., race and gender inquiries prior to tests)." Results for females and for Latinos were reported for Schmader and Johns (2003) Study 2, but the design was described in that manuscript as a "2 (Latino or White) × 2 (stereotype threat or control) factorial design", with participants asked to indicate ethnicity before the test (p. 445); therefore, there is a clear reason to exclude that study from the female stereotype threat set of studies. However, NR2008 considered Stricker and Ward (2004) Study 2 to be only a race/ethnicity stereotype threat study and not a female stereotype threat study, but the design was the same as the included Stricker and Ward (2004) Study 1: some participants marked demographic information before the test, and other participants marked demographic information after the test. Therefore, I added an observation for a female stereotype threat study from Stricker and Ward (2004) Study 2, with a sample size of 694 and a *d* of 0.0575, based on the average of the four tests: Elementary Algebra *d* of −0.152, Arithmetic *d* of 0.051, Reading Comprehension *d* of 0.264, and Sentence Skills *d* of 0.067.

[3] Pellegrini (2005) had a sample of Hispanic females, and the study made both ethnicity and sex salient by asking respondents to identify as Hispanic or white and as male or female. However, ethnicity was made more salient in the consent form, with a reference to White populations ("The purpose of the current study is to assess your intellectual ability as a Hispanic female using a test of intelligence that has been standardized on White populations."). NR2008 coded this only as a race/ethnicity stereotype threat study, and this coding was retained.

Table 1
Studies in the Revised Data Set

Study no.	Study name	Study	Female	*N*	*d*	STA	STR	TD
1	Ambady et al. (2004)	1 of 2	1	20	−0.95	1		3
2	Ambady et al. (2004)	2 of 2	1	20	−0.9	1		3
3	Anderson (2001)	1 of 1	1	604	−0.07	1		
4	Aronson et al. (1999)	1 of 2	0	23	−1.03	3		
5	Aronson et al. (1999)	2 of 2	0	26	−0.85	3		3
6	Aronson et al. (1999)	2b of 2	0	23	−1.02	3		3
7	Bailey (2004)	1 of 1	1	44	−0.6	3		3
8	J. L. Brown et al. (n.d.)	2 of 2	0	28	−0.898	1	2	3
9	R. P. Brown & Day (2006)	1 of 1	0	34	−0.94	1	1	2
10	R. P. Brown & Josephs (1999)	1 of 3	1	65	−0.05	1	1	2
11	R. P. Brown & Josephs (1999)	2 of 3	1	35	−0.95	1	1	2
12	R.P. Brown & Pinel (2003)	1 of 1	1	46	−0.37	2	2	
13	Cadinu et al. (2003)	1 of 2	1	25	−1.79	3	2	3
14	Cadinu et al. (2003)	1b of 2	1	38	0.17	3	2	3
15	Cadinu et al. (2003)	2 of 2	0	50	−0.59	3	2	
16	Cadinu et al. (2005)	1 of 1	1	60	−0.55	2	2	2
17	G. L. Cohen & Garcia (2005)	2 of 3	0	41	−0.92	3	1	
18	Cotting (2003)	1 of 1	1	51	−0.13	2	1	
19	Cotting (2003)	1b of 1	0	55	−0.5	2	1	
20	Davies et al. (2002)	1 of 2	1	25	−0.89	1	1	3
21	Davies et al. (2002)	2 of 2	1	34	−0.94	1		2
22	Dinella (2004)	1 of 1	1	232	0.47	2	2	
23	Dodge et al. (2001)	1 of 1	0	93	−0.22	1	1	2
24	Edwards (2004)	1 of 1	0	79	−0.1	2	1	
25	Elizaga & Markman (n.d.)	1 of 1	1	145	−0.36	1	1	
26	Foels (1998)	1 of 1	1	33	0.24	1	2	1
27	Foels (1998)	1b of 1	1	32	0.76	1	2	3
28	Foels (2000)	1 of 1	1	71	0.17	1	2	3
29	Ford et al. (2004)	2 of 2	1	31	−0.75	2	2	
30	Gamet (2004)	1 of 1	1	51	−0.25	3		
31	Gresky et al. (n.d.)	1 of 1	1	23	−1.35	3	1	3
32	Gresky et al. (n.d.)	1b of 1	1	37	0.19	3	1	3
33	Guajardo (2005)	1 of 2	1	56	0.16	1	2	
34	Guajardo (2005)	2 of 2	1	30	−0.16	1		
35	Harder (1999)	1 of 2 (pilot)	1	36	−0.89	1	2	3
36	Harder (1999)	2 of 2	1	19	0.27	1	2	3
37	Johns et al. (2005)	1 of 1	1	46	−0.92	2	1	3
38	Josephs et al. (2003)	1 of 1	1	39	−0.235	3		
39	Keller (2002)	1 of 1	1	37	−0.42	3		
40	Keller (2007)	1 of 1	1	19	−0.04	2	2	1
41	Keller (2007)	1b of 1	1	18	0.21	2	2	3
42	Keller (2007)	1c of 1	1	18	0.24	2	2	1
43	Keller & Bless (n.d.)	2 of 3	1	66	−0.21	2	2	
44	Keller & Dauenheimer (2003)	1 of 1	1	33	−0.47	2	2	2
45	Lewis (1998)	1 of 1	0	71	−0.06	1		3
46	Martens et al. (2006)	1 of 2	1	22	−1.38	1	1	
47	Martens et al. (2006)	2 of 2	1	38	−0.76	3	1	
48	Martin (2004)	2 of 2	0	100	−0.11	1	1	
49	Martin (2004)	2b of 2	0	102	−0.57	1	1	
53	Marx et al. (2005)	3 of 4	1	27	−0.99	1	1	3
54	Marx et al. (2005)	3b of 4	1	25	−2.74	1	1	3
55	Marx et al. (2005)	4 of 4	1	25	−0.09	1	1	3
56	McFarland, Kemp, et al. (2003)	1 of 1	1	126	0.38	3	1	2
57	McFarland, Lev-Arey, & Ziegert (2003)	1 of 1	0	50	−0.09	1	1	3
58	McIntyre et al. (2003)	1 of 2	1	116	−0.53	3	2	3
59	McIntyre et al. (2003)	2 of 2	1	74	−0.62	3	2	3
60	McIntyre et al. (2005)	1 of 1	1	81	−1.17	3	1	3
61	McKay (1999)	1 of 1	0	103	−0.1	1	1	2
62	Nguyen et al. (2003)	1 of 1	0	80	0.02	1		2
63	Nguyen et al. (2004)	1 of 1	1	114	−0.19	3	1	2

