Laypeople Can Predict Which Social Science Studies Replicate: Online Supplemental Material

In this online supplement of the manuscript "Laypeople Can Predict Which Social Science Studies Replicate" we provide details about the Bayesian reanalyses of the original findings, the sampling plan, the statistical models and prior specifications, and two additional exploratory analyses. Finally, we included an image to illustrate how the material was presented to the participants in Figure 3 on page 7 and present a table with all study descriptions in English and Dutch as presented to the participants. Note that these were inspired by the descriptions provided in the prediction markets surveys of the SSRP (Camerer et al., 2018) and the ML2 project (Forsell et al., 2018), but rephrased to make them comprehensible for laypeople.

Bayesian Reanalyses of Original Studies

The Bayes factors for the original studies were based on Bayesian reanalyses of the original findings conducted using the Summary Stats module in JASP. Two of the selected studies turned out to require a more elaborate reanalysis for which we requested the original data from the respective authors: For the Balafoutas and Sutter (2012) study, we conducted a Bayesian contingency table analysis (cf. the frequentist Chi-square test) to obtain the Bayes factor for the original effect. For the Derex, Beugin, Godelle, and Raymond (2013) study, we conducted an ordered binomial test using the encompassing prior approach (Klugkist, Kato, & Hoijtink, 2005) as proposed by Marsman in the Bayesian Supplement of Camerer et al. (2018, see: `https://osf.io/z9mwq/`).

Sampling Plan

The sampling plan for our study was based on Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018; Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019), a recently-developed method to help balance informativeness and efficiency of planned experiments within a Bayesian framework. We used the BFDA Shinyapp (`http://shinyapps.org/`

apps/BFDA/) to compute the required sample size for an expected medium effect size of $\delta = .5$, as preregistered. The sampling plan was constructed for the hypothesis that the inclusion of Bayes factors – in addition to study descriptions – would enhance laypeople's prediction performance.

Model & Prior Specifications

Difference Between Research Condition. The difference between the Description Only and the Description Plus Evidence condition was assessed by means of the Brier score. The Brier Score is determined per individual and is calculated as follows:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{c}_i - \mathrm{o}_i)^2 \tag{1}$$

where N is the number of predictions made by each individual, c_i is the rated confidence in replicability $(0-1)$ for the ith study, and o_i the actual outcome (0 = did not replicate, 1 = replicated). For hypothesis testing, we assigned a one-sided Cauchy prior distribution with scale 0.707 to the effect size (i.e., $\delta \sim \text{Cauchy}^{-}(0, 0.707)$).[1]

Hierarchical Accuracy Model. To estimate the prediction accuracy rate and make subsequent inferences, we set up a Bayesian hierarchical model. In the model $\boldsymbol{\theta}$ denotes the vector of accuracy parameters with θ_k being the accuracy of the kth participant, and $\boldsymbol{\theta} \in (0, 1)$. The Bayesian hierarchical model assumes that θ_k is drawn from a group-level distribution which can be characterized using a Beta distribution with mode ω, and a concentration parameter κ, which characterizes the homogeneity of the accuracy across participants (see e.g., Kruschke, 2015, Chapter 9.2 for a formalization of the model). For the quality check (see below) a Beta$(1, 1)$ prior distribution was used for the parameter ω. For the confirmatory hypothesis that laypeople's accuracy $\omega > 0.5$ and the exploratory hypothesis that experts' accuracy in the SSRP and ML2 project $\omega > 0.5$, we used a Beta$(10, 10)$

[1] In our preregistration, we proposed a one-sided hypothesis; however, the corresponding JASP file tested a two-sided hypothesis. Both analyses yield the same conclusion and are included in the JASP file of the final analysis on the OSF (accessible via https://osf.io/dfwmv/). Here, we report the outcome for the one-sided test.

prior distribution for ω. This prior places most density around 0.5 as these values are assumed to be more likely than values in the extremes, but does not influence the likelihood substantially. Note that we assigned an unrestricted prior distribution to the parameter ω for parameter estimation, however, we truncated the parameter space of the prior distribution (i.e., we assigned to all $\omega < 0.5$ a prior density of zero) for hypothesis testing. For the concentration parameter κ a Gamma$(0.01, 0.01)$ prior distribution was used for all hypotheses. The Bayes factors were computed using an approximation of the Savage-Dickey density ratio using the logspline nonparametric density estimate (Dickey & Lientz, 1970; Stone, Hansen, Kooperberg, & Truong, 1997).

Spearman Correlations. The aggregated effect size estimates of the replication studies were estimated as the correlation coefficients r, and were retrieved from Camerer et al. (2018) and Klein et al. (2018). For parameter estimation, a Uniform$(-1, 1)$ distribution was assigned to ρ. For hypothesis testing, we assigned a Uniform$(0, 1)$ distribution to the parameter ρ which truncates the parameter space of ρ to solely allow positive correlations. The Bayes factor was computed using the Savage-Dickey density ratio on the posterior logspline estimate (Stone et al., 1997). Since the posterior distribution for the parameter ρ is not available in closed form, we used data augmentation to draw the posterior samples (cf., van Doorn, Ly, Marsman, & Wagenmakers, 2018).

Signal Detection Model. The Bayesian hierarchical signal detection theory model is based on the model provided in M. D. Lee and Wagenmakers (2013, pp. 164–166). The parameters of interest are the group-level means of the respective Gaussian distributions, that is μ_d for discriminability, and μ_c for bias. In the model, d stands for discriminability or d' (d-prime) and c for criterion. These terms are most often used in the SDT literature to denote discriminability and bias, respectively. We assigned a Gaussian distribution to the group-level discriminability parameter μ_d and the group-level bias parameter μ_c, that is $\mu_d \sim$ Normal$(0, 0.001)$ and $\mu_c \sim$ Normal$(0, 0.001)$. The model file "SDT_JagsModel.txt" is included on the OSF page (`https://osf.io/97aur/`). The AUC was computed based on Wickens (2002, p.68, equation 4.6), assuming equal variances of the signal and noise

distribution (i.e., $\sigma = 1$):

$$AUC = \Phi\left(\frac{\mu_d}{\sqrt{2}}\right). \tag{2}$$

Additional Exploratory Analyses

Here we describe two additional exploratory analyses. The first assess the accuracy of predictions derived from the Bayes factors alone, without human evaluation. This analysis was added as it provides an interesting comparison to the results of our laypeople. At the same time, due to some aspects of the data, the results from this analysis should be interpreted with great caution – as will be explained below. In the second exploratory analysis we estimate the prediction accuracy by means of a hierarchical logistic regression model, since it allows us to model -in addition to the preregistered random intercept for participants- also a random effect of condition and random intercepts for studies.

