No conclusive evidence that difficult general knowledge questions cause a "Google Stroop effect". A replication study

Guido Hesselmann [Corresp. 1]

1 Department of General and Biological Psychology, Psychologische Hochschule Berlin (PHB), Berlin, Germany

Corresponding Author: Guido Hesselmann
Email address: g.hesselmann@phb.de

Access to the digital "all-knowing cloud" has become an integral part of our daily lives. It has been suggested that the increasing offloading of information and information processing services to the cloud will alter human cognition and metacognition in the short and long term. A much-cited study published in *Science* in 2011 provided first behavioral evidence for such changes in human cognition. Participants had to answer difficult trivia questions, and subsequently showed longer response times in a variant of the Stroop task with internet-related words ("Google Stroop effect"). The authors of this study concluded that the concept of the Internet is automatically activated in situations where information is missing (e.g., because we might feel the urge to "google" the information). However, the "Google Stroop effect" could not be replicated in two recent replication attempts as part of a large replicability project. After the failed replication was published in 2018, the first author of the original study pointed out some problems with the design of the failed replication. In our study, we therefore aimed to replicate the "Google Stroop effect" with a research design closer to the original experiment. Our results revealed no conclusive evidence in favor of the notion that the concept of the Internet or internet access (via computers or smartphones) is automatically activated when participants are faced with hard trivia questions. We provide recommendations for follow-up research

1 **No conclusive evidence that difficult general**

2 **knowledge questions cause a "Google Stroop effect".**

3 **A replication study**

5 Guido Hesselmann[1]

6 [1] Psychologische Hochschule Berlin (PHB), 10179 Berlin, Germany

9 Corresponding author:

10 Guido Hesselmann

11 Am Köllnischen Park 2

12 10179 Berlin

13 Germany

14 Email: g.hesselmann@phb.de

Abstract

Access to the digital "all-knowing cloud" has become an integral part of our daily lives. It has been suggested that the increasing offloading of information and information processing services to the cloud will alter human cognition and metacognition in the short and long term. A much-cited study published in *Science* in 2011 provided first behavioral evidence for such changes in human cognition. Participants had to answer difficult trivia questions, and subsequently showed longer response times in a variant of the Stroop task with internet-related words ("Google Stroop effect"). The authors of this study concluded that the concept of the Internet is automatically activated in situations where information is missing (e.g., because we might feel the urge to "google" the information). However, the "Google Stroop effect" could not be replicated in two recent replication attempts as part of a large replicability project. After the failed replication was published in 2018, the first author of the original study pointed out some problems with the design of the failed replication. In our study, we therefore aimed to replicate the "Google Stroop effect" with a research design closer to the original experiment. Our results revealed no conclusive evidence in favor of the notion that the concept of the Internet or internet access (via computers or smartphones) is automatically activated when participants are faced with hard trivia questions. We provide recommendations for follow-up research.

Introduction

It seems intuitively plausible that today's ubiquitous 24/7 access to the Internet via smartphones and computers will affect our cognitive functioning and strategies. Specifically, it has been suggested that different types of "cognitive offloading" (i.e., the use of our bodies, objects, and technology to alter the processing requirements of a task to reduce cognitive demand) may alter human cognition and metacognition in the short and long term (Risko & Gilbert, 2016). In the memory domain, one idea is that the Internet is taking the place not just of other humans as external sources of memory ("transactive memory"), but also of our own cognitive faculties (Ward, 2013; Wegner & Ward, 2013). We increasingly offload information to "the cloud", as almost all information today is readily available through a quick Internet search.

44 Evidence for such "Google effects on memory" has been presented in a much-cited[1] and
45 influential[2] paper published in *Science* (Sparrow, Liu & Wegner, 2011). Across four experiments,
46 the authors showed that a) when people expect to have future access to information, they have
47 lower rates of recall of the information itself and enhanced recall instead for where to access it,
48 and b) when faced with difficult general knowledge (or, trivia) questions, people are primed to
49 think about the Internet and computers. The latter effect was demonstrated in one experiment
50 (Exp.1) where participants answered easy or difficult trivia questions, and then completed a
51 variant of the Stroop task (MacLeod, 1991). In a paradigm conceptually similar to the emotional
52 Stroop paradigm (Algom, Chajut & Lev, 2004), participants responded to the ink color of written
53 words, which were either related or unrelated to the Internet. Stroop-like interference from words
54 relating to computers and Internet search engines was increased after participants answered
55 difficult compared with easy questions, consistent with those terms being "primed" in participants'
56 minds (Doyen et al., 2014). Briefly, the results seem to suggest that whenever information is
57 needed and lacking, the concept of the Internet (including computer-related terms) is activated
58 and can interfere with our behavior in subsequent tasks (e.g., because we might feel the urge to
59 "google" the information).

60 In 2018, Camerer and colleagues published a meta-analysis of 21 replications of social science
61 and psychology experiments published in *Science* or *Nature* between 2010 and 2015 (Camerer
62 et al., 2018). This large replicability project included a replication of the "Google Stroop effect" by
63 Sparrow and colleagues (2011). All materials and data from this replication – led by Holzmeister
64 & Camerer - are freely available on OSF (https://osf.io/wmgj9/). In two separate experiments, the
65 authors tested the hypothesis that, after answering a block of hard trivia questions, color-naming
66 reaction times (RTs) are longer for computer-related terms than for general words. Neither
67 experiment showed a significant effect despite adequate statistical power (see
68 https://osf.io/4rfme/ for a short summary of this replication). However, as the original authors of
69 Sparrow et al. (2011) did not provide Holzmeister & Camerer with any materials or feedback on
70 their inquiries, it was difficult to replicate the experimental design of the original study. After the
71 replication had been completed and published, Sparrow noted some design differences compared
72 to the original study (Sparrow, 2018). As Holzmeister & Camerer point out (Camerer et al., 2018),
73 it cannot be ruled out that these design differences, including the manipulation of cognitive load

---

[1] 1296 citations on Google Scholar, as of October 7 2020.
[2] The study is mentioned in a 2019 publication by the EU on harmful internet use:
https://op.europa.eu/en/publication-detail/-/publication/fb2d58ea-8e58-11e9-9369-01aa75ed71a1/language-en/format-PDF/source-127484707

74 (see below), influenced the replication result. Therefore, it was our aim to investigate the "Google

75 Stroop effect" in a further study, based on the original experiment, the materials provided by

76 Holzmeister & Camerer, as well as the critical comments by the original authors.

77

Materials & Methods

Sample

80 This work is based on an undergraduate student research project ("Experimentell-Empirisches

81 Praktikum - ExPra") at the Psychologische Hochschule Berlin (PHB). A total of 117 participants

82 were tested. The sample consisted of students at the Psychologische Hochschule Berlin (PHB),

83 as well as friends and families of the student experimenters (see acknowledgment). All

84 participants provided written informed consent. The experiment was approved by the ethics

85 committee of the PHB (approval number PHB10032019).

