Stereotype Lift

Gregory M. Walton* and Geoffrey L. Cohen
Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06520, USA
Received 25 September 2001; revised 9 October 2002

Abstract

When a negative stereotype impugns the ability or worth of an outgroup, people may experience stereotype lift—a performance boost that occurs when downward comparisons are made with a denigrated outgroup. In a meta-analytic review, members of non-stereotyped groups were found to perform better when a negative stereotype about an outgroup was linked to an intellectual test than when it was not ($d = .24$, $p < .0001$). Notably, people appear to link negative stereotypes to evaluative tests more or less automatically. Simply presenting a test as diagnostic of ability was thus sufficient to induce stereotype lift. Only when negative stereotypes were explicitly invalidated or rendered irrelevant to the test did the lift effect disappear.

© 2003 Elsevier Science (USA). All rights reserved.

Keywords: Stereotype; Stereotype threat; Performance; Downward comparison; White privilege; Academic achievement; Prejudice; Race; Gender

Stereotype lift is the performance boost caused by the awareness that an outgroup is negatively stereotyped. People may benefit from stereotype lift when the ability or worth of an outgroup is explicitly called into question. But they may also benefit even when there is no specific reference to a stereotyped outgroup, if the performance task is linked to a widely known negative stereotype. Because stereotypes about the intellectual abilities of different social groups are pervasive in American society (Devine, 1989), people may link negative stereotypes to intellectual tests more or less automatically. As research on stereotype threat suggests (Spencer, Steele, & Quinn, 1999; Steele & Aronson, 1995; for a review see Steele, 1997), negative stereotypes about women’s ability in math, and about racial minorities’ intellectual abilities in general, are so ubiquitous that evaluative tests can trigger among members of these groups the fear that, should they do poorly, they could confirm a negative stereotype about their gender or racial group. This fear, in turn, can cause their performance to suffer (see also Cohen & Steele, 2002; Gonzales, Blanton, & Williams, 2002). We argue that negative stereotypes are linked to intellectual tests in the minds of members of non-stereotyped groups too (e.g., men and Whites), but that here their effect is to enhance performance rather than to undermine it.

How do negative outgroup stereotypes improve performance? They do so, we suggest, by encouraging downward social comparisons with a denigrated outgroup (see Blanton, Buunk, Gibbons, & Kuyper, 1999; see also Fein & Spencer, 1997). By comparing themselves with a socially devalued group, people may experience an elevation in their self-efficacy or sense of personal worth (Bandura, 1986), which may, in turn, improve performance. Particularly for difficult tests where one must persist in the face of frustration, the extra boost in feelings of efficacy and worth may be important to maintaining confidence and motivation. Knowing that another group is stereotyped as inferior to their own, people may also expect to be viewed with respect rather than suspicion (see Cohen, Steele, &

*Corresponding author. Fax: 1-203-432-7172.
E-mail address: gregory.walton@yale.edu (G.M. Walton).
Ross, 1999; Walton & Cohen, 2003). Non-stereotyped test-takers may feel that even failure is unlikely to bring them the low status or social rejection faced by the outgroup, and this assurance may help them to maintain optimal performance (see also Baumeister, Twenge, & Nuss, 2002). In sum, stereotype-inspired social comparison may alleviate the self-doubt, anxiety, and fear of rejection that could otherwise hamper performance on important intellectual tests (see also Sarason, 1991).

As past research suggests, certain individuals are more likely than others to engage in downward social comparison. For example, people who value a given achievement domain as a source of self-esteem (who are identified; Steele, 1997) may view the possibility of failure on a test as particularly self-threatening and may thus welcome the opportunity to buttress their self-worth by comparing themselves with a stereotyped outgroup (see Fein & Spencer, 1997; Hogg, 2000; Tesser, 1988; Wills, 1981). In addition, people who are prejudiced may also be particularly likely to experience stereotype lift, because they view the outgroup more negatively than do people low in prejudice (Wills, 1981).

We provide evidence for stereotype lift using meta-analysis. Relevant studies encompass those that assess the performance of members of a non-stereotyped group in at least two conditions—one condition where a negative outgroup stereotype is linked or made relevant to a test, and another condition where such a stereotype is not linked or made irrelevant to that test. The data are drawn largely from the non-stereotyped participants used as controls in stereotype threat research. We statistically combine results from these studies to assess whether performance on the part of non-stereotyped people is higher in the stereotype-relevant condition than in the stereotype-irrelevant condition.