Estimating prediction accuracy of the original Bayes factors. We exploratorily assessed accuracy of predictions derived from the Bayes factors alone, without human evaluation. We dichotomized the Bayes factors' predictions using a cut-off value of BF = 10 in favor of the alternative hypothesis, since Bayes factors of 10 or higher are often considered compelling (see e.g., Etz & Vandekerckhove, 2016). That is, we generated predictions from the Bayes factor where BF $\leq$ 10 reflects the Bayes factor's prediction that a given study will replicate and BF $<$ 10 the prediction that the study will not replicate. A Bayesian binomial model was used to estimate the accuracy rate θ based on these predictions and test if this was higher than chance. We assigned a Beta$(1, 1)$ prior distribution to the accuracy parameter θ with $\theta \in (0, 1)$. For hypothesis testing, we truncated the parameter space of the prior to range from 0.5 to 1. The analysis demonstrated a median accuracy rate of $0.69\,[0.51, 0.84]$. Note that this median is higher than that of laypeople, but that the estimation is far less certain due to the low number of observations (i.e., only one per study). This is reflected in the wide credible interval and the fact that the data provide only moderate evidence for the hypothesis that the predictive accuracy of original Bayes

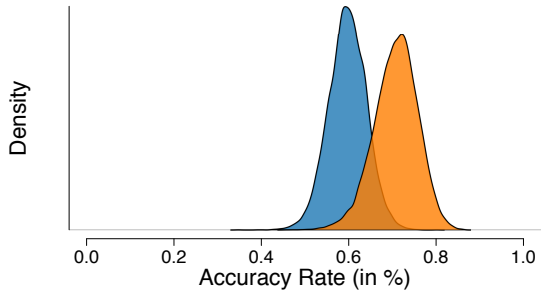


Figure 1. Accuracy rates of laypeople in both conditions estimated using the logistic regression model with random effects for participants and for studies. Posterior distribution of the accuracy in the Description Only condition is depicted in blue and that of accuracy in the Description Plus Evidence condition is depicted in orange.

factors for replication outcomes is above 50%, i.e., $\text{BF}_{+0} = 4.2$.

Logistic regression model with random effects. In the main analysis for the prediction accuracy, we constructed hierarchical models to estimate the group-level accuracy rates, as the primary unit of interest was laypeople's overall accuracy. In an exploratory fashion, we also used the R package `brms` (Bürkner, 2017) to analyze the accuracy data using Bayesian logistic regression models with a random intercept for participants (as before), a random effect of condition, and random intercepts for studies. The logit link function and weakly informative priors on the parameters were used for this analysis (i.e., a Student's t-distribution with 3 degrees of freedom, a location of 0 and a scale of 2.5; Betancourt, Vehtari, & Gelman, 2015).

The posterior distribution for the overall intercept (i.e., the intercept in the Description Only condition) has a median of 0.40 with a 95% credible interval of $[0.05, 0.76]$, which translates to $0.60\,[0.51, 0.68]$ on the accuracy rate scale. The posterior distribution for the coefficient for condition (i.e., Description Plus Evidence condition) gives a median of 0.50 and a 95% credible interval of $[0.08, 0.92]$. This corresponds to an estimated accuracy of $0.71\,[0.60, 0.80]$ for the Description Plus Evidence condition. Note that the median accuracy rates are slightly higher yet credible intervals are wider compared to the main confirmatory analysis. Figure 1 displays the posterior distributions of the accuracy rates per condition and Figure 2 displays the estimated accuracy per study (random effects).

