Black and White as Valence Cues:

A Large Scale Replication Effort of Meier, Robinson, and Clore (2004)

Brian P. Meier, Adam K. Fetterman, & Michael D. Robinson

Word Count: 2,643

Abstract

Replication efforts involving large samples are recommended in helping to determine the reliability of an effect. This approach was taken for a study from Meier, Robinson, and Clore (2004), one of the first papers in social cognition guided by conceptual metaphor theory, which reported that evaluations were faster when word valence metaphorically matched (e.g., a word with a negative meaning in black) rather than mismatched (e.g., a word with a negative meaning in white) font color. The present investigation was a direct large scale replication attempt involving 980 participants who completed an experiment using web-based software and were diverse in terms of race, age, and geographical location. Words with a positive meaning were evaluated faster when font color was white rather than black and words with a negative meaning were evaluated faster when font color was black rather than white, replicating the main results of Meier et al. (2004).

KEYWORDS: Evaluation, Affect, Metaphor, Black, White, Speed, Embodiment

Black and White as Valence Cues:

A Large Scale Replication Effort of Meier, Robinson, and Clore (2004)

Embodiment is the idea that mental representations are grounded in perceptual and bodily experiences (Kiefer & Barsalou, 2013; Semin & Smith, 2013). Conceptual metaphor theory (CMT; Lakoff & Johnson, 1999) is a particularly interesting theory related to embodiment in that it contends that this grounding can sometimes be predicted by metaphor. For example, in infancy and adolescent, people likely experience powerful people and animals as being high in vertical position. As a child, more powerful people are typically taller and higher and less powerful people are typically shorter and lower. Thus, we physically experience powerfulness as high and powerlessness as low. In adulthood, such experiences may lead to conceptual metaphors that pair powerfulness with high and powerlessness with low (e.g., rising to the top).

CMT proponents (Lakoff & Johnson, 1999) suggest that conceptual metaphors help us represent abstract concepts like power because they link them to more physical domains that are easier to comprehend like space. If such contentions are accurate, than thinking about power should be affected by manipulations involving vertical space. Indeed, research has shown that people are faster at judging stimuli as powerful and powerless (e.g., words like king and slave) when stimulus location matches (e.g., powerful/high & powerless/low) rather than mismatches (e.g., powerful/low & powerless/high) stimuli meaning (Schubert, 2005).

A large number of papers are consistent with the basic idea that seemingly irrelevant perceptual inputs (e.g., warmth) can influence judgments and behavior (e.g., person perception) in a metaphor-consistent fashion (Landau, Meier, & Keefer, 2010; Landau, Robinson, & Meier, 2014). In other words, common metaphors may reveal something basic about our thought processes and behaviors. Priming and related effects of this type can be surprising and perhaps

counterintuitive (Bower, 2012). Furthermore, there are legitimate concerns about replication in this area (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012), in part because the field has tended to focus on the generation of new knowledge to a disproportionate degree relative to efforts at replication (Koole & Lakens, 2012). As a consequence, whether published findings in this area will replicate is often unknown and uncertain (Pashler & Wagenmakers, 2012; Simmons, Nelson, & Simonsohn, 2011). Replication issues have been expressed in other scientific disciplines as well (Begley & Ellis, 2012; Ioannidis, 2005).

We have been influenced by these discussions and by suggestions that researchers should consider engaging in high-powered replication efforts as well as more controlled study design and reporting (Brandt et al., 2014). We conducted a large scale replication of a study from Meier, Robinson, and Clore (2004), one of the first papers to report experimental support for CMT in social cognition. Following the CMT analysis of Lakoff and Johnson (1999) and considering the prevalence of light/dark linguistic metaphors for affect (e.g., a bright day or a dark time), Meier et al. (2004) hypothesized that if abstract concepts like affect are partially structured via metaphors, then the manner in which people encode or represent affective stimuli should be biased by the metaphor-consistent physical aspects of those stimuli (e.g., positive stimuli should be encoded faster if they are white rather than black). It was found that participants were faster to categorize words with a negative meaning if they were shown in black versus white font, but participants were faster to categorize words with a positive meaning if they were shown in a white versus black font. Meier et al. (2004) suggested that such results reveal that metaphoric linkages may partially guide judgment processes related to evaluation. In other words, Meier et al. (2004) were able to predict evaluative behavior based on common metaphors that pair good with light and bad with dark. Although there have been some conceptual replications of these

results (e.g., Peña & Yoo, 2014; Sherman & Clore, 2009), we are unaware of any direct replication attempts.