We have argued that people link negative stereotypes to intellectual tests more or less automatically, and that they thus experience stereotype lift even when there is no explicit reference to an outgroup. To test this idea, we examine the size of the lift effect as a function of the types of experimental manipulations used in different studies. Almost all the studies included in our sample used an evaluative test in the stereotype-relevant condition, that is, one presented as diagnostic of ability (the two exceptions were Skaquaptewa & Thompson, 2002, Study 2, and Steele & Aronson, 1995, Study 4). If people automatically link such evaluative tests to negative stereotypes, then there should be little if any additional benefit of making that link explicit in the stereotype-relevant condition. That is, studies that explicitly validate the stereotype (e.g., by stating that women perform worse than men on a math test) should show no greater lift effect than studies that do not explicitly validate the stereotype (e.g., by simply presenting a test as diagnostic of math ability). Making the link between the test and the stereotype explicit should have little if any effect if that link is already implicit. Rather, what should be predictive of stereotype lift is whether the stereotype-irrelevant condition adequately refutes the assumed link between the test and the negative stereotype—for example, by presenting the test as not diagnostic of ability or by stating that the test yields no group differences in performance.

Method

Retrieval of studies and inclusion criteria

To retrieve relevant studies, we first conducted a January 2002 search of the PsychINFO database using the words “stereotype,” “threat,” and “performance” or “test.” Second, we solicited additional studies by emailing the discussion list of the Society for Personality and Social Psychology (spsp@stolaf.edu). Finally, we emailed the first author of every study obtained through each of these two methods.

Our inclusion criteria required that studies assess the test performance of members of a real-world, self-identified, and non-stereotyped group (e.g., men and Whites). Participants had to be randomly assigned to one of at least two conditions—one “stereotype-relevant” condition and one “stereotype-irrelevant” condition. The performance test had to be linked to a negative stereotype more in the stereotype-relevant condition than in the stereotype-irrelevant condition. The manipulation could be accomplished through explicit instructions that implied the relevance or validity (or irrelevance or invalidity) of a negative stereotype. Alternatively, the manipulation could be accomplished through features of the test-taking environment that could increase (or decrease) the perceived relevance of a stereotype. For example, male students completing a math test could plausibly infer from the presence of predominantly female test-takers that one purpose of the study involved examining gender differences in math. Studies that used implicit manipulations (e.g., with subliminal stimuli) or manipulations embedded in tasks ostensibly unrelated to the test were thus excluded. Such studies either (a) activate the content of the stereotype without linking that content to a specific outgroup (e.g., Bargh, Chen, & Burrows, 1996; see Steele, Spencer, & Aronson, 2002), or (b) activate the stereotype without making it psychologically relevant to the test.

Finally, because stereotype lift is assumed to alleviate the doubt, anxiety, or fear of rejection that accompanies the threat of failure, it was also required that each included study use a difficult test rather than an easy one. Ultimately, 43 studies meeting these criteria were
identified.\(^1\) For each study, we calculated the standard deviation (SD), the size of the stereotype lift effect (d), the t test statistic (t), and the p value corresponding to the t test.\(^2\)

We also used a coding system to test the claim that people link intellectual tests to negative stereotypes automatically. If this claim is accurate, then studies that use a stereotype-relevant treatment condition where an evaluative test is explicitly linked to a negative stereotype will not yield a greater lift effect than studies that use a stereotype-relevant treatment condition where the test is merely presented as evaluative in nature. Adding further relevance to the stereotype will have little if any effect if its relevance is assumed. By contrast, studies that use a stereotype-irrelevant condition where the link between the test and the stereotype is refuted (either by presenting the test as not-diagnostic of ability or by presenting it as insensitive to group differences) will yield a substantially larger lift effect than studies that use a stereotype-irrelevant condition where that link is not refuted.

To test our reasoning, we coded each study on two dichotomous scales to assess: (a) whether the stereotype-relevant condition reinforced the link between the test and the stereotype, and (b) whether the stereotype-irrelevant condition refuted that link. The stereotype-relevant condition was coded as a 1 if the link between the test and the stereotype was explicit and thus reinforced (i.e., if the test was presented not only as being diagnostic of ability, but also as yielding group differences in performance) and as a 0 if that link was implicit (e.g., if the test was presented as only diagnostic of ability). The stereotype-irrelevant control condition was coded as a 1 if the link between the test and the stereotype was refuted (e.g., if the test was presented as not diagnostic of ability or if it was characterized as not yielding group differences in performance) and as a 0 if that link was not refuted (i.e., if the test was presented as diagnostic of ability).