86

Paradigm

88 Our version of the experiment was based on two previous studies: Sparrow et al.'s original Exp.1

89 (Sparrow, Liu & Wegner, 2011), and the replication study (Camerer et al., 2018). We incorporated

90 comments provided by the first author of the original study (Sparrow, 2018), published in response

91 to the failed replication[3].

92 The original experiment tested the hypothesis that participants are primed to think about the

93 Internet when faced with difficult trivia questions (e.g., "Did Benjamin Franklin give piano

94 lessons?"). Participants first had to answer a block of either hard or simple question followed by

95 a modified Stroop Task. In this task, Internet-related and neutral words were presented in random

96 order, and participants were instructed to indicate the word's color (blue or red) via button press.

97 RTs to the words were measured as the dependent variable. In order to manipulate cognitive

98 load, a random six-digit number was presented, and participants were instructed to memorize it

99 for delayed retrieval (Figure 1). After their response in the modified Stroop Task, participants were

100 asked to enter the six-digit number. The results from the modified Stroop task showed the

101 predicted pattern: RTs to computer terms (e.g., Google) were longer than RTs to neutral terms

---

[3] Our own request for further details about the experimental design remained unanswered (email from March 2019).

102   (e.g., Target), especially after participants were faced with difficult trivia questions ("question type

103   x word type" interaction; F(1,66) = 5.02, p < .03).

104

105

106   *Figure 1 about here*

107   **Figure 1.** Schematic illustration of the modified Stroop paradigm. Participants first answer a list

108   of 16 easy or hard trivia questions, followed by a block of color-naming trials (modified Stroop

109   task). Before each Stroop block, or before each Stroop trial, a random six-digit number is

110   presented for later retrieval (i.e., delayed retrieval at the end of a block of several Stroop trials, or

111   at the end of each Stroop trial).

112

113   Although Sparrow et al. (2011) reported that they used eight target words related to computers

114   and search engines, and 16 unrelated words, the total number of trials presented in each "hard

115   question" and "easy question" block remained unclear, and also whether words were repeated.

116   In their replication, Holzmeister & Camerer used the 24 words originally reported by Sparrow et

117   al. (2011), and decided to run 48 trials per block, resulting in the repetition of words (i.e., each

118   participant saw each word four times, twice in the "hard question" block, and twice in the "easy

119   question" block.) In her response to the failed replication (Sparrow, 2018), Sparrow then strongly

120   argues against the repetition of words, and reported the full set of 16 Internet-related words used

121   in the original study.

122   The cognitive load manipulation is described as follows in the original study: "Participants are

123   presented with words in either blue or red, and were asked to press a key corresponding with the

124   correct color. At the same time, they were to hold a 6 digit number in memory, creating cognitive

125   load" (Sparrow et al., 2011, supplement). In their replication, Holzmeister & Camerer decided to

126   manipulate cognitive load by presenting a six-digit number before each Stroop task block. After

127   the block (involving several trials), participants were asked to enter the memorized number. In her

128   response, Sparrow strongly argues against this block-wise procedure and provided more detail

129   about the original study: "[…] before each word, they were shown a different six digit number,

130   which was reported back by them after each Stroop word".

131   Finally, Sparrow mentions that the original experiment was run in 2006, which is why she believes

132   that the computer terms used originally are obsolete and should not have been used in a

133   replication (Sparrow, 2018). In her response, she writes that she "would focus primarily on target
134   words about phones (as they seem most ubiquitous), but would also be sure to pre-test many
135   possible words to ensure their contextual relevance before putting them into the modified Stroop".

136   For our replication study, we used the material provided by Holzmeister & Camerer on the open
137   science platform OSF (https://osf.io/wmgj9/), and revised it according to Sparrow's comments. All
138   materials and data are available at OSF (https://osf.io/cjgea/). The experiment was implemented
139   using oTree (Chen, Schonger & Wickens, 2016). We translated the trivia questions and Stroop
140   words into the German language, and adjusted some questions to the German context (e.g., "Was
141   Cat in the Hat written by J.D. Salinger?"). The experiment consisted of two blocks of 16 either
142   hard or easy questions, followed by 24 Stroop words of which eight were Internet-related (e.g.,
143   WLAN, Google, Website) and 16 were unrelated (e.g., frame, bottle, bamboo). The words were
144   pre-tested for contextual relevance (see below), and randomly assigned to the easy and hard
145   question condition. Words were presented in random order, and word color (blue or red) was
146   randomly chosen on each trial. Participants were instructed to indicate the word color as quickly
147   and as accurately as possible via button press ("e" for blue and "i" for red, using the index fingers
148   of both hands). Participants were asked to place their fingers on the keys of the computer
149   keyboard before the start of the Stroop task. Before the presentation of each Stroop word,
150   participants had to memorize a random six-digit number for delayed retrieval after their response
151   in the color-naming task.

152   After the main experiment, participants provided information about their age, level of education,
153   color blindness, and what they thought the purpose of the experiment was. Finally, participants
154   were asked what they would normally do when faced with a hard general knowledge question in
155   their daily life: a) look up the answer on the Internet, b) ask someone who might know the answer,
156   c) leave it at that. Including instructions, debriefing and signing the informed consent forms, the
157   experiment lasted approximately 45 minutes in total.