(*table continues*)

Table 1 (*continued*)

Study no.	Study name	Study	Female	*N*	*d*	STA	STR	TD
64	O'Brien & Crandall (2003)	1 of 1	1	58	0.015	2	2	2
65	Oswald & Harvey (2000)	1 of 1	1	34	0.74	1	2	2
66	Pellegrini (2005)	1 of 1	0	60	−0.53	2		
67	Philipp & Harton (2004)	1 of 1	1	38	0.44	3		
68	Ployhart et al. (2003)	1 of 1	0	48	−0.71	1	1	3
69	Ployhart et al. (2003)	1b of 1	0	48	0.27	1	1	3
70	Prather (2005)	1 of 1	1	114	0.11	1	1	2
71	Rivadeneyra (2001)	1 of 1	0	116	0.045	1	1	3
72	H. E. S. Rosenthal & Crisp (2006)	2 of 3	1	24	−0.78	1	1	1
73	H. E. S. Rosenthal & Crisp (2006)	3 of 3	1	29	−0.38	2	2	1
74	H. E. S. Rosenthal & Crisp (2006)	3b of 3	1	27	−0.78	2	2	1
75	Salinas (1998)	1 of 2	0	27	−0.3	2		
76	Salinas (1998)	2 of 2	0	56	−0.7	2		
77	Sawyer & Hollis-Sawyer (2005)	1 of 1	0	66	−1.7	1	1	
78	Sawyer & Hollis-Sawyer (2005)	1b of 1	0	47	−1.51	1	1	
79	Schimel et al. (2004)	2 of 3	1	46	−0.32	1	1	
80	Schmader (2002)	1 of 1	1	32	−0.66	1		3
81	Schmader & Johns (2003)	1 of 3	1	28	−0.45	2	1	2
82	Schmader & Johns (2003)	2 of 3	0	33	0.03	1	1	2
83	Schmader & Johns (2003)	3 of 3	1	28	−0.52	1	1	2
84	Schmader et al. (2004)	2 of 2	1	68	−0.04	2	1	3
85	Schneeberger & Williams (2003)	1 of 1	1	61	0.27	3	2	1
86	Schultz et al. (n.d.)	1 of 2	0	44	−0.79	3	1	3
87	Schultz et al. (n.d.)	2 of 2	0	40	−0.57	3	1	3
88	Seagal (2001)	6 of 6	0	101	−0.38	3		3
89	Sekaquaptewa & Thompson (2003)	1 of 1	1	80	−0.62	2	2	
90	C. E. Smith & Hopkins (2004)	1 of 1	0	160	−0.1	3	1	
91	J. L. Smith & White (2002)	1 of 2	0	47	−0.95	3	2	2
92	J. L. Smith & White (2002)	2 of 2	1	23	−1.11	3	2	2
93	S. J. Spencer et al. (1999)	2 of 3	1	30	−0.67	2	2	3
94	S. J. Spencer (2005)	1 of 1	1	40	−0.12	3		
95	Spicer (1999)	2 of 2	0	39	−0.11	1	1	3
96	Spicer (1999)	2b of 2	0	39	0.54	1	1	1
97	Steele & Aronson (1995)	1 of 4	0	38	−0.93	1	1	3
98	Steele & Aronson (1995)	2 of 4	0	20	−1.07	1	1	3
99	Steele & Aronson (1995)	4 of 4	0	22	−1.22	1	1	3
100	Sternberg et al. (n.d.)	1 of 2	1	27	−1.24	3	2	3
101	Sternberg et al. (n.d.)	2 of 2	1	96	0.56	3	2	3
102	Stricker & Ward (2004)	1 of 2	0	122	−0.16	1		
103	Stricker & Ward (2004)	1b of 2	1	730	−0.14	1		
104	Stricker & Ward (2004)	2 of 2	0	468	−0.035	1		
105	Tagler (2003)	1 of 1	1	136	−0.22	2	1	2
106	van Dijk et al. (n.d.)	1 of 1	1	38	−0.98	1	1	
107	van Dijk et al. (n.d.)	1b of 1	1	38	−0.52	1	1	
108	von Hippel et al. (2005)	4 of 4	0	56	−0.49	3	2	3
109	Walsh et al. (1999)	2 of 2	1	96	0.91	3	1	3
110	Walters (2000)	1 of 2	0	49	0.05	1	1	3
111	Wicherts et al. (2005)	1 of 3	0	138	0.06	1		2
112	Wicherts et al. (2005)	3 of 3	1	95	−0.305	3	2	2
113	Wout et al. (n.d.)	1 of 4	0	57	−0.06	1	1	
114	Wout et al. (n.d.)	2 of 4	0	29	−1.03	2	2	3
115	Wout et al. (n.d.)	3 of 4	0	24	−1.21	2	2	2
116	Wout et al. (n.d.)	4 of 4	0	26	−0.59	2	2	2
117	Stricker & Ward (2004)	2b of 2	1	694	0.0575	1		

Note. Study numbers 1 to 116 are from NR2008. Studies 50 to 52 were retracted after the publication of NR2008, and Study 117 has been added. STA is a code for stereotype threat activating cues (1 for subtle, 2 for moderately explicit, 3 for blatant), STR is a code for stereotype threat removal strategies (1 for subtle, 2 for explicit), and TD is a code for test difficulty (1 for easy, 2 for moderately difficult, 3 for difficult). See the "Dataset Adjustment" section and the online supplemental material for more detail on the construction of the table.