For example, studies that manipulated the perceived diagnosticity of the test (e.g., Steele & Aronson, 1995, Studies 1 and 2) were coded as a 0 for the stereotype-relevant condition (because the test in this condition was presented as diagnostic of ability but was not said to yield group differences in performance) and as a 1 for the stereotype-irrelevant condition (because the test in this condition was presented as non-diagnostic of ability and hence as irrelevant to a negative stereotype). In contrast, studies that manipulated the perception that men outperformed women on a math test (e.g., Spencer et al., 1999) were coded as a 1 for the stereotype-relevant condition (because the test in this condition was said to yield gender differences) and as a 1 for the stereotype-irrelevant condition (because the test in this condition was presented as yielding no gender differences and hence as irrelevant to a negative stereotype). Finally, studies that manipulated stereotype relevance by asking participants to note their race or gender before an evaluative test (i.e., Anderson, 2002; Stricker, 1998; Stricker & Ward, 1998) were coded as a 0 for the stereotype-relevant condition (because the test in this condition was not said to yield group differences in performance) and as a 0 for the stereotype-irrelevant condition (because the test in this condition was presented as diagnostic of ability and hence as relevant to a negative stereotype).

**Results**

**Overview of sample**

The key characteristics of each study included in the meta-analysis are presented in Table 1. The summary statistics and condition codes are presented in Table 2. As noted in Table 2, 30 of the 43 studies showed the predicted pattern of results (in four studies the results were statistically significant). The stereotype-relevant condition yielded higher performance on the part of members of non-stereotyped groups than did the stereotype-irrelevant condition. The effect sizes were normally distributed (skewness = .06, SE = .36, ns; kurtosis = .26, SE = .71, ns).

