Psychon Bull Rev (2014) 21:771–776
DOI 10.3758/s13423-013-0548-3

BRIEF REPORT

# Failure to replicate the Mehta and Zhu (2009) color-priming effect on anagram solution times

**Kenneth M. Steele**

Published online: 13 November 2013


**Abstract** Mehta and Zhu (*Science, 323*, 1226–1229, 2009) hypothesized that the color red induces avoidance motivation and that the color blue induces approach motivation. In one experiment, they reported that anagrams of avoidance motivation words were solved more quickly on red backgrounds and that approach motivation anagrams were solved more quickly on blue backgrounds. Reported here is a direct replication of that experiment, using the same anagrams, instructions, and colors, with more than triple the number of participants used in the original study. The results did not show the Mehta and Zhu color-priming effects, even though statistical power was sufficient to detect the effect. The results call into question the existence of their color-priming effect on the solution of anagrams.

**Keywords** Color · Priming · Approach · Avoidance · Motivation · Anagram

Few people would object to the hypothesis that specific colors may produce strong effects on a person. For instance, I may have a preference for, or react with pride to, a dark blue because of my association with a university years ago. I may be more or less aware of the cause of the emotion. An important question, however, is whether we should expect that color to induce a common affective or motivational state in a randomly selected sample of people. Recently, several articles have reported that certain colors prime specific motivational or affective states in particular contexts and that these priming effects change behavior predictably (Elliot & Maier, 2012). This article will focus on one of those studies.

K. M. Steele (✉)
Department of Psychology, Appalachian State University, Boone, NC 28608, USA
e-mail: steelekm@appstate.edu

In an article reported in *Science*, Mehta and Zhu (2009) hypothesized that exposure to the color red activates a state of avoidance motivation because of its association with danger and mistakes. This state causes people to become more vigilant and risk-averse, resulting in better performance on a task that requires attention to details. In contrast, exposure to blue activates a state of approach motivation due to its association with openness, peace, and tranquility. These associations signal a benign environment and facilitate performance on tasks that require innovative solutions.

Mehta and Zhu (2009) reported the results of six studies that investigated the ability of red to improve performance on detail-oriented tasks and blue to improve performance on creative tasks. The general procedure was a between-groups comparison among people assigned to either a red, a blue, or a neutral (white) condition over a wide range of tasks. Mehta and Zhu reported that exposure to the color red decreased the solution times of avoidance-related anagrams, increased preference for brands that stressed safety, and produced greater free recall in a memory task, greater proofreading accuracy, more practical toy designs, and higher favorability ratings for a camera ad centered on product details. In contrast, exposure to blue decreased solution times for approach-motivation anagrams, increased preference for brands that stressed style or adventure, increased the number of creative uses for a brick, produced toy designs that were more original, and increased the favorability ratings for a camera ad with a travel and adventure theme.

The Mehta and Zhu (2009) article has inspired subsequent studies. Rutchick, Slepian, and Ferris (2010) reported that people using red pens marked more errors and awarded lower grades than did people using blue pens. Genschow, Reutner, and Wänke (2012) investigated the effect of color on snack food consumption and reported that red inhibited consumption of both food and drink relative to blue. Smeesters and Liu (2011) reported that red produced contrast away from an

exemplar, whereas blue produced assimilation toward an exemplar. (The Smeesters and Liu report has been shown to be fraudulent and has been retracted: Simonsohn, 2013b; Smeesters & Liu, 2013.)

Despite the potential importance of the Mehta and Zhu (2009) findings, no studies since have replicated their effects. The purpose of the experiment reported here was to replicate directly the procedure of one of the Mehta and Zhu (2009) experiments. Their first experiment, concerning the effect of color on anagram solution times, was chosen because the effect sizes were large—Cohen's *d* values ranged from 0.81 to 1.1 (Mehta & Zhu, 2009, p. 1227)—and the experiment used a simple design that would be easy to replicate. The participants ($n = 69$) in the original experiment were randomly assigned to one of three color-background conditions (red, neutral [white], or blue) and asked to solve 12 anagrams. Mehta and Zhu reported a significant interaction effect: Words that activated avoidance motivation were solved more quickly on red backgrounds, and words that activated approach motivation were solved more quickly on blue backgrounds. The purpose of the experiment reported here was to replicate those findings.

