Processing the Word *Red* and Intellectual Performance: Four Replication Attempts

Timo Gnambs [1,2,*], Carrie Kovacs [3], & Barbara Stiglbauer [1]

[1] Johannes Kepler University Linz

[2] Leibniz Institute for Educational Trajectories

[3] University of Applied Sciences Upper Austria

[*] Corresponding author

Author Note

Correspondence concerning this article should be addressed to Timo Gnambs, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: timo.gnambs@lifbi.de.

Acknowledgment

We gratefully acknowledge the financial support of the Leibniz Institute for Psychology Information (ZPID) for funding experiment 4. The ZPID had no involvement in the study design, data analyses, or writing of the manuscript.

The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Accepted for publication in *Collabra: Psychology*.

Author Contributions

TG devised the replications and collected the data for replications 1, 3, and 4. TG, CK, and BS collected the data for replication 2. TG analyzed the data and wrote the first draft of the manuscript. CK and BS revised the manuscript. All authors approved the submitted version for publication.

Data Accessibility

The data for all experiments, including the statistical syntax to reproduce our findings, can be found at https://osf.io/5ckvd/; methodological details on all four experiments are available in the electronic supplemental material.

Abstract

Colors convey meaning and can impair intellectual performance in achievement situations. Even the processing of color words can exert similar detrimental effects. In four experiments, we tried to replicate previous findings regarding the processing of the word "red" (as compared to a control color) on cognitive test scores. Experiments 1 and 2 (*N*s = 69 and 104) are direct replications of Lichtenfeld, Maier, Elliot, and Pekrun (2009). Both experiments failed to uncover a red color effect on verbal reasoning scores among high school students and undergraduates (Cohen's d = 0.04 and -0.23). Experiments 3 and 4 (N = 103 and 1,149) failed to identify an effect of processing red on general knowledge test scores (Cohen's d = 0.19) and 0.01) among undergraduates and adults. Together, these results do not corroborate the assumption that processing red impairs intellectual performance.

Keywords: red color; cognitive performance; intelligence; general knowledge

Processing the Word *Red* and Intellectual Performance: Four Replication Attempts

Biologically inherited and socially learned color associations can affect psychological functioning and observed behaviors (cf. Elliot & Maier, 2014). According to color-in-context theory (Elliot & Maier, 2012), colors may have different, potentially even opposite effects, depending on the prevalent situational conditions. For example, in an achievement context, red color tends to convey a negative meaning due to its implicit association with failure and danger and, thus, induces avoidance motivation. In contrast, in mating contexts, the same color has a positive meaning, evoking approach motivation because red is typically associated with romance and sexual desire. Several empirical findings have supported key propositions in this respect: viewing red impaired cognitive performance on standardized achievement tests (Elliot, Maier, Moller, Friedman, & Meinhardt, 2007) and reduced risky choices to avoid financial losses (Gnambs, Appel, & Oeberst, 2015), whereas it increased interpersonal attraction (Lehmann, Elliot, & Calin-Jageman, 2018) and signaled social status (Wu, Lu, Dijk, Li, & Schnall, 2018). After several studies demonstrated the effect of viewing the color red in achievement contexts (for reviews see Elliot, 2015, and Elliot & Maier, 2014), it has been suggested that even simply processing the word *red* is sufficient to yield effects that are comparable to actually seeing a red stimulus (Lichtenfeld, Maier, Elliot, & Pekrun, 2009). Referring to the well-known Stroop effect indicating a close connection between color words and actual color representations (e.g., DeHouwer, 2003; Richter & Zwaan, 2009) and to supportive neuropsychological evidence (e.g., Teichmann, Grootswagers, Carlson, & Rich, 2019), Lichtenfeld and colleagues (2009) demonstrated in four experiments that presenting the word *red* before a reasoning test resulted in significantly lower test scores as compared to reading the word *gray* or *green*. The observed effects of Cohen's $d = 0.57$, 0.73, 0.64, and 0.99 suggested a substantial impact of processing color words, despite the subtle color manipulations. For example, in two experiments the authors manipulated a small copyright notice including seven words (10 point font size) at the bottom of the cover pages of the test

booklet. In another experiment, an example item containing the word red or gray was placed before a reasoning test. In all experiments, reading the word red consistently led to poorer test performance as compared to reading another color word. These findings could have important implications for psychological and educational assessments. If reading the word red in an exam item influences subsequent test performance, this might bias estimates of students' proficiency and even threaten test fairness if students are differentially affected by color cues. Thus, it is important to scrutinize whether these effects can be robustly substantiated.

Despite the substantial effects previously triggered by the word red, the effect sizes reported in Lichtenfeld et al. (2009) were highly uncertain. The respective confidence intervals included large effects up to Cohen's $d = 1.60$ as well as vanishingly small effects close to zero (see Table 1). Thus, the available findings encompass effects of clear practical importance as well as effects that are unlikely to impact applied assessments. Similar, recent replications of viewing the color red have failed to find consistent effects on, for example, perceived attractiveness (Lehmann & Calin-Jageman, 2017; Peperkoorn, Roberts, & Pollet, 2016) or cognitive test performance (Gnambs, 2019). These results suggest that red color effects might be overestimated in published studies and that actual effects are more modest. Therefore, the present study sought to replicate and extend the study by Lichtenfeld and colleagues (2009). Importantly, we sought to narrow the range of compatible effect sizes to derive a more precise estimate of red color effects. Thus, we present two direct replications and two conceptual replications (cf. Schmidt, 2009) of Experiment 2 in Lichtenfeld et al. (2009) testing the hypothesis that respondents reading the word *red* before working on an intelligence test would solve fewer items correctly as compared to respondents reading a control color word. The data, including the statistical syntax to reproduce our findings, can be found at https://osf.io/5ckvd/; methodological details on all four experiments are available in the electronic supplemental material.