Meier et al. (2004) had five studies in their paper, yet the key finding came from a study with a sample size of 21. Much larger samples and direct replications are preferable in investigating the reliability of a phenomenon (Brandt et al., 2014). We chose to focus our replication efforts on Study 1b in combination with a very large sample size of 980, which gave us an observed power for the predicted effect of 1 (compared to .53 in the original study). Study 1b is most representative of the general findings of the original manuscript. Furthermore, the other studies are similar to Study 1b and provide only modest incremental changes (e.g., presenting stimuli twice rather than once or focusing on accuracy rates rather than reaction times; RTs). In our replication study, we expected evaluations of positive words to be faster when shown in a white versus black font color, but we expected evaluations of negative words to be faster when shown in a black versus white font color.

Method

Participants. The initial sample consisted of 1,007 people recruited through Amazon's Mechanical Turk who received $1 in compensation. Fourteen data files were incomplete and 13 participants had accuracy rates more than 3 *SD*s below the mean. These data files were discarded according to a priori criteria, resulting in a sample size of 980 (494 females; *M* age = 35.04 years, with an *SD* of 11.67 years). Participants were located in 48 different U.S. states and were somewhat diverse with respect to self-reported race (78.8% Caucasian, 6.8% African American or Black, 6.1% Asian or Pacific Islander, 4.8% Hispanic, 2.6% mixed race, .8% American Indian or Alaskan Native, .1% unknown).

Materials and Procedure. Mechanical Turk was programmed to end the study after 1,000 participants completed the task. The evaluation task was programmed using Inquisit software, which performs well in web-based studies according to the company's online literature (www.millisecond.com). Some stimulus manipulations (e.g., chromatic color, screen position, or subliminal presentation) may be noisy in a context in which different participants complete the experiment using different computers in different environments. The present manipulation, however, involves achromatic stimuli that are centered on the computer screen, a manipulation that was likely similar from computer to computer. In addition, the noisiness that may have occurred was likely offset by the large sample size.

We followed the procedures of Study 1b of Meier et al. (2004) as closely as possible and carried out a 2 (color: black vs. white) by 2 (valence: negative vs. positive) word evaluation design. After giving informed consent, participants were told the following information regarding the evaluation task:

> This is a short study in which we want you to categorize words as negative or positive. If a word is negative, hit the 1 key at the top of the keyboard; if it is positive, hit the 9 key. We want you to be as quick as possible in making your decisions. However, it is even more important that you are accurate. If you make an error in classifying a particular word, you will get an error message. If you see several of these in a short period of time, you should slow down a bit.
>
> To remind you which key is which, the words 'negative' (1 key) and 'positive' (9 key) will be presented throughout the trials. The word itself will sometimes be presented in different colors. You should ignore this factor, making the classifications both quickly and accurately as possible.

There were 100 trials, each involving a different word presented at center screen. Fifty were positive (e.g., wise) and 50 were negative (e.g., foolish). See Meier et al. (2004) and the supplemental materials for a list of the stimuli. The program generated a different random order of stimuli for each participant. It also randomly assigned words to black or white font colors,

subject to the constraint that there were 25 trials for each cell of the 2 (color) by 2 (valence) design. The background was gray (50% grayscale) and the words "negative = 1" and "positive = 9" were presented in blue font below and to the left and right of the screen, respectively, as response mapping reminders. Correct evaluations were followed by a blank interval for 150 milliseconds (ms) and incorrect evaluations were followed by the world "INCORRECT" in yellow font for 1,000 ms. The program recorded the accuracy and latency (in ms) of responses. Demographics were collected after the task, and participants also completed a self-report measure of metaphor usage included for an exploratory purpose not related to the replication attempt.

Results

Words were relatively clear in evaluative meaning and errors were penalized. As a consequence, accuracy was high (M = 95.01%; SD = 3.79%). Standard (Robinson, 2007) procedures were used to prepare the latencies. Inaccurate trials were dropped, a log transformation was performed, and logged latencies more than 2.5 *SD*s from the grand mean (across participants and trials) were replaced with the 2.5 *SD* value. Analyses focused on logged latency means, but results are reported in terms of raw values (ms) to aid interpretation. As a brief note, Meier et al. (2004) did not transform their evaluation time data, but the present transformation procedures are preferable and we have consistently used them since this early paper. The reaction time results reported below are identical when the raw data is analyzed.

Evaluation times were analyzed in a 2 (font color: black versus white) by 2 (valence: positive versus negative) repeated-measures ANOVA. The main effect of color was not significant, $F(1, 979) = 2.64$, $p = .11$. The main effect of valence was significant, $F(1, 979) = 165.74$, $p < .01$, partial eta squared = .14; evaluations were faster for positive (M = 870 ms; SD =

158) rather than negative (*M* = 899 ms; *SD* = 170) words. Valence main effects are typical in the social cognition literature and often result in positive words being categorized faster than negative words because of their greater density in memory (Unkelbach, Fiedler, Bayer, Stegmüller, & Danner, 2008; Unkelbach et al., 2010).