**Overall tests of homogeneity of effect sizes and statistical significance**

We followed procedures outlined by Hedges and Olkin (1985) to calculate all meta-analytic statistics. The homogeneity test assesses whether the sampled effect sizes are likely to derive from the same population. They were not, \(Q_7 = 59.65, \chi^2_6, \ p = .038\), indicating variance in effect sizes among the sampled studies (a point to
<table>
<thead>
<tr>
<th>Study No.</th>
<th>Study</th>
<th>Participants</th>
<th>Stereotype</th>
<th>Manipulation of stereotype relevance</th>
<th>Dependent measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Anderson (2002)</td>
<td>Male students</td>
<td>Gender</td>
<td>Before ability-diagnostic test, participants either indicated race or not</td>
<td>Math and science general education instrument</td>
</tr>
<tr>
<td>2</td>
<td>Blascovich, Spencer, Quinn, and Steele (2001)</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as diagnostic of “intelligence” or as “culturally unbiased” (i.e., yielding no race differences)</td>
<td>Remote Associates Test</td>
</tr>
<tr>
<td>3</td>
<td>Broadnax, Crocker, and Spencer (1997)</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as an “excellent indicator of English competence” or as a “culture fair” test yielding “no race differences”</td>
<td>English literature test</td>
</tr>
<tr>
<td>4</td>
<td>Brown, Steele, and Atkins (2001), Study 1</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as diagnostic of verbal skills or as “racially and ethnically unbiased”</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>5</td>
<td>Brown et al. (2001), Study 2</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as diagnostic of verbal skills or as “racially and ethnically unbiased”</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>6</td>
<td>Croizet and Claire (1998)</td>
<td>High SES students</td>
<td>SES</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of “intellectual ability”</td>
<td>GRE-like verbal questions</td>
</tr>
<tr>
<td>7</td>
<td>Danso and Esses (2001), Pilot Study</td>
<td>White students</td>
<td>Race</td>
<td>Either Black or White experimenter administered ability-diagnostic test</td>
<td>Necessary Arithmetic Operations</td>
</tr>
<tr>
<td>8</td>
<td>Danso and Esses (2001), Main Study</td>
<td>White students</td>
<td>Race</td>
<td>Either Black or White experimenter administered ability-diagnostic test</td>
<td>Necessary Arithmetic Operations</td>
</tr>
<tr>
<td>9</td>
<td>Davies et al. (2001), Pilot Study</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of mathematical ability</td>
<td>Math GRE</td>
</tr>
<tr>
<td>10</td>
<td>Ewing and Smith (2001)</td>
<td>Young students</td>
<td>Age</td>
<td>Participants either told aging is related to memory loss or told aging is unrelated to memory loss</td>
<td>Weschler Adult Intelligence Scale-III</td>
</tr>
<tr>
<td>11</td>
<td>Foels (2000)</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as a math test or as a math test yielding “no gender differences”</td>
<td>Math GMAT</td>
</tr>
<tr>
<td>12</td>
<td>Gonzales et al. (2002), analyses by gender</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of mathematical and spatial abilities</td>
<td>Wonderlic Personnel Test</td>
</tr>
<tr>
<td>13</td>
<td>Gonzales et al. (2002), analyses by race</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of mathematical and spatial abilities</td>
<td>Wonderlic Personnel Test</td>
</tr>
<tr>
<td>14</td>
<td>Inzlicht and Ben-Zeev (2000), Study 2</td>
<td>Male students</td>
<td>Gender</td>
<td>Groups composed either of one man and two women or of three men completed ability-diagnostic test</td>
<td>Math GRE</td>
</tr>
<tr>
<td>15</td>
<td>Josephs, Newman, Brown, and Beer (2001), Study 1</td>
<td>Male students</td>
<td>Gender</td>
<td>Before ability-diagnostic test, participants either completed questions about gender and math or completed questions about school</td>
<td>Math GRE</td>
</tr>
<tr>
<td>16</td>
<td>Keller (2002)</td>
<td>Male adolescents</td>
<td>Gender</td>
<td>Test characterized either as yielding “gender differences” or as yielding no gender differences</td>
<td>26 math problems</td>
</tr>
<tr>
<td>17</td>
<td>Keller and Dauenheimer (2002)</td>
<td>Male adolescents</td>
<td>Gender</td>
<td>Test characterized either as yielding “gender differences” or as yielding no gender differences</td>
<td>26 math problems</td>
</tr>
<tr>
<td>18</td>
<td>Martens, Johns, Greenberg, and Schimmel (2002)</td>
<td>Male students</td>
<td>Gender</td>
<td>Test either said to assess “math and reasoning abilities” or said to be a pilot test for future studies</td>
<td>Math GMAT</td>
</tr>
<tr>
<td>19</td>
<td>Marx and Roman (2002), Study 1</td>
<td>Male students</td>
<td>Gender</td>
<td>Either female or male experimenter administered ability-diagnostic test</td>
<td>Math GRE</td>
</tr>
<tr>
<td>20</td>
<td>Marx and Roman (2002), Study 2</td>
<td>Male students</td>
<td>Gender</td>
<td>Either math-incompetent or math-competent female experimenter administered ability-diagnostic test</td>
<td>Math GRE</td>
</tr>
<tr>
<td>Study No.</td>
<td>Study</td>
<td>Participants</td>
<td>Stereotype</td>
