Replication of “The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency” by JR Schmidt and D Besner (2008, *JEPLMC*)

Mark D. Cloud and Megan M. Kyc  
Department of Psychology

Lock Haven University

Contact: MCloud@lhup.edu

In their original paper, “The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency,” Schmidt and Besner (2008) suggest that some errors made by participants during a Stroop task can be explained by the contingency of the trial. Their hypothesis seeks to explain the item-specific proportion congruent (ISPC) effect, wherein participants display a stronger Stroop effect for words that usually are presented in their congruent color and a weaker Stroop effect for words that are usually presented in non-congruent colors.

Schmidt and Besner (2008) suggest that participants implicitly learn contingencies (i.e., correlations) between a word and a particular color response. Thus in high contingency trials, when a word such as *TABLE* is presented most often in *yellow*, participants after reading the word will unconsciously predict the key response for *yellow*. This prediction facilitates more rapid and accurate responding by priming the response to the correct key. In low contingency trials, reading the word *TABLE* in a color more rarely paired with the word (e.g., *blue*) results in interference by the tendency of the word to evoke the more common color response. In medium contingency trials, a word is equally paired with every color so that a color predicts no specific response.

In this replication of Schmidt and Besner (2008), we sought to exactly reproduce the methodology used in the original paper. As the present work was conducted as part of the Reproducibility Project: Psychology, which asks volunteers to replicate the final study in their chosen paper, this is a direct replication of Experiment 2. We particularly focus on the finding that low contingency trials produce more errors than medium contingency trials, *t*(94) = 1.929, *p*= .028, indicating that there is greater response interference when a trial predicts the wrong answer than a trial that makes no prediction.

# Methods

## Power Analysis

The key effect in Experiment 2 is not the one being replicated in this study. This is because that key effect, *t*(94) = .113, *p* = .910—the finding that once the influence of a low threshold of responding to the wrong answer was removed from low contingency trials, low contingency trials errors would be equal to that of medium contingency trials—was a null effect. Thus to be adequately powered, it would require a sample size that was too large to be feasible for replication.

Instead, we chose to focus on the comparison of error rates between low and medium contingency trials *before* accounting for the low contingency trial’s increased number of errors. This effect, *t*(94) = 1.929, *p* = .028, showed that low contingency trials have a higher rate of errors than medium contingency trials.

To calculate power, we used G\*Power, which required Cohen’s *d* as an effect size measurement. Using Daniel Lakens’s effect size calculating spreadsheet, we calculated *d* = 0.197. Then, plugging this into G\*Power, we set alpha to .05 and inserted first .8, then .9, and .95 to receive suggested sample sizes of 160, 220, and 278 respectively.

## Planned Sample

Participants will be 242 voluntary undergraduate college students from Lock Haven University who have normal vision. Participants will be offered a choice of receiving a $5 Amazon Gift e-card or a Research Participation Slip, which can be used for course points in some psychology classes. All participants will give informed consent prior to their data collection.

## Materials

A computer and monitor with a QWERTY keyboard were used to present a four-choice task through E-Prime software (Psychology Software Tools, 2012). Participants viewed six colored display words (*LOOP, FINS, MEET, SLID, CALL, TUBE)* and indicated their color choices of blue, green, orange, and yellow by pressing the A, Z, K, and M keys respectively. Attached to the bottom of the computer monitor was an index card that provided a review of ink color and key matches.

The following setup from Schmidt and Besner (2008) was mimicked exactly:

For each participant, one of the words was presented most often (6 out of 12 times per block) in blue [(e.g., LOOP)], another most often in green [(e.g., FINS)], another most often in yellow, and another most often in orange. These words were presented equally often in the remaining colors. The remaining two words were presented equally often (3 times each per block) in all colors. Assignment of words to colors was counterbalanced across participants” (The RGB values for the new stimulus colors were 255, 255, 0 (yellow) and 255, 125, 0 (orange). There were three contingency levels in the experiment: high (.50) [(eg. LOOPblue)], medium (.25) [(eg. MEETyellow)], and low (.17) [(eg. FINSblue)] (p. 521).