The interaction between color and valence was significant, $F(1, 979) = 69.40$, $p < .01$, partial eta squared = .07 (95% CIs: .04 & .10). This significant interaction replicates Meier et al. (2004) Study 1b, which had a partial eta squared of .18 (95% CIs: .00 & .44). The effects of font color and valence were similar in that participants were faster to evaluate words with a negative meaning when presented in a black versus a white font, $F(1, 979) = 47.28$, $p < .01$, partial eta squared = .05 (95% CIs: .02 & .07; Meier et al., 2004 partial eta squared = .03; 95% CIs: .00 & .26), but they were faster to evaluate words with a positive meaning when presented in a white versus a black font, $F(1, 979) = 24.73$, $p < .01$, partial eta squared = .03 (95% CIs: .01 & .05; Meier et al., 2004 partial eta squared = .25; 95% CIs: .01 & .50). The means for the interaction along with means from Study 1b of Meier et al. (2004) are shown in Table 1.

As in Meier et al. (2004), accuracy rates were also analyzed in a supplemental 2 (font color: black versus white) by 2 (valence: positive versus negative) repeated-measures ANOVA. The main effect of color was not significant, $F < 1$. The main effect of valence was significant, $F(1, 979) = 94.86$, $p < .01$, partial eta squared = .09; accuracy was higher for positive (*M* = 95.76%; *SD* = 4.06%) rather than negative (*M* = 94.27%; *SD* = 4.86%) words. The interaction between color and valence was significant, $F(1, 979) = 57.54$, $p < .01$, partial eta squared = .06 (95% CIs: .03 & .09). The interaction for accuracy rates (partial eta squared = .02; 95% CIs: .00 & .24) was not significant in Meier et al. (2004). In the current study, the effects of font color and valence were similar in that participants were more accurate in evaluating words with a

negative meaning when presented in a black versus a white font, $F(1, 979) = 31.12$, $p < .01$, partial eta squared = .03 (95% CIs: .01 & .06; Meier et al., 2004 partial eta squared = .00; 95% CIs: .00 & .09), but they were more accurate in evaluating words with a positive meaning when presented in a white versus a black font, $F(1, 979) = 30.66$, $p < .01$, partial eta squared = .03 (95% CIs: .01 & .05; Meier et al., 2004 partial eta squared = .02; 95% CIs: .00 & .24). The means for the interaction along with means from Study 1b of Meier et al. (2004) are shown in Table 2. Although an accuracy rate effect was not central to the original hypothesis because the instructions encouraged speed and accuracy but stressed accuracy, the results here suggest that the effect of accuracy rates nonetheless reliably support the hypothesis. The smaller effect size in the original study could be due to the much smaller sample size. The current sample size of 980 likely creates a more accurate estimate of effect size.

Discussion

Valence and brightness are often paired in linguistic metaphors such that negative valence is dark and positive valence is light. CMT further contends that these mappings are likely to extend beyond language to the manner in which affect is mentally represented (Lakoff & Johnson, 1999). Meier et al. (2004) provided experimental support for this idea in a RT task in which it was shown that evaluations were faster when color and valence were metaphorically congruent (negative/black & positive/white) rather than incongruent (negative/white & positive/black). Yet, sample sizes were small and there have been concerns about the replicability of CMT and embodiment findings. We sought to address these issues and concerns by conducting a large scale replication of one study from that paper with a more diverse sample with respect to race, age, and geography. The same crossover interaction occurred and was

similar in nature to Study 1b of Meier et al. (2004). Although there are many reasons why one may or may not replicate a study, it appears that the original effect is reliable.

The basic nature of the current paradigm, materials, and results (see supplemental information for the raw data and materials) renders the effect a useful one to build upon. For example, past work has extended Meier et al.'s (2004) findings by examining moderators and contextual factors (e.g., Sherman & Clore, 2009) and more "macro" effects of light versus dark environmental cues (e.g., Chiou & Cheng, 2013; Webster, Urland, & Correll, 2012). Future work could use large-scale studies to further extend this work.

The current paradigm could also be used to examine alternative or more specific mechanisms involved in the effects. For example, Lakens (2012) contends that effects like the current ones have little to do with metaphor and are better explained through congruency principles related to polarity theory and word frequency. The current effects certainly involve congruency mechanisms like the ones shown in associative priming or the Stroop effect, which is why we chose to use a categorization task in the Meier et al. (2004) paper in the first place. However, while we believe that conceptual metaphors likely enhance these congruency effects, Lakens (2012) does not. Although a lengthy explanation of such issues is beyond the scope of this direct replication paper, the current methods could allow creative researchers to examine such alternative explanations using high-powered studies. In reality, it is likely that both congruency/polarity mechanisms and conceptual metaphors contribute to such results.