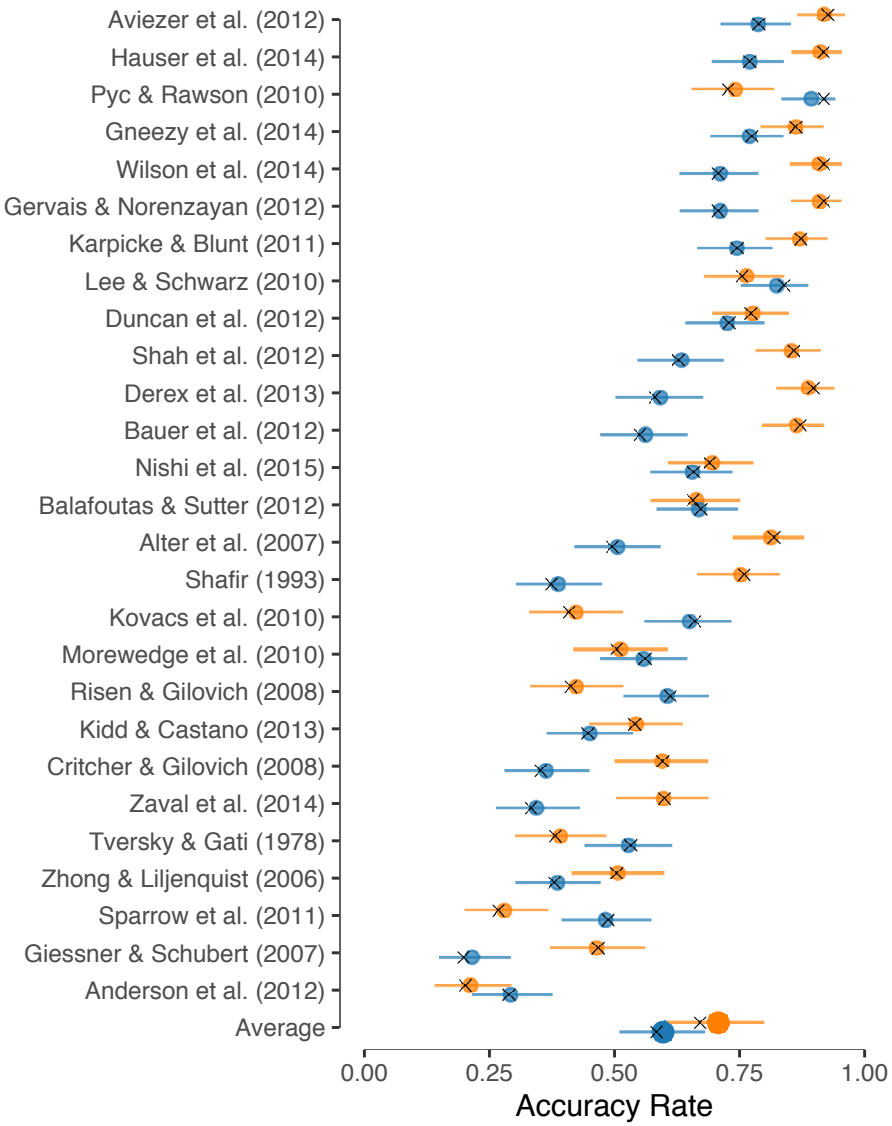
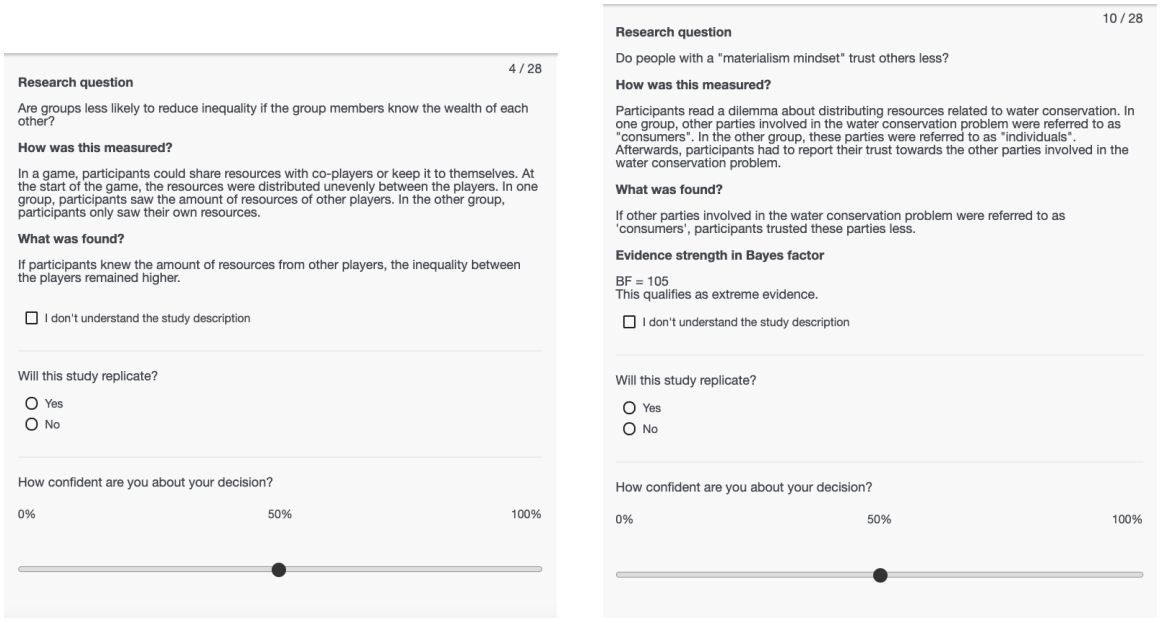


Figure 2. Estimated accuracy of laypeople's predictions for each study and for both conditions (random effects per study). For each study, the median accuracy and 95% credible interval are depicted. Estimates for the Description Only condition are given in blue and estimates for the Description Plus Evidence condition are given in orange. The x's are the sample means for accuracy per study and condition. The bottom row of the figure shows the average per condition.



(a) Description Only condition (b) Description Plus Bayes Factor condition

Figure 3. Examples of the visual display of the survey as presented to the participants. The left image shows an item from the Description Only condition and the right image shows an item from the Description Plus Bayes Factor condition and hence includes additional information on the evidence strength in the original study.

Study Descriptions

Table 1 and 2 provide the descriptions of all included studies. Figure 3 illustrates the set-up for one of the included studies in each condition. The full Qualtrics .qsf file can be found on our OSF page (`https://osf.io/x72cy/`).

Table 1
English Descriptions of Included Studies as Presented to the Participants.

Study	**Description**
Aviezer, Trope, and Todorov (2012)	Can people distinguish between positive and negative emotions based on the body alone? Participants looked at images of athletes that either just scored or just lost a point in a competition. Without seeing any facial expressions, they then judged whether the image depicted a positive or negative emotion. Participants correctly recognized the underlying emotions based on the image of the body alone. That is, they rated the images of winning athletes as more positive then the images of losing athletes. *BF = larger than 1 million. This qualifies as extreme evidence.*
Balafoutas and Sutter (2012)	Does preferential treatment encourage women to enter competition with others? In a game, male and female participants chose whether they wanted to compete against each other, or whether they wanted to get judged based on their individual performance. In one group, women received a preferential treatment if they entered the competition, that is, one point was automatically added to their score. In the other group, women did not receive a preferential treatment. If women received a preferential treatment, then they would choose to enter a competition with others more often. *BF = 4.6. This qualifies as moderate evidence.*
Derex et al. (2013)	Are bigger groups characterized by more cultural diversity? A computer game was set up to be a multiplayer game, consisting of either 2, 4, 8, or 16 players. In this game the players had to collect points by either building arrowheads (a simple task), or building fishing nets (a difficult task). The more players were playing the game, the more likely it was that they increased the cultural diversity within the game (i.e., by building both items). *BF = 21111. This qualifies as extreme evidence.*
Duncan, Sadanand, and Davachi (2012)	Does the detection of new objects improve people's ability to detect subtle changes in similar objects? Participants looked at a series of object images. Afterwards, they had to look again at a series of objects and judged whether they were 'old' (seen before), 'new', or 'similar' (looked like something seen before, but not the same). If 'new' objects were shown before 'similar' objects, then more participants recognized the 'similar' objects correctly. *BF = 11.4. This qualifies as strong evidence.*

Table 1 (continued)

Study	**Description**
Gervais and Norenzayan (2012)	Does analytic thinking trigger religious disbelief? Participants looked at an image. One group looked at an image of "The Thinker", that is a sculpture of a person in a thinking position. The other group looked at an image which depicted a sculpture of a discus thrower. Afterwards, they answered questions about their religiosity. Participants who had seen "The Thinker" reported lower religious belief compared to people who had seen the sculpture of a discus thrower. *BF = 2.1. This qualifies as weak evidence.*
Gneezy, Keenan, and Gneezy (2014)	Do people avoid charities that dedicate a high percentage of donations to administrative and fundraising costs? People chose between donating to either a drinking water charity, or a volunteering charity. One group was told that 50% of the donations to the water charity were used to cover administrative and fundraising costs. The other group was told that all costs for the water charity were already covered. If administrative and fundraising costs were covered, people were more likely to donate to the water charity than to the volunteering charity, compared to when the costs were not covered. *BF = 11.9. This qualifies as strong evidence.*
Hauser, Rand, Peysakhovich, and Nowak (2014)	Do people preserve common resources for future generations if they make collective decisions? In a game, participants chose how many resources they would extract from a pool, and how many they wanted to preserve for future game players. In one group, each participant made their own decision on how many resources they want to extract from the pool. In the other group, participants voted for the combined amount of resources they would extract. If people had to vote on the combined amount of resources they would extract from the pool, more resources were preserved for future generations of game players. *BF = larger than 1 million. This qualifies as extreme evidence.*

