Two Bayesian Tests of the GLOMO^{sys} Model

Sarahanne M. Field[1, 5], Eric-Jan Wagenmakers[2], Ben R. Newell[3], René Zeelenberg[4]

and Don van Ravenzwaaij[5]

[1]University of Newcastle, Australia, [2]University of Amsterdam, Netherlands, [3]University of New South Wales, Australia, [4]Erasmus University Rotterdam, [5]University of Groningen

This research was supported by a European Research Council grant (Bayes or Bust) to E-JW, and two grants from the Australian Research Council to BN (DP140101145) and DvR (DE140101181). We are indebted to Şule Güney for her assistance in Experiment 1.

Abstract

Priming is arguably one of the key phenomena in contemporary social psychology. Recent retractions and failed replication attempts have led to a division in the field between *proponents* and *skeptics*, and reinforce the importance of confirming certain priming effects through replication. In this study, we describe the results of two preregistered replication attempts of one experiment by Förster and Denzler (2012). In both experiments, participants first processed letters either globally or locally, then were tested using a typicality rating task. Bayes factor hypothesis tests were conducted for both experiments: Experiment 1 (N=100) yielded an indecisive Bayes factor of 1.38, indicating that the in-lab data are 1.38 times more likely to have occurred under the null hypothesis than under the alternative. Experiment 2 (N=908) yielded a Bayes factor of 10.84, indicating strong support for the null hypothesis that global priming does not affect participants' mean typicality ratings. The failure to replicate this priming effect challenges existing support for the GLOMO^{sys} model.

Keywords: Priming, Replication, GLOMO^{sys}, Bayesian Statistics

Two Bayesian Tests of the GLOMO^{sys} Model

Over the past years, the reproducibility of psychological science has become a topic of much debate (e.g., Carey, 2015; Lindsay, 2015; Pashler and Wagenmakers, 2012; Spellman, 2015). This debate has prompted constructive efforts to improve the validity of published findings in psychology. Research groups worldwide are collaborating to produce guidelines for transparent research (Nosek et al., 2015) and to conduct massive replication studies such as the Many Labs and Open Science Collaboration projects (Klein et al., 2014; Open Science Collaboration, 2015). These efforts are bolstered by the commitment of several high-profile journals to reinforce transparency in experimentation and to improve protocols for scholarly publication. For instance, the *Journal of Experimental Psychology: General* publishes replication studies following peer review of intended research protocols. By peer-reviewing research methodologies before the data are collected and by committing to publish the associated results regardless of the outcome, journals are attempting to counter hindsight bias and publication bias (Chambers, 2013). This commitment follows C. S. Peirce's third rule of empirically assessing scientific hypotheses: "The failures as well as the successes of the predictions must be honestly noted." (Peirce, 1878; as reprinted in Hartshorne and Weiss, 1932, p. 2635).

In this article we present two preregistered replication studies that form a part of the field-wide effort to assess the reproducibility of key research findings (e.g., Alogna et al., 2014; Klein et al., 2014; Nosek & Lakens, 2014). Our specific interest is in a series of twelve experiments designed by Förster and Denzler (henceforth FD2012; 2012, retracted).[1] The primary goal of this article is to replicate the empirical priming effect reported in the first of the FD2012 studies, and the experimental procedure designed to yield this effect. As FD2012 originally existed in the priming literature as a key empirical demonstration of the 'Global

[1] During preparation of this manuscript, FD2012 was retracted from the literature.

versus Local processing Model' (Förster and Dannenberg, 2010; GLOMO^{sys}), this article seeks to assess whether and to what extent the empirical support in favor of the GLOMOsys model can be reproduced, by closely replicating Experiment 1 from FD2012.

GLOMO^{sys} is a dual-system model in which global and local processing styles create different yet related mental representations of the world in procedural memory. These representations allow people to extract different kinds of information from the environment, depending on which of the two systems is activated.

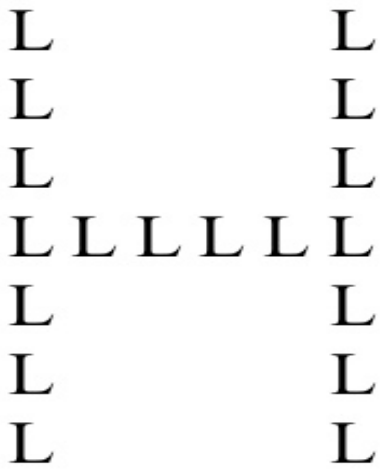

Figure 1. One of six stimulus letters presented, recreated from the description that accompanied Experiment 1 of FD2012. In the global condition, the correct answer is 'H', whereas in the local condition, the correct answer is 'L'.

Among other predictions, the model hypothesizes that activation of the global system in GLOMO^{sys} broadens "concepts in memory" to facilitate holistic and creative thinking. In the words of FD2012, the "activation of remote exemplars…should further support among others the generation of creative ideas…" (p. 109). In contrast, activation of the local system in GLOMO^{sys} is hypothesized to narrow semantic category breadth, facilitating analytical and

detail-oriented processing. In FD2012 this concept is explored through priming participants via the Navon task for either a global or local processing style. These processing styles are thought to then shift to a conceptual plane, influencing performance in the word typicality-rating task used in FD2012.

In Experiment 1 of FD2012, the hypotheses of GLOMO^{sys} were tested through the use of an unrelated-task paradigm. Either global or local processing styles were induced through a Navon-style visual task (for an example, see Figure 1), in which participants viewed a series of large letters comprised of small letters. After the Navon task was administered, creativity was measured by mean responses to fringe word exemplars in Rosch's Breadth of Categorization task (BOC; Rosch, 1975), a simple word typicality-rating task in which participants assess how typical words are to given word categories. One example is: "How typical is the word 'wheelchair' for the category of 'vehicle'?" The results of FD2012 supported their hypotheses: compared to a control condition, inducing global perceptual processing increased typicality ratings in the BOC task, whereas inducing local perceptual processing decreased typicality ratings. FD2012 concluded that (1) global processing enhances creativity, whereas local processing suppresses creativity; (2) breadth of processing in perceptual mechanisms such as vision influences breadth of processing in conceptual mechanisms such as creativity. FD2012 apparently demonstrate that priming for global versus local processing styles can be induced in perception and then successfully shift to influence conceptual processes. These findings provide support for Posner's (1987) argument that perceptual and conceptual attention share a common underlying mechanism.

Within the field of priming, the development of GLOMO^{sys} has quickly become foundational to a number of relevant and recent research findings relating to attention, general information processing mechanisms (Markman & Dyczewski, 2010), affect and its relation to self-regulation, self-construal (Kühnen & Hannover, 2010), and social judgments

and decision-making processes (Dijkstra, van der Pligt, van Kleef, & Kerstholt, 2012). Nevertheless, there has been a recent debate about the reproducibility of the results (e.g., Klauer & Singmann, 2015).

We sought to replicate the experimental protocol of FD2012, to examine what is arguably the former publication's key empirical effect. To fulfill our aims, we conducted two replications of Experiment 1 from FD2012: one in-lab replication study, and one online replication study conducted via Amazon's Mechanical Turk (MTurk). The methodology for both studies closely follows that of FD2012, except that no experimenter was present for the administration of the tasks of the online study. The stimuli and questionnaires used in Experiments 1 and 2 are identical to one another, yet deviate slightly from the original FD2012 stimuli. These small deviations mean our study is not an exact replication, but we believe that the differences are sufficiently minor that they will not be the main cause for divergent results. Both replications are preregistered on the Open Science Framework (OSF) at https://osf.io/ynr2q/. Further, the preregistration of the first experiment in this paper has been peer-reviewed and accepted by the Editorial committee at the *Journal of Experimental Psychology: General*, and is in line with their stipulations for replication articles.

Experiment 1

In accordance with the method of Experiment 1 in FD2012, in our first experiment we set out to measure the typicality ratings for in-lab participants on the BOC task, after priming them for either a global or local processing style with a version of the Navon task. As with the original experiment, the tasks were administered in the presence of an experimenter, and were conducted in a manner true to the original procedure. Our experiment was a preregistered replication of the original Experiment 1 in FD2012. The preregistration details and materials for this experiment are freely available on the OSF at https://osf.io/ynr2q/. The experiment does not deviate from the OSF Preregistration Document.

Method

Intended Sampling Plan

The following sampling plan has been adapted from Wagenmakers et al. (2014). A frequentist analysis would start with an assessment of the effect size that would then form the basis for a power calculation that seeks to determine the number of participants that yields a specific probability for rejecting the null hypothesis when it is false. This frequentist analysis plan is needlessly constraining and potentially wasteful: the experiment cannot continue after the planned number of participants has been tested, and it cannot stop even when the data yield a compelling result earlier than expected (e.g., Wagenmakers, 2007). Here we circumvent these frequentist limitations by calculating and monitoring the Bayes factor (e.g., Rouder, Morey, Speckman, & Province, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Berger & Mortera, 1999; Edwards, Lindman, & Savage, 1963). For the interpretation of evidence in the Bayesian paradigm, the intention with which the data are collected is irrelevant; hence, the Bayes factor can be monitored as the data come in, and data collection may be terminated at any point (Berger & Wolpert, 1988; Rouder, 2014). Appendix A provides a brief description of the motivation and implementation of the Bayes factor hypothesis test used in this article.