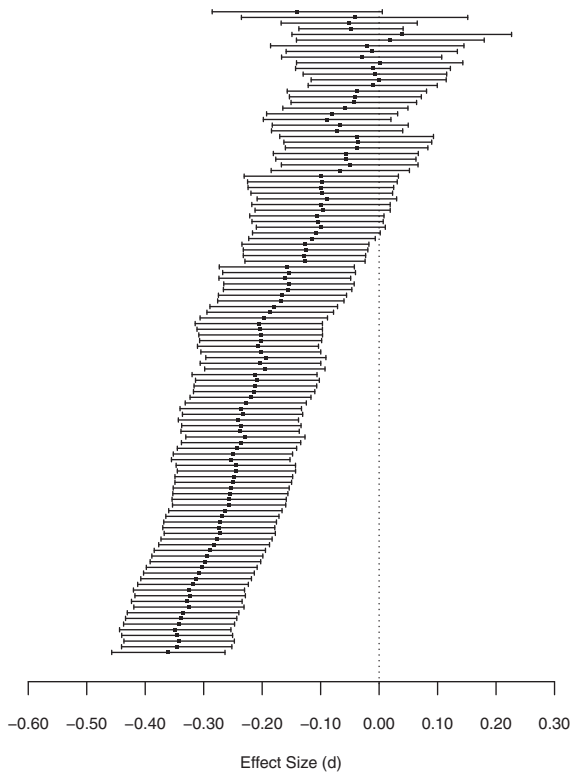

Figure 2. Cumulative meta-analysis plot. The top dot and line indicate the effect size estimate and 95% confidence interval for a random-effects meta-analysis based on only the most precise study, as measured by estimated standard error; each successive line below indicates the effect size estimate and 95% confidence interval for a random-effects meta-analysis with the addition of the next most precise study. The leftward drift of the figure illustrates how the less precise studies pull the effect size estimate in the negative direction and away from zero.

interval for a random-effects meta-analysis with the addition of the next most precise study. This plot illustrates how the less precise studies pull the effect size estimate in the negative direction and away from zero.

Evidence of Publication Bias?

Small study effects can be produced by multiple causes, such as effect size moderators and true heterogeneity in effects as a function of sample size. But small study effects can also reflect publication bias in which some studies are conducted but not reported. Direct evidence that this type of publication bias occurs in social science literature is available from admissions of researchers (e.g., Galak & Meyvis, 2012) and from observations of a registry of survey experiments in Franco et al. (2014) in which many null results were not written up in a manuscript.

However, available evidence for publication bias in the stereotype threat literature and in the NR2008 sample is only circumstantial:

- Begg's and Egger's tests indicate funnel plot asymmetry for NR2008 for the original sample, the revised sample, the original and revised sample restricted to race/ethnicity stereotype threat studies, and the original and revised sample restricted to female stereotype threat studies.
- Evidence of publication bias was detected across multiple methods in prior research on a stereotype threat sample (Flore & Wicherts, 2015).
- In the NR2008 sample, effect sizes from studies from manuscripts coauthored by stereotype threat originators Steele and Aronson were consistently and substantially larger than the average effect size of residual studies, with sample Steele or Aronson-coauthored studies effect sizes ranging from −0.67 to −1.22, with sample sizes from 20 to 38 and a weighted mean from a random effects meta-analysis of −0.94 [−1.23, −0.64], compared with residual included studies, which had corresponding a corresponding mean of −0.33 [−0.42, −0.23]. See the top right panel in Figure 1.
- The recent Finnigan and Corker (2016) preregistered replication "indicate[d] a failure to replicate Chalabaev et al. (2012), with no evidence suggesting the presence of significant stereotype threat main effects, nor any moderation by performance avoidance goals, in spite of the fact that the current replication study had a much larger sample size than the original study" (p. 40).

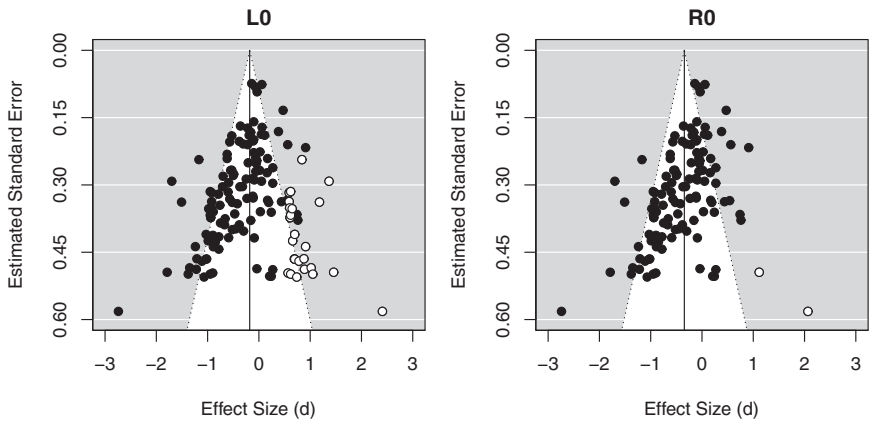

Figure 3. Trim-and-fill results. The black dots represent the sample studies. The white dots represent studies that have been added by the trim-and-fill. The left panel is the trim-and-fill with the L0 estimator, and the right panel is the trim-and-fill with the R0 estimator.