Table 1 (continued)

Study	Description
Bauer, Wilkie, Kim, and Bodenhausen (2012)	Do people with a "materialism mindset" trust others less? Participants read a dilemma about distributing resources related to water conservation. In one group, other parties involved in the water conservation problem were referred to as "consumers". In the other group, these parties were referred to as "individuals". Afterwards, participants had to report their trust towards the other parties involved in the water conservation problem. If other parties involved in the water conservation problem were referred to as 'consumers', participants trusted these parties less. *BF = 105. This qualifies as extreme evidence.*
Critcher and Gilovich (2008)	Are people's estimations of numbers influenced by unrelated numbers that are incidentally present in the environment? Participants read a description about a new cell phone. In one group, the cell phone was called P17. In the other group, the cell phone was called P97. Afterwards, participants predicted its proportion of sales. If the cell phone was called P97, participants predicted a higher proportion of sales. *BF = 1.2. This qualifies as weak evidence.*
Karpicke and Blunt (2011)	Is the learning strategy 'retrieval practice' more efficient than the learning strategy 'concept mapping'? Participants studied a science text using a certain learning strategy. In one group, participants used the learning strategy 'retrieval practice', which consists of learning - testing - learning. In the other group, participants used the learning strategy 'concept-mapping', in which participants create a diagram with nodes and links to connect different concepts. One week later participants were tested on this text. If participants used the learning strategy 'retrieval practice', they performed better on a memory test. *BF = 484. This qualifies as extreme evidence.*
Anderson, Kraus, Galinsky, and Keltner (2012)	Do social comparisons influence people's well-being? Participants read a description of a person and were instructed to think about similarities and differences between them. In one group, the person was described as respected and liked by all their social groups. In the other group, the person was described as not respected and liked in any of their social groups. Afterwards, participants answered questions about their well-being. If participants compared themselves with a highly respected and liked person, participants reported lower well-being. *BF = 11.8. This qualifies as strong evidence.*

Table 1 (continued)

Study	Description
Giessner and Schubert (2007)	Is perceived power related to the vertical location of a person? Participants studied a schematic display of the hierarchy within an organization, including a manager and his team. In one group, the vertical line connecting the manager to the team was long (i.e., 7 cm). In the other group, the vertical line was short (i.e., 2 cm). Afterwards, participants estimated how much power they thought the manager held within the organization. If the vertical line connecting the manager to the team was long, participants estimated that the manager held more power within the organization. *BF = 1.9. This qualifies as weak evidence.*
Risen and Gilovich (2008)	Do people belief that tempting fate leads to negative consequences? Participants imagined a scenario in which they would come to a lecture in which the professor picks out one student to answer a difficult question in front of the entire class. In one group, participants imagined that they tempted fate by coming to the lecture unprepared. In the other group, participants imagined that they came to the lecture prepared. Afterwards, participants had to estimate how likely it was that they would get chosen. If participants imagined that they tempted fate, they thought it was more likely that they would get chosen by the professor to answer a difficult question in front of the entire class. *BF = 1.5. This qualifies as weak evidence.*
Shafir (1993)	Do people find positive characteristics more important in decisions in which they are awarding something, and find negative characteristics more important in decisions in which they are denying something? Participants imagined that they were members of a jury, who had to decide which parent would obtain custody for a child. One parent was described as having average features concerning, for instance, income, work-related absence, and relationship to the child. The other parent had both extreme positive characteristics (e.g., very close relationship to the child), but also extreme negative characteristics (e.g., much work-related absence). In one group, participants were asked to which parent they would award custody for the child. In the other group, participants were asked to which parent they would deny custody for the child. In both groups, participants selected the extreme parent more often than the average parent. *BF = 1.9. This qualifies as weak evidence.*

Table 1 (continued)

Study	Description
Zaval, Keenan, Johnson, and Weber (2014)	Do people's concerns about global warming change when the concept of heat or cold is active in their minds? Participants performed a task in which they had to unscramble sentences. In one group, participants read sentences that contained words related to hot temperatures. In the other group, people read sentences that contained words related to cold temperatures. Afterwards, participants reported the degree in which they were concerned about global warming. If participants read sentences that contained words related to hot temperatures, participants were more concerned about global warming. *BF = 1.3. This qualifies as weak evidence.*
Tversky and Gati (1978)	Are people's judgements of how similar two concepts are influenced by the order in which they were mentioned? Participants rated how similar two countries were to each other (e.g., "How similar is the USA to Lebanon?"). One of the countries was well-known to the participants (e.g., the USA). The other country was less familiar to the participants (e.g., Lebanon). In one group, the well-known country was mentioned first. In the other group, the less known country was mentioned first. If the well-known country was mentioned first, participants judged the two countries as less similar. *BF = 6.4. This qualifies as moderate evidence.*
Kidd and Castano (2013)	Can reading literary fiction improve people's understanding of other people's emotions? Participants read a short text passage. In one group, the text passage was literary fiction. In the other group, the text passage was non-fiction. Afterwards, participants had to identify people's expressed emotion (e.g., happy, angry) based on images of the eyes only. Participants were better at recognizing the correct emotion after reading literary fiction. *BF = 3.5. This qualifies as moderate evidence.*
Shah, Mullainathan, and Shafir (2012)	Does poverty drain people's attention? Participants played the game "Wheel of Fortune", a game in which people have to guess letters in word puzzles. In one group, participants were given 6 chances per round to guess letters (i.e., 'poor' players). In the other group, participants were given 20 chances per round to guess letters (i.e., 'rich' players). Afterwards, they completed an attention task. If participants were given few chances per round to guess letters, they performed worse in the subsequent attention task. *BF = 1.5. This qualifies as weak evidence.*

Table 1 (continued)