Based on the above considerations, our intended sampling plan was as follows: We planned to collect a minimum of 20 participants in each between-subject condition (i.e., the global and local condition, for a minimum of 40 participants in total). We would then start to monitor the Bayes factor and stop the experiment whenever the critical hypothesis test (detailed below) reached a Bayes factor that would be considered "strong" evidence (Jeffreys, 1961); this meant that the Bayes factor is either 10 in favor of the null hypothesis, or 10 in favor of the alternative hypothesis. The experiment would also stop whenever we reached the

maximum number of participants, which was set to 50 participants per condition (i.e., a maximum of 100 participants in total).[2]

Participants

We recruited 112 students from the University of Newcastle and the University of New South Wales. Participants were remunerated with course credit or shopping vouchers. We screened participants on their ability to speak English and on having normal or corrected-to-normal vision. Participants were randomly allocated to either the global or local condition as they signed up for the study.

Materials/Stimuli

Instructions for each task were programmed to appear on screen before participants began. The Navon task was speeded, as this was the case in FD2012's procedure; participants were explicitly instructed to complete the task as quickly as possible. All other tasks were self-paced. The wording of each task's instructions can be found in Appendix B.

Navon Task. Participants completed a computerized version of the Navon task used in FD2012. In this task, participants saw a series of large letters, comprised of a number of smaller letters, one at a time. For instance, participants were presented with a number of small capital Ls that were visually arranged in the shape of a large capital H.

The experiment included two conditions: the global and the local condition. The control condition featured in FD2012 was omitted from the current experiment in order to focus all statistical power on a comparison of conditions that were maximally different. In the global condition, participants were required to respond to the identity of the large letter (H or

[2] Figure 2a in FD2012 shows that F (2, 57) = 8.93. Using that data to get our best estimate of the obtained effect size between the global and local condition, we obtain d = 0.952. For 100 participants, this yields a power of 0.999. A perhaps more realistic effect size estimate of d = 0.5 yields a power of 0.799.

L), while ignoring the identity of the small letters. Participants completed 48 trials in which each of four kinds of stimuli are presented 12 times each in random order: a large H comprised of small Ls, a large L comprised of small Hs, a large H comprised of small Fs, and a large L comprised of small Fs. An example of a stimulus letter can be found in Figure 1, and a list of the six letters used can be found in Appendix C.

In the local condition, participants were required to respond to the identity of the small letters (Hs or Ls), while ignoring the identity of the large letters. Participants completed 48 trials in which each of four kinds of stimuli are presented 12 times each in random order: a large L comprised of small Hs, a large F comprised of small Hs, a large H comprised of small Ls, or a large F comprised of small Ls.

For example, when responding to the stimulus displayed in Figure 1, global participants were required to respond with key "H", indicating that they judged the large letter to be an H, while ignoring that the small letters are Ls. In contrast, local participants presented with this same stimulus responded using key "L", indicating that the local letters were Ls, while ignoring the identity of the global H.

In both conditions, the global letter measured 2.5cm x 2.5cm in dimension, as specified in FD2012. Before each trial, a fixation cross of 22 x 22 pixels was displayed on the screen for 500 milliseconds, followed directly by a random stimulus letter. The first fixation cross and subsequent stimulus letter immediately followed the instruction screen. In both conditions, participants responded to targets using key 'L' (if target was an L) or 'L' (if target was an H). The six letters that were used in the study (two letters are used in both conditions) can be found in Appendix C.

The Navon task used in this experiment (and in Experiment 2) deviates from the one employed in the original FD2012 study in two ways: first, we instructed participants to attend to either the small or large stimulus letters, and second, we used response-incompatible

stimulus letters (i.e., an H made of Ls, or an L made of Hs in which the participant must suppress a competing response to correctly identify the target letter), as well as response compatible ones (i.e., an H made of Hs, in which the participant does not need to override a competing response). Our version of the task is consistently used by other researchers (e.g., Brand & Johnson, 2014; De Dreu, Baas & Nijstad, 2008; De Dreu, Nijstad & Baas, 2011; Gervais, Guinote, Allen & Slabu, 2012; Sligte, de Dreu & Nijstad, 2011) and generally results in robust effects. Therefore we do not expect these minor deviations to cause a discrepancy between our results and those of FD2012.

Mood Questionnaire. Participants also completed three administrations of the same computerized mood questionnaire: the Positive and Negative Affect Schedule (PANAS; Watson, Clark & Tellegen, 1988) as reported in FD2012. A copy of the paper version can be found in Appendix D.

BOC Task. Rosch's BOC task (Rosch, 1975) was administered, as in the original study. This task, as outlined in FD2012, seeks to measure participants' typicality ratings on exemplars of semantic categories. This task was administered via a computerized survey, and closely followed the task administration in the original publication.

Specifically, the BOC task used by FD2012 featured four semantic categories: 'furniture', 'vehicles', 'vegetables', and 'clothing'. For each of these categories, nine exemplars were presented, which fit into the categories to three varying degrees: three words were said to be 'fringe exemplars' of the category in question, three were 'good exemplars', and three were 'moderately good'. In FD2012 participants gave typicality ratings on a 10-point rating scale, with 0 denoting an exemplar that is 'not typical' of the category, and 9 denoting a rating of 'very typical'. FD2012 state that ratings for the three fringe exemplar words in each category "…reflect changes in perceptual breadth" (p. 112), whereas ratings of

good and moderately good exemplar words should not be associated with any effect of priming, as they are thought to be 'expected' of participants.

The task we used is that of Rosch (1975); however, note that one of the four categories we used differs from what was reported by FD2012 (i.e., we replaced the category 'clothing' used by FD2012 with the category 'sport', as the word category of clothing in Rosch's publication only contained two words per exemplar classification). We used Rosch's (1975) word lists as they appear. A complete list of these words, separated into their categories can be found in Appendix E. The words were presented in a randomized order to each participant. The entire browser-based experiment was programmed using the Qualtrics Survey software suite (http://www.qualtrics. com/).

Procedure

After signing a consent form, participants followed the experimental procedure as outlined in FD2012: Participants started by filling out the PANAS; after this, participants completed the Navon task. When the Navon task was completed, participants filled out the PANAS for the second time. Then, participants completed the BOC task. Finally, participants filled out the PANAS for a third time. The PANAS in the present study was administered three times (at the same points in the procedure as were reported in FD2012), to assess whether mood had a significant relationship to global or local processing styles in their experiments. Their analysis concluded that mood did not have a significant role in participant typicality ratings or their evaluation of the tasks; however, we nonetheless included the PANAS to avoid any differences in results that may arise from changing the sequence or duration of the procedure from that of FD2012.

FD2012 debriefed participants after completion of the study, and it was reported that none of the participants noticed a relationship between the Navon task, BOC, and the mood questionnaires. We also probed for bias, asking participants to ascertain the extent to which

they thought their behavior was affected by response or expectation biases at the time of their testing.[3]

Additionally, FD2012 discussed global versus local processing with participants after they completed the study, and asked them to rate to what extent they focused on details as opposed to the gestalt of the visual stimuli. FD2012 reported that these ratings did not vary between conditions, and therefore we refrained from collecting these ratings for the current replication attempt. After the experiment was completed, participants were thanked and remunerated for their participation.

Intended Analysis

The analysis partly follows FD2012: We calculated a creativity score for each participant, averaging their ratings for each of the 12 fringe word exemplars in the BOC task (recall that for each of the four word categories, three out of nine words are considered fringe exemplars). This produced a single "creativity score" with a range from zero to nine as the dependent variable. Data of participants whose average typicality rating falls outside of 2.5 standard deviations from their group (condition) mean were excluded from the final analysis. For the Navon task, all data of participants with an error rate >25% were excluded from analysis.

We compared the global and local conditions on the typicality ratings, using an independent samples one-sided Bayesian *t*-test as outlined in Rouder, Speckman, Sun, Morey, and Iversen (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009).[4] The

[3] The wording of these debriefing questions can be found in Appendix F.

[4] We use a Cauchy prior distribution for effect size with scaling parameter r, where $r = \sqrt{2}/2$ (i.e., 0.707; for details see Rouder et al., 2009; for the R BayesFactor package see Morey and Rouder, 2015, and see Appendix A).

resulting metric of this, and any other, Bayesian hypothesis test is a *Bayes factor*. Bayes factors are "...the primary tool used in Bayesian inference for hypothesis testing and model selection..." (Berger, 2006, p. 378), and unlike conventional *p*–values, allow for the quantification of evidence in favor of the null hypothesis relative to the alternative. For example, a Bayes factor of BF_{01} = 10 indicates that the observed data are 10 times more likely to have occurred under the null hypothesis than under the alternative hypothesis. In contrast, when a Bayes factor of BF_{01} = 1/5 is reported, the observed data are 5 times more likely to have occurred under the alternative hypothesis than the null hypothesis.