158

159   **Word stimuli**

160   In the original study, the color naming task contained "8 target words related to computers and
161   search engines (e.g., Google, Yahoo, screen, browser, modem, keys, Internet, computer), and
162   16 unrelated words (e.g., Target, Nike, Coca Cola, Yoplait, table, telephone, book, hammer, nails,
163   chair, piano, pencil, paper, eraser, laser, television)" (supplement, p.2). According to Sparrow
164   (2018), the computer-related target words were selected from a larger set of the following 16

165 words: "Google, Yahoo, mouse, keys, Internet, browser, computer, screen, Altavista, Wikipedia,
166 disk, Lycos, Netscape, modem, router, online" (p.1). The full list of words included in the set of
167 general terms is not provided. Thus, at least in the case of computer terms, the number of target
168 words used in the experiment was smaller than the number of available target words (eight and
169 16, respectively). It remains unclear how the target words were chosen for each participant. What
170 seems to be clear is that the words Target, Nike, Google, and Yahoo were chosen for each
171 participant, because a within-subject comparison was calculated using this subset of words (see
172 below). Furthermore, the authors of the original study write that the sets "were matched for
173 frequency to the target words (11)" (supplement, p.2). Without the full word sets, this claim is hard
174 to evaluate. A brief search in the referenced word corpus (reference 11) revealed no hit for the
175 computer terms "Altavista" and "Google" (Nelson, McEvoy & Schreiber, 2004).

176 Based on Sparrow's reply to the first replication (Sparrow, 2018), we prepared two new sets of
177 German words, one set with general terms (32 words), and one set with computer terms (16
178 words). As suggested by Sparrow (2018), we validated the contextual relevance of these words.
179 31 naïve participants (mostly undergraduate students who did not participate in the main
180 experiment) took part in an online survey (https://www.surveymonkey.de/). For each of the 48
181 words, participants rated the contextual relevance (i.e., "Internet-relatedness") on a 5-point scale
182 (statement: "I think of the Internet when reading this word"; Rating: 1 = Strongly agree; 2 = Agree;
183 3 = Neutral; 4 = Disagree; 5 = Strongly disagree). For computer words, the mean rating (i.e., the
184 mean rating across the median ratings per participant) turned out to be 1.06, while it was 4.94 for
185 the neutral terms.

186 Our list of computer (or, Internet-related) terms contained: website, data volume, email, search
187 engine, Wikipedia, WLAN, app, Google, smartphone, hotspot, online, blog, Spotify, Firefox,
188 Whatsapp, Chrome. The list of general (or, unrelated) terms contained words like nail or car, but
189 also names of grocery stores (all words in German; full list on OSF). After completion of the study,
190 we used the online DlexDB database to estimate word frequency in the two sets
191 (www.dlexdb.de/query/kern/typposlem/). The median absolute type frequency (corpus frequency)
192 for the general terms was 374, while it was 23.5 for the computer terms. However, half of the
193 words in the list of computer terms were not part of the data base (Wikipedia, WLAN, Google,
194 smartphone, blog, Spotify, Firefox, Whatsapp). To what degree this difference in word frequency
195 between the two sets could be problematic for the current experiment, is hard to say. Under the
196 assumption that participants read the words when responding to their color, one would expect
197 longer RTs for words in the set of computer terms due to the lower word frequency (Larsen,
198 Mercer & Balota, 2006). The main effect of "word type" reported by Sparrow et al. (2011) could

199 parsimoniously be explained by differences in word frequency, if word sets were not matched.
200 However, differences in word frequency between the two sets do not seem to preclude the
201 proposed "question type x word type" interaction, which is driven by priming (according to the
202 dual-route model). In a control analysis, we did not find evidence for an effect of word frequency
203 on RTs in our data set[4].

204 As Sparrow (2018) in her reply to the replication strongly argues against presenting each target
205 word more than once (referring to "active thought suppression"), we decided to randomly select
206 words for the easy and hard blocks. From our pool of 48 words, eight computer terms and 16
207 general terms were presented in each block, so that each word was presented only once in the
208 experiment. The words were randomly chosen for each participant. Of note, each word was
209 presented twice in the original study, once in the easy block, and once in the hard block. Since
210 we did not intend to select single words for post hoc pairwise tests, we did consider a strict "no
211 word repetition" approach to be more in line with Sparrow's (2018) suggestions. Table 1
212 summarizes the known differences between the original study and the replication studies.

213

214 *Table 1 about here*

215

216 **Table 1.** Differences between the original study by Sparrow et al. (2011) and the replication studies. (*)
217 Original sample size is based on the supplementary information. (**) Total number of trials is an informed
218 guess based on the original study and the response to the failed replication (Sparrow, 2018).

219

220 **Data preprocessing**

221 We preregistered our analyses including confirmatory and exploratory statistical tests
222 (https://aspredicted.org/z3xt4.pdf), and later decided to restrict the analysis to confirmatory tests.
223 All data are available on OSF (https://osf.io/cjgea/). Data were preprocessed and analyzed using
224 R version 3.3.2 (www.r-project.org), and RStudio version 1.0.136 (www.rstudio.com). The original
225 "csv" data files were exported from oTree, imported into R, converted into the long format and
226 merged using custom R scripts. Each student experimenter contributed one "csv" file containing
227 the data from multiple participants. The resulting data file in the long format thus contains blocks

---

[4] Using a Bayes Factor approach (https://richarddmorey.github.io/BayesFactor/), we found that a model not containing word frequency as predictor was preferred to a model containing word frequency by a factor of 16. Details can be found in the R analysis script on OSF.

228   of data from different student experimenters, because we did not sort the data according to
229   recording time and date.

230   The age distribution of participants turned out to be heavily skewed (mean age: 31 years; range:
231   18 – 73). This was because the participant sample included students' families and friends, and
232   we originally did not set a maximum age. Since the sample in the original paper (Sparrow, Liu &
233   Wegner, 2011) consisted of undergraduate students (supplement, p.2), we decided to deviate
234   from our preregistered protocol, and set the maximum age as the 75% quantile of the age
235   distribution (44 years). In the remaining sample of 89 participants, the mean age was 24 years.

236   A second deviation from the preregistered protocol was due to the fact that we originally assumed
237   that trials with incorrect responses (i.e., when participants pressed the key for the wrong color)
238   should be excluded from the analysis. According to the replication report (https://osf.io/84fyw/),
239   which includes personal communication with the original authors, this was not the case in the
240   original study. We therefore included trials with correct and incorrect responses in the color-
241   naming task. (Note that trials with incorrect responses in the memory task were not excluded,
242   either.)

