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*

Abstract

Behavioral 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 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, participant first processed letters either globally or locally, then were tested for creativity using a word categorization 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 10.35, indicating strong support for the null hypothesis that global priming does not affect participant creativity.

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

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 Guney for her assistance in data collection for Experiment 1.

Over the past years, the reproducibility of psychological science has become a topic of much debate (e.g., Carey, 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 some high-profile journals to reinforce transparency in experimentation, as well as improving protocols for scholarly publication. For instance, the *Journal of Experimental Psychology: General* publishes replication studies following peer review of intended research protocols (see also Chambers, 2014). 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).

In this article we present two preregistered replication studies which 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 (2012; henceforth FD2012) to explore the predictions of the GLObal versus LOcal processing Model (Förster and Dannenberg, 2010; GLOMO^{sys}). GLOMO^{sys} is a dual-process model in which global and local processing styles create different yet related mental representations of the world. These representations allow people to extract different kinds of information from the environment, depending on which of the two systems is activated. Among other predictions, the model hypothesizes that activation of the global system in GLOMO^{sys} broadens semantic categories in memory to facilitating holistic and creative thinking. 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 then thought to shift to a conceptual plane, influencing performance in the word categorization task, used in FD2012 as a measure of creativity.

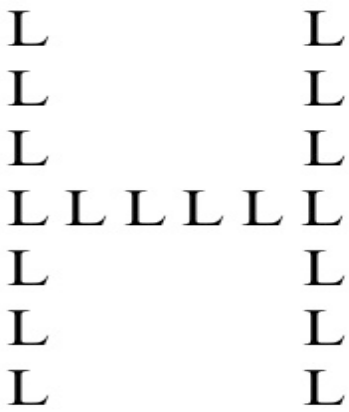


Figure 1. One of six stimulus letters presented, recreated from the description that accompanied Experiment 1 of FD2012. In the global condition, the correct answer is 'H', whereas in the local condition, the correct answer is 'L'.

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 categorization task in which participants assess how typical words are to given semantic categories. One example is: "How typical is the word 'wheelchair' for the category of 'vehicle'?" The results of FD2012 supported their hypotheses: participants primed with a global processing style indicated fringe exemplars to be significantly more typical category members than participants primed with a local processing style (see Figure 2). 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 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 some debate about the reproducibility of the results (e.g., LOWI, 2012; Klauer & Singmann, in press).

We planned to validate the GLOMO^{sys} model and its predictions by examining the replicability of the key empirical effect. To fulfill our aims, we conducted two replications of Experiment 1 from FD2012: one direct 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 both replications are identical.

Both replications are preregistered on 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.

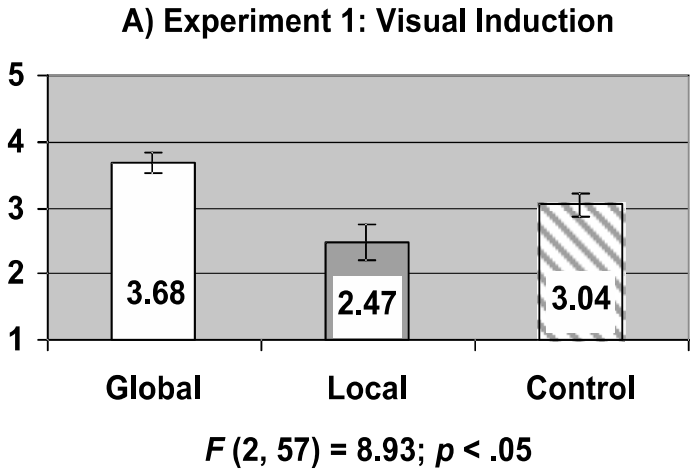


Figure 2. Mean typicality ratings of fringe exemplars for the global, control, and local conditions in Experiment 1 from FD2012, consistent with the predictions of GLOMOsys (Förster & Dannenberg, 2010). Compared to a control condition, inducing global perceptual processing increased typicality ratings in the BOC task, whereas inducing local perceptual processing decreased typicality ratings. Figure taken directly from FD2012 (see their Figure 2 A-E, pg. 113).