Table 2
Meta-Regression Results

Predictors	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Constant	−.11 (.36)	−.44* (.08)	−.41*(.09)	−.33* (.07)	−.57* (.19)	−.55*(.19)	−.08 (.25)	−.06 (.21)
Stereotype threat type								
Female	.17 (.17)	.14 (.10)	.15 (.11)		.15 (.14)		.03 (.14)	
Stereotype threat activating cue								
Moderately explicit	−.27 (.22)		−.07 (.13)	−.04 (.13)				
Blatant	−.11 (.19)		−.08 (.12)	−.06 (.12)				
Stereotype threat removal strategy								
Explicit	.02 (.17)				.07 (.13)	.12 (.13)		
Test difficulty								
Moderately difficult	−.24 (.27)						−.23 (.25)	−.24 (.24)
Difficult	−.38 (.26)						−.40 (.24)	−.40 (.23)
N	65	114	114	114	88	88	76	76

Note. Outcome variable for the meta-regression is study effect size. Omitted categories for the categorical variables are: race/ethnicity (stereotype threat type), subtle (stereotype threat activating cue), subtle (stereotype threat removal strategy), and easy (test difficulty). Data were available for 114 stereotype threat type codes, for 114 stereotype threat activating cue codes, for 88 stereotype threat removal strategy codes, and for 76 test difficulty codes.
* $p < .05$ (two-tailed).

Adjusted Effect Size Estimates

Several methods are available to assess how sensitive meta-analysis effect sizes are to small study effects or potential publication bias. Stanley et al. (2010) proposed restricting the meta-analysis to the top 10% of studies by precision. A random-effects meta-analysis on the 11 studies from the revised sample with the lowest estimated standard error produced an estimated effect size of −0.01 [−0.14, 0.12]. This trivial estimated effect size for the most precise studies differs substantially from the estimated effect size for the 11 least precise studies: −0.84 [−1.38, −0.29].

The PET-PEESE method proposed in Stanley and Doucouliagos (2014) can be used to adjust effect size estimates for potential publication bias: in this two-part test, a regression weighted by the inverse of the variance predicts the effect size using the standard error (precision-effect test, PET); if the constant from this regression does not differ from zero at a statistically significant level, the constant is used as the estimated effect size; otherwise, the effect size is estimated as the constant from a regression weighted by the inverse of the variance predicting the effect size using the variance (precision-effect estimate with standard error, PEESE). For the revised NR2008 sample, the PET-PEESE method produced an estimate of $d = 0.03$, with a 95% confidence interval of [−0.07, 0.12].

A third method to adjust effect sizes for potential publication bias is trim-and-fill (Duval & Tweedie, 2000), which imputed only two studies using the R0 estimator, as indicated in the right side of Figure 3, but imputed 26 studies using the L0 estimator, as indicated in the left side of Figure 3.[4] The unadjusted estimated effect size of −0.36 [−0.46, −0.26] was reduced to −0.34 [−0.44, −0.24] with the imputation of the two studies but was reduced to −0.18 [−0.29, −0.07] with the imputation of the 26 studies. Trim-and-fill was thus sensitive to the choice of estimator, with the effect size reduced by 5% with the R0 estimator and by 50% with the L0 estimator. Visual inspection of Figure 3 indicates that the L0 trim-and-fill produced a more symmetric funnel plot that the R0 estimator, with the mean L0 effect size closer than the mean R0 effect size to the mean effect size of the most precise studies.

Adjusted effect sizes were similar for the top 10% method ($d = -0.01$) and for PET-PEESE (0.03). Trim-and-fill estimates with the L0 and R0 estimators ($d = -0.18$ and $d = -0.34$) differ substantially from each other and from the top 10% and PET-PEESE estimates, but this difference might largely reflect the assumption of the trim-and-fill method "that it is the most 'negative' or 'undesirable' studies which are missing" (Duval, 2006; p. 130). This assumption is incorrect if Simonsohn et al. (2014) is correct that "in many disciplines, including psychology, publication bias operates primarily through statistical significance rather than effect size (Fanelli, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995)" (p. 545). Figure 3 indicates that trim-and-fill retained a relatively empty space on either side of zero at the bottom of the graphs, with the L0 estimator trim-and-fill producing the implausible scenario of 26 missing studies in which zero studies had an effect size less than $d = 0.50$.

Moderator Analyses

NR2008 reported meta-analysis results disaggregated by test difficulty, domain identification, stereotype threat activating cues, and stereotype threat removal strategies. Using codes from the dataset that Dr. Nguyen provided, I conducted a meta-regression predicting the effect size using a dichotomous variable for whether the study concerned female stereotypes (with race/ethnicity stereotypes as the omitted category), dichotomous variables for moderately explicit and blatant stereotype threat activating cues (with subtle as the omitted category), a dichotomous variable for explicit stereotype threat removal strategies (with subtle as the omitted category), and dichotomous variables for moderately difficult and difficult test difficulties (with easy as the omitted category).[5]

[4] Duval (2006) notes: "Both [estimators] have low bias, and as n gets larger the estimator R0 becomes preferable to L0 in terms of having a relatively smaller variance. Initial simulations also show that L0 is more robust than R0 against certain data configurations that might occur under some circumstances" (p. 132). Kepes et al. (2012) noted, "The L estimator is generally preferred and the most commonly used approach. It is more robust, especially when the number of samples in the distribution is small" (p. 633, citations omitted).