## Method

### Participants

The recruitment goal was to obtain 270 participants. That count had been identified as producing a 95 % chance of detecting a small size interaction effect by the program G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007). A total of 269 students participated, but four of the students were excluded for failing a red–green color deficiency test, and the data from two students were unusable, leaving a total of 263 participants (80 men, 183 women). The participants received course credit. The study was approved by the Appalachian State University Institutional Review Board.

### Apparatus

Sessions were run on four Windows-based personal computers using E-Prime control software. Each computer had a Dell P2210 monitor that used a 4:3 aspect ratio mode, which resulted in a 49-cm diagonal display. Each monitor had been calibrated with a Spyder 4 colorimeter. Each computer was visually isolated from the others.

### Procedure

Participants were told that the experiment was an investigation of "how people solve puzzles" and that they were to be shown words in which the order of letters had been scrambled. The participant's task was to unscramble the letters and enter the original word. Nothing was said about color in the instructions. The experimenter then began the session and left the experimental room.

Two procedural aspects are noted here because they were not described in the published report or in the supplemental online materials (www.sciencemag.org/cgi/content/full/1169144/DC1) for Mehta and Zhu (2009), but they were described in an e-mail from the senior author (R. J. Zhu, personal communication, 2/20/12)[1]: The first one or two letters of the correct solution were underlined, and participants were instructed to type "don't know" and to advance to the next item if the anagram was not solved within 2 min.

The session began with a reminder of the nature of the task. The instructions were presented in black letters on the assigned background color (red, white, or blue) for that session. Mehta and Zhu (2009) used a hue–saturation–lightness (HSL) color model to report their color values for red and blue. These values were converted to hexadecimal red–green–blue (RGB) notation. "Red" was defined as RGB = #FF0000, "blue" was defined as RGB = #0000FF, and "white" was defined as RGB = #FFFFFF. Mehta and Zhu did not report the HSL values for white. The assigned background color remained in effect for the session. Pressing the space bar advanced the participant to the next instruction screen. The instructions informed the participant of the hint provided by the underlined letter(s) and to type "don't know" and proceed to the next anagram after 2 min. The final instruction screen indicated that the presentation of the anagrams would begin with the next press of the space bar.

Table 1 shows the exact anagrams used by Mehta and Zhu (2009) and their motivational classifications. No specific empirical work on these words was cited to explain their classifications. All anagrams appeared in uppercase 2-cm-tall letters. Keypresses caused letters of the same size to appear about 2/3 of the distance from the top of the screen, underneath the anagram. Typed letters could be erased, and pressing the Enter key terminated the trial. No feedback on accuracy was provided, and the next anagram was presented immediately.

Note in Table 1 that word length is confounded with motivational classification. Approach ($M = 8$ letters, $SD = 1.00$) and avoidance ($M = 8.67$ letters, $SD = 1.53$) words are longer than the neutral ($M = 5.67$ letters, $SD = 1.21$) words. Typically, anagrams with more letters take a longer time to solve and result in more incorrect solutions (Johnson, 1966; Mayzner & Tresselt, 1963).

Participants were asked to rank their agreement with three statements about their speed–accuracy strategies after being presented with the anagrams. These statements were taken from Study 3 of Mehta and Zhu (2009): (1) *I focused on*

[1] A prior experiment had used the published procedure. Neither the interaction effect nor the predicted color effects were obtained.

**Table 1** Anagrams, solutions, and classifications from Mehta and Zhu (2009)

| ANAGRAM | SOLUTION | CLASSIFICATION |
|---|---|---|
| LONIVI | VIOLIN | NEUTRAL |
| HONPE | PHONE | NEUTRAL |
| NIRDK | DRINK | NEUTRAL |
| TCUON | COUNT | NEUTRAL |
| PUTCOMER | COMPUTER | NEUTRAL |
| RCHNA | RANCH | NEUTRAL |
| VANCEAD | ADVANCE | APPROACH |
| PICSMOLY | OLYMPICS | APPROACH |
| ADTRUEVEN | ADVENTURE | APPROACH |
| GOBILIATION | OBLIGATION | AVOIDANCE |
| TEENPRV | PREVENT | AVOIDANCE |
| ANTGUAREE | GUARANTEE | AVOIDANCE |

The underlined letters in the anagrams indicate the underlined letters that appeared on the screen

*completing the tasks as quickly as possible*, (2) *I was concerned about making mistakes*, and (3) *I was more concerned about accuracy than speed*. Their bipolar scale was used, from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). Finally, participants were administered a brief version of the Ishihara test for red–green discrimination deficiency.