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).

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).

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 A.

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 L), while ignoring the identity of the small letters. Participants completed 48 trials in which different stimulus letters are presented 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 can be found in Figure 1, and a list of the six letters used can be found in Appendix B.

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 'h' (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 B.

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 C.

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" (pp. 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, in English, as it is unclear whether the original study conducted this task in English or Dutch. A complete list of these words, separated into their categories can be found in Appendix D. 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.[1]

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 these creativity scores, 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)[2]. The 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

[1]The wording of these debriefing questions can be found in Appendix E.

[2]We use a Cauchy prior distribution for effect size with scaling parameter r, where $r = \sqrt{2}/2$ (i.e., 0.707 or "medium"; for details see Rouder et al., 2009; for the R BayesFactor package see Morey, 2014).

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.

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 C. 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, Love et al., 2015), the newly developed Bayesian statistical software program. Figures 3-6 were produced using this software.

Results

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 strength of evidence for increasing sample size is shown in Figure 3. We report only the Bayes factor at 100 data sets; however Figure 3 reveals that at no point do the data demonstrate even moderate support for the alternative hypothesis after we begin monitoring Bayes factor (at 40 participants).

In summary, the Bayesian analysis yielded inconclusive support for the proposed hypothesis; Experiment 1 was not able to replicate the results of FD2012.

Exploratory Analyses: Experiment 1

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 participant's creativity.