[5] Domain identification was not included as a predictor because the NR2008 Table 4 pattern for domain identification did not match the pattern from the dataset domain identification codes. See the online supplemental materials for more detail on how well dataset codes matched patterns reported in NR2008.

Table 3
Selected Results for Analyses on the Full Sample and Disaggregated Samples

Sample	*N*	Unadjusted effect size estimate	Egger's test *t* score	Adjusted effect size estimate				
				Top 10% by precision	PET	PEESE	Trim-and-fill (L0 estimator)	Trim-and-fill (R0 estimator)
Full sample	114	−.36 [−.46, −.26]	−5.92	−.01 [−.14, .12]	.19 [.04, .34]	.03 [−.07, .12]	−.18 [−.29, −.07]	−.34 [−.44, −.24]
Minority sample	39	−.39 [−.54, −.24]	−4.21	−.05 [−.18, .08]	.29 [.002, .59]	.03 [−.14, .21]	−.20 [−.38, −.02]	−.37 [−.53, −.21]
Female sample	70	−.31 [−.44, −.18]	−3.86	.01 [−.20, .22]	.15 [−.03, .33]	.04 [−.09, .16]	−.13 [−.28, .02]	−.29 [−.43, −.16]
Easy test	8	−.04 [−.39, .31]	−1.46	.27 [−.24, .78]	1.10 [−.77, 2.97]	.54 [−.50, 1.58]	.05 [−.32, .40]	.05 [−.32, .40]
Moderately difficult test	24	−.27 [−.44, −.10]	−3.57	−.08 [−.35, .19]	.52 [.09, .95]	.14 [−.09, .37]	−.09 [−.29, .12]	−.27 [−.44, −.10]
Difficult test	44	−.47 [−.67, −.28]	−3.77	−.08 [−.55, .39]	.65 [.10, 1.20]	.15 [−.15, .45]	−.27 [−.48, −.07]	−.47 [−.67, −.28]

Note. The minority sample is the set of race/ethnicity samples excluding the five white samples. Sample sizes for the top 10% by precision estimates are 11 (full sample), 4 (minority sample), 7 (female sample), 1 (easy test), 2 (moderately difficult test), and 4 (difficult test). Square brackets indicate 95% confidence intervals. PET = Precision-Effect Test; PEESE = Precision-Effect Estimate with Standard Error.

Meta-regression results presented in Table 2 for various combinations of these predictors indicated that none of these moderators reached statistical significance in any of the meta-regressions.

Point estimates for the test difficulty variables did not reach statistical significance in the meta-regression but were substantively large, so Table 3 reports results disaggregated by test difficulty. Table 3 also reports results for the full sample and for the sample disaggregated into minority and female test taker samples. Egger's test indicated funnel plot asymmetry for each disaggregated analysis except for the easy test sample that had only eight cases. The adjusted effect size estimate patterns in Table 3 for the disaggregated analyses generally reflect the pattern for the full sample: trivial negative adjusted effect sizes for the top 10% by precision method, adjusted effect sizes that do not differ from zero at conventional levels of statistical significance for the PET-PEESE method, adjusted effect sizes for the trim-and-fill method with the R0 estimator close to the unadjusted effect size estimates, and adjusted effect sizes for the trim-and-fill method with the L0 estimator that are substantially lower than for the R0 estimator.

Conclusion

The present comment reports evidence from a revised Nguyen and Ryan (2008) sample of stereotype threat studies that more precise studies produced a smaller effect size estimate than less precise studies. For the full sample, the top 10% method and PET-PEESE produced trivially small adjusted effect size estimates, trim-and-fill with the L0 estimator produced an adjusted effect size estimate that was half as large as the unadjusted estimate, and trim-and-fill with the R0 estimator produced an adjusted effect size estimate that was nearly identical to the unadjusted estimate.

None of these methods for adjusting effect size estimates is perfect. Simulations have indicated that trim-and-fill can impute missing studies even in full samples (Sterne & Egger, 2000; Terrin, Schmid, Lau, & Olkin 2003; p. 2121), that PET-PEESE can perform poorly (Gervais, 2015), and that the top 10% method is not effective for small non-nil true effect sizes (Inzlicht et al., 2015).[6] Therefore, caution is warranted before claiming a zero or trivial negative effect size based on these adjustments, especially because a better coding of moderators or different moderators might explain some or all of the small study effects. But caution is also warranted when citing Nguyen and Ryan (2008) as evidence for a meaningfully large stereotype threat effect that hinders the real-world test-taking of stereotyped groups. Given recent failures to replicate in social psychology (Engber, 2016) and in the stereotype threat literature (Finnigan & Corker, 2016), the best way to estimate the effect of stereotype threat might be to develop a literature of preregistered studies in authentic situations that avoids concern about estimates being contaminated by publication bias.

[6] Discussing the performance of adjustment techniques based on simulations, Inzlicht et al. (2015) recommended not using PET and indicated that "PEESE, Top10, and (to our surprise) Trim and Fill might be decent, but not excellent, all-purpose corrections. PEESE was good for medium and large effects, acceptable for small effects, but woeful for nils. Top10 was good for nils and large effects, decent for medium effects, but ineffective for small effects. Trim and Fill was good for small and large effects, middling for medium effects, and atrocious for nils" (p. 14). However, as noted by Inzlicht et al. (2015; p. 13), the true effect size is not known.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*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,* 401–408. http://dx.doi.org/10.1016/j.jesp.2003.08.003

*Anderson, R. D. (2001). *Stereotype threat: The effects of gender identification on standardized test performance*. (Unpublished doctoral dissertation). James Madison University, Harrisonburg, Virginia.