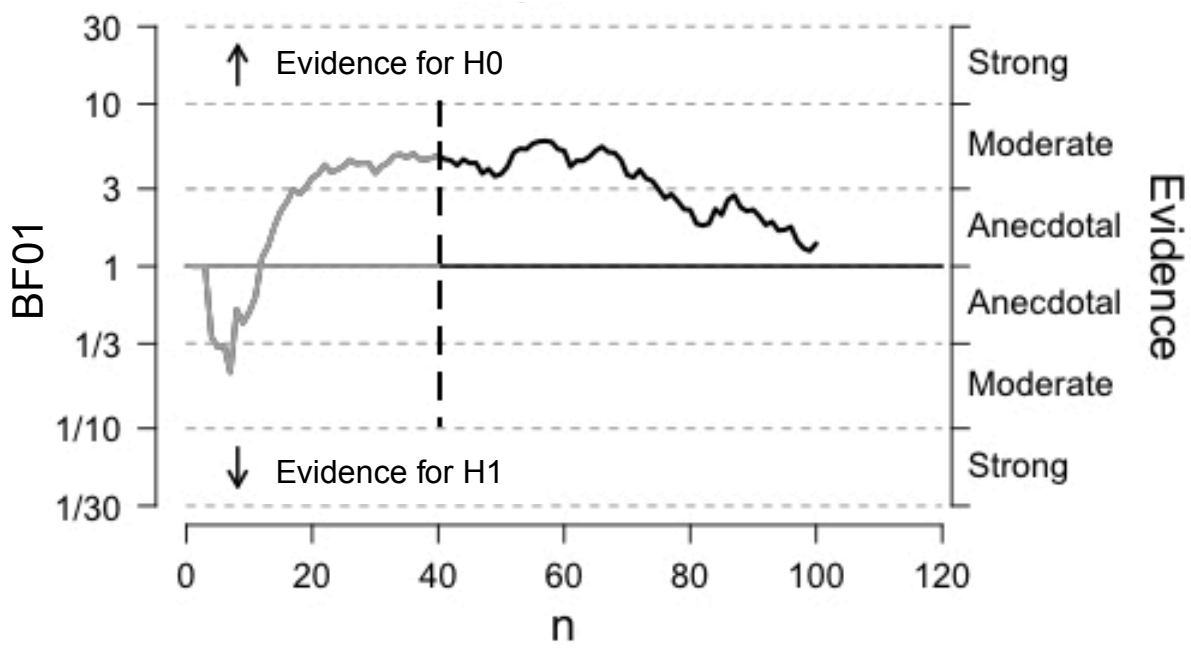

Figure 2. Development of the Bayes factor for Experiment 1 as evidence accumulates. The Bayes factor in this analysis demonstrates anecdotal evidence in favor of the null hypothesis, based on the categories defined by Jeffreys (1961). The vertical dotted line indicates the point at which we began to monitor the Bayes factor, per our sampling plan. Figure adjusted from the JASP output (jasp-stats.org).

Support in favor of the alternative hypothesis constitutes support in favor of the effect reported by FD2012 in their experiments. A Bayes factor lower than 1 indicates support for the hypothesis that ratings on the fringe exemplars will be higher in the global condition than in the local condition. Conversely, a Bayes factor higher than 1 indicates support for the null hypothesis: that there is no difference in these ratings between the priming conditions.

Analogous to FD2012, we examined the extent to which PANAS scores were related to typicality ratings. A frequentist regression analysis was conducted to establish whether a relationship exists between these typicality ratings (the dependent variable) and participant mood, whether positive or negative. A positive and negative score for each participant was calculated using the scoring guidelines that can be found at the bottom of Appendix D. Note that FD2012 did not find a relationship between PANAS scores and typicality ratings. Our analysis must accommodate this hypothesis, given that we too expect a null result for the regression.

The R code for this analysis can be found on the OSF page for this project: https://osf.io/ynr2q/. Our analysis was also conducted in JASP (Version 0.7, The JASP Team, 2016), a recently developed statistical software program that supports both classical and Bayesian analyses. Figures 2-5 were produced using JASP.

Results

BOC Measurement Check

The mean typicality ratings for high, moderate, and fringe stimuli in the BOC were 8.42 (SD = 0.63), 6.34 (SD = 1.30), and 5.45 (SD = 1.19), respectively. Moderate stimuli received lower typicality ratings than did high stimuli ($t(206) = 14.68$, $p<.05$, $BF_{01} < 10^{-30}$, which indicates the evidence in favor of the alternative hypothesis relative to the null hypothesis is over 10^{30} to 1) and fringe stimuli received lower typicality ratings than did

moderate stimuli ($t(206) = 5.12$, $p<.05$, $BF_{01} = 2.6 * 10^{-5}$, which indicates the evidence in favor of the alternative hypothesis relative to the null hypothesis is about 38,000 to 1).

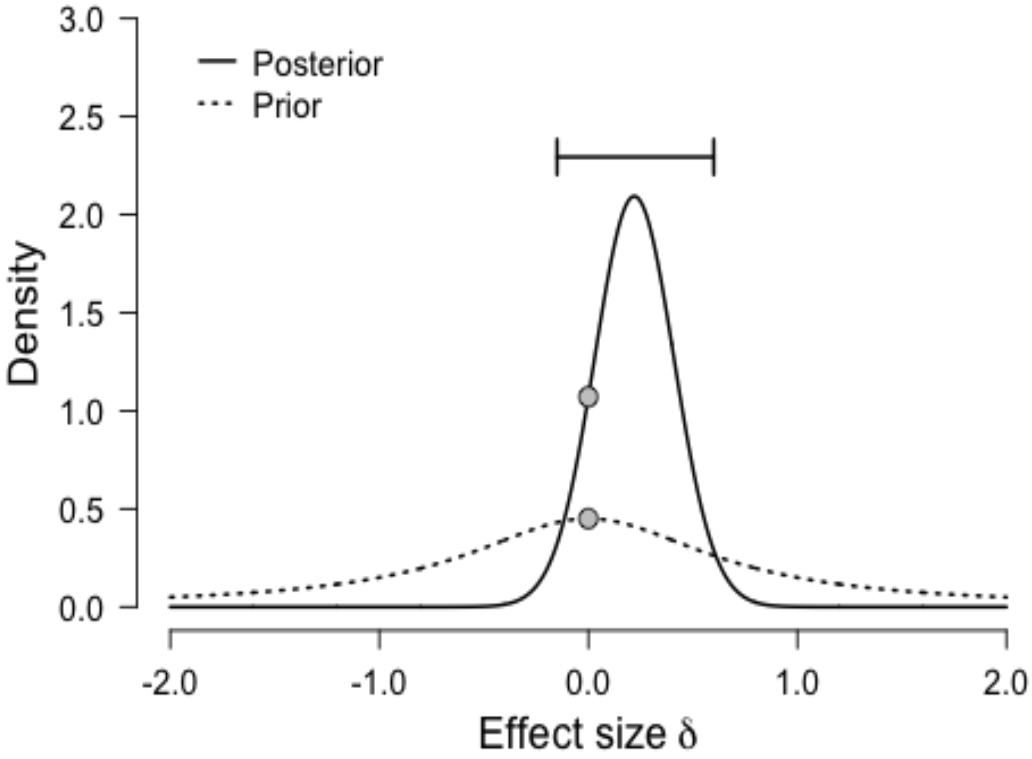

Figure 3. Prior and posterior distribution for effect size under the two-sided alternative hypothesis. The posterior 95% central credible interval ranges from -0.149 to 0.591; the dots indicate that the posterior density at δ=0 is 2.416 times higher than the prior density.

Exclusions

In total, 8 participants were excluded from the final analysis due to having an error rate >25% in the Navon task. Of the 104 data sets remaining, we used only the first 100 for the final analysis (as per our preregistered protocol). The data files with and without outlier exclusions are available on the OSF webpage for this replication project at https://osf.io/ynr2q/.

Confirmatory Analysis

To test our primary hypothesis, that of FD2012, we conducted a Bayesian independent samples one-tailed *t*-test, comparing the mean typicality ratings on BOC scores in the global and the local condition. The mean typicality rating for the global condition was 5.640 (*SD* = 1.192) and the mean typicality rating for the local condition was 5.347 (*SD* = 1.186). The Bayesian *t*-test yielded BF_{01} = 1.38, indicating that the data is 1.38 times more likely to have occurred under the null hypothesis than under the alternative hypothesis.

A Bayes factor <3 and >1/3 is conventionally considered to be anecdotal or "...not worth more than a bare mention." (Jeffreys 1961, p. 432). The evidential trajectory for increasing sample size is shown in Figure 2. We report only the Bayes factor at 100 data sets; however Figure 2 reveals that at no point do the data demonstrate even moderate support for the alternative hypothesis after we began monitoring the Bayes factor (at 40 participants).

Exploratory Analyses

Moderate and high typicality stimuli

For moderate stimuli, the mean typicality rating was 6.500 (*SD* = 1.296) in the global condition and 6.258 (*SD* = 1.329) in the local condition. The Bayesian *t*-test yielded BF_{01} = 2.02 after including all participants, indicating that the data is 2.02 times more likely to have occurred under the null hypothesis than under the alternative hypothesis which states that priming a global processing style increases typicality ratings for moderate stimuli.