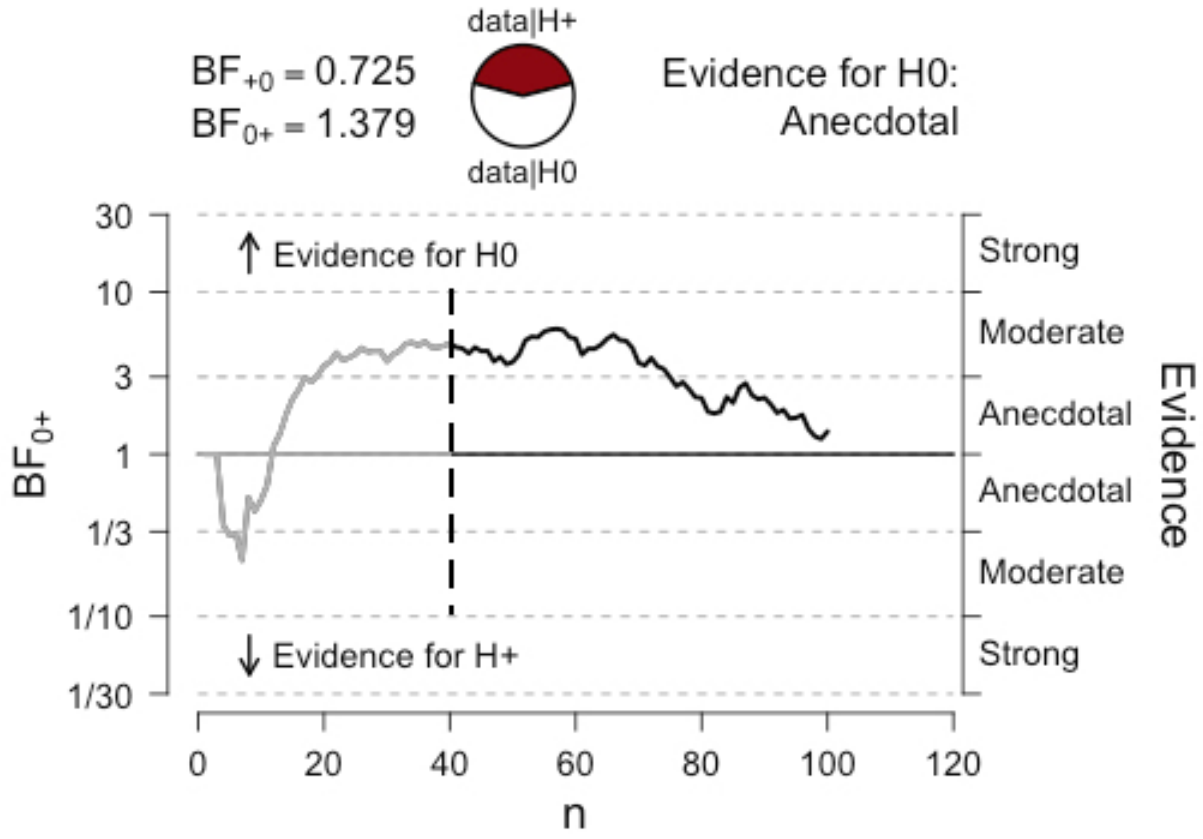


Figure 3. A sequential graph of the activity of the Bayes factor for Experiment 1 as evidence accumulates. The Bayes factor in this analysis demonstrates anecdotal evidence in favor of the null hypothesis, based on the categories defined by Jeffreys (1961). The vertical dotted line indicates the point at which we began to monitor the Bayes factor, per our sampling plan. Figure adjusted from the JASP output (jasp-stats.org).

Figure 4 shows the posterior distribution of effect size 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.

Bayes Factor for All 104 Participants

Should we have analyzed all the available data, that is, beyond the pre-specified threshold, the final Bayes factor yielded by these data would have been BF_{01} = 1.28. Using the same categorical classification as in our confirmatory analysis, this Bayes factor too is merely anecdotal evidence in favor of the null hypothesis.

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 collected more data per condition than was collected in FD2012, it is plausible that this effect is theoretically real, but of a small effect size, and thus not detected. As such, 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. Conversely, if the effect is not real, a sample size of 100 may not have been great enough to detect a null effect of the magnitude needed to reach our prescribed threshold.

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.