Experiment 1

Power Analysis

The sample size was determined based upon *a priori* power analyses to identify an effect of Cohen's $d = 0.70$ (as in Experiment 2 in Lichtenfeld et al., 2009) for a one-tailed *t*-test at a significance level of 5% and a power of 80%. This resulted in a minimum sample size of 52.

Method

Sixty-nine students (41 female and 28 male) from two upper secondary schools ("*Gymnasium*") in Austria with a median age of 17 years (*Min* = 16, *Max* = 19) were randomly assigned to a red color ($n = 30$) and a gray color ($n = 39$) condition. The participants were informed that they were about to work on a short intelligence test. Then the verbal analogy subtest of the *Intelligence Structure Test 2000 R* (Liepmann, Beauducel, Brocke, & Amthauer, 2007) also used by Lichtenfeld and colleagues (2009) was administered in both groups. The test included 20 multiple-choice items forming a word pair and the first word of a second pair (e.g., "fast : slow = young : ?"). For each item, five response options were presented, one of which correctly completed the analogy (e.g., "quick, long, tall, tardy, old"). The number of correct answers was the dependent variable. Missing responses were scored as incorrect. Before the actual test, two example items were presented to explain the logic of the test. Following Experiment 2 in Lichtenfeld and colleagues (2009), the second example item included the experimental manipulation: "animal : hound = plant : ?" with five response options "branch, red/gray-alder, root, tree, organism". The manipulation was instigated by the correct solution being presented either as "red-alder" (red color condition) or "gray-alder" (gray color condition; both these trees are quite common in Austria). Moreover, a description below the item (one sentence) explained that "red/gray-alder" was the correct solution. The study was not preregistered.

Statistical Analyses

We expected that reading the word *red* in the example item presented before the analogy test would result in lower test scores as compared to reading the word *gray*. This hypothesis was tested with a one-sided *t*-test for independent groups using an alpha level of 5%. The color effect was quantified using Cohen's *d* coded in such a way that positive effects fell in line with our hypothesis and indicated lower scores in the red condition. Because the original study considered the respondents' sex as a potential moderator, we replicated these analyses and also report the results of a 2 (color: red versus gray) x 2 (gender: girls versus boys) analysis of variance (ANOVA). To test the directed hypothesis for the main effect of color, we made use of the equivalence between the *F* distribution with one degree of freedom and the t^2 distribution to derive the one-sided *p*-value. The effect size for these analyses was partial eta squared η_p^2. Finally, to determine the replication success we conducted an equivalence test (Lakens, 2017) and examined whether the observed effect could be distinguished from the effect of $d = 0.73$ reported in Lichtenfeld et al. (2009). The smallest effect size of interest for this analysis was set to the effect size that the original study had a power of 33% to detect (cf. Simonsohn, 2015), that is, at a *d* of 0.47. A significant test result would indicate an observed effect that was equivalent to the original effect.

Results

Participants in the red color condition solved fewer analogy items correctly ($M = 11.20$, $SD = 2.68$) as compared to participants in the gray color condition ($M = 11.30$, $SD = 2.47$). However, an independent samples *t*-test (one-tailed) showed no significant ($p < .05$) difference between the two color conditions, $t(67) = 0.14$, $p = .443$, and an effect size close to zero, $d = 0.04$, 95% CI [-0.44, 0.51]. Moreover, an equivalence test, $t(67) = -1.90$, $p_{\text{one-tailed}} = .969$, indicated that the observed effect was not equivalent to the original effect and, thus, showed no replication success. Following Lichtenfeld and colleagues (2009), the analyses were repeated considering the respondents' sex as a potential moderator. The ANOVA

showed no significant main effects for the color condition, $F(1, 65) = 0.21$, $p_{\text{one-tailed}} = .326$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.08], resulting in a Cohen's *d* of 0.11, 95% CI [-0.38, 0.61]. Moreover, neither the main effect of sex, $F(1, 65) = 0.02$, $p_{\text{two-tailed}} = .888$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.04], or the respective interaction, $F(1, 65) = 2.89$, $p_{\text{two-tailed}} = .094$, $\eta_p^2 = 0.04$, 95% CI [0.00, 0.17], were significant. In summary, these analyses failed to support the hypothesis that reading the word red would impair analogy test performance.

Experiment 2

Power Analysis

Following the social science replication project (Camerer et al., 2018), we aimed at identifying a smaller effect that was only 75% of the original effects size. Thus, the sample size was determined based upon *a priori* power analyses to identify an effect of Cohen's $d = 0.50$ for a one-tailed *t*-test at a significance level of 5%, and a power of 80%. This resulted in a minimum sample size of 102.