Aronson, J., Dweck, C. S., Erman, S., Good, C., Inzlicht, M., Logel, C., . . . Yeager, D. (2015). Brief of experimental psychologists as amici curiae in support of respondents, Fisher v. University of Texas at Austin, 579 US__(2016).

*Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35,* 29–46. http://dx.doi.org/10.1006/jesp.1998.1371

*Bailey, A. A. (2004). *Effects of stereotype threat on females in math and science fields: An investigation of possible mediators and moderators of the threat-performance relationship*. (Unpublished doctoral dissertation) Georgia Institute of Technology, Atlanta, Georgia.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. http://dx.doi.org/10.2307/2533446

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. http://dx.doi.org/10.1002/9780470743386

*Brown, J. L., Steele, C. M., & Atkins, D. (n.d.). *Performance expectations are not a necessary mediator of stereotype threat in African American verbal test performance*. Unpublished manuscript.

*Brown, R. P., & Day, E. A. (2006). The difference isn't black and white: Stereotype threat and the race gap on Raven's Advanced Progressive Matrices. *Journal of Applied Psychology, 91,* 979–985. http://dx.doi.org/10.1037/0021-9010.91.4.979

*Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76,* 246–257. http://dx.doi.org/10.1037/0022-3514.76.2.246

*Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39,* 626–633. http://dx.doi.org/10.1016/S0022-1031(03)00039-8

*Cadinu, M., Maass, A., Frigerio, S., Impagliazzo, L., & Latinotti, S. (2003). Stereotype threat: The effect of expectancy on performance. *European Journal of Social Psychology, 33,* 267–285. http://dx.doi.org/10.1002/ejsp.145

*Cadinu, M., Maass, A., Rosabianca, A., & Kiesner, J. (2005). Why do women underperform under stereotype threat? Evidence for the role of negative thinking. *Psychological Science, 16,* 572–578. http://dx.doi.org/10.1111/j.0956-7976.2005.01577.x

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,* 130–142. http://dx.doi.org/10.1007/s11031-011-9241-x

*Cohen, G. L., & Garcia, J. (2005). "I am us": Negative stereotypes as collective threats. *Journal of Personality and Social Psychology, 89,* 566–582. http://dx.doi.org/10.1037/0022-3514.89.4.566

*Cotting, D. I. (2003). *Shedding light in the black box of stereotype threat: The role of emotion*. (Unpublished doctoral dissertation) City University of New York, New York, NY.

*Davies, P. G., Spencer, S. J., Quinn, D. M., & Gerhardstein, R. (2002). Consuming images: How television commercials that elicit stereotype threat can restrain women academically and professionally. *Personality and Social Psychology Bulletin, 28,* 1615–1628. http://dx.doi.org/10.1177/014616702237644

*Dinella, L. M. (2004). *A developmental perspective on stereotype threat and high school mathematics*. (Unpublished doctoral dissertation) Arizona State University, Tempe, Arizona.

*Dodge, T., Williams, K., & Blanton, H. (2001, April). *Motivational mediators of the stereotype threat effect*. Paper presented at the 16th annual conference for the Society of Industrial and Organizational Psychology, San Diego, CA.

Duval, S. (2006). The trim and fill method. *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Advance onilne publication. http://dx.doi.org/10.1002/0470870168.ch8

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

*Edwards, B. D. (2004). *An examination of factors contributing to a reduction in race-based subgroup differences on a constructed response paper-and-pencil test of achievement*. (Unpublished doctoral dissertation) TX A&M University, College Station, Texas.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal (Clinical Research Ed.), 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

*Elizaga, R. A., & Markman, K. D. (n.d.). *Peers and performance: How in-group and out-group comparisons moderate stereotype threat effects*. Unpublished manuscript. http://dx.doi.org/10.1007/s12144-008-9041-y

Engber, D. (2016, March 6). Everything is crumbling. *Slate*. Retrieved from http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.html?utm_source=nextdraft&utm_medium=email

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891–904. http://dx.doi.org/10.1007/s11192-011-0494-7

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. http://dx.doi.org/10.1016/j.jrp.2016.05.009

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,* 25–44. http://dx.doi.org/10.1016/j.jsp.2014.10.002

*Foels, R. (1998, June). *Women's math ability: An investigation of stereotype threat*. Poster presented at the Society for the Psychological Study of Social Issues conference, Ann Arbor, MI.

*Foels, R. (2000, February). *Disidentification in the face of stereotype threat*. Paper presented at the Society for Personality and Social Psychology conference, Nashville, TN.

*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,* 643–653. http://dx.doi.org/10.1177/0146167203262851

Franco, A., Malhotra, N., & Simonovits, G. (2014, September 19). Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science, 345,* 1502–1505. http://dx.doi.org/10.1126/science.1255484

Galak, J., & Meyvis, T. (2012). You could have just asked: Reply to Francis (2012). *Perspectives on Psychological Science, 7,* 595–596. http://dx.doi.org/10.1177/1745691612463079

*Gamet, M. M. (2004). *Stereotype threat and the effects on women in mathematical tasks*. Unpublished manuscript.