Study	Description
Zhong and Liljenquist (2006)	Does feeling morally dirty increase people's need to wash themselves? Participants hand copied a story written in the first person. In one group, participants rewrote an unethical short story about sabotaging a co-worker. In the other group, participants rewrote an ethical short story about helping a co-worker. Afterwards, participants expressed their desire for cleaning products (e.g., soap, toothpaste). If participants rewrote an unethical story, they had a higher desire for cleansing products. *BF = 4. This qualifies as moderate evidence.*
Kovacs, Téglás, and Endress (2010)	Do beliefs of others influence people's actions, even if these beliefs are irrelevant? Participants watched a short cartoon. In the cartoon, a character places a ball behind a box, which then rolls away. Then, the character lifts the box and reveals the ball, which returned unexpectedly. Participants were instructed to press a button as soon as they detected the ball. In one group, the character saw the ball rolling away. In the other group, the character did not see the ball rolling away. If the character believed that the ball was behind the box, participants detected the ball faster. *BF = 2.4. This qualifies as weak evidence.*
S. W. S. Lee and Schwarz (2010)	Does washing hands weaken people's urge to justify their choice for a non-preferred item? Participants ranked 10 CDs based on how much they would like to own them. Then, the participants could choose to keep their 5th or 6th choice. After the choice, participants were asked to evaluate a soap and then to rank the CDs again. In one group, participants evaluated the soap after they washed their hands with it. In the other group, participants evaluated the soap without washing their hands with it. If participants evaluated the soap without washing their hands with it, they increased their preference for their chosen CD. *BF = 4. This qualifies as moderate evidence.*
Morewedge, Huh, and Vosgerau (2010)	Do people want to eat less food, after they repeatedly imagined eating it? Participants had to imagine 33 repetitive actions, one at a time. In one group, participants imagined eating 30 M&Ms and then inserting 3 coins in a laundry machine. In the other group, participants imagined inserting 33 coins in a laundry machine. Afterwards, participants could eat from a bowl containing M&Ms. If participants imagined eating 30 M&Ms, they ate fewer M&Ms from the bowl. *BF = 5.4. This qualifies as moderate evidence.*

Table 1 (continued)

Study	Description
Nishi, Shirado, Rand, and Christakis (2015)	Are groups less likely to reduce inequality if the group members know the wealth of each other? In a game, participants could share resources with co-players or keep it to themselves. At the start of the game, the resources were distributed unevenly between the players. In one group, participants saw the amount of resources of other players. In the other group, participants only saw their own resources. If participants knew the amount of resources from other players, the inequality between the players remained higher. *BF = 7.1. This qualifies as moderate evidence.*
Pyc and Rawson (2010)	When learning new words, does taking a test help people to remember links between the words and their meaning? Participants learned 48 Swahili words by writing down links between each Swahili word and its meaning. In one group, the learning strategy of the participants involved learning - testing - learning. In the other group, the learning strategy of the participants involved learning - relearning, without testing. One week later, participants did a memory test in which they got the instruction to write down the meaning of the Swahili word, and their self-generated link between the word and its meaning. If participants used a learning strategy that involved learning - testing - learning, they remembered more of their self-generated links between the words and their meaning. *BF = 2.6. This qualifies as weak evidence.*
Sparrow, Liu, and Wegner (2011)	Do difficult questions activate the concepts of "Google" and computers in people's minds? Participants answered knowledge questions. In one group, participants answered difficult questions. In the other group, participants answered easy questions. Afterwards, they performed a reaction time task where they had to ignore computer-related words that were presented on the screen. If participants answered difficult questions, their reaction times were slower when computer-related words were present on the screen (which indicates that the concept of computers was active in their minds). *BF = 15.5. This qualifies as strong evidence.*

Table 1 (continued)

Study	Description
Wilson et al. (2014)	Do people enjoy doing nothing? Participants spent a short amount of time by themselves in an empty room. In one group, participants were instructed to spend their time on a non-social activity (e.g., listening to music, reading a book, surfing the Web). In the other group, participants were instructed to entertain themselves with their thoughts (without any external activity). Afterwards participants answered questions about how enjoyable this experience was. If participants had just their thoughts to entertain themselves, they enjoyed themselves less. *BF = 422. This qualifies as extreme evidence.*
Alter, Oppenheimer, Epley, and Eyre (2007)	Do people's logic skills improve if the concept of analytic thinking is active in their minds? Participants solved difficult logic problems. In one group, the logic problems were printed in a font that was hard to read (activating the analytical mindset). In the other group, the logic problems were printed in a font that was easy to read. If the logic problems were printed in a font that was hard to read, participants solved more of the logic problems correctly. *BF = 1.5. This qualifies as weak evidence.*

Note. Information about the Bayes factor (BF; in italics) was only added in the Bayes factor condition. The original term 'anecdotal evidence' was changed into 'weak evidence' to make it more understandable and comparable to the other labels. As all Bayes factors were in the direction of support for the alternative hypothesis, we did not distinguish between BF_{10} and BF_{01}, but only report BF_{10}.

Table 2
Dutch Descriptions of Included Studies as Presented to the Participants.

Study	**Description**
Aviezer et al. (2012)	Kunnen mensen onderscheid maken tussen positieve en negatieve emoties op basis van lichaamshouding? Deelnemers keken naar foto's van atleten die net een punt hebben gescoord of net een punt hebben verloren in een competitie. Zonder gezichtsuitdrukkingen te zien, beoordeelden zij vervolgens of de foto een positieve of negatieve emotie verbeeldde. De deelnemers herkenden de onderliggende emoties correct op basis van de foto van het lichaam alleen. Dat wil zeggen, ze beoordeelden de foto's van winnende atleten als positiever dan de foto's van verliezende atleten. *BF = groter dan 1 miljoen. Dit geldt als extreem bewijs.*
Balafoutas and Sutter (2012)	Stimuleert een voorkeursbehandeling vrouwen om de concurrentie met anderen aan te gaan? In een spel kozen mannelijke en vrouwelijke deelnemers of ze tegen elkaar wilden strijden, of dat ze beoordeeld wilden worden op basis van hun individuele prestaties. In één groep kregen vrouwen een voorkeursbehandeling als ze meededen aan de wedstrijd, dat wil zeggen dat één punt automatisch aan hun score werd toegevoegd. In de andere groep kregen vrouwen geen voorkeursbehandeling. Als vrouwen een voorkeursbehandeling kregen, dan kozen ze vaker om deel te nemen aan een wedstrijd met anderen. *BF = 4.6. Dit geldt als matig bewijs.*
Derex et al. (2013)	Worden grotere groepen gekenmerkt door meer culturele diversiteit? Deelnemers speelden een computerspel met 2, 4, 8 of 16 medespelers. In dit spel moesten de spelers punten verzamelen door ofwel pijlpunten te bouwen (een eenvoudige taak), ofwel visnetten te bouwen (een moeilijke taak). Hoe meer spelers het spel speelden, hoe waarschijnlijker het was dat ze de culturele diversiteit binnen het spel vergrootten (d.w.z. door beide items te bouwen). *BF = 21111. Dit geldt als extreem bewijs.*
Duncan et al. (2012)	Verbetert de detectie van nieuwe objecten het vermogen van mensen om subtiele veranderingen in soortgelijke objecten te detecteren? De deelnemers keken naar een serie afbeeldingen van objecten. Daarna moesten ze opnieuw naar een reeks objecten kijken en beoordeelden ze of ze 'oud' (eerder gezien), 'nieuw' of 'soortgelijk' waren (ze leken op iets wat eerder gezien werd, maar niet hetzelfde). Als 'nieuwe' objecten direct voor 'soortgelijke' objecten werden getoond, dan classificeerden deelnemers de 'soortgelijke' objecten vaker correct. *BF = 11.4. Dit geldt als sterk bewijs.*