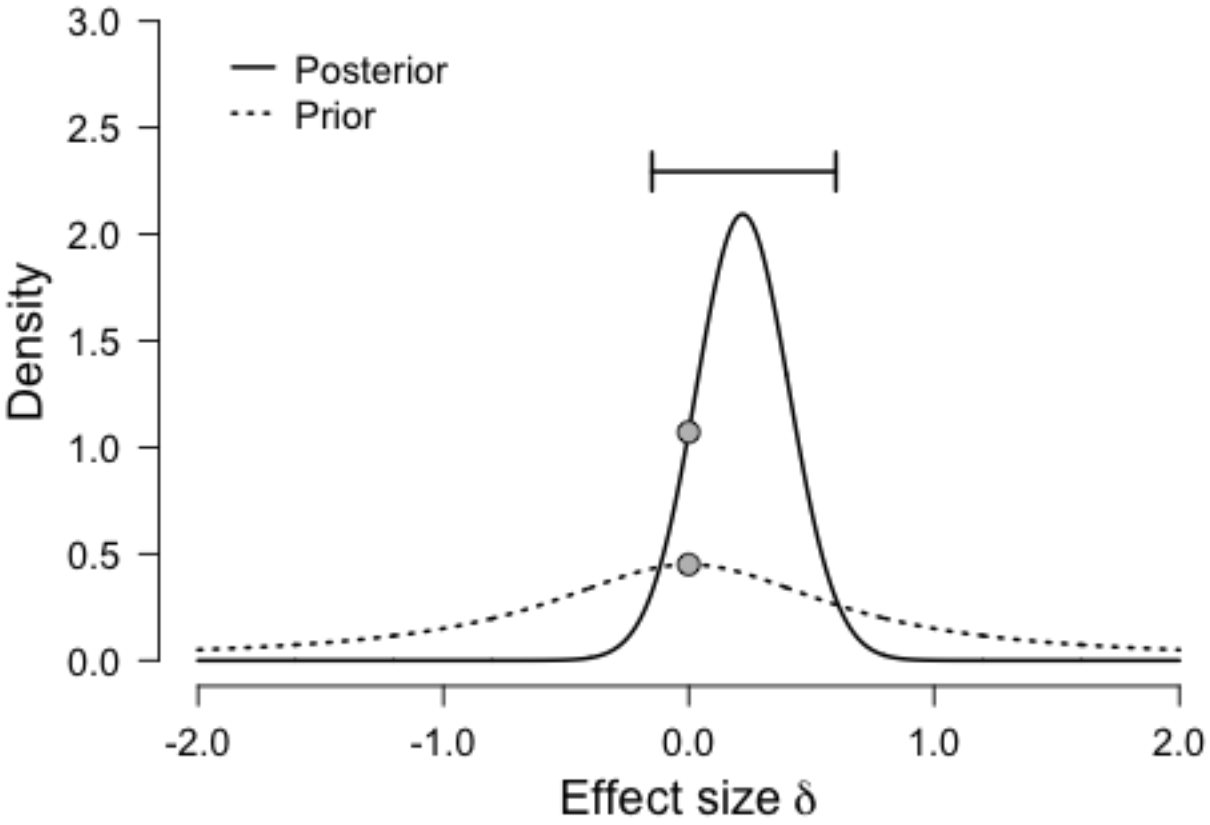


Figure 4. The posterior distribution of effect size δ over a two-sided prior distribution for Experiment 1. The posterior distribution density over an effect size of 0 is 1.38 times greater than the prior distribution indicated before the data had been collected.

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 addition '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 United States 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, 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

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 creativity scores 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 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

*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 strength of evidence for increasing sample size is shown in Figure 5.