Gervais, W. (2015). *Putting PET-PEESE to the test*. Retrieved from http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

*Gresky, D. M., Eyck, L. L. T., & Lord, C. G. (n.d.). *Effects of salient multiple identities on women's performance under mathematics stereotype threat*. Unpublished manuscript. http://dx.doi.org/10.1007/s11199-005-7735-2

*Guajardo, G. A. (2005). *Modifying stereotype relevance and altering affect attributions to reduce performance suppression on cognitive ability selection tests*. (Unpublished master's thesis) Northern Illinois University, DeKalb, Illinois.

*Harder, J. A. (1999). *The effect of private versus public evaluation on stereotype threat for women in mathematics*. (Unpublished doctoral dissertation) University of Texas at Austin.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412985031

Inzlicht, M., Gervais, W. M., & Berkman, E. T. (2015). Bias-correction techniques alone cannot determine whether ego depletion is different from zero: Commentary on Carter, Kofler, Forster, & McCullough, 2015. *Social Science Research Network*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2659409

*Johns, M., Schmader, T., & Martens, A. (2005). Knowing is half the battle: Teaching stereotype threat as a means of improving women's

math performance. *Psychological Science, 16,* 175–179. http://dx.doi.org/10.1111/j.0956-7976.2005.00799.x

*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,* 158–163. http://dx.doi.org/10.1111/1467-9280.t01-1-01435

*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, 47,* 193–198. http://dx.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. *The British Journal of Educational Psychology, 77,* 323–338. http://dx.doi.org/10.1348/000709906X113662

*Keller, J., & Bless, H. (n.d.). *When positive and negative expectancies disrupt performance: Regulatory focus as a catalyst*. Unpublished manuscript. http://dx.doi.org/10.1002/ejsp.452

*Keller, J., & Dauenheimer, D. (2003). Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance. *Personality and Social Psychology Bulletin, 29,* 371–381. http://dx.doi.org/10.1177/0146167202250218

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15,* 624–662. http://dx.doi.org/10.1177/1094428112452760

*Lewis, P. B. (1998). *Stereotype threat, implicit theories of intelligence, and racial differences in standardized test performance*. (Unpublished doctoral dissertation) Kent State University, Kent, Ohio.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

*Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology, 42,* 236–243. http://dx.doi.org/10.1016/j.jesp.2005.04.010

*Martin, D. E. (2004). *Stereotype threat, cognitive aptitude measures, and social identity*. (Unpublished doctoral dissertation) Howard University, Washington, DC.

Marx, D. M., & Stapel, D. A. (2005). It's all in the timing: Measuring emotional reactions to stereotype threat before and after taking a test. *European Journal of Social Psychology, 36,* 687–698.

Marx, D. M., & Stapel, D. A. (2006a). Distinguishing stereotype threat from priming effects: On the role of the social self and threat-based concerns. *Journal of Personality and Social Psychology, 91,* 243–254.

Marx, D. M., & Stapel, D. A. (2006b). Retraction: It's all in the timing: Measuring emotional reactions to stereotype threat before and after taking a test. *European Journal of Social Psychology, 47,* 933.

Marx, D. M., & Stapel, D. A. (2013). Retraction: Distinguishing stereotype threat from priming effects: On the role of the social self and threat-based concerns. *Journal of Personality and Social Psychology, 91,* 196.

*Marx, D. M., Stapel, D. A., & Muller, D. (2005). We can do it: The interplay of construal orientation and social comparisons under threat. *Journal of Personality and Social Psychology, 88,* 432–446. http://dx.doi.org/10.1037/0022-3514.88.3.432

*McFarland, L. A., Kemp, C. F., Viera, L., Jr., & Odin, E. P. (2003, April). *Stereotype threat and male–female differences in test performance*. Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

*McFarland, L. A., Lev-Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16,* 181–205. http://dx.doi.org/10.1207/S15327043HUP1603_2

*McIntyre, R. B., Lord, C. G., Gresky, D. M., Eyck, L. L. T., Frye, G. D. J., & Bond, C. F., Jr. (2005). A social impact trend in the effects of role models on alleviating women's mathematics stereotype threat. *Current Research in Psychology, 10,* 116–136.

*McIntyre, R. B., Paulson, R. M., & Lord, C. G. (2003). Alleviating women's mathematics stereotype threat through salience of group achievements. *Journal of Experimental Social Psychology, 39,* 83–90. http://dx.doi.org/10.1016/S0022-1031(02)00513-9

*McKay, P. F. (1999). *Stereotype threat and its effect on the cognitive ability test performance of African-Americans: The development of a theoretical model*. (Unpublished doctoral dissertation) University of Akron, Akron, Ohio.

*Nguyen, H-H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16,* 261–293. http://dx.doi.org/10.1207/S15327043HUP1603_5

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,* 1314–1334. http://dx.doi.org/10.1037/a0012702

*Nguyen, H. H. D., Shivpuri, S., Ryan, A. M., & Langset, K. (2004, April). *Relations of stereotype threat effects to assessment domains and self-identity*. Paper presented at the 19th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

*O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29,* 782–789. http://dx.doi.org/10.1177/0146167203029006010

*Oswald, D. L., & Harvey, R. D. (2000). Hostile environments, stereotype threat, and math performance among undergraduate women. *Current Psychology, 19,* 338–356. http://dx.doi.org/10.1007/s12144-000-1025-5

*Pellegrini, A. V. (2005). *The impact of stereotype threat on intelligence testing in Hispanic females*. (Unpublished doctoral dissertation) Carlos Albizu University, Miami, Florida.

*Philipp, M. C., & Harton, H. C. (2004, January). *The role of social dominance in stereotype threat effects*. Paper presented at the annual meeting of the Society for Personality and Social Psychology, Austin, TX.

*Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16,* 231–259. http://dx.doi.org/10.1207/S15327043HUP1603_4

*Prather, H. M. (2005). *Controlling the threat of stereotypes: The effectiveness of mental control strategies in increasing female math ability test performance*. (Unpublished doctoral dissertation) George Washington University, Washington, DC.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

*Rivadeneyra, R. (2001). *The influence of television on stereotype threat among adolescents of Mexican descent*. (Unpublished doctoral dissertation) University of Michigan, Ann Arbor, Michigan.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

*Salinas, M. F. (1998). *Stereotype threat: The role of effort withdrawal and apprehension on the intellectual underperformance of Mexican-Americans*. (Unpublished doctoral dissertation) University of Texas at Austin.

*Sawyer, T. P., Jr., & Hollis-Sawyer, L. A. (2005). Predicting stereotype threat, test anxiety, and cognitive ability test performance: An examination of three models. *International Journal of Testing, 5,* 225–246. http://dx.doi.org/10.1207/s15327574ijt0503_3

*Schimel, J., Arndt, J., Banko, K. M., & Cook, A. (2004). Not all self-affirmations were created equal: The cognitive and social benefits of affirming the intrinsic (vs. extrinsic) self. *Social Cognition, 22,* 75–99. http://dx.doi.org/10.1521/soco.22.1.75.30984

*Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology, 38,* 194–201. http://dx.doi.org/10.1006/jesp.2001.1500

*Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85,* 440–452. http://dx.doi.org/10.1037/0022-3514.85.3.440

*Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles, 50,* 835–850. http://dx.doi.org/10.1023/B:SERS.0000029101.74557.a0

Schmidt, F. L., & Le, H. A. (2005). The Hunter–Schmidt meta-analysis programs package (Version 1.1) [Computer software]. Retrieved from http://www.testpublishers.org/Documents/FrankSchmidtSoftware.pdf

*Schneeberger, N. A., & Williams, K. (2003, April). *Why women "can't" do math: The role of cognitive load in stereotype threat research*. Paper presented at the 18th meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

*Schultz, P. W., Baker, N., Herrera, E., & Khazian, A. (n.d.). *Stereotype threat among Hispanic-Americans and the moderating role of ethnic identity*. Unpublished manuscript.

*Seagal, J. D. (2001). *Identity among members of stigmatized groups: A double-edged sword*. (Unpublished doctoral dissertation) University of Texas at Austin.

*Sekaquaptewa, D., & Thompson, M. (2002). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39,* 68–74. http://dx.doi.org/10.1016/S0022-1031(02)00508-5

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143,* 534–547. http://dx.doi.org/10.1037/a0033242

*Smith, C. E., & Hopkins, R. (2004). Mitigating the impact of stereotypes on academic performance: The effects of cultural identity and attributions for success among African American college students. *The Western Journal of Black Studies, 28,* 312–321.

*Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47,* 179–191. http://dx.doi.org/10.1023/A:1021051223441

*Spencer, S. L. (2005). *Stereotype threat and women's math performance: The possible mediating factors of test anxiety, test motivation and self-efficacy*. (Unpublished doctoral dissertation) Rutgers, The State University of New Jersey, New Brunswick, New Jersey.

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

*Spicer, C. V. (1999). *Effects of self-stereotyping and stereotype threat on intellectual performance*. (Unpublished doctoral dissertation) University of Kentucky, Lexington, Kentucky.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5,* 60–78. http://dx.doi.org/10.1002/jrsm.1095

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64,* 70–77. http://dx.doi.org/10.1198/tast.2009.08205

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49,* 108–112.

*Sternberg, R. J., Jarvin, L., Leighton, J., Newman, T., Moon, T., Callahan, C., & Grigorenko, E. L. (n.d.). *Girls can't do math?: The disidentification effect and gifted high school students' math performance*. Unpublished manuscript.

Sterne, J. A. C., & Egger, M. (2000). High false positive rate for trim and fill method. *British Medical Journal, 320,* 1574–1577.

*Stricker, L. J., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and gender, and standardized test performance. *Journal of Applied Social Psychology, 34,* 665–693. http://dx.doi.org/10.1111/j.1559-1816.2004.tb02564.x

*Tagler, M. J. (2003). *Stereotype threat: Prevalence and individual differences*. (Unpublished doctoral dissertation) Kansas State University, Manhattan, Kansas.

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113–2126. http://dx.doi.org/10.1002/sim.1461

*van Dijk, A., Koenders, H., Korenhof, I. H., Mulder, H. R., & de Vries, H. (n.d.). *The moderating role of group membership activation on stereotype lift and threat*. Unpublished manuscript.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48. http://dx.doi.org/10.18637/jss.v036.i03

*von Hippel, W., von Hippel, C., Conway, L., Preacher, K. J., Schooler, J. W., & Radvansky, G. A. (2005). Coping with stereotype threat: Denial as an impression management strategy. *Journal of Personality and Social Psychology, 89,* 22–35. http://dx.doi.org/10.1037/0022-3514.89.1.22

*Walsh, M., Hickey, C., & Duffy, J. (1999). Influence of item content and stereotype situation on gender differences in mathematical problem solving. *Sex Roles, 41,* 219–240. http://dx.doi.org/10.1023/A:1018854212358

*Walters, A. M. (2000). *Stereotype threat: An examination of process*. (Unpublished doctoral dissertation) University of Florida, Gainesville, Florida.

*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,* 696–716. http://dx.doi.org/10.1037/0022-3514.89.5.696

*Wout, D. A., Shih, M. J., Jackson, J. S., & Sellers, R. M. (n.d.). *Targets as perceivers: How Blacks determine if they will be stereotyped*. Unpublished manuscript.

Received March 2, 2016
Revision received November 7, 2016
Accepted November 9, 2016 ■