Table 2 (continued)

Study	Description
Gervais and Norenzayan (2012)	Leidt analytisch denken tot religieus ongeloof? Deelnemers keken naar een afbeelding. Een groep keek naar een foto van 'De Denker', dat is een standbeeld van een persoon in een denkpositie. De andere groep keek naar een foto van een standbeeld van een discuswerper. Daarna beantwoordden ze vragen over hun religiositeit. Deelnemers die 'De Denker' hadden gezien, rapporteerden lager religieus geloof in vergelijking met mensen die het beeld van een discuswerper hadden gezien. *BF = 2.1. Dit geldt als zwak bewijs.*
Gneezy et al. (2014)	Mijden mensen liefdadigheidsinstellingen die een hoog percentage van de donaties besteden aan administratieve en fondsenwervingskosten? Deelnemers kozen tussen het doneren aan een liefdadigheidsinstelling voor schoon drinkwater, of een liefdadigheidsinstelling voor vrijwilligerswerk. Een groep kreeg te horen dat 50% van de donaties aan het waterfonds werd gebruikt om de administratieve en fondsenwervingskosten te dekken. De andere groep kreeg te horen dat alle kosten voor het goede doel water al gedekt waren. Als de administratieve en fondsenwervingskosten gedekt waren, waren deelnemers meer geneigd om te doneren aan het waterfonds dan aan het goede doel voor vrijwilligers, in vergelijking met wanneer de kosten niet gedekt waren. *BF = 11.9. Dit geldt als sterk bewijs.*
Hauser et al. (2014)	Bewaren mensen gemeenschappelijke hulpbronnen voor toekomstige generaties als ze collectieve beslissingen nemen? In een spel kozen de deelnemers hoeveel grondstoffen ze uit een gezamenlijke pot zouden halen, en hoeveel ze wilden behouden voor toekomstige spelers. In één groep kon elke deelnemer zijn eigen beslissing nemen over het aantal middelen dat hij of zij uit de pot wilde halen. In de andere groep stemden de deelnemers voor de totale hoeveelheid middelen die zij zouden onttrekken. Als mensen moesten stemmen over de totale hoeveelheid grondstoffen die ze uit de pot haalden, werden er meer grondstoffen bewaard voor toekomstige generaties spelers. *BF = groter dan 1 miljoen. Dit geldt als extreem bewijs.*

Table 2 (continued)

Study	Description
Bauer et al. (2012)	Vertrouwen mensen met een 'materialistische mentaliteit' minder op anderen? Deelnemers lazen een dilemma over de verdeling van hulpbronnen in verband met waterbehoud. In één groep werden andere partijen die betrokken zijn bij het waterbesparingsprobleem aangeduid als "consumenten". In de andere groep werden deze partijen aangeduid als "individuen". Na afloop moesten de deelnemers hun vertrouwen rapporteren over de andere partijen die betrokken waren bij het waterbesparingsprobleem. Als andere partijen die betrokken waren bij het waterbesparingsprobleem 'consumenten' werden genoemd, vertrouwden de deelnemers minder op deze partijen. *BF = 105. Dit geldt als extreem bewijs.*
Critcher and Gilovich (2008)	Worden de schattingen van aantallen beÃŕnvloed door niet-verwante getallen die incidenteel aanwezig zijn in de omgeving? Deelnemers lazen een beschrijving van een nieuwe mobiele telefoon. In één groep werd de mobiele telefoon P17 genoemd. In de andere groep werd de mobiele telefoon P97 genoemd. Daarna voorspelden de deelnemers de verkoopomzet. Als de mobiele telefoon P97 werd genoemd, voorspelden de deelnemers een hogere verkoopomzet. *BF = 1.2. Dit geldt als zwak bewijs.*
Karpicke and Blunt (2011)	Is de leerstrategie 'retrieval practice' efficiënter dan de leerstrategie 'concept mapping'? Deelnemers bestudeerden een wetenschappelijke tekst aan de hand van een bepaalde leerstrategie. In één groep maakten de deelnemers gebruik van de leerstrategie 'retrieval practice', die bestaat uit leren - testen - leren. In de andere groep maakten de deelnemers gebruik van de leerstrategie 'concept-mapping', waarbij de deelnemers een schema maken met knooppunten en koppelingen om verschillende concepten met elkaar te verbinden. Een week later werden de deelnemers getest op deze tekst. Als de deelnemers de leerstrategie 'retrieval practice' gebruikten, presteerden ze beter op een geheugentest. *BF = 484. Dit geldt als extreem bewijs.*

Table 2 (continued)