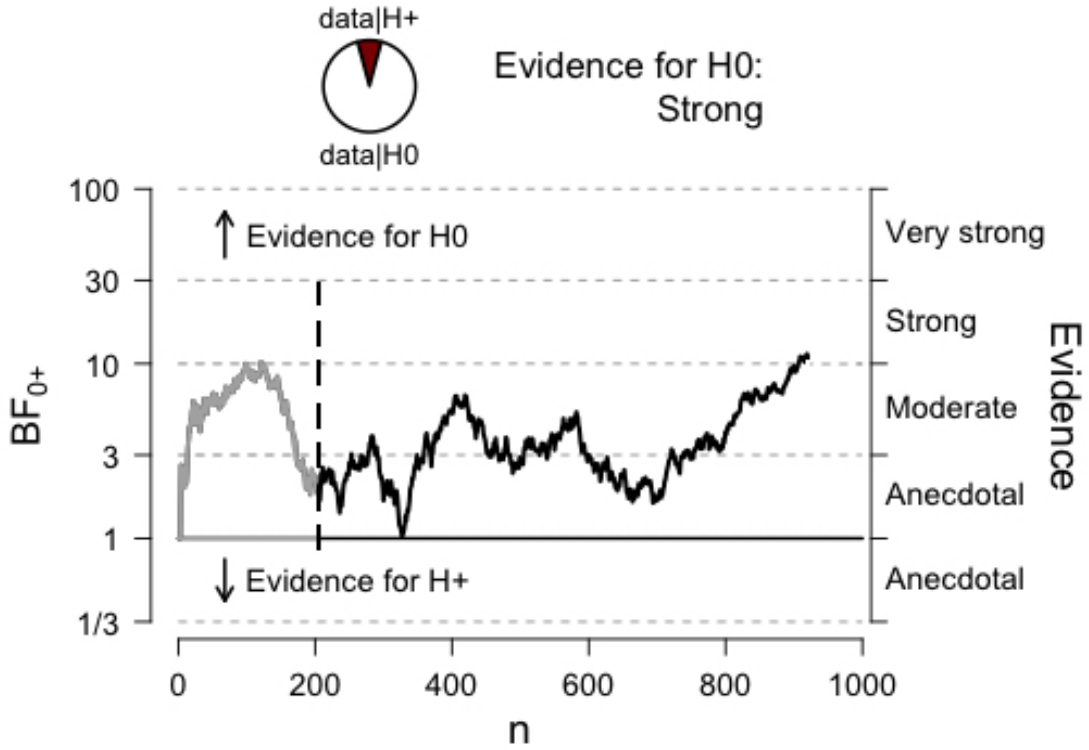


Figure 5. A sequential graph of the activity of the Bayes factor for Experiment 2 as evidence accumulates. The Bayes factor in this analysis demonstrates strong evidence in favor of the null hypothesis, based on the categories defined by Jeffreys (1961). The vertical dotted line indicates the point at 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 creativity scores relative to local priming.

Exploratory Analyses: Experiment 2

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.

Posterior Distribution of Effect Size

Figure 6, like Figure 4 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 $\delta = 0$, 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 $BF_{01} = 10.84$, and when assessed under the criteria of Jeffreys, also qualifies as strong evidence in favor of the null hypothesis.

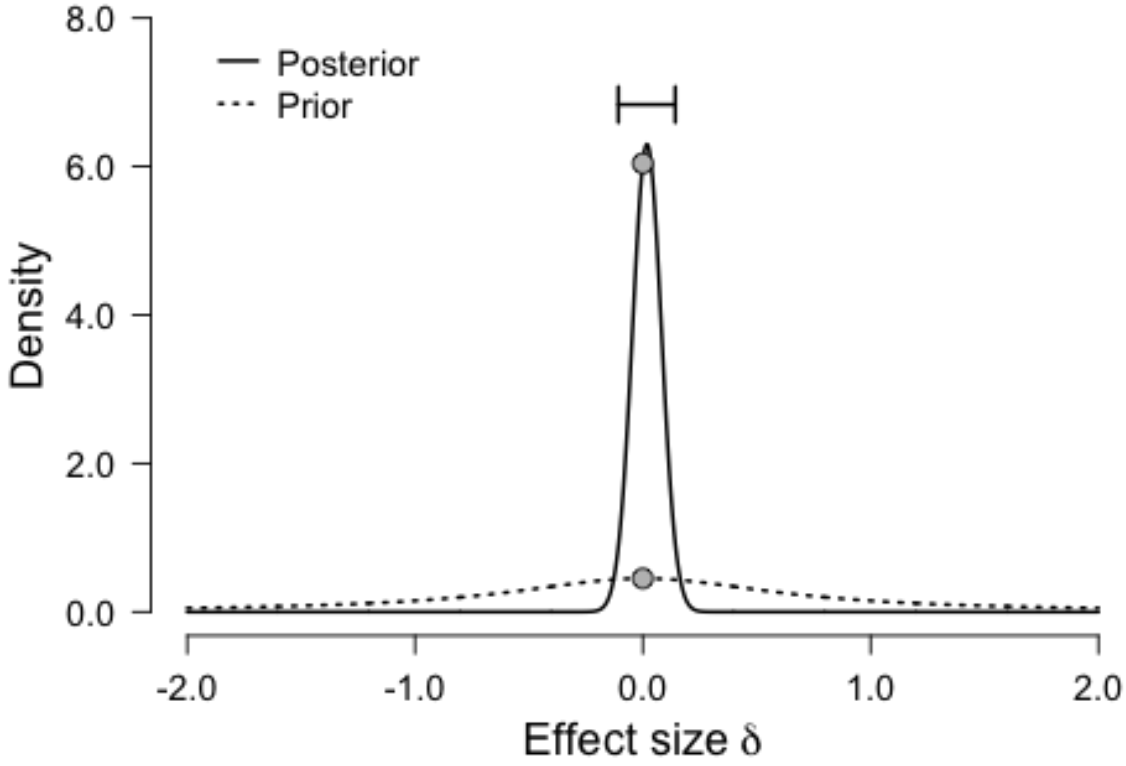


Figure 6. The posterior distribution of effect size δ over a two-sided prior distribution for Experiment 1. The posterior distribution density over an effect size of 0 is 10.35 times greater than the prior distribution indicates.