243   Although not specified in the original study, and neither in the replication study, we preregistered
244   an additional exclusion of trials with RT outliers and anticipatory responses (< 0.1s). We defined
245   RT outliers using the interquartile range (IQR) method (Tukey, 1977), separately for each
246   participant. These criteria resulted in the exclusion of 7 ± 3% of trials (mean percentage ± standard
247   deviation). For our final analysis, we did not exclude RT outliers, but also report the results of the
248   analysis including the exclusion of RT outliers.

249

250   **Data analysis**

251   For each participant, we calculated the performance in the easy and hard questionnaires, the
252   performance in the number memory task, and the mean RT in each of the four conditions. The
253   condition averages from all participants were then exported into JASP 0.10.2 (https://jasp-
254   stats.org/) for frequentist and Bayes Factor (BF) analysis, using default Cauchy priors (scale
255   0.707). To test the predicted "question type x word type" interaction, we calculated the following
256   difference and tested it against zero using a two-sided paired test: $[RT(computer) -$
257    $RT(general)]_{hard} - [RT(computer) - RT(general)]_{easy}$. Performance in the number memory task was
258   calculated for all trials (i.e., including trials with RT outliers in the color-naming task).

259

Results

261 Participants found the easy questions to be answerable (97 ± 4%, mean accuracy ± standard
262 deviation), but had difficulty finding the correct answers to the hard questions (60 ± 11%). As in
263 the original study (98% versus 47%), this difference was significant (t_{88} = 48.31, p < .001). Mean
264 accuracies in the color-naming task and number memory task were high (98 ± 3% and 81 ± 16%,
265 respectively).

266 Figure 2A plots the RT data from the color-naming task. The dual-route model predicts that
267 computer terms create more interference, and thus are associated with longer RTs than general
268 terms, in particular in hard question blocks. The results show that in easy question blocks, mean
269 RT was 830 ± 366 ms for general terms, and 901 ± 620 ms for computer terms (mean ± standard
270 deviation). In hard question blocks, mean RT was 825 ± 453 ms for general terms, and 821 ±
271 379 ms for computer terms. Thus, the RT data do not show the predicted pattern. Accordingly,
272 the sequential Bayes Factor (BF) analysis yielded a final BF_{01} of 5.07 in favor of the null over the
273 alternative hypothesis, after the data from N=89 participants were taken into account (Figure 2B).
274 The observed data are thus 5 times more likely under the null model than under the alternative
275 model (i.e., the "question type x word type" interaction). When the specific data pattern reported
276 in the original study was considered as alternative model in an exploratory analysis, our data were
277 16 times more likely under the null model (BF_{0+} = 16.43). The preregistered two-sided one-sample
278 t-test was not significant (t_{88} = -1.04, p = .301).

279 The pattern of results remained the same when incorrect responses in the color-naming task were
280 excluded. When RT outliers were excluded from data analysis, overall mean RTs decreased, but
281 the predicted interaction could still not be observed (BF_{01} = 3.74; t_{88} = 1.31, p = .194). In easy
282 question blocks, mean RT was 737 ± 296ms for general terms, and 723 ± 295ms for computer
283 terms (mean ± standard deviation). In hard question blocks, mean RT was 707 ± 268ms for
284 general terms, and 712 ± 261ms for computer terms (data not plotted in a figure).

285 In the debriefing after the main experiment, 11 out of 89 participants did not provide a response
286 (e.g., closed the browser before answering). 73% (57/78) of all responders said that they would
287 look up the answer on the Internet when faced with a hard general knowledge question in their
288 daily life; 18% (14/78) said that they would ask someone who might know the answer, and 9%
289 (7/78) responded that they would "leave it at that". When we restricted the RT data analysis to
290 participants consulting the Internet (N=57), the pattern of results was very similar to the one

291 reported above (BF_{01} = 5.01). Finally, when all participants (N=117) were included in the data
292 analysis, the pattern of results turned out to be 9 times more likely under the null model than under
293 the alternative model (BF_{01} = 9.31).

294

295 *Figure 2 about here*

296

297 **Figure 2.** Main results (N=89). (A) Reaction time (RT) data from the color-naming task. Grey
298 squares represent the mean RT in the "easy question" blocks, white squares represent the mean
299 RT in the "hard question" blocks. Errors bars represent the standard error of the mean. (B)
300 Sequential Bayes Factor (BF) analysis of the "question type x word type" interaction. User prior
301 refers to a Cauchy prior with scale 0.707, wide prior to a Cauchy prior with scale 1, and ultrawide
302 prior to a Cauchy prior with scale √2.

303

304 **Statistical issues in the original study**

305 During our work on this project, we noticed three statistical issues in the original paper by Sparrow
306 et al. (2011). In the following paragraph, we address these observations in turn.

307 First, it remains unclear whether N=69 or N=46 participants were tested in Experiment 1 of the
308 original paper (Sparrow, Liu & Wegner, 2011). In the supporting online material[5], the authors write
309 that "Forty-six undergraduate students (28 female, 18 male) at Harvard University were tested in
310 a within subjects experiment" (p.2). In their 2x2 within-subject design (easy/hard questions;
311 computer/general words), the correct degrees of freedom (df) would then be 45 for paired t-tests,
312 as well as for main effects, the interaction and simple main effects in the rm-ANOVA. Under the
313 assumption that N=69 participants were tested, the correct df would be 68. In the main text and
314 supplement of the original paper, the reported df is 68 for paired t-tests, and 66 for the rm-ANOVA.
315 According to the authors of the replication, the original authors initially confirmed that the sample
316 size was 46 participants, and that dfs were misreported in the paper, but after the publication of
317 the replication the original authors pointed out that the number of participants was 69

---

[5] https://science.sciencemag.org/content/suppl/2011/07/13/science.1207745.DC1

318   (https://osf.io/84fyw/). While the reported t-, F-, and p-values appear to be in better agreement

319   with the N=69 scenario[6], the reported dfs are incorrect in both scenarios.