It is useful to engage in large scale direct replication efforts of other findings reported in the metaphor and embodiment literature (as well as psychology as a whole). Some will likely replicate and some will likely not, resulting in a more nuanced picture of the processes involved.

For certain purposes at least, the web-based procedures of the current investigation can be useful (also see Schubert, Murteira, Collins, & Lopes, 2013).

References

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., et al. (2014). The Replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50,* 217-224.

Begley, C. G. & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483,* 531-533.

Bower, B. (2012). The hot and cold of priming: Psychologists are divided on whether unnoticed cues can influence behavior. *Science News, 181*, 26-29.

Chiou, W. B., & Cheng, Y. Y. (2013). In broad daylight, we trust in God! Brightness, the salience of morality, and ethical behavior. *Journal of Environmental Psychology, 36,* 37-42.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012) Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7,* e29081. doi:10.1371/journal.pone.0029081

Ioannidis, J. P. A. (2005). Why most published findings are false. *PLOS Med 2,* e124. doi:10.1371/journal.pmed.0020124

Kiefer, M. & Barsalou, L. W. (2013). Grounding the human conceptual system in perception, action, and internal states. In: W. Prinz, M. Beisert, & A. Herwig (Eds.), *Action science: Foundations of an emerging discipline* (pp. 381-407). Cambridge, MA: MIT Press.

Koole, S. L. & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science, 7,* 608-614.

Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 726-736.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenges to western thought.* New York: Basic Books.

Landau, M. J., Meier, B. P., & Keefer, L. A. (2010). A metaphor-enriched social cognition. *Psychological Bulletin, 136,* 1045-1067.

Landau, M. J., Robinson, M.D., & Meier, B. P. (Eds.). (2014). *The power of metaphor: Examining its influence on social life.* Washington, D.C.: American Psychological Association.

Meier, B. P., Robinson, M. D., & Clore, G. L. (2004). Why good guys wear white: Automatic inferences about stimulus valence based on brightness. *Psychological Science, 15,* 82-87.

Pashler, H. & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science,* 7, 528-530.

Peña, J., & Yoo, S.C. (2014). Under pressure: Avatar appearance and cognitive load effects on attitudes, trustworthiness, bidding, and interpersonal distance in a virtual store. *Presence, 23*, 18-32.

Robinson, M. D. (2007). Lives lived in milliseconds: Using cognitive methods in personality research. In: R. W. Robbins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 345-359). New York: Guilford Press.

Schubert, T. W. (2005). Your highness: Vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology, 89*, 1-21.

Schubert, T. W., Murteira, C., Collins, E. C., Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PLoS ONE, 8*, e67769. doi:10.1371/journal.pone.0067769

Semin, G. R., & Smith, E. (2013). Socially situated cognition in perspective. *Social Cognition*, 31, 125-146.

Sherman, G. D., & Clore, G. L. (2009). The color of sin: White and black are perceptual symbols of moral purity and pollution. *Psychological Science, 20,* 1019-1025.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 11,* 1359-1366.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology, 95,* 36-49.

Unkelbach, C., von Hippel, W., Forgas, J. P., Robinson, M. D., Shakarchi, R. J., & Hawkins, C. (2010). Good things come easy: Subjective exposure frequency and the faster processing of positive information. *Social Cognition, 28,* 538-555.

Webster, G. D., Urland, G. R., & Correll, J. (2012). Can uniform color "color" aggression? Quasi-experimental evidence from professional ice hockey. *Social Psychological and Personality Science, 3,* 274-281.

Table 1

Means and Standard Deviations for the Stimulus Color by Stimulus Valence Interaction on Reaction Times, Current Experiment and Meier et al. (2004) Experiment 1b

Condition	Current Experiment	Meier et al. (2004) Study 1b
Negative Words/Black Font	890 ms (176 ms)	881 ms (202 ms)
Negative Words/White Font	909 ms (180 ms)	897 ms (181 ms)
Positive Words/Black Font	876 ms (165 ms)	896 ms (208 ms)
Positive Words/White Font	863 ms (165 ms)	856 ms (185 ms)

Table 2

Means and Standard Deviations for the Stimulus Color by Stimulus Valence Interaction on Accuracy Rates, Current Experiment and Meier et al. (2004) Experiment 1b

Condition	Current Experiment	Meier et al. (2004) Study 1b
Negative Words/Black Font	94.87% (5.46%)	98.10% (2.60%)
Negative Words/White Font	93.66% (6.36%)	98.00% (3.00%)
Positive Words/Black Font	95.24% (5.41%)	97.10% (4.10%)
Positive Words/White Font	96.27% (4.55%)	97.90% (2.80%)