Method

One hundred and four Austrian university students (35 female, 68 male, 1 without information on sex) with a median age of 22 years (*Min* = 18, *Max* = 57) were randomly assigned to a red color ($n = 53$) and a gray color ($n = 51$) condition. The procedure was identical to the first experiment, with participants completing the questionnaire and intelligence test voluntarily during the first 15 minutes of a class period. The study was not preregistered.

Results

In contrast to our hypothesis, participants in the red color condition solved more analogy items correctly ($M = 12.10$, $SD = 2.82$) as compared to participants in the gray color condition ($M = 11.50$, $SD = 3.09$). An independent samples *t*-test (one-tailed) showed no significant ($p < .05$) difference between the two color conditions, $t(102) = -1.14$, $p = .872$, and an effect size in the wrong direction, $d = -0.23$, 95% CI [-0.61, 0.16]. Again, an equivalence

test, $t(102) = -2.68$, $p_{\text{one-tailed}} = .996$, showed no replication success. Follow-up analyses for moderating effects of sex revealed no significant main effects for the color condition, $F(1, 99) = 1.73$, $p_{\text{one-tailed}} = .096$, $\eta_p^2 = 0.02$, 95% CI [0.00, 0.10] (corresponding to Cohen's *d* of -0.27, 95% CI [-0.68, 0.14]) or sex, $F(1, 65) = 3.76$, $p_{\text{two-tailed}} = .055$, $\eta_p^2 = 0.04$, 95% CI [0.00, 0.13], and no significant interaction, $F(1, 65) = 0.615$, $p_{\text{two-tailed}} = .435$, $\eta_p^2 = 0.01$, 95% CI [0.00, 0.07]. These results mirrored Experiment 1 and gave no support to the red color hypothesis.

Experiment 3

Power Analysis

As in Experiment 2, the sample size was determined based upon *a priori* power analyses to identify an effect of Cohen's $d = 0.50$ for a one-tailed *t*-test at a significance level of 5%a power of 80%. This resulted in a minimum sample size of 102.

Method

One hundred and seven students (81 female and 26 male) from a German university with a median age of 21 years (*Min* = 19, *Max* = 44) were randomly assigned to a red color ($n = 56$) and a green color ($n = 51$) condition. In line with previous research (Gnambs, Appel, & Batinic, 2010; Gnambs, Appel, & Kaspar, 2015), it was expected that reading the word *red* would impair performance on an indicator of crystallized intelligence. Therefore, participants were administered a short version of the *General Knowledge Test – German* (GKT-D; Lynn, Wilberg, & Margraf-Stiksrud, 2004). The test included 37 items from different domains that were answered in open response fields. The experimental manipulation required respondents to write down the word red (or a control color word) to allow for a deeper processing of the color word. Thus, the wording of item 19 of the GKT-D was changed to ask either about the color of a ripe tomato (correct answer: red) or about the color of a ripe cucumber (correct answer: green). The number of correct responses after the experimental manipulation was the dependent variable. Missing responses were scored as incorrect. The study was not preregistered.

Results

Participants in the red color condition answered fewer knowledge items correctly (M = 8.86, SD = 2.53) as compared to participants in the green color condition (M = 9.31, SD = 2.36). An independent sample t-test (one-tailed) showed no significant ($p < .05$) difference between the two color conditions, $t(105) = 0.96$, $p = .169$, $d = 0.19$, 95% CI [-0.19, 0.57]. Moreover, an equivalence test, $t(105) = -2.17$, $p_{\text{one-tailed}} = .984$, indicated no replication success. An ANVOA found no significant main effects for the color condition, $F(1, 103) = 0.20$, $p_{\text{one-tailed}} = .328$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.05] (corresponding to Cohen's d of 0.10, 95% CI [-0.35, 0.55]) or sex, $F(1, 103) = 0.02$, $p_{\text{two-tailed}} = .880$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.03], and also no interaction between color and sex, $F(1, 103) = 0.52$, $p_{\text{two-tailed}} = .472$, $\eta_p^2 = 0.01$, 95% CI [0.00, 0.06]. Finally, these analyses were also repeated controlling for the pre-experimental knowledge test scores. However, this analysis identified neither a main effect of the color condition, $F(1, 102) = 0.02$, $p_{\text{one-tailed}} = .448$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.03], nor an interaction with sex, $F(1, 102) = 0.97$, $p_{\text{two-tailed}} = .328$, $\eta_p^2 = 0.01$, 95% CI [0.00, 0.08]. In summary, these analyses failed to corroborate an effect of reading the word *red* on knowledge test scores.

Experiment 4

Power Analysis

Although Lichtenfeld and colleagues (2009) reported effect sizes between Cohen's d = 0.57 and 0.99 for their color manipulations, other research on behavioral priming has typically identified substantially smaller effects. For example, meta-analytic estimates for action and goal priming using incidentally presented words have been about $d = 0.35$ (Weingarten et al., 2016). In order to increase statistical power to detect even such a small effect, the present study used a more conservative effect size estimate of $d = 0.30$ (i.e., less than half the effect reported in Lichtenfeld et al., 2009). Moreover, to guard against type II error, the power was

set to 95%. An *a priori* power analysis estimated a required sample size of $N = 1{,}180$ (for details see the supplement material).