## Results

The primary prediction of Mehta and Zhu (2009) was a significant Background Color × Word Type interaction, such that avoidance words were solved more quickly on a red background and approach words were solved more quickly on a blue background. Table 2 shows correct solution times as a function of the background color and motivational classification of the words. Table 3 shows accuracy as a function of the background color and motivational classification of the words.

**Table 2** Anagram correct solution times, in seconds, as a function of background color and word type

| | | Word type | | |
|---|---|---|---|---|
| Color | | Avoidance | Neutral | Approach |
| Red | *M* | 13.1 | 9.4 | 10.6 |
| | *(SD)* | (7.4) | (5.4) | (6.) |
| White | *M* | 15.7 | 9.9 | 13.4 |
| | *(SD)* | (11.4) | (8.5) | (8.7) |
| Blue | *M* | 15.6 | 10.2 | 15.6 |
| | *(SD)* | (11.3) | (6.9) | (15.0) |

**Table 3** Anagram correct solution percentages as a function of background color and word type

| | | Word type | | |
|---|---|---|---|---|
| Color | | Avoidance | Neutral | Approach |
| Red | *M* | 79.5 | 88.3 | 64.5 |
| | *(SD)* | (26.8) | (16.6) | (25.6) |
| White | *M* | 78.8 | 86.3 | 65.6 |
| | *(SD)* | (25.0) | (16.7) | (25.4) |
| Blue | *M* | 81.1 | 89.3 | 65.6 |
| | *(SD)* | (26.6) | (17.9) | (26.6) |

A repeated measures analysis of variance (ANOVA) on correct solution times was performed, with Color as the between-subjects factor and Word Type as the within-subjects factor. The effect of background color was significant, $F(2, 248) = 3.47$, $p = .03$, $\eta^2 = .015$, and the effect of word type was also significant, $F(2, 496) = 25.8$, $p < .001$, $\eta^2 = .045$. The Color × Word Type interaction failed to reach significance, $F(4, 496) = 1.84$, $p = .12$, $\eta^2 = .007$. The interaction effect explained less than 1 % of the variance.

Additionally, *t* tests were performed to see whether the crucial predictions were supported. Avoidance words had been predicted to be solved more quickly on red backgrounds. Instead, avoidance words were solved *more slowly* than neutral words, $t(85) = 5.0$, $p < .001$, $d = 0.58$, and approach words, $t(84) = 2.9$, $p = .005$, $d = 0.37$, when the words occurred on a red background. A similar failure of the prediction was seen when words appeared on blue backgrounds. Approach words were solved more slowly than neutral words, $t(84) = 4.6$, $p < .001$, $d = 0.44$, and were not statistically different from avoidance words, $t(82) = 0.07$, $p = .94$, $d < 0.001$.

The effect of word type was influenced by the differences in word length among the groups. Pairwise comparisons showed that neutral words were solved more quickly than either avoidance words, $t(255) = 7.9$, $p < .001$, $d = 0.56$, or approach words, $t(255) = 4.7$, $p < .001$, $d = 0.36$. Approach words were solved more quickly than avoidance words, $t(250) = 2.1$, $p = .04$, $d = 0.15$. A similar pattern occurred for accuracy. Neutral words were solved more accurately than either avoidance words, $t(262) = 5.5$, $p < .001$, $d = 0.37$, or approach words, $t(262) = 15.0$, $p < .001$, $d = 1.05$, but avoidance words were solved more accurately than approach words, $t(262) = 7.9$, $p < .001$, $d = 0.56$. The main effect of color was investigated by collapsing across the word type variable. Anagrams were solved more quickly on red than on blue backgrounds, $t(166) = 2.66$, $p = .009$, $d = 0.41$. The color background differences between red and white ($p = .16$) and white and blue ($p = .75$) failed to reach significance.

A repeated measures ANOVA was performed on the log-transformed solution times to see whether extreme scores were either masking or producing significant differences.

The Color × Word Type interaction again failed to reach significance, $F(4, 496) = 2.2$, $p = .06$, $\eta^2 = .008$. (Note that the interaction accounted for less than 1 % of the variance, even if it had been statistically significant.) The significant main effect of color was eliminated, $F(2, 248) = 2.3$, $p = .11$, $\eta^2 = .013$. The significant main effect of word type remained, $F(2, 496) = 44.8$, $p < .001$, $\eta^2 = .066$. Pairwise comparisons among the types of words showed that neutral words continued to be solved more quickly than either approach ($p < .001$) or avoidance ($p < .001$) words, and approach words were solved more quickly than avoidance ($p < .001$) words.