Study	Description
Anderson et al. (2012)	Hebben sociale vergelijkingen invloed op het welbevinden van mensen? Deelnemers lazen een beschrijving van een persoon en kregen de opdracht om na te denken over overeenkomsten en verschillen tussen hen. In één groep werd de persoon beschreven als gerespecteerd en geliefd bij al hun sociale groepen. In de andere groep werd de persoon beschreven als niet gerespecteerd en in geen van hun sociale groepen geliefd. Na afloop beantwoordden de deelnemers vragen over hun welbevinden. Als de deelnemers zichzelf vergelijken met een zeer gerespecteerde en gewaardeerde persoon, rapporteerden de deelnemers een lager welbevinden. *BF = 11.8. Dit geldt als sterk bewijs.*
Giessner and Schubert (2007)	Is de waargenomen macht gerelateerd aan de verticale locatie van een persoon? De deelnemers bestudeerden een schematische weergave van de hiërarchie binnen een organisatie, inclusief een manager en zijn team. In één groep was de verticale lijn die de manager met het team verbond lang (7 cm). In de andere groep was de verticale lijn kort (2 cm). Daarna schatten de deelnemers in hoeveel macht zij dachten dat de manager in de organisatie had. Als de verticale lijn die de manager met het team verbindt lang was, schatten de deelnemers dat de manager meer macht had binnen de organisatie. *BF = 1.9. Dit geldt als zwak bewijs.*
Risen and Gilovich (2008)	Geloven mensen dat het lot tarten negatieve gevolgen heeft? De deelnemers stelden zich een scenario voor waar ze naar een les kwamen waarin de professor één student zou uitkiezen om een moeilijke vraag voor de hele klas te beantwoorden. In één groep stelden de deelnemers zich voor dat ze het lot tartten door onvoorbereid naar de les te komen. In de andere groep stelden de deelnemers zich voor dat ze voorbereid naar de les kwamen. Na afloop moesten de deelnemers inschatten hoe groot de kans was dat ze gekozen zouden worden. Als de deelnemers zich voorstelden dat ze het lot zouden tarten, dachten ze dat het waarschijnlijker was dat ze door de professor zouden worden uitgekozen om een moeilijke vraag voor de hele klas te beantwoorden. *BF = 1.5. Dit geldt als zwak bewijs.*

Table 2 (continued)

Study	Description
Shafir (1993)	Vinden mensen positieve kenmerken belangrijker in beslissingen waarbij ze iemand iets toekennen en negatieve kenmerken belangrijker in beslissingen waarbij ze iemand iets weigeren? De deelnemers stelden zich voor dat ze lid waren van een jury, die moest beslissen welke ouder de voogdij over een kind zou krijgen. Één ouder werd beschreven als een ouder met gemiddelde kenmerken met betrekking tot bijvoorbeeld inkomen, werkgerelateerde afwezigheid en de relatie met het kind. De andere ouder had zowel extreem positieve kenmerken (bijv. zeer nauwe relatie met het kind), maar ook extreem negatieve kenmerken (bijv. veel werkgerelateerde afwezigheid). In één groep werd de deelnemers gevraagd aan welke ouder zij de voogdij over het kind zouden toekennen. In de andere groep werd de deelnemers gevraagd aan welke ouder zij de voogdij over het kind zouden weigeren. In beide groepen kozen de deelnemers vaker voor de extreme ouder dan de gemiddelde ouder. *BF = 1.9. Dit geldt als zwak bewijs.*
Zaval et al. (2014)	Verandert de bezorgdheid over de opwarming van de aarde wanneer het concept van warmte of kou geactiveerd is? Deelnemers voerden een taak uit waarbij ze de zinnen moesten ontcijferen. In één groep lazen de deelnemers zinnen die woorden bevatten die te maken hebben met hoge temperaturen. In de andere groep lazen mensen zinnen die woorden bevatten die te maken hebben met lage temperaturen. Na afloop gaven de deelnemers aan in welke mate zij zich zorgen maakten over de opwarming van de aarde. Als de deelnemers zinnen lazen die woorden bevatten die te maken hebben met de hoge temperaturen, waren de deelnemers meer bezorgd over de opwarming van de aarde. *BF = 1.3. Dit geldt als zwak bewijs.*
Tversky and Gati (1978)	Worden oordelen over in hoeverre twee concepten overeenkomen beÃŕnvloed door de volgorde waarin ze worden genoemd? De deelnemers beoordeelden de gelijkenis tussen twee landen (bijvoorbeeld: “Hoe vergelijkbaar is de VS met Libanon?”). Een van de landen was bekend bij de deelnemers (bijv. de VS). Het andere land was minder bekend bij de deelnemers (bijvoorbeeld Libanon). In één groep werd als eerste het bekende land genoemd. In de andere groep werd als eerste het minder bekende land genoemd. Als het bekende land als eerste werd genoemd, beoordeelden de deelnemers de twee landen als minder overeenkomstig. *BF = 6.4. Dit geldt als matig bewijs.*

Table 2 (continued)

Study	**Description**
Kidd and Castano (2013)	Kan het lezen van literaire fictie het herkennen van de emoties van anderen verbeteren? Deelnemers lazen een korte tekstpassage. In één groep was de tekstpassage literaire fictie. In de andere groep was de tekstpassage non-fictie. Na afloop moesten de deelnemers de emoties die mensen uitdrukten (bv. gelukkig, boos) identificeren aan de hand van foto's van alleen de ogen. Deelnemers waren beter in het herkennen van de juiste emotie na het lezen van literaire fictie. *BF = 3.5. Dit geldt als matig bewijs.*
Shah et al. (2012)	Zorgt armoede voor verminderde aandacht? Deelnemers speelden het spel "Wheel of Fortune", een spel waarin mensen letters moeten raden in woordpuzzels. In één groep kregen de deelnemers 6 kansen per ronde om letters te raden (d.w.z. 'arme' spelers). In de andere groep kregen de deelnemers 20 kansen per ronde om letters te raden (d.w.z. 'rijke' spelers). Daarna voerden ze een aandachtstaak uit. Als de deelnemers per ronde weinig kans kregen om letters te raden, presteerden ze slechter in de daaropvolgende aandachtstaak. *BF = 1.5. Dit geldt als zwak bewijs.*
Zhong and Liljenquist (2006)	Verhoogt het gevoel van moreel besmetting de behoefte van mensen om zich te wassen? Deelnemers kopieerden met de hand een verhaal geschreven in de eerste persoon. In één groep herschreven de deelnemers een onethisch kort verhaal over het saboteren van een collega. In de andere groep herschreven de deelnemers een ethisch kort verhaal over het helpen van een collega. Na afloop rapporteerden de deelnemers hun behoefte aan reinigingsproducten (bijv. zeep, tandpasta). Als deelnemers een onethisch verhaal herschreven, hadden ze meer behoefte aan reinigingsmiddelen. *BF = 4. Dit geldt als matig bewijs.*
Kovacs et al. (2010)	Hebben overtuigingen van anderen invloed op het handelen van mensen, zelfs als deze overtuigingen irrelevant zijn? Deelnemers keken naar een korte cartoon. In de cartoon plaatst een personage een bal achter een doos, die dan wegrolt. Vervolgens tilt het personage de doos op en onthult de bal, die onverwacht terugkeerde. Deelnemers kregen de opdracht om op een knop te drukken zodra ze de bal waarnamen. In één groep zag het personage de bal wegrollen. In de andere groep zag het personage de bal niet wegrollen. Als het personage geloofde dat de bal achter de doos lag, ontdekten de deelnemers de bal sneller. *BF = 2.4. Dit geldt als zwak bewijs.*