Method

The study was conducted as an unproctored, web-based test. A sample of $N = 1{,}149$ participants from a German online access panel (596 female, 552 male, and 1 without specified gender) with a median age of 38 years (*Min* = 16, *Max* = 85) were randomly assigned to a red color ($n = 563$) or a gray color ($n = 586$) condition. The respondents were administered a short knowledge test measuring crystallized intelligence from the *Berlin Test of Fluid and Crystallized Intelligence – Short Scale* (Schipolowski, Wilhelm, & Schroeders, 2013). The test included 12 multiple-choice items with four response options each (one of which was correct). The number of correct answers was the dependent variable. Missing responses were scored as incorrect. The experimental manipulation was implemented in a similar way as in Experiment 2 of Lichtenfeld et al. (2009). Before the knowledge test, the following example item explaining the logic of the test was presented: "Which of these trees is a leaf tree?" with four response options "Nordmann-fir, red/gray-alder, Sargent-spruce, mountain-pine". The experimental manipulation again consisted of the correct solution shown either as "red-alder" (red color condition) or "gray-alder" (gray color condition). In addition, a description below the item (one sentence) explained that "red/gray-alder" was the correct solution. To enforce processing of the color word, respondents had to give the correct response to the manipulated example item before being able to proceed to the knowledge test. Because Lichtenfeld and colleagues (2009) assumed that worries about test performance would mediate the color effect on that performance, three worry items (e.g., "I am not satisfied about my performance in the test.") based on Morris, Davis and Hutchings (1981) were presented after the knowledge test with seven-point response scales from 1 (*does not apply at all*) to 7 (*strongly applies*). The study was preregistered at https://doi.org/10.23668/psycharchives.2102.

Results

Participants in the red color condition answered fewer knowledge items correctly (M = 8.71, SD = 2.22) as compared to participants in the gray color condition (M = 8.76, SD = 2.22). However, an independent sample t-test (one-tailed) showed no significant ($p < .05$) difference between the two color conditions, $t(1147) = 0.35$, $p = .365$, and an effect size close to zero, $d = 0.02$, 95% CI [-0.10, 0.14]. Furthermore, an equivalence test, $t(1147) = -7.88$, $p_{\text{one-tailed}} > .999$, indicated that the observed effect was not statistically equivalent to the original effect. Examining respondents' sex revealed a significant gender difference, $F(1, 1144) = 13.64$, $p_{\text{two-tailed}} < .001$, $\eta_p^2 = 0.01$, 95% CI [0.00, 0.03]. Women (M = 8.51, SD = 2.17) solved fewer items correctly than men (M = 8.98, SD = 2.25), resulting in a Cohen's d = -0.22, 95% CI [-0.33, -0.10]. However, neither the main effect of color, $F(1, 1144) = 0.05$, $p_{\text{one-tailed}} = .414$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.00], nor the interaction between color and sex, $F(1, 1144) = 1.97$, $p_{\text{two-tailed}} = .161$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.01], were significant.

Participants in the red color condition voiced more worries (M = 4.18, SD = 1.39) as compared to participants in the gray color condition (M = 4.17, SD = 1.42). However, an independent sample t-test (one-tailed) revealed no significant difference between the experimental conditions, $t(1142) = -0.11$, $p = .546$, d = -0.01 95% CI [-0.12, 0.11]. The ANOVA controlling for sex identified a significant ($p < .05$) gender difference, $F(1, 1139) = 36.71$, $p_{\text{two-tailed}} < .001$, $\eta_p^2 = 0.03$, 95% CI [0.01, 0.05], with women (M = 4.41, SD = 1.37) reporting more worries as compared to men (M = 3.92, SD = 1.40). However, neither the main effect of color, $F(1, 1139) = 0.00$, $p_{\text{one-tailed}} = .499$, $\eta_p^2 = 0.00$, 95% CI [0.00, 1.00], nor the interaction between color and sex, $F(1, 1139) = 0.57$, $p_{\text{two-tailed}} = .450$, $\eta_p^2 = 0.00$, 95% CI [0.00, 0.01], was significant. In conclusion, despite the high power of the study, we found no support for an effect of reading the word *red* on knowledge test performance or self-reported worries.

Discussion

Colors are assumed to convey meaning that can influence cognitive functioning in achievement contexts (cf. Elliot, 2015; Elliot & Maier, 2014). Even the mere processing of color words without actually seeing any color stimuli has been reported to exert such effects (Lichtenfeld et al., 2009). Thus, reading the word *red* supposedly impairs performance on verbal and numeric intelligence tests. The present study tested this assumption by trying to replicate Experiment 2 in Lichtenfeld et al. (2009). In two direct replications and two conceptual replications, red color effects were examined for different outcomes (i.e., verbal analogy and general knowledge tests) covering different age groups (from high school students to adults). However, across the four experiments no effect of processing red was observed (see Table 1). Notably, the largest effect size of Cohen's d = -0.23 fell in the wrong direction. Despite the large effects identified in the original study and the substantially larger sample sizes in our replication attempts, we were unable to corroborate that simply processing the word *red* impairs intellectual performance.