Table 4 shows the mean response scores to the speed–accuracy questions. The label for each column summarizes the strategy, and a higher score indicates greater agreement with the strategy. Mehta and Zhu (2009) predicted that red should evoke stronger concerns about avoiding mistakes and maintaining accuracy. There appears to be no pattern of differences in the responses to statements across the color backgrounds. Separate one-way (color) ANOVAs for each question failed to reach statistical significance. Mehta and Zhu averaged the scores on the three questions (reverse-coding where appropriate) to create an avoidance index and an approach index in their Study 3. The same analysis was applied here, and no main effect of color was obtained, $F(2, 258) = 0.51$, $p = .60$, $\eta^2 = .004$. (Mathematically, the ANOVA results are identical for both indexes, since one index is the reversed values of the other.)

Figure 1 shows participants' solution times for avoidance, neutral, and approach words under red, white, and blue background conditions for Mehta and Zhu (2009), on the left, and the present results, on the right side of the figure. Qualitatively, the results for neutral words look very similar in Mehta and Zhu's study and the present one. The differences appear with the avoidance and approach words. Avoidance words, in the Mehta and Zhu study, showed both a decrease in solution times on a red background and an increase on both white and blue backgrounds, relative to the present study. Approach words showed the reverse pattern: In the Mehta and Zhu study, they showed a decrease in solution times on the blue background and an increase in solution times on the red and white backgrounds, relative to the results in the present study.

**Table 4** Mean response scores to the speed–accuracy strategy questions

| | | Question | | |
|---|---|---|---|---|
| Color | | Fast as possible | Avoid mistakes | Be accurate |
| Red | *M* | 4.8 | 4.7 | 4.6 |
| | *(SD)* | (1.6) | (1.6) | (1.6) |
| White | *M* | 5.0 | 4.6 | 4.6 |
| | *(SD)* | (1.6) | (1.6) | (1.6) |
| Blue | *M* | 4.6 | 4.4 | 4.7 |
| | *(SD)* | (1.9) | (1.7) | (1.6) |

## Discussion

Mehta and Zhu (2009) reported that anagrams of words hypothesized to invoke avoidance motivation were solved more quickly on a red background and that words hypothesized to invoke approach motivation were solved more quickly on a blue background. The purpose of this experiment was to replicate their results using their own procedure. Specifically, the same colors, instructions, and anagrams were used. A conservative approach was taken in the assumption of the size of the effect. Mehta and Zhu obtained effect sizes that would be classified as large. The target sample size in this study had a 95 % chance of detecting a small-size interaction effect.

The first prediction of Mehta and Zhu (2009) was the occurrence of a significant Background Color × Word Type interaction. No significant interaction was found when using raw correct solution times or log-transformed solution times. Effect size measures showed that the interaction accounted for less than 1 % of the variance in both analyses. Tests of Mehta and Zhu's specific predictions were conducted. Contrary to the previous results, avoidance words were solved more slowly on red backgrounds than were neutral and approach words. Approach words, on the other hand, were solved more slowly than neutral words on a blue background and were not significantly different from avoidance words.

It is difficult to interpret the effects with the different word types, because word length was confounded with word classification. The shorter-length neutral words were solved most quickly, an effect consistent with findings in the previous literature (Johnson, 1966; Mayzner & Tresselt, 1963). Approach words were solved more quickly than avoidance words, but approach words were solved less often than avoidance words. This difference could have been due to the properties of the individual words themselves (such as word familiarity) or the structure of the anagram. The factors that predict anagram difficulty are still contested (Knight & Muncer, 2011; Muncer & Knight, 2011). One suggestion for future research would be to use equal-length anagrams to see whether the word-type effect remained.

Another concern about the words used in Mehta and Zhu (2009) is their hypothesized effects on motivation. The classifications of the words were not validated empirically, and many of the classifications can be argued. "Guarantee," for instance, was classified as producing a state of avoidance motivation, but products that come with a guarantee would seem to be more attractive that products without a guarantee, suggesting that "guarantee" could be an approach word. Also, "ranch" was classified as a neutral word, but going to a ranch could be an adventure, and "ranch" could thus be classified as an approach word. A suggestion for future research would be to validate the motivational status of each word empirically.