For high stimuli, the mean typicality rating was 8.508 (*SD* = 0.533) in the global condition and 8.348 (*SD* = 0.706) in the local condition. The Bayesian *t*-test yielded BF_{01} = 1.30 after including all participants, indicating that the data is 1.30 times more likely to have occurred under the null hypothesis than under the alternative hypothesis which states that priming a global processing style increases typicality ratings for high stimuli.

PANAS

As in FD2012, we chose to assess whether affect meaningfully interfered with our results. There were no differences in PANAS scores between the global and local conditions for all three administrations of the task (all *p*s>.05). Regressing BOC scores on PANAS scores yielded no associations (all *p*s>.05).

Frequentist Analysis

In a frequentist one-tailed independent samples *t*-test, a comparison of global and local group means was not significant ($t(98) = 1.233$, $p = .110$). This analysis indicates that global priming did not significantly enhance participants' creativity.

Posterior Distribution of Effect Size

Figure 3 shows the posterior distribution of effect size for fringe exemplars using a two-sided prior distribution, as specified in the 'Intended Analysis' section. Most mass under the posterior density curve lies over and to the right of $\delta = 0$, indicating that although the direction of the results are consistent with the hypothesis of FD2012, the magnitude of the effect detected in Experiment 1 remains close to 0. A central 95% credible interval for effect sizes ranges from -0.149 to 0.591.

Bayes Factor for All 104 Participants

Should we have analyzed all available data, that is, beyond the pre-specified threshold, the final Bayes factor yielded by these data would have been $BF_{01} = 1.28$, qualitatively consistent with the weak evidence obtained in our confirmatory analysis.

Interim Conclusion

Experiment 1 did not conclusively demonstrate that global priming elicited an increase in creativity; as such, the results of Experiment 1 do not match those of FD2012. Although we tested more participants per condition than did FD2012 (i.e., 50 versus 20), it is possible that the true effect is non-zero, but smaller than reported by FD2012. In this case, a

Bayes factor threshold of 10 or 1/10 may not be a realistic level of evidence to expect when testing this effect with a sample size of 100 participants.

To conclusively address the question of a small sample size, we decided to carry out a second experiment. This experiment was conducted on MTurk to allow for testing a much larger batch of participants. This experiment was also preregistered on the OSF site (https://osf.io/ynr2q/), and methodologically identical to the first experiment, except that the tasks were administered via MTurk instead of in person.

Experiment 2

Method

Intended Sampling Plan

The sampling plan for the online study, also preregistered with the OSF and found at https://osf.io/ynr2q/, was similar in both rationale and theory to the in-lab study (see the methods section of Experiment 1, above for these considerations). However, there were a few practical differences. We planned to initially collect a minimum of 100 participants per condition, for a minimum of 200 participants in total. We would then begin monitoring the Bayes factor, stopping data collection until the sequential analysis yields a Bayes factor that would be considered 'strong' evidence (Jeffreys, 1961), as was the plan for Experiment 1.

In the event of not reaching either Bayes factor threshold after testing 100 participants in each condition, 20 more participants would be tested per condition (i.e., 40 further participants). A step size of 20 participants per condition is practical, given that testing participants on MTurk is relatively easy and fast.

The experiment would terminate when a maximum number of 300 participants per condition (i.e., 600 participants maximum, in total) had been reached, regardless of whether a threshold either in favor of the null or of the alternative hypothesis had been reached.

Furthermore, we planned to cease data collection if we were unable to recruit 600 participants before the date of July 6, 2015.

Outliers were determined in the same way as in Experiment 1. In the case of the sampling plan for Experiment 2, however, note that the classification of participants as outliers was continually reassessed after testing each additional 'step' of 20 participants per condition. That is, participants who were classified as outliers after testing *n* participants per condition may no longer be classified as outliers after *n* + 20 participants per condition. Similarly, participants who were not classified as outliers after testing *n* participants per condition may be classified as outliers after testing *n* + 20 participants per condition.

Participants

A total of 1307 participants were recruited via MTurk, and received $1.50 US as remuneration. Participants were randomly assigned to either the global or local conditions, within the restrictions imposed by the sampling plan. On MTurk we set the worker requirements such that only people who were in the US could participate.

Materials/Stimuli

The materials used in Experiment 2 are identical to those used in Experiment 1, except that during the Navon task, participants saw explicit instructions on screen about the speeded nature of the task. MTurk participants used their own computers for the experiment, and used the MTurk user interface online to access the experiment.

Procedure

Participants, once logged in to their MTurk participant accounts, were guided by the MTurk interface to the initial screen of the experiment program. From this point in the experiment, the procedure was identical to that of Experiment 1. Remuneration occurred when participants entered in a code word once the browser had redirected them back to the MTurk website, at the completion of the final task.

Contrary to Experiment 1, participants were not probed for bias or debriefed upon study completion. Results of Experiment 1 indicated that the tasks were perceived to be unrelated, so we opted to leave out the bias probing in Experiment 2.

Intended Analysis

The intended analysis of Experiment 2 was identical to that of Experiment 1.

Deviations from OSF Preregistration Document

We declare three minor deviations from the OSF Preregistration Document, all of which involve our protocol's sampling plan. Firstly, we stated that after 200 participants (100 per condition), we would increase the participant number by 'jumps' of 20 participants per condition. We stated that we would collect a maximum of 300 participants per condition, unless the Bayes factor reached the conventional upper threshold of 10, or the lower threshold of 1/10 (0.10) before 600 data sets were collected. Finally, we stated that we would cease data collection if we would be unable to recruit 600 participants before June 6, 2015.

Once testing using MTurk began, it became apparent that the attrition rate for both conditions was very high (>25%, as opposed to the in-lab study attrition rate, which was approximately 13%), which delayed collecting an acceptable amount of data in a timely fashion. In addition, we experienced significant delays with data collection due to unforeseen difficulties with updates in the survey software used. Approximately halfway through data collection, it was decided that due to these issues it would be reasonable to collect extra data, while still ceasing sequential analysis if either BF threshold was met. We decided to collect the data of a further ~800 participants, thus increasing the total viable sample size to 1307. This final stage did not commence until July 2015, however, due to a temporary shortage of funds in the MTurk account used. We decided to test this final batch of participants all at once, rather than in 20-participant batches, in order to minimize any further time delay with data collection. In this sample, some participant IP addresses were not consistent across the

five task stages due to a software problem. In an attempt to be as thorough as possible, and to ensure that the participants were reliably tested, an R script was written to check that the IP addresses of each of the five test phases matched only with the other four test phases, and only once. After this check, 221 data sets were discarded, leaving 1086 viable data sets. The R script written for this checking process can be found on the project's OSF page at https://osf.io/ynr2q/.

Results

BOC Measurement Check

The mean typicality ratings for high, moderate, and fringe stimuli in the BOC were 8.33 (SD = 0.69), 6.15 (SD = 1.18), and 4.68 (SD = 1.31), respectively. Moderate stimuli received lower typicality ratings than did high stimuli ($t(1969) = 50.20$, $p<.05$, $BF_{01} < 10^{-30}$, which indicates the evidence in favor of the alternative hypothesis relative to the null hypothesis is over 10^{30} to 1) and fringe stimuli received lower typicality ratings than did moderate stimuli ($t(1970) = 26.04$, $p<.05$, $BF_{01} < 10^{-30}$, which indicates the evidence in favor of the alternative hypothesis relative to the null hypothesis is over 10^{30} to 1)

Exclusions

In total, of the 1086 viable data sets, 90 were excluded from final analysis due to error rates >25% in the Navon task, and 10 data sets were excluded due to individual mean typicality ratings lying more than 2.5 standard deviations outside of the group (condition) mean. A further 10 data sets were excluded from the final analysis, as we had 10 more local data sets than global due to uneven attrition across the conditions. Of the 976 data sets remaining, we used only the first 908 for the final analysis (as per our preregistered stopping rule). The data files with and without outlier exclusions are available on the OSF webpage for this replication project at https://osf.io/ynr2q/.

Confirmatory Analysis

The main analysis strategy was identical to that of Experiment 1. The mean typicality rating for the global condition was 4.695 (*SD* = 1.333) and the mean typicality rating for the local condition was 4.670 (*SD* = 1.287). The Bayesian *t*-test yielded BF_{01} = 10.35 after including 908 participants (N = 454 per condition), indicating that the data are more than ten times more likely to have occurred under the null hypothesis than under the alternative hypothesis. The evidential trajectory for increasing sample size is shown in Figure 4.