320   Second, two computer-related words (Google/Yahoo), and two unrelated words (Target/Nike)

321   were selected for analysis in the original paper. In the main text, RT data are reported only for

322   this subset, together with the "question type x word type" interaction ($F(1,66) = 5.02$, $p < .03$). The

323   reasoning behind this post hoc selection (i.e., selection of four words out of a pool of Stroop

324   words) remains unclear. Without further information, it remains possible that the choice was made

325   after seeing the data. Such post hoc choices and data-contingent analyses are misleading when

326   they are not presented as exploratory analysis. The potential impact of this post hoc selection

327   should not be underestimated. As pointed out by Sparrow in a reply to the failed replication, "each

328   Stroop word was seen only once by participants" (Sparrow, 2018). Hence, the RTs for

329   Google/Yahoo and Target/Nike reported in the original paper are based on single trials per

330   participant. Given such noisy measurements with low precision (Smith & Little, 2018), the authors

331   might have capitalized on chance by selecting a subset of words for their main analysis.

332   Third, two paired t-tests on word type are reported for the complete data set ("hard questions"

333   condition, computer words versus general words: $t(68) = 3.26$, $p < .003$; "easy questions"

334   condition, $t(68) = 2.98$, $p < .005$). The "question type x word type" interaction is reported only for

335   the 4-word-subset ($F(1,66) = 5.02$, $p < .03$). Based on visual inspection of the reported average

336   RTs, it seems plausible that the "question type x word type" interaction should be smaller for the

337   complete data set (Figure 3A) than for the 4-word-subset (Figure 3B). Sparrow et al. (2011) report

338   the following result in the supplement: "Taking out the 4 terms (Google/Yahoo and Target/Nike)

339   which yielded an interaction with easy/hard questions ($F(1,66) = 5.52$ [sic], $p < .03$), the interaction

340   between computer and general terms and easy/hard questions *remains* significant $F(1,66) = 9.49$,

341    $p < .004$" (p.3, italics added). In fact, the interaction for the reduced data set (i.e., all data minus

342   the 4-word-subset) turned out to be *larger* than the interaction reported for the subset (F-values

343   9.49 and 5.52, respectively). This result seems difficult to reconcile with the reported data, but

344   access to the original data would be necessary to clarify this point.

345

346   *Figure 3 about here*

347

---

[6] In the main text of Sparrow et al. (2011), the simple effect is reported as follows: $F(1,66) = 4.44$, $p < .04$ (exact p-value: .03891. For N=46, the result would be: $F(1,45) = 4.44$, $p = .040712$, and thus $p > .04$.

348 **Figure 3.** Reaction time (RT) data from Exp.1 in Sparrow et al. (2011). A) RTs for the complete
349 data set (16 general words, 8 computer-related words). B) RTs for the 4-word-subset (2 general
350 words: Target and Nike; 2 computer-related words: Google and Yahoo). Grey squares represent
351 the mean RT in the "easy question" condition, white squares represent the mean RT in the "hard
352 question" condition. Errors bars represent the standard error of the mean (for N=69).

353

Discussion

355 Although the majority of participants in our study reported that they would normally look up the
356 answer to hard general knowledge questions on the Internet, we did not find evidence for the
357 "Google Stroop effect", as originally published by Sparrow et al. (2011). Thus, our data are more
358 in line with the results from Camerer et al. (2018) who failed to replicate the original effect in two
359 independent experiments. Importantly, the design of our study considered the suggestions for
360 improvement provided by Sparrow (2018) in response to the failed replications. Based on her
361 commentary, we carefully updated and validated the computer-related terms, strictly avoided
362 word repetitions, and manipulated cognitive load as in the original study.
363 It seems worthwhile to take a closer look at the hypothetical cognitive model underlying the
364 "Google Stroop effect". Sparrow et al. (2011) describe their Exp.1 as a modified Stroop task, as
365 follows: "People who have been disposed to think about a certain topic typically show slowed
366 reaction times (RTs) for naming the color of the word when the word itself is of interest and is
367 more accessible, because the word captures attention and interferes with the fastest possible
368 color naming" (p.776). Classic and modified Stroop tasks, such as the emotional Stroop task,
369 have been discussed in much detail elsewhere (MacLeod, 1991; Algom, Chajut & Lev, 2004). In
370 Exp.1 by Sparrow et al. (2011), there are two colors (red, blue), and two types of words: general
371 terms (e.g., sport) and computer terms (e.g., Google). The authors propose a priming mechanism
372 that specifically affects response times in trials with computer terms: "not knowing the answer to
373 general knowledge questions primes the need to search for the answer, and subsequently
374 computer interference is particularly acute" (p.776).
375 According to this model, computer terms become more accessible when the concept of
376 knowledge is activated, or when information necessary for answering a trivia question is lacking,
377 which results in more interference. We think that at least two points need to be made about this
378 model. First, what remains unspecified is the duration of the proposed priming effect (i.e., for how
379 long the computer terms are more accessible than general terms due following the activation of
380 concepts). Follow-up studies could focus more on the temporal dynamics of the proposed priming

381 effect. Perceptual and semantic priming effects are typically short-lasting, within the range of
382 hundreds of milliseconds, while priming effects from social psychology (e.g., the "Florida effect")
383 are substantially longer lasting, but have turned out to be not robust (Harris et al., 2013; Doyen et
384 al., 2014). Second, the exact role of cognitive load (i.e., working memory load) for the "Google
385 Stroop effect" remains somewhat unclear. In her response to the failed replications, Sparrow
386 (2018) links the working memory task to the paradigm of active thought suppression: "when
387 people are asked explicitly not to think about a single target word, they must engage in active
388 suppression" (p.1). Sparrow (2018) cites an earlier study (Wegner & Erber, 1992) in which
389 participants performed a color-naming task similar to Exp.1 from Sparrow et al. (2011). In this
390 experiment, strongest Stroop-like interference was observed when participants were suppressing
391 a specific target word under cognitive load, and when they were asked to name the color of this
392 target word. In contrast to the "Google Stroop effect", however, participants were asked to actively
393 suppress a single word, so that the importance of high cognitive load in one task might not tell us
394 much about the role of high cognitive load in the other task. Alternatively, it could be argued the
395 word's meaning (in contrast to its color) acts as distracting information in the color-naming task,
396 and that high working memory load increases distractor processing (Lavie, 2005). Therefore, the
397 manipulation of cognitive load might be crucial for the processing along the word reading pathway,
398 and thus for the emergence of the "Google Stroop effect".