The results failed to confirm the interaction effect and the specific contrasts reported by Mehta and Zhu (2009). When do

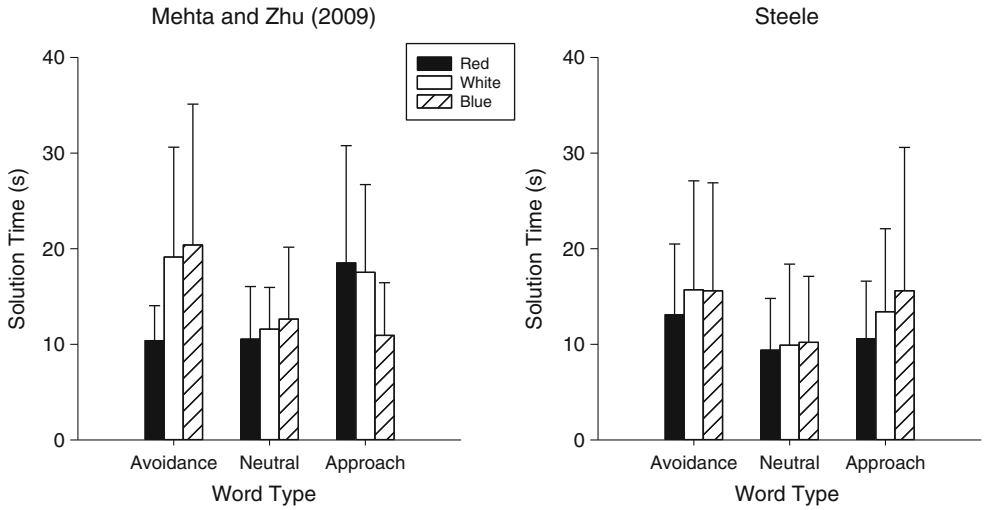


**Fig. 1** Comparison of the results of Mehta and Zhu (2009) to those from the present study: participants' mean solution times to avoidance, neutral, and approach words on red, white, or blue backgrounds. Error bars indicate 1 *SD*

results count as a failure to replicate? Simonsohn (2013a) made the following argument. An effect is convincingly not replicated when the results suggest that the underlying effect, if it exists, is so small that a study based on the original sample size would have been unable to detect the effect. Simonsohn suggested the following rule: Appropriate levels of statistical power can be achieved by using 2.5**n*, where *n* is the sample size of the original study. The multiplicative ratio here was 3.8 (263/69 participants). The present results suggest that the Mehta and Zhu study did not have the statistical power to detect the interaction, given the small effect size obtained in this study.

This study examined one of the experiments from Mehta and Zhu (2009). A question of interest is whether the results in their other studies would replicate. Mehta and Zhu reported six studies, and some of them contained more than one experiment. Some studies would be difficult to replicate, especially the experiments on creative uses of a brick or differences in toy design, since other participants had to be trained to serve as judges. But results like red producing an increase in remembered words in a free-recall task (Mehta & Zhu, 2009, Study 2) should be easy to examine.

One might think that the combination of statistically significant results in all of their studies would support their conclusions about red and blue. However, Francis (2013) has a different view of this issue. Francis pointed out that separate experiments in an article can be considered independent samples and that the probability of *all* studies rejecting the null hypothesis is the *product* of the statistical powers of the individual studies. The probability of throwing six heads *in a row* is a coin-flip analogy that could apply to the Mehta and Zhu (2009) article.

Empirical science is supposed to be self-correcting, as initial findings and explanations are supplanted later by the accumulation of more extensive work and better explanations. A study by Greitemeyer (2013) has pointed to a depressing alternative. Greietmeyer found that undergraduate psychology students were likely still to believe in an effect even after they had been informed that the result had been retracted, a *continued-influence effect*. The history of the Mozart effect supports Greitemeyer's concern. Rauscher, Shaw, and Ky (1993) reported that listening to a Mozart sonata produced a temporary increase in scores using measures from a standardized intelligence test. By 1999, enough work had accumulated that the Mozart effect was described best as a lab effect, with most confirmatory results being connected with a specific research group (Chabris, 1999; Rauscher, 1999; Steele et al., 1999). Two subsequent meta-analyses have agreed with the lab-effect conclusion (Hetland, 2000; Pietschnig, Voracek, & Forman, 2010). But one can review the professional literature and still find the Mozart effect being described as a "scientifically well-known effect" to be used as "a probe into the fundamental cognitive function of music" (Perlovsky, Cabanac, Bonniot-Cabanac, & Cabanac, 2013, p. 10). One must consider the remaining issue of how best to undo the harm of a widely publicized study.