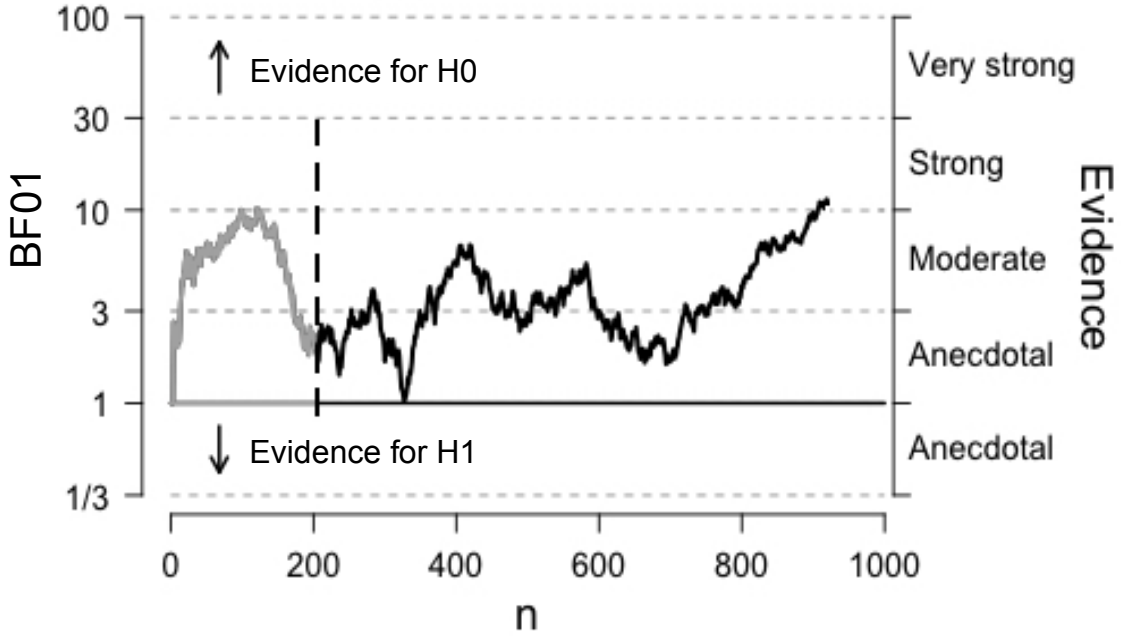

Figure 4. Development of the Bayes factor for Experiment 2 as evidence accumulates. The Bayes factor in this analysis demonstrates strong evidence in favor of the null hypothesis, based on the categories defined by Jeffreys (1961). The vertical dotted line indicates the point at which we began to monitor the Bayes factor, per our sampling plan. Figure adjusted from the JASP output (jasp-stats.org).

In summary, the Bayesian analysis provides strong evidence against the notion that globally primed participants gave higher typicality ratings in the BOC task than their locally

primed counterparts. We conclude that the results of Experiment 2 support the null hypothesis that, in the present context, global priming does not elicit inflated mean typicality ratings relative to local priming.

Exploratory Analyses

Moderate and high typicality stimuli

For moderate stimuli, the mean typicality rating was 6.159 (*SD* = 1.179) in the global condition and 6.133 (*SD* = 1.188) in the local condition. The Bayesian *t*-test yielded BF_{01} = 10.07 after including 422 participants (N = 211 per condition), indicating that the data are more than ten times more likely to have occurred under the null hypothesis than under the alternative hypothesis.

For high stimuli, the mean typicality rating was 8.334 (*SD* = 0.681) in the global condition and 8.335 (*SD* = 0.695) in the local condition. The Bayesian *t*-test yielded BF_{01} = 10.27 after including 216 participants (N = 108 per condition), indicating that the data are more than 10 times more likely to have occurred under the null hypothesis than under the alternative hypothesis.

PANAS

As in Experiment 1, we assessed the possible role of mood in the main analysis. There were no differences in PANAS scores between the global and local conditions for all three measurements (all *p*s > .05). Regressing BOC scores on PANAS scores yielded a positive relationship between positive mood and BOC scores on both the first and third PANAS measure (*p*s < .05). No other relationships between BOC scores and PANAS scores were found (all other *p*s > .05).

Frequentist Analysis

A frequentist one-tailed independent samples *t*-test conducted in Experiment 2 was not significant, $t(906) = 0.309$, $p = .379$, as was found in Experiment 1. These exploratory

results also echo those of the confirmatory analysis, and indicate that priming for global processing does not enhance participant's creativity.

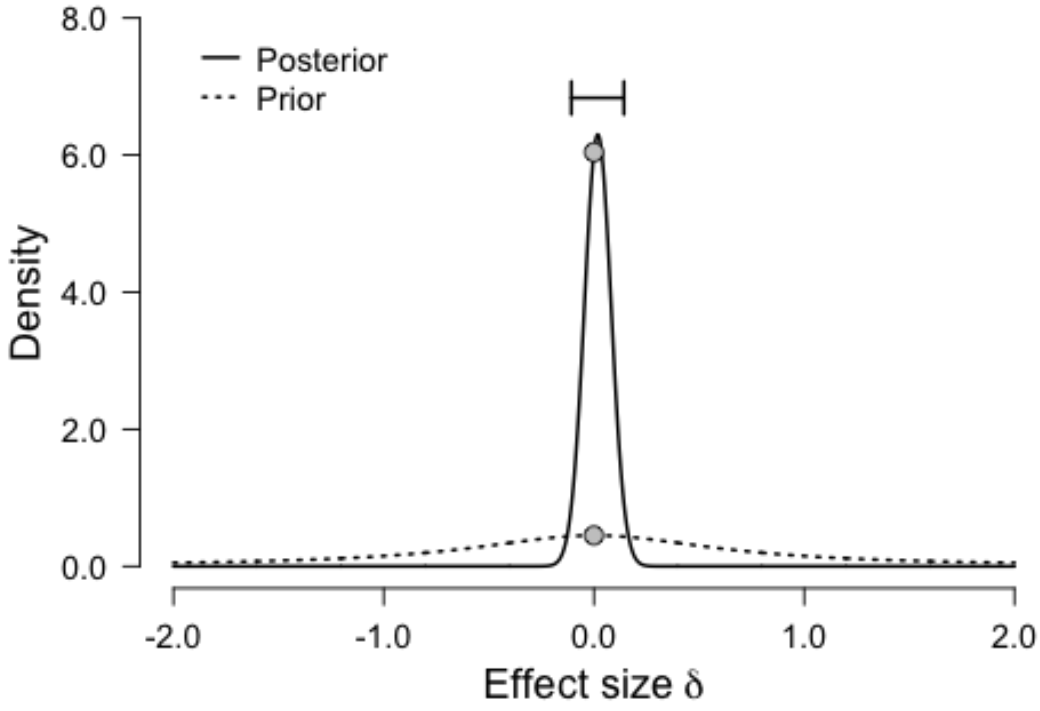

Figure 5. Prior and posterior distribution for effect size under the two-sided alternative hypothesis. The posterior 95% central credible interval ranges from -0.109 to 0.148; the dots indicate that the posterior density at δ=0 is 12.847 times higher than the prior density.

Posterior Distribution of Effect Size

Figure 5 shows the posterior distribution for effect size using a two-sided prior distribution. In this case, however, nearly all posterior mass is tightly concentrated around δ = 0, indicating that -compared to local priming- global priming does not result in a marked increase in creativity.

Bayes Factor for All 976 Participants

Here we report the final Bayes factor should we have analyzed all of the data available, as in Experiment 1. The final Bayes factor yielded by these data is BF_{01} = 12.847, which is almost the same as the Bayes factor reported in our confirmatory analysis.

General Discussion

In this study, we attempted to replicate the results of Experiment 1 as carried out by Förster and Denzler (2012). In our two experiments, we were unable to demonstrate support for the hypothesis that global priming enhances creativity relative to local priming. In Experiment 1, our Bayesian analysis yielded a Bayes factor of 1.38, which indicates that both the null and the alternative hypothesis are about equally supported by the data. In contrast, the Bayesian analysis of Experiment 2 yielded a Bayes factor of 10.35, indicating that the data are over 10 times more likely to have occurred under the null hypothesis than under the alternative. Jeffreys (1961) categorizes this as strong evidential support for the null hypothesis. Consistent with this assessment, the posterior distribution for effect size in Experiment 2 is tightly concentrated on zero. In both of our experiments, our exploratory frequentist analysis failed to produce a significant *p*-value (all *p*s > .05).

What have we learned from these results? One may argue that there exists a plausible theory that predicts that global perception increases creativity, whereas there exists no theory that predicts the absence of such a relationship. However, this line of reasoning overlooks the fact that the onus is on new scientific theories to prove themselves in light of empirical data. The mere fact that a theory has been proposed to account for a hypothesized relationship does not mean that this relationship is actually present. We agree with Peirce (1878) that an honest assessment of a hypothesis requires that both successes and failures be reported. In this specific example, it stands to reason that the original article was published in large part

because its results supported the GLOMO^{sys} model; consequently, the non-replication of these results is also informative.

In addition, one may argue that we have demonstrated nothing more than that the original effect is malleable. This can of course be said about all failures to replicate. In terms of this replication, it is possible that the methodology may have somehow introduced moderator variables that may have diluted FD2012's original effect. Given the malleability of the GLOMO^{sys} model and the complexity of priming tasks, these moderators may have been introduced through slight cultural and linguistic differences between the original and replication sample, differences in the distance between the participant and the screen, and variability in the focus of participants during the tasks. Such judgments are difficult to make in the absence of qualifying evidence. Regardless, the present data demonstrate that the proposed effect is relatively brittle. Indeed, one may wonder what we can learn from the original finding if the effect cannot be reproduced in a different lab with the same experimental set-up.