399 Our data, however, do not support the notion that the concept of the Internet (together with
400 computer-related terms) becomes automatically activated when participants need to answer
401 difficult general knowledge questions. We are aware that our study does not provide a definite
402 answer, and the potential effects of "cognitive offloading" on human cognition (Risko & Gilbert,
403 2016) are definitely worth further attention and investigation. As mentioned above, the original
404 study by Sparrow et al. (2011) avoided word repetitions, so that the RT data from each participant
405 were based on single presentations of target words. Given such noisy RT measurements, in
406 combination with the suboptimal analysis of variance on the sample mean (Whelan, 2008), post
407 hoc selection of target words can easily lead to the wrong impression that an effect exists (e.g.,
408 we might find a "Wikipedia Stroop effect" or "Firefox Stroop effect" in our data set). Therefore, we
409 recommend using the complete data set in future studies when testing for the crucial "question
410 type x word type" interaction, preferably with linear mixed-effects models that can better account
411 for stimulus-driven variability in RTs than repeated measures ANOVA (Baayen, Davidson &
412 Bates, 2008). Fitting more complex linear models, e.g., models with random slopes, would also
413 require more data than in the present experimental design in which a limited set of words is
414 presented only once per participant (Meteyard & Davies, 2020).

415

Conclusion

416

417 Our results revealed no evidence in favor of the notion that the concept of the Internet or internet
418 access (via computers or smartphones) becomes automatically activated whenever participants
419 are faced with hard trivia questions. Thus, the "Google Stroop effect" might be much smaller than
420 previously thought, and less robust to variations in the experimental design. What else have we
421 learned from this second replication of the "Google Stroop effect"? Our work on this project
422 revealed that the original research design might not be a good starting point to further investigate
423 this effect. To study the effects of the digital "all-knowing cloud" on our cognition, research designs
424 with more precision and power are clearly necessary. Finally, as has been so adequately pointed
425 out by Holzmeister & Camerer, this set of replications "illustrates the importance of open access
426 to all of the materials of published studies for conducting direct replications and accumulating
427 scientific knowledge" (Camerer et al., 2018).

428

429

Acknowledgments

430

431 This work is based on an undergraduate student research project ("Experimentell-Empirisches
432 Praktikum - ExPra") at the Psychologische Hochschule Berlin (PHB) in summer 2019. The
433 following students took part in this project and collected the data: Franzceska Cubela, Luisa
434 Engel, Isabella Ermel, Eran Fortus, Lena Jaeschke, Sarah Koch, Morgane Kroeger, Nikita Kühn,
435 Fabio Kunze, Maximilian Kurtz, Clara Kursawe, Can-Pascal Scheftlein, Alexandra Seidel, and
436 Nora Teichmann. We thank Leona Hammelrath for her invaluable support in helping us set up the
437 oTree experiment and the R scripts, and Marcus Rothkirch for his helpful comments on an earlier
438 version of the manuscript.

439

440

441

References

442

443   Algom D, Chajut E, Lev S. 2004. A rational look at the emotional stroop phenomenon: a generic

444   slowdown, not a stroop effect. *Journal of Experimental Psychology. General* 133:323–

445   338. DOI: 10.1037/0096-3445.133.3.323.

446   Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed random

447   effects for subjects and items. *Journal of Memory and Language* 59:390–412. DOI:

448   10.1016/j.jml.2007.12.005.

449   Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, Kirchler M, Nave G,

450   Nosek BA, Pfeiffer T, Altmejd A, Buttrick N, Chan T, Chen Y, Forsell E, Gampa A,

451   Heikensten E, Hummer L, Imai T, Isaksson S, Manfredi D, Rose J, Wagenmakers E-J,

452   Wu H. 2018. Evaluating the replicability of social science experiments in Nature and

453   Science between 2010 and 2015. *Nature Human Behaviour* 2:637–644. DOI:

454   10.1038/s41562-018-0399-z.

455   Chen DL, Schonger M, Wickens C. 2016. oTree—An open-source platform for laboratory,

456   online, and field experiments. *Journal of Behavioral and Experimental Finance* 9:88–97.

457   DOI: 10.1016/j.jbef.2015.12.001.

458   Doyen S, Klein O, Simons DJ, Cleeremans A. 2014. On the Other Side of the Mirror: Priming in

459   Cognitive and Social Psychology. *Social Cognition* 32:12–32.

460   Grange JA, Houghton G. 2014. Models of cognitive control in task switching. In: *Task switching*

461   *and cognitive control*. New York, NY, US: Oxford University Press, 160–199. DOI:

462   10.1093/acprof:osobl/9780199921959.003.0008.

463   Harris CR, Coburn N, Rohrer D, Pashler H. 2013. Two Failures to Replicate High-Performance-

464   Goal Priming Effects. *PLOS One* 8. DOI: UNSP e72467 10.1371/journal.pone.0072467.

465   Larsen RJ, Mercer KA, Balota DA. 2006. Lexical characteristics of words used in emotional

466   Stroop experiments. *Emotion (Washington, D.C.)* 6:62–72. DOI: 10.1037/1528-

467   3542.6.1.62.

468   Lavie N. 2005. Distracted and confused?: Selective attention under load. *Trends in Cognitive*

469   *Sciences* 9:75–82. DOI: 10.1016/j.tics.2004.12.004.

470   MacLeod CM. 1991. Half a century of research on the Stroop effect: An integrative review.

471   *Psychological Bulletin* 109:163–203.

472   Meteyard L, Davies RAI. 2020. Best practice guidance for linear mixed-effects models in

473   psychological science. *Journal of Memory and Language* 112:104092. DOI:

474   10.1016/j.jml.2020.104092.

475   Nelson DL, McEvoy CL, Schreiber TA. 2004. The University of South Florida free association,

476   rhyme, and word fragment norms. *Behavior Research Methods, Instruments, &*

477   *Computers* 36:402–407. DOI: 10.3758/BF03195588.

478   Risko EF, Gilbert SJ. 2016. Cognitive Offloading. *Trends in Cognitive Sciences* 20:676–688.

479   DOI: 10.1016/j.tics.2016.07.002.

480   Smith PL, Little DR. 2018. Small is beautiful: In defense of the small-N design. *Psychonomic*

481   *Bulletin & Review* 25:2083–2101. DOI: 10.3758/s13423-018-1451-8.