Pashler and Harris (2012) asked rhetorically whether the replicability crisis was overblown, and they concluded that the crisis was genuine. One important distinction that they made was the difference in interpretation between a conceptual replication and a direct replication. A researcher is likely to question his or her own methods when a conceptual replication fails, instead of the original effect. Pashler and Harris thus called for more direct replications. The results of the direct replication here have called into question the Mehta and Zhu (2009) color-priming effect on anagram solution times.

**Author Note** I thank Dan Simons for comments on an earlier version of the manuscript, and the three reviewers for their comments. I thank Alex Butts, Tyler Erath, Jennifer Gray, Natsumi Kimura, Sarah Marger, Patrick Tobin, and Alannah Wray for able assistance in data collection.

## References


Chabris, C. F. (1999). Prelude or requiem for the Mozart effect? *Nature, 400,* 826–827.

Elliot, A. J., & Maier, M. A. (2012). Color-in-context theory. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology, vol. 45* (pp. 61–125). San Diego: Academic.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. doi:10.3758/BF03193146

Francis, G. (2013). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General, 142,* 292–296.

Genschow, O., Reutner, L., & Wänke, M. (2012). The color red reduces snack food and soft drink intake. *Appetite, 58,* 699–702.

Greitemeyer, T. (2013). Article retracted, but the message lives on. *Psychonomic Bulletin and Review*. Advance online publication. doi:10.3758/s13423-013-0500-6

Hetland, L. (2000). Listening to music enhances spatial reasoning: Evidence for the "Mozart effect. *Journal of Aesthetic Education, 34,* 105–148.

Johnson, D. M. (1966). Solution of anagrams. *Psychological Bulletin, 66,* 371–384.

Knight, D., & Muncer, S. J. (2011). Type and token bigram frequencies for two-through nine-letter words and the prediction of anagram difficulty. *Behavior Research Methods, 43,* 491–498.

Mayzner, M. S., & Tresselt, M. E. (1963). Anagram solution times: A function of word length and letter position variables. *Journal of Psychology: Interdisciplinary and Applied, 55,* 469–475.

Mehta, R., & Zhu, R. J. (2009). Blue or red? Exploring the effect of color on cognitive task performances. *Science, 323,* 1226–1229.

Muncer, S. J., & Knight, D. (2011). The syllable effect in anagram solution: Unrecognized evidence from past studies. *Journal of Psycholinguistic Research, 40,* 111–118.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536.

Perlovsky, L., Cabanac, A., Bonniot-Cabanac, M., & Cabanac, M. (2013). Mozart effect, cognitive dissonance, and the pleasure of music. *Behavioural Brain Research, 244,* 9–14.

Pietschnig, J., Voracek, M., & Forman, A. K. (2010). Mozart effect-shmozart effect: A meta-analysis. *Intelligence, 38,* 314–323.

Rauscher, F. H. (1999). Prelude or requiem for the 'Mozart effect'? *Nature, 400,* 827–828.

Rauscher, F. H., Shaw, G. L., & Ky, K. N. (1993). Music and spatial task performance. *Nature, 365,* 611.

Rutchick, A. M., Slepian, M. L., & Ferris, B. (2010). The pen is mightier than the word: Object priming of grading standards. *European Journal of Social Psychology, 40,* 704–708. doi:10.1002/ejsp.753

Simonsohn, U. (2013a). *Evaluating replication results* (May 6, 2013). Retrieved from http://ssrn.com/abstract=2259879

Simonsohn, U. (2013b). *Just post it: The lesson from two cases of fabricated data detected by statistics alone*. Retrieved from http://pss.sagepub.com on September 3, 2013.

Smeesters, D., & Liu, J. E. (2011). The effect of color (red vs. blue) on assimilation vs. contrast in prime-to-behavior effects [Retracted article]. *Journal of Experimental Social Psychology, 47,* 653–656.

Smeesters, D., & Liu, J. E. (2013). Retraction notice to "The effect of color (red vs. blue) on assimilation vs. contrast in prime-to-behavior effects" [Journal of Experimental Social Psychology, (2011) 47, 653–656]. *Journal of Experimental Social Psychology, 49,* 315.

Steele, K. M., Dalla Bella, S., Peretz, I., Dunlop, T., Dawe, L. A., Humphrey, G. K., . . . Olmsted, C. G. (1999). Prelude or requiem for the "Mozart effect"? *Nature, 400*, 827–828.