These replication failures do not necessarily invalidate the basic premise of red color effects in achievement situations (Elliot et al., 2007) or color-in-context theory (Elliot & Maier, 2012). Rather, they raise doubts regarding the robustness and generalizability of previously reported results. For example, it could be the case that red color impairs intellectual performance, but these effects are so small (and practically negligible) that they require huge sample sizes to be reliably identified. After all, the original studies reported rather imprecise effect estimates (see Table 1) ranging from substantial red color effects (exceeding d = 1.00) to negligible effects close to zero. Our results suggest that the latter seems more likely. On the other hand, Elliot (2015, 2019) emphasized that only a precise combination of luminance, chroma, and hue is expected to produce intellectual impairment. According to this line of reasoning, simply reading the word red should not elicit any cognitive effects. Moreover, subtle differences in the experimental procedure by Lichtenfeld

et al. (2009) and our replications could have introduced unknown confounds (e.g., regarding the precise instructions or respondent incentives; see supplementary materials for details) that led to different results. Therefore, we concur with Elliot (2019) that further high-quality studies are needed "to serve as a cornerstone on which a solid empirical foundation can be built" (p. 16). A large-scale collaborative research project including different research teams (cf. Moshontz et al., 2018) might help devise a study to uncover robust color effects with sufficient power. On a positive note, the present results suggest that color effects are unlikely to exert systematic biases in applied settings. If red color effects require highly standardized settings to be observable, a robust impact on, for example, school exams or cognitive testing in personnel selection seems improbable – at least in the case of purely text-based color effects. Thus, it would seem premature to recommend considering such effects in standard psychological and educational assessment.

References

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*, 637-644. https://doi.org/10.1038/s41562-018-0399-z

DeHouwer, J. (2003). On the role of stimulus–response and stimulus–stimulus compatibility in the Stroop effect. *Memory & Cognition, 31*, 903–904. https://doi.org/10.3758/BF03194393

Elliot, A. J. (2015). Color and psychological functioning: a review of theoretical and empirical work. *Frontiers in Psychology, 6*, 368. https://doi.org/10.3389/fpsyg.2015.00368

Elliot, A. J. (2019). A historically based review of empirical work on color and psychological functioning: Content, methods, and recommendations for future research. *Review of General Psychology, 23*, 177-200. https://doi.org/10.1037/gpr0000170

Elliot A. J., & Maier, M. A. (2012). Color-in-context theory. *Advances in Experimental Social Psychology, 45*, 61-126. https://doi.org/10.1016/B978-0-12-394286-9.00002-0

Elliot, A. J., & Maier, M. A. (2014). Color psychology: Effects of perceiving color on psychological functioning in humans. *Annual Review of Psychology, 65*, 95-120. https://doi.org/10.1146/annurev-psych-010213-115035

Elliot, A. J., Maier, M. A., Moller, A. C., Friedman, R., & Meinhardt, J. (2007). Color and psychological functioning: The effect of red on performance attainment. *Journal of Experimental Psychology: General, 136*, 154-168. https://doi.org/10.1037/0096-3445.136.1.154

Gnambs, T. (2019). Limited evidence for the effect of red color on cognitive performance: A meta-analysis. *Manuscript submitted for publication*.

Gnambs, T., Appel, M., & Batinic, B. (2010). Color red in web-based knowledge testing. *Computers in Human Behavior, 26*, 1625-1631. https://doi.org/10.1016/j.chb.2010.06.010

Gnambs, T., Appel, M., & Kaspar, K. (2015). The effect of the color red on encoding and retrieval of declarative knowledge. *Learning and Individual Differences, 42*, 90-96. https://doi.org/10.1016/j.lindif.2015.07.017

Gnambs, T., Appel, M., & Oeberst, A. (2015). Color red and risk-taking in online environments. *PLoS ONE, 10*(7). https://doi.org/10.1371/journal.pone.0134033

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological & Personality Science, 8*, 355-362. https//doi.org/10.1177/1948550617697177

Lehmann, G. K., & Calin-Jageman, R. J. (2017). Is red really romantic? *Social Psychology, 48*, 174-183. https://doi.org/10.1027/1864-9335/a000296

Lehmann, G. K., Elliot, A. J., & Calin-Jageman, R. J. (2018). Meta-analysis of the effect of red on perceived attractiveness. *Evolutionary Psychology, 16*. Advance online publication. https://doi.org/10.1177/1474704918802412

Lichtenfeld, S., Maier, M. A., Elliot, A. J., & Pekrun, R. (2009). The semantic red effect: Processing the word red undermines intellectual performance. *Journal of Experimental Social Psychology, 45*, 1273-1276. https://doi.org/10.1016/j.jesp.2009.06.003

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* [Intelligence structure test 2000 R]. Göttingen, Germany: Hogrefe.

Lynn, R., Wilberg, S., & Margraf-Stiksrud, J. (2004). Sex differences in general knowledge in German high school students. *Personality and Individual Differences, 37*, 1643-1650. https://doi.org/10.1016/j.paid.2004.02.018

Morris, L. W., Davis, M. A., & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale.