482   Sparrow B. 2018. The importance of contextual relevance. *Nature Human Behaviour* 2:607.

483   DOI: 10.1038/s41562-018-0411-7.

484   Sparrow B, Liu J, Wegner DM. 2011. Google effects on memory: cognitive consequences of

485   having information at our fingertips. *Science (New York, N.Y.)* 333:776–778. DOI:

486   10.1126/science.1207745.

487   Tukey JW. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

488   Ward AF. 2013. Supernormal: How the Internet is changing our memories and our minds.

489   *Psychological Inquiry* 24:341–348. DOI: 10.1080/1047840X.2013.850148.

490   Wegner DM, Erber R. 1992. The hyperaccessibility of suppressed thoughts. *Journal of*

491   *Personality and Social Psychology* 63:903–912. DOI: 10.1037/0022-3514.63.6.903.

492   Wegner DM, Ward AF. 2013. How Google is changing your brain. *Scientific American* 309:58–

493   61. DOI: 10.1038/scientificamerican1213-58.

494   Whelan R. 2008. Effective analysis of reaction time data. *The Psychological record* 58:475–482.

495

Figure 1

Schematic illustration of the modified Stroop paradigm.

Participants first answer a list of 16 easy or hard trivia questions, followed by a block of color-naming trials (modified Stroop task). Before each Stroop block, or before each Stroop trial, a random six-digit number is presented for later retrieval (i.e., delayed retrieval at the end of a block of several Stroop trials, or at the end of each Stroop trial).
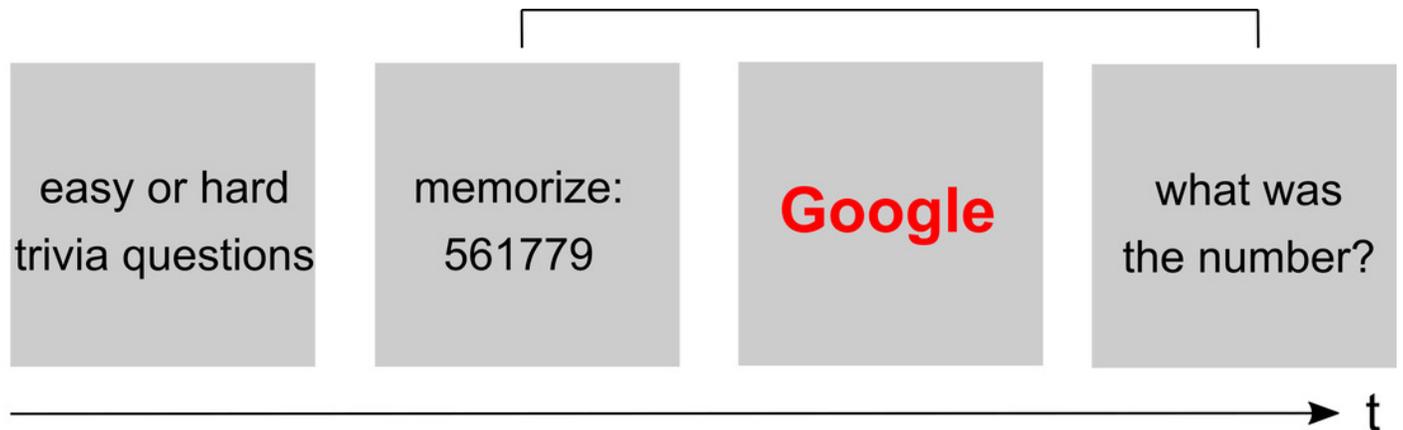
Figure 2

Main results (N=89)

(A) Reaction time (RT) data from the color-naming task. Grey squares represent the mean RT in the "easy question" blocks, white squares represent the mean RT in the "hard question" blocks. Errors bars represent the standard error of the mean. (B) Sequential Bayes Factor (BF) analysis of the "question type x word type" interaction. User prior refers to a Cauchy prior with scale 0.707, wide prior to a Cauchy prior with scale 1, and ultrawide prior to a Cauchy prior with scale √2.
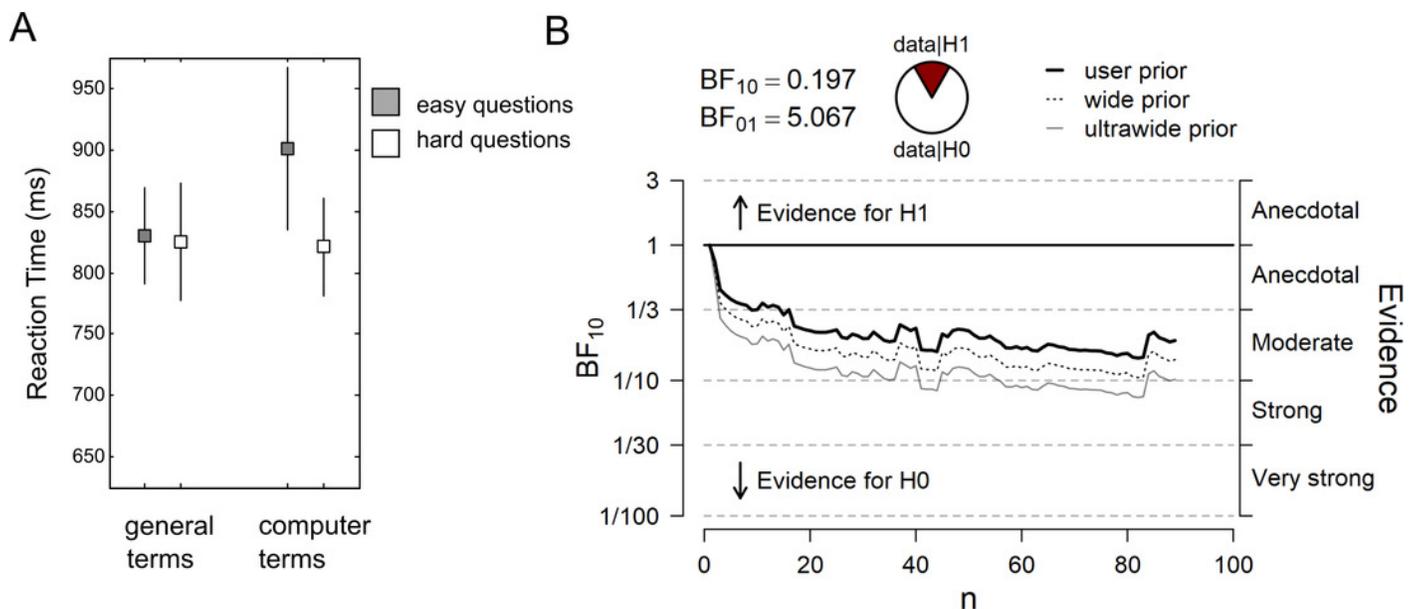
Figure 3