This replication also emphasizes the need for clearly operationalized variables, as well as the use of psychometrically construct-valid measures. In their title, FD2012 suggest 'creativity' as a dependent variable. As the nature of creativity is multidimensional in nature, there are many possible ways to operationalize the construct for use as a dependent variable in empirical settings (Davis & Belcher, 1971). Unfortunately, the authors did not explicitly operationalize creativity. This is not uncommon for publications in the creativity literature: an estimated 62% of articles fail to operationalize creativity as a dependent variable (Plucker, Beghetto & Dow, 2004). FD2012 do not motivate their use of the BOC task for their study, nor do they provide any information on its psychometric properties. It is possible that between the lack of clear operationalization and the use of an unvalidated measure, the

construct being manipulated by either global or local priming in FD2012 is not reliable, or may simply have resulted from Type-1 error.

We wish to acknowledge that the methodology of our Navon task differs from that of FD2012 in two subtle and unintended ways, as was kindly pointed out to us by a reviewer. First, in FD2012, participants were asked to press a response key if the stimulus contained an "L" or an "H". In the global (local) condition, the letter was always the global (local) one. Participants were not explicitly asked to look at 'big' or 'small' letters. One might argue that this subtle difference in instructions may have made our participants aware of the priming procedure, diluting or even eliminating the effect. We believe this possibility is extremely unlikely; for any specific participant, the link between the between-participant manipulation and the creativity test is virtually impossible to discern. In our lab setting we confirmed immediately after the experiment that none of the participants divined the connection between the priming task and the typicality-rating task.

Second, we used response-incompatible stimulus letters as well as response compatible ones, in contrast to FD2012 in which only response compatible trials were featured. As we discuss in our Experiment 1 method section, our version of the task has been used successfully in various other experiments. As such, we do not believe this deviation has resulted in a discrepancy between the findings of the current article and those of FD2012.

On the other hand, there is some evidence in the literature to suggest that the Navon task may not be suitable for induction purposes. Perfect, Weston, Dennis and Snell (2008), for example, raise concerns about the use of the Navon task to induce global and local processing styles. The study shows, among other things, that manipulating the distance between the local (small) letters relative to their size affects the extent to which participants naturally perceive the global or local features of the stimulus. The authors state that "The original Macrae and Lewis (2002) study provides no detail as to the construction of the

Navon stimuli, and subsequent studies do not provide much information either. Given this lack of information, researchers have tended to create the Navon stimuli themselves, and this may lead to discrepancies between studies." (p.1485). This indicates that the stimulus may be sufficiently complex that minor deviations between studies may cause major differences in experimental outcomes. This highlights the potentially limited utility of the Navon task in protocols such as that of FD2012, and by extension, the current one.

What do our findings mean for the GLOMO^{sys} model? One may argue that the GLOMO^{sys} model has already been validated in many conceptual replication studies (e.g., de Dreu, Baas & Nijstad, 2008; Gervais, Guinote, Allen & Slabu, 2012; Jia, Hirt & Karpen, 2009; Liberman, Polack, Hameiri & Blumenfeld, 2011; Sligte, de Dreu & Nijstad, 2011). In a field that is beset with publication bias, however, conceptual replication studies alone do not provide strong evidence (Pashler & Harris, 2012; Yong, 2012). In the words of Brian Nosek, "…psychology would suffer if [conceptual replication] wasn't practiced but it doesn't replace direct replication. To show that 'A' is true, you don't do 'B'. You do 'A' again." (Nosek, in Yong, 2012). Recent investigations have caused doubt about popular phenomena such as power posing (Ranehill et al., 2015; Simmons and Simonsohn, 2015), ego-depletion (Xu et al., 2014), the impact of disfluent fonts on numerical tasks (Alter, Oppenheimer, Epley & Eyre, 2007), the Mozart effect (Steele, Bass & Crook, 1999; Pietschnig, Voracek & Formann, 2010), and the Macbeth effect (Earp, Everett, Madva, & Hamlin, 2014; Fayard, Bassi, Bernstein & Roberts, 2009). We note that the original finding of FD2012 has only been *directly* tested by one party (Klauer & Singmann, 2015). This replication was not successful.

Finally, one might wonder whether our statistical framework of hypothesis testing hinges on the null-hypothesis being true. The question of whether or not a point-null hypothesis is ever exactly true has entertained statisticians and philosophers for many decades. In contrast to classical inference, however, the interpretation of the Bayes factor

does not require the null hypothesis (or the alternative hypothesis) to be true in some absolute sense; instead, the Bayes factor quantifies the relative predictive adequacy of the competing hypotheses (Wagenmakers, Grünwald, & Steyvers, 2006). For the data sets reported here the null hypothesis outpredicts the alternative hypothesis. A similar result will be obtained when the point null hypothesis is replaced with a distribution that is tightly centered on zero (Berger & Delampady, 1987).

We failed to replicate the target study despite both experiments containing a much larger sample size than the original. Furthermore, Experiment 2 was administered without experimenter-to-participant contact of any sort. As such, our results are not vulnerable to the argument of skeptics that our studies were underpowered, nor could they have been caused by experimenter bias or lack of skill in test administration (e.g., Bargh, 2012).

The reader may worry that data collected via MTurk is of lower quality and as such caused a failure to replicate. Many recent large-scale studies, however, demonstrate that MTurk data compares favorably to in-lab data, and to data collected via other online sources (see Bartneck, Duenser, Moltchanova & Zawieska, 2015; Buhrmester, Kwang & Gosling, 2011). Furthermore, any participants with high error rates in the Navon task were excluded from final analysis. This diminishes the possibility of participants' lack of focus or lack of understanding of the task interfering with the prime and contaminating the experimental data as a result. In an online study, however, the experimenter is unable to control for other extraneous factors relating to the experimental environment, as they would otherwise in an in-lab setting. The brightness of the participant's environment, or the distance of the participant from the on-screen stimuli cannot be guaranteed, for example. The relevance of such factors can never be excluded, but we deem these explanations unlikely a priori.

FD2012 argue that global priming enhances creativity in people due to broadening of semantic categories in memory. Unfortunately, the results of the current study would suggest

that until further confirmatory studies are conducted, these results and their corollaries should be regarded with caution, as should the broader predictions of the GLOMO^{sys} model. However, no single failure to replicate dismantles an entire study or body of work. We intend for the current replication attempt to act as part of a wide-scale and constructive research effort in psychological science.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Bornstein, B., ... & McCoy, S. (2014). Contribution to Alonga et al (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556-578.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136,* 569-576.

Bargh, J. A. (2012). Nothing in their heads. Psychology Today. Retrieved from: http://web.archive.org/web/20120307100648/http://www.psychologytoday.com/blog/the-natural-unconscious/201203/nothing-in-their-heads

Bartneck, C., Duenser, A., Moltchanova, E., & Zawieska, K. (2015). Comparing the similarity of responses received from studies in Amazon's Mechanical Turk to studies conducted online and with direct recruitment. *PLoS One, 10,* e0121595

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1,* 385-402.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2,* 317-335.

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94,* 542-554.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle.* Hayward, CA: Institute of Mathematical Statistics.

Brand, J., & Johnson, A. P. (2014). Attention to local and global levels of hierarchical Navon figures affects rapid scene categorization. *Frontiers in Psychology, 5,* 1-19.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3-5.

Carey, B. (2015, August 28). Many psychology findings not as strong as claimed, study says. New York Times. Retrieved from: http://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html?_r=0

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Davis, G. A., & Belcher, T. L. (1971). How Shall Creativity be Measured? Torrance Tests, RAT, Alpha Biographical, and IQ*. *The Journal of Creative Behavior, 5,* 153-161.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at *Cortex*. *Cortex, 49*, 609-610.

de Dreu, C. K., Baas, M., & Nijstad, B. A. (2008). Hedonic tone and activation in the mood-creativity link: Towards a dual pathway to creativity model. *Journal of Personality and Social Psychology, 94,* 739–756.

de Dreu, C. K., Nijstad, B. A., & Baas, M. (2011). Behavioral activation links to creativity because of increased cognitive flexibility. *Social Psychological and Personality Science, 2,* 72-80.

Dijkstra, K. A., van der Pligt, J., van Kleef, G. A., & Kerstholt, J. H. (2012). Deliberation versus intuition: Global versus local processing in judgment and choice. *Journal of Experimental Social Psychology, 48,* 1156-1161.

Earp, B. D., Everett, J. A., Madva, E. N., & Hamlin, J. K. (2014). Out, damned spot: Can the "Macbeth Effect" be replicated? *Basic and Applied Social Psychology, 36,* 91-98.

Edwards, W., Lindman, H., & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193-242.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2016). How to become a Bayesian in eight easy steps: An annotated reading list. Manuscript submitted for publication.