Journal of Educational Psychology, 73, 541-555. https://doi.org/10.1037//0022-0663.73.4.541

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1,* 501-515. https://doi.org/10.1177/2515245918797607

Peperkoorn, L. S., Roberts, S. C., & Pollet, T. V. (2016). Revisiting the red effect on attractiveness and sexual receptivity: No effect of the color red on human mate preferences. *Evolutionary Psychology, 14*, 1-13. https://doi.org/10.1177/1474704916673841

Richter, T., & Zwaan, R. A. (2009). Processing of color words activates color representations. *Cognition*, *111*, 383-389. https://doi.org/10.1016/j.cognition.2009.02.011

Schipolowski, S., Wilhelm, O. & Schroeders, U. (2013). BEFKI GC-K. Berliner Test zur Erfassung fluider und kristalliner Intelligenz – GC-Kurzskala. In C. J. Kemper, E. Brähler, & M. Zenger (Eds.), *Psychologische und sozialwissenschaftliche Kurzskalen –Standardisierte Erhebungsinstrumente für Wissenschaft und Praxis* [Short scales for psychology and the social sciences – Standardized assessment instruments for science and practice] (pp. 30-34). Berlin, Germany: Medizinisch Wissenschaftliche Verlagsgesellschaft.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*, 90-100. https://doi.org/10.1037/a0015108

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569. https://doi.org/10.1177/0956797614567341

Teichmann, L., Grootswagers, T., Carlson, T., & Rich, A. N. (2019). Seeing versus knowing: The temporal dynamics of real and implied colour processing in the human brain. *NeuroImage*. Advance online publication. https://doi.org/10.1016/j.neuroimage.2019.06.062

Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., & Albarracin, D. (2016). On priming action: conclusions from a meta-analysis of the behavioral effects of incidentally-presented words. *Current Opinion in Psychology, 12*, 53-57. https://doi.org/10.1016/j.copsyc.2016.04.015

Wu, Y., Lu, J., van Dijk, E., Li, H., & Schnall, S. (2018). The color red is implicitly associated with social status in the United Kingdom and China. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01902

Table 1.

Summary of Results for Processing Red Color and Intellectual Performance

Experiment	Outcome	*N*	Country	% Female	Age group	Cohen's *d*	95% CI
Lichtenfeld et al. (2009)							
Experiment 1	Verbal analogies	49	Germany	33%	Highschool students	0.57	[-0.01, 1.14] [a]
Experiment 2	Verbal analogies	44	Germany	100%	Highschool students	0.73	[0.14, 1.32]
Experiment 3	Numeric reasoning	40	Germany	65%	Highschool students	0.64	[0.05, 1.22]
Experiment 4	Numeric reasoning	20	Germany	30%	Highschool students	0.99	[0.36, 1.60]
Present study: Direct replications of Experiment 2							
Experiment 1	Verbal analogies	69	Austria	59%	Highschool students	0.04	[-0.44, 0.51]
Experiment 2	Verbal analogies	104	Austria	34%	Undergraduates	-0.23	[-0.61, 0.16]
Present study: Conceptual replications of Experiment 2							
Experiment 3	General knowledge	103	Germany	76%	Undergraduates	0.19	[-0.19, 0.57]
Experiment 4	General knowledge	1,149	Germany	47%	Adults	0.01	[-0.11, 0.13]

Note. The effect sizes were coded in such a way that positive values indicate lower scores in the red color condition as compared to the control condition. [a] The effect was reported as "$F(1, 47) = 3.91$, $p \leq .05$" in Lichtenfeld et al. (2009, p. 1274), which is not strictly significant ($p = .054$) at the conventional alpha level of 5%.

Electronic Supplement Material for

Processing the Word Red and Intellectual Performance: Four Replications Attempts

Detailed Methods for Experiment 1

Power Analysis

The sample size was determined based upon *a priori* power analyses in *pwr* version 1.2-2 (Champely, 2018) to identify an effect of Cohen's $d = 0.70$ (as in Experiment 2 in Lichtenfeld et al., 2009) for a one-tailed *t*-test at a significance level of 5% and a power of 80%. This resulted in a minimum sample size of 52.

Participants

Sixty-nine students (41 female and 28 male) from two upper secondary schools ("*Gymnasium*") in Austria with a median age of 17 years (*Min* = 16, *Max* = 19) were randomly assigned to a red color ($n = 30$) and a gray color ($n = 39$) condition. The study was conducted in small groups at the students' respective schools. All participants gave written informed consent and had good or very good German proficiency. None of the students guessed the purpose of the experiment after finishing the test. In exchange for their participation, all students received a candy bar. All data was collected in the year 2017.

Materials

The participants were informed that they were about to work on a short intelligence test. Then the verbal analogy subtest of the *Intelligence Structure Test 2000 R* (Liepmann, Beauducel, Brocke, & Amthauer, 2007) also used in Lichtenfeld and colleagues (2009) was administered in both groups. The test included 20 multiple-choice items forming a word pair and the first word of a second pair (e.g., "fast : slow = young : ?"). For each item, five response options were presented, one of which was correct (e.g., "quick, long, tall, tardy, old"). The number of correct answers was the dependent variable. Missing responses were scored as incorrect. Before the actual test, two example items were presented to explain the logic of the test. After the test, socio-demographic information (sex, age) was assessed and students' proficiency in German ("How well do you understand German?") was measured on

a four-point response scale (1 = *very badly*, 4 = *very well*). Finally, students indicated their assumptions about the purpose of the study in an open response field.