Reaction time (RT) data from Exp.1 in Sparrow et al. (2011)

A) RTs for the complete data set (16 general words, 8 computer-related words). B) RTs for the 4-word-subset (2 general words: Target and Nike; 2 computer-related words: Google and Yahoo). Grey squares represent the mean RT in the "easy question" condition, white squares represent the mean RT in the "hard question" condition. Errors bars represent the standard error of the mean (for N=69).
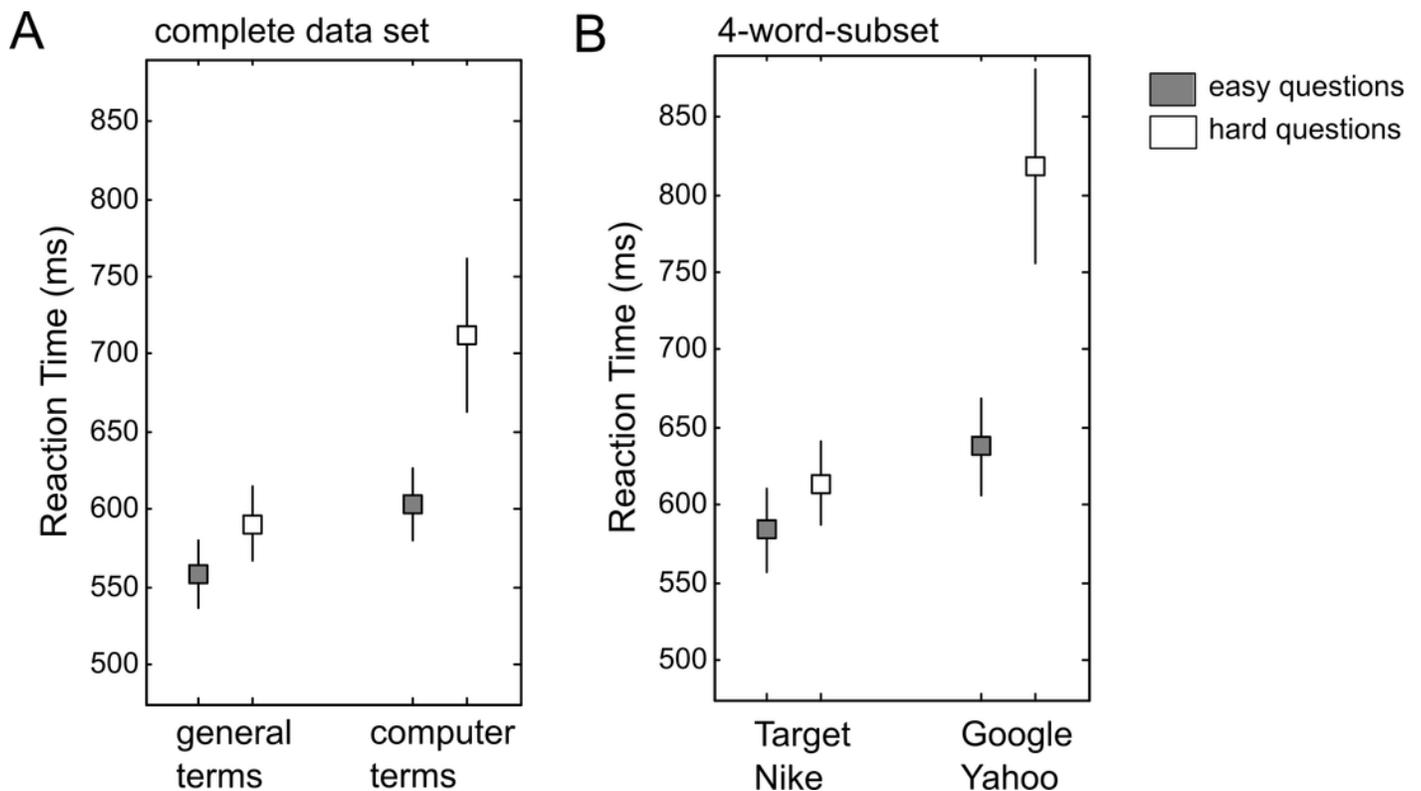
Table 1

Differences between the original study by Sparrow et al. (2011) and the replication studies. (*) Original sample size is based on the supplementary information. (**) Total number of trials is an informed guess based on the original study and the response to the failed replication (Sparrow, 2018).

1 **Table 1.** Differences between the original study by Sparrow et al. (2011) and the replication studies. (*)

2 Original sample size is based on the supplementary information. (**) Total number of trials is an informed

3 guess based on the original study and the response to the failed replication (Sparrow, 2018).

	Sparrow et al. (2011), Exp.1	Camerer et al. (2018)	Hesselmann (2020)
Data collection	2006	2017	2019
Presentation software	DirectRT	oTree	oTree
Participant sample	Undergraduate students (USA)	Students and non-students (USA)	Students and non-students (Germany)
Sample size	N=46*	N=104 & N=130	N=117
Trivia questions	Original set (16 easy, 16 hard)	Original set (16 easy, 16 hard)	Revised & translated set (16 easy, 16 hard)
Stroop words	Original set (incl. 16 internet words)	Original subset (incl. 8 internet words)	Revised, translated & validated set (incl. 16 internet words)
Number memory task	On each Stroop trial	1x per block of trials (easy/hard)	On each Stroop trial
Number of presentations per Stroop word	1x per block of trials (easy/hard)**	2x per block of trials (easy/hard)	1x per experiment
Total number of trials	48 (24 easy, 24 hard)**	96 (48 easy, 48 hard)	48 (24 easy, 24 hard)
Participant debriefing	Unknown	Unknown	Yes