<td>Manipulation of stereotype relevance</td>
<td>Dependent measure</td>
</tr>
<tr>
<td>----------</td>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
<td>-------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>21</td>
<td>McKay, Doverspike, Bowen-Hilton, and Martin (2002)</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of intellectual ability</td>
<td>Raven Advanced Progressive Matrices</td>
</tr>
<tr>
<td>22</td>
<td>Quinn and Spencer (2001) Study 2</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as “math test” or as “gender-fair” math test (i.e., yielding no gender differences)</td>
<td>Math SAT</td>
</tr>
<tr>
<td>23</td>
<td>Salinas (1998), Study 1</td>
<td>White students</td>
<td>Race</td>
<td>Midway through the ability-diagnostic test, participants either had the opportunity to label the test as biased or did not have that opportunity</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>24</td>
<td>Salinas (1998), Study 2</td>
<td>White students</td>
<td>Race</td>
<td>Midway through the ability-diagnostic test, participants either had the opportunity to label the test as biased or did not have that opportunity</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>25</td>
<td>Schmader (2002)</td>
<td>Male students</td>
<td>Gender</td>
<td>Study said to assess either gender differences in math ability or individual differences in math ability</td>
<td>Math GRE</td>
</tr>
<tr>
<td>26</td>
<td>Schultz, Baker, Herrera, and Khazian (2002), Study 1</td>
<td>White students</td>
<td>Race</td>
<td>Study said to assess either racial differences in verbal skills or individual differences in verbal skills</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>27</td>
<td>Schultz et al. (2002), Study 2</td>
<td>White students</td>
<td>Race</td>
<td>Study said to assess either racial differences in verbal skills or individual differences in verbal skills</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>28</td>
<td>Sekaquaptewa and Thompson (2001)</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as a “traditional” math test or as “special kind of math material” (i.e., yielding no gender differences)</td>
<td>12 math and math-related questions</td>
</tr>
<tr>
<td>29</td>
<td>Sekaquaptewa and Thompson (2002), Study 2</td>
<td>White students</td>
<td>Race</td>
<td>Groups composed either of one White and three Blacks or of four Whites completing non-diagnostic test</td>
<td>16 questions about classifications of animal species</td>
</tr>
<tr>
<td>30</td>
<td>Spencer (1993), Study 4</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as yielding “gender differences” or as yielding “no gender differences”</td>
<td>Math GRE</td>
</tr>
<tr>
<td>31</td>
<td>Spencer (1993), Study 5</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as a math test, as a math test yielding “gender differences,” or as a math test yielding “no gender differences”</td>
<td>Math GRE</td>
</tr>
<tr>
<td>32</td>
<td>Spencer, Iserman, Davies, and Quinn (2002)</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as “diagnostic” of math ability or as yielding “no gender differences”</td>
<td>Math GRE under cognitive load</td>
</tr>
<tr>
<td>33</td>
<td>Spencer et al. (1999), Study 2</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as yielding “gender differences” or as yielding “no gender differences”</td>
<td>Math GRE</td>
</tr>
<tr>
<td>34</td>
<td>Spencer et al. (1999), Study 3</td>
<td>Male students</td>
<td>Gender</td>
<td>Test characterized either as yielding “gender differences” or as yielding “no gender differences”</td>
<td>Math GMAT</td>
</tr>
<tr>
<td>35</td>
<td>Steele and Aronson (1995), Study 1</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of verbal ability</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>36</td>
<td>Steele and Aronson (1995), Study 2</td>
<td>White students</td>
<td>Race</td>
<td>Test characterized either as “diagnostic” or as “non-diagnostic” of verbal ability</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>37</td>
<td>Steele and Aronson (1995), Study 4</td>
<td>White students</td>
<td>Race</td>
<td>Before non-diagnostic test, participants either indicated race or not</td>
<td>Verbal GRE</td>
</tr>
<tr>
<td>38</td>
<td>Sternberg et al. (2002), Study 1</td>
<td>Male students</td>
<td>Gender</td>
<td>Either men either said to perform better than women, or gender not mentioned; or women said to perform better than men, or men and women said to perform equally well</td>
<td>Math GRE</td>
</tr>
</tbody>
</table>
which we return shortly). After correcting for small sample sizes (Hedges & Olkin, 1985), we conducted the initial meta-analytic test to assess the overall effect size and significance level of stereotype lift. The effect size was $d = .10$, and it differed significantly from $d = 0$, $z = 3.40$, $p < .01$. We also calculated the fail-safe $n$ (Rosenthal, 1984), and found that an additional 122 studies, each one yielding a null stereotype lift effect (i.e., $d = 0$), would be required to render this primary result non-significant. However, the heterogeneity in effect sizes suggests that one or more variables moderate stereotype lift.