General Discussion

In this study, we attempted to replicate the results of Experiment 1 as carried out by Förster and Denzler (2012). In 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 translates to anecdotal evidence in favor of the null hypothesis (Jeffreys, 1961). 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. In both of our experiments, our exploratory frequentist analysis failed to produce a significant *p*-value (all *p*s > .05).

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 Buhrmester, Kwang & Gosling, 2011; Bartneck, Duenser, Moltchanova & Zawieska, 2015). Furthermore, any participants with high error rates in the Navon task (potentially participants who did not attend to the task or instructions) were excluded from final analysis. We therefore deem this explanation unlikely to account for the current failure to replicate.

Other potential explanations may be considered plausible for our results, including the question that perhaps the replication methodology might have somehow introduced moderator variables or boundary conditions that have diluted the original

effect found in FD2012. While such judgments are difficult to make in the absence of qualifying evidence, Cesario (2014) suggests that it may be premature to expect to exactly replicate findings when we have yet to first establish the potential moderators and variables that may be interacting with the priming effects. Bargh (2006) argues that the wealth of knowledge that empirical testing has generated far outweighs the field's resources to be able to conceptualize the mechanisms of priming, and understand the conditions under which priming occurs.

One might conversely question whether the effect initially reported in FD2012 is sufficiently robust and valid when tested under conditions of the replication laboratory. Indeed, an effect may offer only limited insight into human cognition if the effect is extremely sensitive to particular experimental conditions, or to small deviations in experimental method. In the frame of the current findings, one might question what contribution this priming effect can make to what we know about the role of global priming on creativity, if it does not generalize beyond the specific and somewhat sterile conditions of the original study laboratory.

If a reported effect is thought to be sensitive to boundary conditions to the extent that it cannot replicate in other labs, these exact conditions should ideally be outlined in the original report, as well as be supported by theoretical rationale where possible. If such claims are not made known a-priori for the consideration of the replicating lab, they carry little defensive weight. Such claims or stipulations were not made in the FD2012 article, and so we assume that the original authors are not aware of potential moderating variables that may explain the results we report.

Another explanation for our results that is commonly discussed in the literature is that the originally reported priming effect was the result of a type-1 error. Recent studies indicate that over a third (36%) of significant research findings are likely to be due to type-1 error (Pashler & Harris, 2012; Simons, 2014). Although theoretically possible, we have no means to assess such an explanation.

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*(5), 556-578.

Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology*, *36*, 147-168.

Bargh, J. A. (2012). Nothing in their heads. *Psychology Today*. Retrieved 15 June, 2012 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. 2015; 10*(4): e0121595

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

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*, 317-352.

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

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 27). 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 on 21 October, 2015.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40-48.

Chambers, C. (2014, November 13). Registered Reports: A step change in scientific publishing. Innovation in Publishing. Retrieved from: http://www.elsevier.-com/reviewers-update/story/innovation-in-publishing/registered-reports-a-step-change-in-scientific-publishing on October 16, 2015.

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.

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

Förster, J. (2014, May 29). Reactions to NRC article, LOWI report, and Science article. [Web blog post]. Retrieved from http://www.socolab.de/main.php?id=66 on 25 August 2014.

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.

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

Klauer, K. C., & Singmann, H. (in press). Does global and local vision have an impact on creative and analytic thought? Two failed replications. *PloS One*.

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*(3), 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.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A.,

Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.7) [Computer software]. Available from https://jasp-stats.org/

LOWI (2012, September 3). Suspicion of scientific misconduct by Dr. Jens Förster. Retrieved from http://retractionwatch.files.wordpress.com/2014/04/report_foerster.pdf on 25 August 2014.

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., (2014). Using the 'BayesFactor' package, version 0.9.2+. Available at http://bayesfactorpcl.r-forge.r-project.org/.

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*(6242), 1422.

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 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.

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

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., 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.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76-80.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*(6), 886-899.