Fayard, J. V., Bassi, A. K., Bernstein, D. M., & Roberts, B. W. (2009). Is cleanliness next to godliness? Dispelling old wives' tales: Failure to replicate Zhong and Liljenquist (2006). *Journal of Articles in Support of the Null Hypothesis, 6,* 21-30.

Förster, J., & Dannenberg, L. (2010). GLOMOsys: A systems account of global versus local processing. *Psychological Inquiry, 21,* 175-197.

Förster, J., & Denzler, M. (2012). Sense creative! The impact of global and local vision, hearing, touching, tasting and smelling on creative and analytic thought. *Social Psychological and Personality Science, 3,* 108-117. (Retraction published 2015, Social Psychological and Personality Science, 6, 118).

Gervais, S. J., Guinote, A., Allen, J., & Slabu, L. (2013). Power increases situated creativity. *Social Influence, 8,* 294-311.

Hartshorne, C., & Weiss, P. (eds.) (1932). *Collected papers of Charles Sanders Peirce: Volume II: Elements of logic.* Cambridge, UK: Harvard University Press.

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Jia, L., Hirt, E. R., & Karpen, S. C. (2009). Lessons from a faraway land: The effect of spatial distance on creative cognition. *Journal of Experimental Social Psychology, 45*, 1127-1131.

Klauer, K. C., & Singmann, H. (2015). Does Global and Local Vision Have an Impact on Creative and Analytic Thought? Two Failed Replications. *PloS one, 10,* e0132885.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Woodzicka, J. A. (2015). Investigating variation in replicability. *Social Psychology. 45,* 142–152

Kühnen, U., & Hannover, B. (2010). Culture, self-construal, and regulatory focus: How and what to promote or prevent? *Psychological Inquiry, 21,* 233-238.

Liberman, N., Polack, O., Hameiri, B., & Blumenfeld, M. (2012). Priming of spatial distance enhances children's creative performance. *Journal of Experimental Child Psychology, 111,* 663-670.

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26,* 1827-1832.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016). Harold Jeffreys' default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72,* 19-32.

Macrae, C. N., & Lewis, H. L. (2002). Do I know you? Processing orientation and face recognition. *Psychological Science*, *13*, 194-196.

Markman, K. D., & Dyczewski, E. A. (2010). Think and act global and local: A portrait of the individual as a flexible information-processor. *Psychological Inquiry, 21,* 239-241.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406-419.

Morey, R. D., & Rouder, J. N. (2015). BayesFactor 0.9.11-1 (with contributions from Tahira Jamil). Comprehensive R Archive Network, http://cran.r-project.org/web/packages/BayesFactor/index.html.

Mulder, J., & Wagenmakers, E.-J. (2016). Editor's introduction to the special issue on "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology, 72,* 1-5.

Nosek, B. A., & Lakens, D. (2015). Registered reports. *Social Psychology, 45,* 137-141.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348,* 1422-1425.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

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

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.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly, 13,* 470-482.

Perfect, T. J., Weston, N. J., Dennis, I., & Snell, A. (2008). The effects of precedence on Navon-induced processing bias in face recognition. *The Quarterly Journal of Experimental Psychology, 61,* 1479-1486.

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

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist, 39,* 83-96.

Posner, M. (1987). *Structures and functions of selective attention. In American Psychology Association Master Lecture Series.* Washington, DC: American Psychological Association.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the Robustness of Power Posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26,* 653-656

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21,* 301-308.

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra, 2,* 1-12.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (in press). Is There a Free Lunch in Inference? *Topics in Cognitive Science.*

Rouder, J. N., Morey, R.D., Speckman, P.L., & Province, J.M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356-374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225-237.

Simmons, J. & Simonsohn, U. (August 5, 2015). Power posing: Reassessing the evidence behind the most popular TED talk [Blog post]. Retrieved from: http://datacolada.org/2015/05/08/37-power-posing-reassessing-the-evidence-behind-the-most-popular-ted-talk

Sligte, D. J., de Dreu, C. K. W., & Nijstad, B. A. (2011). Power, stability of power, and creativity. *Journal of Experimental Social Psychology, 47,* 891–897.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10,* 886-899.

Steele, K. M., Bass, K. E., & Crook, M. D. (1999). The mystery of the Mozart effect: Failure to replicate. *Psychological Science, 10,* 366-369.

The JASP Team (2016). JASP (Version 0.7.5.6)[Computer software]. https://jasp-stats.org/.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General, 143,* 1457-1475.

Wagenmakers, E. -J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review, 14,* 779-804.

Wagenmakers, E.-J., Beek, T., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., Verhagen, A. J., Selker, R., Sasiadek, A., & Pinto, Y. (2014). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. Accepted preregistration document, available at https://osf.io/y7j4h/.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50,* 149-166.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25,* 169-176.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632-638.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of Positive and Negative Affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review, 19,* 1057-1064.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review, 16,* 752-760.

Xu, X., Demos, K. E., Leahey, T. M., Hart, C. N., Trautvetter, J., Coward, P., Middleton, K. R., & Wing, R. R. (2014). Failure to replicate depletion of self-control. *PloS one, 9,* e109950.

Yong, E. (May 16, 2012). Replication studies: Bad copy. Nature News. Retrieved from http://www.nature.com/news/replication-studies-bad-copy-1.10634.

Appendix A
Motivation and Implementation of the Bayes Factor Hypothesis Test

Below we provide a short description of the Bayes factor hypothesis test; a detailed treatment is available elsewhere (e.g., Edwards et al., 1963; Jeffreys, 1961; Ly, Verhagen, & Wagenmakers, 2016; Rouder et al., 2009; Rouder, Wagenmakers, & Morey, 2016; Wagenmakers, Morey, & Lee, 2016). In addition, Etz, Gronau, Dablander, Edelsbrunner, and Baribault (2016) provide a Bayesian reading list, Mulder & Wagenmakers (2016) introduce a recent special issue on Bayes factors for psychology, and both Morey and Rouder (2015) and The JASP Team (2016) provide free and user-friendly software for obtaining Bayes factors in statistical scenarios that are common in psychological research.

Consider the case of two competing hypotheses: (1) H_0, the skeptics' null hypothesis, stipulates that effect size δ equals zero, that is, H_0: δ=0; (2) H_1, the alternative hypothesis, relaxes the restriction that δ=0 and instead assigns it a prior distribution, allowing δ to vary, that is, H_1: $\delta \sim f(\theta)$.

Note that both a traditional frequentist approach and a Bayesian approach refer to H_0 and H_1. However, only the Bayesian approach explicitly includes the alternative hypothesis when drawing statistical inference. However, the Bayesian approach requires that H1 is specified in more detail – it is not sufficient to say that H1: delta ≠ 0; rather, the analysts must specify a distribution for delta, as this allows H1 to make concrete predictions, the adequacy of which can be contrasted against those made by H0 (e.g., Rouder et al., in press).

An application of Bayes' rule gives:

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior plausibility} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior plausibility} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\substack{\text{Predictive} \\ \text{updating factor}}} \quad (1)$$

In words, the relative posterior plausibility of H_0 and H_1 equals the relative prior plausibility of H_0 and H_1 multiplied by the relative predictive adequacy of H_0 and H_1 for the observed data. Equation 1 showcases three important properties of belief updating. First, the posterior plausibilities are a compromise between prior plausibilities and relative predictive performance. This means that very different prior beliefs will initially lead to divergent posterior beliefs. Concretely, proponents of the GLOMO^{sys} account may store considerable faith in H_1, whereas skeptics may feel that H_0 merits serious attention. The only way to bring these divergent initial opinions into rough agreement is to collect data and apply a rational updating process as specified by Equation 1. We adhere to statistical tradition and focus on the predictive updating factor -the Bayes factor- that is, on the evidence that the data provide for the competing hypotheses. Proponents and skeptics may then adjust their individual beliefs accordingly.

A second property that is evident from Equation 1 is that the Bayes factor hypothesis test is symmetric in that it pits predictive performance of H_0 against that of H_1. Neither H_0 nor H_1 has a special status or is privileged in any way. This property allows the Bayes factor to quantify evidence in favor of the null hypothesis.

A third property is that the updating factor is solely based on relative predictive performance – the notion of a "true" model is absent. Thus, when the Bayes factor equals 4.5, this means that H_1 predicted the observed data 4.5 times better than H_0. This point is particularly important as it has sometimes been asserted that hypothesis tests are useless because the null hypothesis is never true (e.g., Cohen, 1994). This argument carries no force as far as the Bayes factor is concerned, as the Bayes factor is determined by out-of-sample prediction errors (e.g., Wagenmakers et al., 2006).

Thus, it may be that many effects are in the strictest sense not exactly zero. However, the effects may be so small that they drown in measurement error. As remarked by Gelman,

"…when effect size is tiny and measurement error is huge, you're essentially trying to use a bathroom scale to weigh a feather---and the feather is resting loosely in the pouch of a kangaroo that is vigorously jumping up and down." (http://andrewgelman.com/2015/04/21/feather-bathroom-scale-kangaroo/). A similar sentiment was expressed by Edwards et al. (1963, p. 215-216): "Convention asks, 'Do these two programs differ at all in effectiveness?' Of course they do. Could any real difference in the programs fail to induce at least some slight difference in their effectiveness? Yet the difference in effectiveness may be negligible compared to the sensitivity of the experiment. In this way, the conventional question can be given meaning, and we shall often ask it without further explanation or apology." In the cases described above, H_0 will predictively outperform H_1, even though H_0 may not be strictly true. As an aside, all statistical models are abstractions of reality and a case can be made that none of our models are true (Rouder et al., 2016). As is evident from Equation 1, this fact does not invalidate the Bayes factor.