The moderating role of stereotype relevance

As discussed in the Introduction, one potential moderator involves whether the stereotype-irrelevant condition severs the link between the test and the stereotype. If people link evaluative tests to negative stereotypes more or less automatically, studies will yield a between-condition difference only if the stereotype-irrelevant condition refutes either the content of the stereotype or its relevance to the test. To test this idea, we conducted an internal analysis using the two dichotomous codes assigned to each study.

As predicted, the size of the stereotype lift effect did not differ between the two types of stereotype-relevant treatment conditions, $Q_B = \chi^2 = .14, n.s.$ Studies that made the link between the test and the stereotype explicit (i.e., by stating that an ability-diagnostic test showed group differences in performance) yielded roughly the same effect size ($d = .11$) as studies that simply presented the test as diagnostic of ability ($d = .07$). The link between negative outgroup stereotypes and evaluative tests is thus so strong that there is no additional benefit of making that link explicit. By contrast, there was a significant effect involving the two types of stereotype-irrelevant control conditions, $Q_B = \chi^2 = 8.66, p = .003$. Studies that refuted either the validity of the stereotype or its relevance to the test yielded a stronger lift effect ($d = .22$) than studies that did not ($d = .00$). Whether the treatment condition reinforces the relevance of a negative stereotype beyond what is normally assumed for the typical evaluative test does not predict larger lift effects. On the contrary, what is predictive is whether (in the control condition) the link between the test and the stereotype is refuted.

The moderating role of identification

We found tentative evidence that people who are identified with the performance domain benefit most from

---

3 For these internal analyses, we excluded the two studies that presented the test as non-evaluative in both conditions (i.e., Serkaquaptewa & Thompson, 2002, Study 2; Steele & Aronson, 1995, Study 4).
There was a trend in the predicted direction for studies that preselected participants for high levels of identification with the subject matter being tested (or that preselected on the basis of a relevant criterion variable such as SAT score) to show a somewhat larger lift effect ($d = .20$) than studies that did not ($d = .07$). $Q_d = \chi^2 = 2.05, p = .15$. Like stereotype threat (Steele, 1997), stereotype lift seems to have a larger effect on people who care most about doing well. While stereotype threat exacerbates pressures felt by people who identify with the relevant category more strongly, stereotype lift may provide a way for people to overcome these pressures through identification with a more relevant domain.
people who have invested their self-worth in the domain of evaluation, stereotype lift ameliorates those pressures.

Estimation of effect size and clinical significance

The results indicate that people link intellectual tests to negative stereotypes more or less automatically. Only studies that refute that link in the stereotype-irrelevant condition thus provide an accurate estimate of the stereotype lift effect size. We recalculated the meta-analytic statistics using only these studies (n = 28). The lift effect was robust, \( d = .24 \), highly significant, \( z = 4.04, p < .0001 \), and yielded a fail-safe n of 132. The homogeneity statistic was not significant, \( QT = \chi^2 = 24.63, ns \), indicating the consistency of the lift effect in this sample of studies.

Stereotype lift can help to explain racial and gender differences in standardized test performance. Several studies included in the meta-analysis used modified SAT tests (or similar standardized tests) as dependent measures. The standard deviation on each section of the SAT is approximately 112 points (College Board, 2001). The obtained lift effect size (\( d = .24 \)) thus indicates that stereotype lift improves performance on each stereotype-relevant section of the SAT by approximately 25 points. Stereotype lift produces a 50-point advantage for White men—a performance boost that, at the most selective colleges, could make the difference between rejection and acceptance.\(^4\)

Comparative analysis of stereotype threat

For exploratory purposes, we performed the same analyses on the effect sizes obtained for stereotype threat.

\(^4\) Additional analyses were undertaken to rule out alternative explanations and to test other potential moderators of the lift effect. Stereotype lift does not appear to be due to heightened effort on evaluative tests relative to non-evaluative ones. When analyses were confined to studies that presented the test as equally evaluative in both conditions (i.e., equally diagnostic of ability, \( n = 35 \)), the overall lift effect was very consistent with that obtained in the entire sample, \( d = .09, z = 2.68, p = .007; QT = \chi^2 = 61.01, p = .003 \). The lift effect was again larger among studies using stereotype-relevant control conditions that refuted the link between the test and the stereotype (\( d = .23 \)) than among studies that did not (\( d = .00 \)), \( QT = \chi^2 = 7.91, p = .005 \). It also seemed possible that statistical dependency among studies from the same laboratory might violate the assumption of independence. Twenty laboratories are represented in the sample (i.e., studies reported by the same researchers or by overlapping groups of researchers). When laboratory was used as the unit of analysis, the overall lift effect was again very consistent with that obtained when study was the unit of analysis, \( d = .11, z = 3.34, p = .001; QT = \chi^2 = 36.86, p = .008 \). Once more, the lift effect was larger among studies using stereotype-relevant control conditions that refuted the link between the test and the stereotype (\( d = .24 \)) than among studies that did not (\( d = .02 \)), \( QT = \chi^2 = 7.93, p = .005 \). Two final moderators were also assessed. Neither publication status, \( QT = \chi^2 = 3.04, p > .05 \), nor type of stereotype (i.e., racial stereotypes or gender stereotypes), \( QT = \chi^2 = 1.50, p > .20 \), predicted stereotype lift.