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

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., 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., 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.

Appendix A

Instruction wording for each of the three tasks, derived from FD2012.

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.

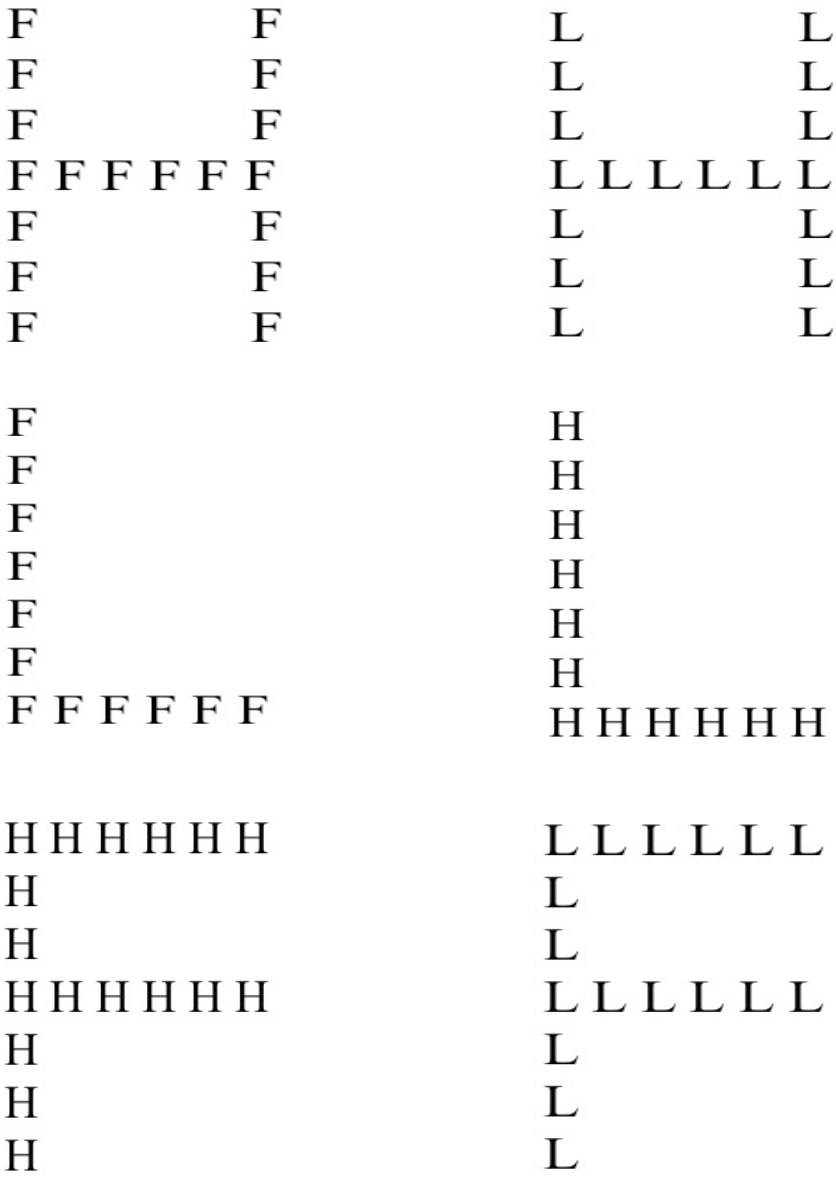
All conditions: BOC instructions
This task is part 4 of 5 in this study. In this task you will be required to rate the typicality of a word to a given category. You may move the sliders below each question to indicate your rating on a scale from 0 to 9, where 0 denotes the lowest typicality rating (i.e., 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 B

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 C

The Positive and Negative Affect Scale as used in FD2012. This questionnaire has been computerized for administration in the proposed 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 D

Word lists for the fringe, good and moderate exemplars for categories of furniture, vehicles, vegetables and sport, taken from Rosch (1975), used in FD2012 as the dependent measure.

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 E

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?"