Experimental Manipulation

Following Experiment 2 in Lichtenfeld and colleagues (2009), the second example item before the analogy test included the experimental manipulation: "animal : hound = plant : ?" with five response options "branch, red/gray-alder, root, tree, organism". The manipulation was instigated by the correct solution being presented either as "red-alder" (red color condition) or "gray-alder" (gray color condition). Moreover, a description below the item (one sentence) explained that "red/gray-alder" was the correct solution.

Statistical Software

The statistical analyses were conducted in *R* version 3.6.1 (R Core Team, 2019) using the packages *car* version 3.0-4 (Fox & Weisberg, 2019), *sjstats* version 0.17.6 (Lüdecke, 2019), and *TOSTER* version 0.3.4 (Lakens, 2017).

Detailed Methods for Experiment 2

Power Analysis

Following the social science replication project (Camerer et al., 2018) we aimed at identifying a smaller effect that was only 75% of the original effects size. Thus, the sample size was determined based upon *a priori* power analyses to identify an effect of Cohen's d = 0.50 for a one-tailed *t*-test at a significance level of 5%, and a power of 80%. This resulted in a minimum sample size of 102.

Participants

From an original sample including 106 students from an Austrian university, two students were excluded because of a suspected color vision deficiency. The remaining N = 104 students (35 female, 68 male, 1 without information on sex) with a median age of 22 years (*Min* = 18, *Max* = 57) were randomly assigned to a red color (n = 53) and a gray color (n = 51) condition. All participants gave written informed consent and had good or very good German proficiency. None of the students guessed the purpose of the experiment after finishing the test. In exchange for their participation, all participants were eligible to enter a lottery for one of three Amazon vouchers worth 50 Euro. All data was collected in 2019.

Materials

The procedure was identical to that of the first experiment. Participants were informed that they were about to work on a short intelligence test and were then administered the analogy subtest of the *Intelligence Structure Test 2000 R* (Liepmann et al., 2007). After the test, socio-demographic information (sex, age) was assessed and students' proficiency in German ("How well do you understand German?") was measured on a four-point response scale (1 = *very badly*, 4 = *very well*). Moreover, participants indicated whether they had a color vision deficiency. Finally, students were asked about the assumed purpose of the study.

Experimental Manipulation

As in Experiment 1, the experimental manipulation was instigated by the second response option (i.e., “red/gray-alder”) of the second example item explaining the test procedure.

Statistical Software

The statistical analyses were conducted in *R* version 3.6.1 (R Core Team, 2019) using the packages *car* version 3.0-4 (Fox & Weisberg, 2019), *sjstats* version 0.17.6 (Lüdecke, 2019), and *TOSTER* version 0.3.4 (Lakens, 2017).

Detailed Methods for Experiment 3

Power Analysis

The sample size was determined based upon *a priori* power analyses to identify an effect of Cohen's $d = 0.50$ for a one-tailed *t*-test at a significance level of 5% and a power of 80%. This resulted in a minimum sample size of 102.

Participants

From an original sample including 111 students attending a German university, four students were excluded because they failed to give the correct response to the item implementing the color manipulation (see below). The remaining $N = 107$ students (81 female and 26 male) with a median age of 21 years (*Min* = 19, *Max* = 44) gave informed consent and were randomly assigned to a red color ($n = 56$) and a green color ($n = 51$) condition. None of the students guessed the purpose of the experiment after finishing the test. The study was conducted in small groups. Students received course credits in exchange for their participation. All data was collected in 2012.

Materials

Participants were told that they were about to work on a short general knowledge test. First, socio-demographic information (sex, age) was collected. Then, a short version of the *General Knowledge Test – German* (GKT-D; Lynn, Wilberg, & Margraf-Stiksrud, 2004) was administered in small groups. The test included 37 items from different domains that had to be answered in open response fields. The number of correct responses after the experimental manipulation (i.e., based on 18 items of the GKT-D) was the dependent variable. Missing responses were scored as incorrect.

Experimental Manipulation

The experimental manipulation was implemented by changing the wording of item 19 of the GKT-D. In the red color condition, the item asked about the color of a ripe tomato (correct answer: "red"), whereas the control color condition inquired about the color of a ripe

cucumber (correct answer: “green”). Four respondents failed to give the correct response to this item (i.e., either “yellow” or no response at all) and, thus, were excluded from the analyses.

Statistical Software

The statistical analyses were conducted in *R* version 3.6.1 (R Core Team, 2019) using the packages *car* version 3.0-4 (Fox & Weisberg, 2019), *sjstats* version 0.17.6 (Lüdecke, 2019), and *TOSTER* version 0.3.4 (Lakens, 2017).