The overall threat effect was significant, \( d = .29, z = 10.44, p < .0001 \). The stereotype-relevant condition yielded lower performance on the part of members of stereotyped groups than did the stereotype-irrelevant condition. The heterogeneity test was also significant, \( Q_T = 84.98 = \chi^2_{28}, p < .0001 \), suggesting that one or more variables moderate stereotype threat.

Like stereotype lift, the stereotype threat effect was larger among studies that refuted the link between the test and the stereotype in the stereotype-irrelevant condition (\( d = .45 \)) than among studies that did not (\( d = .20 \)), \( QT = \chi^2 = 9.61, p = .002 \). This result indicates that, like their non-stereotyped peers, Blacks, women, and members of other historically excluded groups link evaluative tests to negative stereotypes more or less automatically. In contrast to stereotype lift, however, the stereotype threat effect was also larger among studies that reinforced the link between the test and the stereotype in the stereotype-relevant condition (\( d = .57 \)) than among studies that did not (\( d = .29 \)), \( QT = \chi^2 = 5.84, p = .016 \), although this result should be viewed tentatively because it is driven entirely by five studies with disproportionately large sample sizes.\(^5\)

Both targets and non-targets thus link intellectually evaluative tests to negative stereotypes (Steele, 1997). Moreover, like stereotype lift, stereotype threat was larger among studies that selected students who were identified with the performance domain (\( d = .68 \)) than among studies that did not (\( d = .22 \)), \( QT = \chi^2 = 22.04, p < .0001 \), a result consistent with the idea that students at the vanguard of achievement suffer stereotype threat most (see Aronson et al., 1999; Spencer et al., 1999; Steele, 1997).

Because intellectual tests appear to be linked to negative stereotypes more or less automatically, we calculated the stereotype threat effect size using only studies that refuted that link in the stereotype-irrelevant condition (\( n = 28 \)). The effect was robust and significant,

\(^5\) These five studies (Anderson, 2002; Stricker, 1998 analyses by gender and by race; Stricker & Ward, 1998, analyses by gender and by race) did not reinforce the link between the test and the stereotype in the stereotype-relevant condition and they yielded small stereotype threat effects (\( ds = 0.00 \) to .19). On average, they used samples sizes 10 times larger than those of the other studies, and thus may have over-determined the mean effect size for this category of studies. To address this problem, we assessed the effect of the two types of stereotype-relevant conditions using an independent samples \( t \) test (an analysis insensitive to the sample size used to derive each effect size). In this analysis, there was no difference in the mean effect size observed in the two types of stereotype-relevant conditions, \( t(34) = .25, ns \). By contrast, the \( t \) test analysis largely revealed the same moderational patterns found using the meta-analytic statistics reported earlier. There was a trend in the predicted direction such that the type of stereotype-irrelevant condition predicted stereotype threat, \( t(37) = 1.45, p = .15 \). Likewise, while there was no difference in the mean stereotype lift effect size observed in the two types of stereotype-relevant conditions, \( t(36) = .35, ns \), there was a difference in the mean effect size observed in the two types of stereotype-irrelevant conditions, \( t(39) = 2.67, p = .011 \).
Danso & Esses (2001) found that people high in social and each result was consistent with this prediction. From stereotype lift. Three studies in our sample assessed (Wills, 1981), they should also be more likely to benefit larger lift effects than studies that did not. Identified with the subject matter being tested to yield performance for members of non-stereotyped groups.

From evidence of moderation to evidence of mediation If stereotype lift results from downward comparison with a denigrated outgroup, then people who are who are apt to engage in downward comparison should benefit most. People who are identified with the skill under evaluation should be more likely to view the prospect of failure as self-threatening and thus be motivated to buttress their self-worth through downward comparison (Fein & Spencer, 1997; see also Hogg, 2000; Tesser, 1988; Wills, 1981). Some support for this prediction was found. There was a trend for studies that selected participants identified with the subject matter being tested to yield larger lift effects than studies that did not.