## Procedure

The following procedure from Schmidt and Besner (2008) was followed exactly:

“There were 432 trials in this experiment, consisting of six blocks of 72 trials each. To increase error frequency, we instructed participants to respond before a 550-ms deadline when the stimulus display terminated.” (p. 521)

“On each trial, a white (255, 255, 255) fixation cross (i.e., “X”) was presented in the middle of a black screen for 250 ms. This was followed by 250 ms of blank screen, followed by the stimulus display. The stimulus display was presented until a response was made or until the trial timed out at [550 ms]. Correct responses were followed by a blank screen for 250 ms before the next fixation cross.” (p. 520)

Timed-out responses were followed by the message *Too Slow* for a duration of 1,000 ms in red (255, 0, 0). wereIncorrect responses were followed by the message *Incorrect* for a duration of 250 ms in red*.* When the screen indicated the completion of the trials, participants were debriefed and either received a research participation slip or $5 Amazon gift e-card.

## Analysis Plan

The key finding identified for replication is the finding that low contingency trials generate higher performance errors than medium contingency trials. The percentage error rate of viable trials is calculated for each contingency level. Although not a key calculation for the replication, replicating a key theoretical finding requires calculating the “low contingency adjusted values” for each participant. We plan to follow Dr. Schmidt’s instructions to use the following formula (J. Schmidt, personal communication, April 27, 2014): (percentage low contingency errors) x (percentage unpredicted errors) x (3/2). The *percentage unpredicted errors* is the total errors for low contingency trials—excluding the high contingency response—divided by the total errors for low contingency trials—including the high contingency response. The product of those two variables is then multiplied by 3/2 to correct for the fact that one is only looking at two of the three possible incorrect responses.

A one-way repeated measures ANOVA of contingency (high, medium, low) on percentage errors will be performed. A planned post-hoc *t*-test comparisons of medium contingency and low contingency error rates as well as medium contingency and low contingency-adjusted trials will be subsequently performed.

## Differences from Original Study

Participants were offered either a course credit slip or a $5 Amazon e-card for participation, while participants in the original study were only offered course credit. Also, we attached to the bottom of the computer monitor an index card that provided a review of ink color and keyboard key matches. There was no such a hint available to the original study participants. There were no other differences between the replication and the original study.

## Actual Sample

Participants were 242 voluntary undergraduate college students from Lock Haven University who had normal vision. Average age of participants was 19.78, with an age range of 18-37, 74 participants were male (30.6%) and 168 were female (69.4%). Racial/ethnic background of participants was 84.3% White/Caucasian, 12.8% Black/African-American, and 2.9% other racial/ethnic backgrounds. All participants gave informed consent prior to their data collection.

## Differences from pre-data collection methods plan

None.

# Results

## Data preparation

Reflecting the difficulty of the task and similar to the original study, 23% of the trials were spoiled. The percentage error rates for viable trials for the four contingency conditions were as follows: High (*M* = 33.3%; *SE* = 0.86%), Medium (*M* = 34.6%; *SE* = 0.86%), Low (*M* = 36.1%; *SE* = 0.86%), Low-Adjusted (*M* = 35.2%; *SE* = 0.86%).

## Confirmatory analysis

A repeated-measures ANOVA revealed statistically reliable main effect for contingency condition, *F*(1, 241) = 41.148, *p* < .001, η2*p* = .146. The analysis of the key comparison for replication revealed low contingency trials generated higher error rates than medium contingency trials,  *t* (241) = 3.955, *p* < .001, *d*  = .255. The key theoretical comparison between medium contingency and low-adjusted contingency error rates was not statistically reliable as expected,  *t* (241) = 1.106, *p* = .270.

# Discussion

## Summary of Replication Attempt

We successfully replicated the original findings of Schmidt and Besner’s (2008) Experiment 2 supporting the view that participants implicitly learn contingencies between words and responses, which lowers the response threshold for high correlated responses. Further, we can conclude that low contingency trials produce greater response interference for wrong answers than in medium contingency trials when no prediction is made. Finally, our replication of the null result for differences in low contingency-adjusted and medium contingency error rates provides additional evidence that the increase in errors for low contingency trials is solely due to an increase in the specific predicted response rather than a general increase in errors (Schmidt & Besner, 2008).

References

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160. [Download PDF](http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPower31-BRM-Paper.pdf)

E-Prime Standard (Version 2.0)[Software]. (2012). Sharpsburg, PA: Psychology Software Tools.

Lakens, D., (2014). Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-tests and ANOVAs. Retrieved from Open Science Framework, [osf.io/ixgcd](https://osf.io/ixgcd/)

Schmidt, J. R., and Besner, D. (2008). The Stroop effect: Why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 514-523.