Detailed Methods for Experiment 4

Power Analysis

Although Lichtenfeld and colleagues (2009) reported effect sizes between Cohen's *d* = 0.57 and 0.99 for their color manipulations, other research on behavioral priming has typically identified substantially smaller effects. For example, meta-analytic estimates for action and goal priming using incidentally presented words have been about *d* = 0.35 (Weingarten et al., 2016). In order to increase statistical power to detect even such small a small effect, the present study used a more conservative effect size estimate of *d* = 0.30 (i.e., less than half the effect reported in Lichtenfeld et al., 2009). Moreover, to guard against type II error, the power was set to 95%. An *a priori* power analysis estimated a required sample size of *N* = 1,180 to identify a Cohen's *d* of 0.30 using a significance level of 5% (two-tailed) and a power of 95% for an experimental setup with three color conditions (red, gray, and green) analyzed with a one-factorial analysis of variance and Tukey's (1949) honest significant difference post-hoc test. The study was conducted as an unproctored, web-based test. Because 10 to 20 percent of the respondents were expected to be screened out according to the exclusion criteria given below (cf. Chandler & Paolacci, 2017), the target sample size was set to *N* = 1,400.

Participants

Participants were members of an online access panel in Germany that received bonus points (that could be exchanged for monetary incentives) for completing the survey. A sample of *N* = 1,492 respondents finished the web-based test. Participants were excluded from the analyses based on six *a priori* specified criteria: (a) respondents with poor German proficiency (*n* = 35), (b) respondents who failed a seriousness check using a self-reported diligence item (*n* = 106), (c) respondents with a suspected color vision deficiency (*n* = 185),

(d) participants taking an unusually short amount of time[1] to complete the test (n = 73), (e) respondents guessing the hypotheses (i.e., mentioning the effect of any color with regard to cognitive abilities) at the end of the study (n = 0), and (f) respondents with missing values on all items of the knowledge test (n = 0). After applying these exclusion criteria, N = 1,149 participants (596 female, 552 male, and 1 without specified gender) with a median age of 38 years (*Min* = 16, *Max* = 85) remained, roughly half having been randomly assigned to either a red color (n = 563) or a gray color (n = 586) condition. All data was collected in 2019.

Materials

After giving informed consent, participants were told that they were about to work on a general knowledge test and receive feedback on their individual performance in reference to a representative norm sample. Then, the *BEFKI GC-K* (Schipolowski, Wilhelm, & Schroeders, 2013), a short instrument for the measurement of crystallized intelligence, was administered. The test includes 12 multiple-choice items with four response options each (with one option being correct). Each item was presented individually on the screen, without the possibility of returning to previous items and changing a response. The number of correct solutions was the dependent variable. Missing responses were scored as incorrect. Lichtenfeld and colleagues (2009) assumed that worries about the test performance would mediate the color effect on test performance. Therefore, after the knowledge test, worries with regard to the test performance were measured with three items (e.g., "I am not satisfied about my performance in the test.") based on Morris, Davis and Hutchings (1981) on seven-point response scales from 1 (*does not apply at all*) to 7 (*strongly applies*). Then, socio-demographic information and respondents' proficiency in German (on a four-point response scale) were measured. After participants indicated the assumed purpose of the study as an

[1] All respondents falling in the lowest five percentile of the testing time, that is, those taking 6.5 minutes or less for the entire test, were excluded.

open response, they completed one item of Ishigara's (1985) test for color blindness by identifying a colored number presented within a colored circle. Finally, a diligence item (see Aust, Diedenhofen, Ullrich, & Musch, 2013) asked respondents whether they had worked on the test in a serious manner (1= *not true at all*, 5 = *completely true*).

Experimental Manipulation

Although we planned to implement three color conditions (red, gray, and green), a programming error resulted in the green color condition not being administered. Therefore, the experiment included only two color conditions (red, gray), identical to the previous studies. The experimental manipulation was implemented in a similar way as in Experiment 2 of Lichtenfeld et al. (2009). Before the knowledge test, the following example item explaining the logic of the test was presented: "Which of these trees is a leaf tree?" with four response options "Nordmann-fir, red/gray-alder, Sargent-spruce, mountain-pine". The manipulation was again instigated by the correct solution being presented either as "red-alder" (red color condition) or "gray-alder" (gray color condition). In addition, a description below the item (one sentence) explained that "red/gray-alder" was the correct solution. To enforce the processing of the color word, respondents had to give the correct response to the manipulated example item before being able to proceed to the knowledge test.

Statistical Software

The statistical analyses were conducted in *R* version 3.6.1 (R Core Team, 2019) using the packages *car* version 3.0-4 (Fox & Weisberg, 2019), *sjstats* version 0.17.6 (Lüdecke, 2019), and *TOSTER* version 0.3.4 (Lakens, 2017).

Additional References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*, 527-535. https://doi.org/10.3758/s13428-012-0265-2

Champely, S. (2018). *pwr: Basic Functions for Power Analysis*. R package version 1.2-2. https://CRAN.R-project.org/package=pwr

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science, 8*, 500-508. https://doi.org/10.1177%2F1948550617698203

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd edition). Thousand Oaks, CA: Sage.

Ishihara, S. (1985). *Ishihara's test for colour deficiency*. Göttingen, Germany: Hogrefe.

Lüdecke, D. (2019). *sjstats: Statistical Functions for Regression Models* (Version 0.17.6). https://doi.org/10.5281/zenodo.1284472

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*, 99-114. https://doi.org/10.2307/3001913%20