In sum, the predictive performance of H_0 and H_1 is assessed by p(data | H_0: $\delta = 0$) and p(data | H_1: $\delta \sim f(\theta)$), respectively. Because the null hypothesis is instantiated as a single point (i.e., $\delta = 0$; for an interval specification see Morey & Rouder, 2011), the computation of p(data | H_0) is straightforward. The computation of p(data | H_1: $\delta \sim f(\theta)$) is only slightly more involved; for an intuitive explanation, consider p(data | H^*_1: $\delta = x$), the prediction for a specific δ under H_1. Such predictions can be computed for all possible values of δ, and the overall predictive performance for H_1 is then obtained by weighting the specific predictions by the prior distribution $f(\theta)$.

Mathematically, the above section can be summarized as follows:

$$\mathrm{BF}_{10} = \frac{p(\mathrm{data} \mid \mathcal{H}_1)}{p(\mathrm{data} \mid \mathcal{H}_0)} = \frac{\int_\Delta p(\mathrm{data} \mid \delta) p(\delta)\, \mathrm{d}\delta}{p(\mathrm{data} \mid \delta = 0)} \qquad (2)$$

In this equation, p(δ) denotes the prior distribution $f(\theta)$.

For the case of the *t*-test, the prior distribution for the effect size parameter is crucial (Ly et al., 2016). Based on general desiderata, Jeffreys (1961) proposed a Cauchy distribution – a t-distribution with one degree of freedom, which is bell-shaped but is more platykurtic than the normal distribution. The location parameter of the Cauchy prior distribution is usually set to zero, consistent with Jeffreys's conceptualization of H_0 as an invariance or general law, and H_1 as the relaxation of that law (for an alternative see Verhagen & Wagenmakers, 2014). The scale parameter that determines the spread of the predictions from H_1 was set to 1 by Jeffreys, but recent software programs have reduced this to 0.707, meaning that the probability is 50% that the true effect size falls in the interval from -0.707 to 0.707 (Morey & Rouder, 2015; The JASP Team, 2016).

In experimental psychology, researchers usually test a hypothesis that has a specific direction. In the case of GLOMO^{sys}, for instance, the pertinent hypothesis states that a global processing style will promote creativity, not hinder it. The proposed preregistration analysis therefore involved an independent samples *t*-test contrasting the predictive performance of H_0: $\delta = 0$ against that of H_1: $\delta \sim$ Cauchy(location = 0, scale = 0.707) $I(0,\infty)$, where the I operator indicates the allowed interval. This analysis is easy to carry out in freely available software programs (e.g., Morey & Rouder, 2015; The JASP Team, 2016). An annotated JASP file for the results reported in this manuscript is available at https://osf.io/hbk2m/.

Appendix B
Task Instructions

All conditions: PANAS instructions

This task is part 1 of 5 in this study. This scale consists of a number of words that describe different feelings and emotions. Read each item, then move the sliders for each word to indicate to what extent you feel this way right now, that is, at the present moment. Move the slider next to each word choice to indicate your choice. Please complete a rating between 1-5 for each word.

1	2	3	4	5
Very slightly or not at all	A little	Moderately	Quite a bit	Extremely

When you have completed this task, you will be presented with a link to click on to proceed to the next part of the study. You may withdraw from this study at any time without needing to provide an explanation.

All conditions: BOC instructions

This task is part 4 of 5 in this study. In this task you will be required to rate the typicality of a word to a given category. You may move the sliders below each question to indicate your rating on a scale from 0 to 9, where 0 denotes the lowest typicality rating (i.e., the word is NOT typical of the category), and 9 denotes the highest typicality rating (i.e., the word is HIGHLY typical of the category). Consider this example: When asked to rate “How typical is a sparrow to the category bird?" you might move the slider to 8 to indicate that you think a sparrow is very typical of the ‘bird’ category. When you have completed this task, you will be presented with a link to click on to proceed to the next part of the study. You may withdraw from this study at any time without needing to provide an explanation.

Global condition: Navon instructions

Please read the following instructions carefully before proceeding to the task. You are about to begin part 2 of 5 in this study. **In this task, you will need to indicate whether the large letters on the screen (made up of small letters) are Hs (with the 'H' key indicated below), or Ls (with the 'L' key, indicated below).** Please do your best to respond as quickly as possible. When you have completed this part of the study, please click on the link presented on the screen to proceed to the next part of the study. You may withdraw from this study at any time without needing to provide an explanation. When you are reading to begin the task, please click on the red 'next' button on the bottom of this screen. During this task, please keep your forefingers on the H and L keys at all times.

Local condition: Navon instructions

Please read the following instructions carefully before proceeding to the task. You are about to begin part 2 of 5 in this study. **In this task, you will need to indicate whether the small letters on the screen (that make up the large letter) are Hs (with the 'H' key indicated below), or Ls (with the 'L' key, indicated below).** Please do your best to respond as quickly as possible. When you have completed this part of the study, please click on the link presented on the screen to proceed to the next part of the study. You may withdraw from this study at any time without needing to provide an explanation. When you are reading to begin the task, please click on the red 'next' button on the bottom of this screen. During this task, please keep your forefingers on the H and L keys at all times.

Appendix C
Navon Stimulus Letters

Six Navon-style letters constructed for use in the proposed replication, based on details in FD2012's procedure. Note that the top left and the middle left stimuli are used in the global condition only, the bottom two stimuli are used in the local condition only, and the top right and the middle right stimuli are used in both conditions.

Appendix D
The Positive and Negative Affect Scale (PANAS)

This questionnaire was computerized for administration in the replication, however the instructions and items remain as below.

Worksheet 3.1 The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988)

PANAS Questionnaire

This scale consists of a number of words that describe different feelings and emotions. Read each item and then list the number from the scale below next to each word. **Indicate to what extent you feel this way right now, that is, at the present moment *OR* indicate the extent you have felt this way over the past week (circle the instructions you followed when taking this measure)**

1	2	3	4	5
Very Slightly or Not at All	A Little	Moderately	Quite a Bit	Extremely

_____ 1. Interested	_____ 11. Irritable
_____ 2. Distressed	_____ 12. Alert
_____ 3. Excited	_____ 13. Ashamed
_____ 4. Upset	_____ 14. Inspired
_____ 5. Strong	_____ 15. Nervous
_____ 6. Guilty	_____ 16. Determined
_____ 7. Scared	_____ 17. Attentive
_____ 8. Hostile	_____ 18. Jittery
_____ 9. Enthusiastic	_____ 19. Active
_____ 10. Proud	_____ 20. Afraid

Scoring Instructions:

Positive Affect Score: Add the scores on items 1, 3, 5, 9, 10, 12, 14, 16, 17, and 19. Scores can range from 10 – 50, with higher scores representing higher levels of positive affect. Mean Scores: Momentary = 29.7 (*SD* = 7.9); Weekly = 33.3 (*SD = 7.2)*

Negative Affect Score: Add the scores on items 2, 4, 6, 7, 8, 11, 13, 15, 18, and 20. Scores can range from 10 – 50, with lower scores representing lower levels of negative affect. Mean Score: Momentary = 14.8 (*SD* = 5.4); Weekly = 17.4 (*SD* = 6.2)

Appendix E
Rosch (1975) Word List

Word lists for the fringe, good and moderate exemplars for categories of furniture, vehicles, vegetables and sport, taken from Rosch (1975).

Furniture	Good	1. Chair 2. Table 3. Bed
	Moderate	1. Lamp 2. Desk 3. Television
	Fringe	1. Rug 2. Stove 3. Fan
Vehicles	Good	1. Car 2. Bus 3. Truck
	Moderate	1. Airplane 2. Bicycle 3. Boat
	Fringe	1. Wheelchair 2. Tractor 3. Wagon
Vegetables	Good	1. Peas 2. Corn 3. Carrots
	Moderate	1. Celery 2. Turnips 3. Tomatoes
	Fringe	1. Mushrooms 2. Potatoes 3. Pumpkin
Sport	Good	1. Football 2. Tennis 3. Baseball
	Moderate	1. Archery 2. Fishing 3. Ping-Pong
	Fringe	1. Chess 2. Horseback-Riding 3. Hunting

Appendix F
Bias Probe Questions

Wording for verbally delivered questions to probe for bias, based on detail in the FD2012 procedure. Participants were requested to answer these questions upon completion of the five tasks.

1. "Did you get the sense or idea that the three types of tasks (mood questionnaire, Navon and Breadth of Categorization) were related to one another in any way?"

2. "Did you feel as though your responses were influenced by what you thought was expected of you, or by anything other than the on-screen instructions?"