Because prejudiced people are more likely to engage in downward comparison with a stereotyped outgroup (Wills, 1981), they should also be more likely to benefit from stereotype lift. Three studies in our sample assessed individual differences in prejudice or in related measures, and each result was consistent with this prediction. Danso & Esses (2001) found that people high in social dominance orientation—the motivation to protect the social hierarchy—performed better than people low in social dominance orientation in the presence of a Black experimenter but not in the presence of a White experimenter. Likewise, Josephs et al. (2001) found that men high in testosterone (a physiological marker of dominance orientation) performed marginally better on a math test than men low in testosterone after being primed on negative gender stereotypes. Finally, Schultz, Baker, Herrera, and Khazian (2002) reported that White students’ racial prejudice correlated positively with performance on a test portrayed as assessing racial differences but that racial prejudice did not correlate with performance on a test portrayed as assessing individual differences. Stereotype lift particularly benefits people who believe either in the validity of negative stereotypes, or in the legitimacy of group-based hierarchy.

If stereotype lift is driven by downward comparison, then other variables beyond identification and prejudice should moderate its effect. For example, people low in self-esteem appear to be particularly likely to make downward comparisons to protect their self-image (Aspinwall & Taylor, 1993; Smith & Insko, 1987; Wills, 1981, 1991), and they may thus experience stereotype lift more. Individuals low in self-certainty may also be more likely to experience stereotype lift, because their less stable self-concepts make them more susceptible to the effects of social comparison (Bui & Pelham, 2000; Pelham & Wachsmuth, 1995). If stereotype lift is mediated by downward social comparison, it should also produce specific psychological outcomes, including cognitive activation of the relevant stereotype and elevations in self-efficacy or in perceived social acceptance.

Stereotype lift is unlikely to be stereotype susceptibility Stereotype lift appears similar to stereotype susceptibility (Shih et al., 1999), the performance boost caused by activation of a positive ingroup stereotype. In contrast to stereotype susceptibility, however, stereotype lift is triggered not by a positive stereotype about one’s own group but by a negative stereotype about another group. While stereotype susceptibility focuses on groups targeted by positive stereotypes (e.g., Asians completing a math test), stereotype lift focuses on groups that are non-stereotyped. Whites, men, and other majority groups are considered normal and typical in most sectors of society (Miller, Taylor, & Buck, 1991). They are thus less likely to be targets of either negative stereotypes or positive ones (Miller et al., 1991). Indeed, Aronson et al. (1999, p. 40) report that among a large, diverse sample of undergraduates asked to describe stereotypes about various groups, not a single participant mentioned a stereotype about the intellectual abilities of Whites. Instead, they cited stereotypes about Blacks, Asians, and women. Thus, we think that stereotype lift is driven not by positive stereotypes about Whites, men, and wealthy people, but by negative stereotypes about Blacks, women, and poor people. This conclusion is buttressed by
the result (noted above) that lift effects are larger among people high in outgroup prejudice (Schultz et al., 2002; see also Danso & Esses, 2001; Josephs et al., 2001).

Implications of stereotype lift

Stereotype lift has at least four implications. First, it may improve intellectual attainment on evaluative tests that scrupulously avoid reference to race or gender. Second, stereotype lift may improve performance among many social groups rather than only one, insofar as each group has higher status than the stereotyped outgroup. Third, stereotype lift implies ingroup advantage indirectly rather than directly, and it is thus unlikely to cause the ironic, performance-debilitating effects that sometimes accompany the direct invocation of a positive ingroup stereotype (see Cheryan & Bodenhausen, 2000). Fourth, although the effects of stereotype lift may be subtle on any given test, its impact on the achievement of the nonstereotyped may be dramatic when its effects accumulate either within a large group of test-takers or across numerous performance opportunities for a single individual.

Stereotype lift complements stereotype threat by providing further evidence for the social-psychological origins of the Black–White test score gap and other group-based differences in achievement (see Jencks & Phillips, 1998; Steele, 1997). Because negative stereotypes are embedded in the social representation of standardized tests, these tests disadvantage certain groups and advantage others.

Acknowledgments

This research was supported by a research grant from the URO at Stanford University to Gregory Walton. We thank Joseph Brown, Peter Hegarty, Julio Garcia, Felicia Pratto, Claude Steele, Ewart Thomas, and Vanessa Robin Weersing for their valuable assistance and insightful comments. We also thank Keisha Burdick, Victoria Brescoll, Stacy Fambro, Elizabeth Levy Paluck, Aaron Sackett, Eric Uhlmann, Shirley Wang and others in the Department of Psychology at Yale University who provided helpful suggestions.

References

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


**Further reading**