Table 2 (continued)

Study	**Description**
S. W. S. Lee and Schwarz (2010)	Verzwakt handenwassen de behoefte om de keuze voor een item dat niet de voorkeur had te rechtvaardigen? Deelnemers rangschikten 10 cd's op basis van hoe graag ze deze wilden hebben. Vervolgens konden de deelnemers ervoor kiezen om hun 5e of 6e keuze te behouden. Na de keuze werd de deelnemers gevraagd om een zeep te evalueren en vervolgens de cd's opnieuw te rangschikken. In één groep evalueerden de deelnemers de zeep nadat ze er hun handen mee gewassen hadden. In de andere groep evalueerden de deelnemers de zeep zonder er hun handen mee te wassen. Als de deelnemers de zeep evalueerden zonder de handen ermee te wassen, verhoogden ze hun voorkeur voor de door hen gekozen cd. *BF = 4. Dit geldt als matig bewijs.*
Morewedge et al. (2010)	Willen mensen minder van een product eten, nadat ze zich herhaaldelijk hebben voorgesteld het te eten? Deelnemers moesten zich 33 repetitieve acties voorstellen, één voor één. In één groep stelden de deelnemers zich voor om 30 M&M's te eten en vervolgens 3 munten in een wasmachine te stoppen. In de andere groep stelden de deelnemers zich voor om 33 munten in een wasmachine te stoppen. Na afloop konden de deelnemers uit een schaal met M&M's eten. Als de deelnemers zich hadden voorgesteld om 30 M&M's te eten, aten ze minder M&M's uit de schaal. *BF = 5.4. Dit geldt als matig bewijs.*
Nishi et al. (2015)	Zijn groepen minder geneigd om de ongelijkheid te verminderen als de groepsleden elkaars rijkdom kennen? In een spel konden deelnemers middelen delen met medespelers of voor zichzelf houden. Aan het begin van het spel waren de grondstoffen ongelijk verdeeld over de spelers. In één groep zagen de deelnemers de hoeveelheid middelen van andere spelers. In de andere groep zagen de deelnemers alleen hun eigen middelen. Als de deelnemers de hoeveelheid middelen van andere spelers wisten, bleef de ongelijkheid tussen de spelers groter. *BF = 7.1. Dit geldt als matig bewijs.*

Table 2 (continued)

Study	Description
Pyc and Rawson (2010)	Helpt het afnemen van een test bij het leren van nieuwe woorden, om mensen te herinneren aan de verbanden tussen de woorden en hun betekenis? Deelnemers leerden 48 Swahili woorden door het opschrijven van verbanden tussen elk Swahili woord en de betekenis ervan. In één groep bestond de leerstrategie van de deelnemers uit leren - testen - leren. In de andere groep ging het bij de leerstrategie van de deelnemers om leren - opnieuw leren, zonder testen. Een week later deden de deelnemers een geheugentest waarin ze de instructie kregen om de betekenis van het Swahili woord op te schrijven, en hun zelfgegenereerde verband tussen het woord en zijn betekenis. Als de deelnemers een leerstrategie gebruikten die het leren - testen - leren omvatte, herinnerden ze zich meer van hun zelfgegenereerde verbanden tussen de woorden en hun betekenis.*BF = 2.6. Dit geldt als zwak bewijs.*
Sparrow et al. (2011)	Leiden moeilijke vragen tot de mentale activatie van de concepten 'Google' en computers? Deelnemers beantwoordden kennisvragen. In één groep beantwoordden de deelnemers moeilijke vragen. In de andere groep beantwoordden de deelnemers eenvoudige vragen. Daarna voerden ze een reactietijdstaak uit waarbij ze computergerelateerde woorden die op het scherm werden gepresenteerd, moesten negeren. Als de deelnemers moeilijke vragen beantwoordden, waren hun reactietijden trager als er computergerelateerde woorden op het scherm stonden (wat aangeeft dat het concept van computers mentaal geactiveerd was). *BF = 15.5. Dit geldt als sterk bewijs.*
Wilson et al. (2014)	Houden mensen ervan om niets te doen? Deelnemers brachten een korte tijd door in een lege ruimte. In één groep werden de deelnemers geïnstrueerd om hun tijd te besteden aan een niet-sociale activiteit (bijv. luisteren naar muziek, een boek lezen, surfen op het web). In de andere groep kregen de deelnemers de opdracht zich te vermaken met hun gedachten (zonder enige externe activiteit). Na afloop beantwoordden de deelnemers vragen over hoe leuk deze ervaring was. Als de deelnemers alleen hun gedachten hadden om zich mee te vermaken, genoten ze minder van de ervaring. *BF = 4̇22. Dit geldt als extreem bewijs.*

Table 2 (continued)

Study	**Description**
Alter et al. (2007)	Worden vaardigheden in logica beter als het concept van analytisch denken mentaal geactiveerd is? Deelnemers losten moeilijke logicaproblemen op. In één groep werden de logicaproblemen afgedrukt in een moeilijk leesbaar lettertype (het activeren van de analytische mindset). In de andere groep werden de logicaproblemen in een gemakkelijk leesbaar lettertype afgedrukt. Als de logicaproblemen in een moeilijk leesbaar lettertype werden afgedrukt, losten de deelnemers meer van de logicaproblemen correct op. *BF = 1.5. Dit geldt als zwak bewijs.*

Note. Information about the Bayes factor (BF; in italics) was only added in the Bayes factor condition. As all Bayes factors were in the direction of support for the alternative hypothesis, we did not distinguish between BF_{10} and BF_{01}, but only report BF_{10}.

References

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

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect: Social status and subjective well-being. *Psychological Science*, *23*, 764–771. doi: 10.1177/0956797611434537

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*, 1225–1229.

Balafoutas, L., & Sutter, M. (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science*, *335*, 579–582.

Bauer, M. A., Wilkie, J. E., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism: Situational materialism undermines personal and social well-being. *Psychological Science*, *23*, 517–523.

Betancourt, M., Vehtari, A., & Gelman, A. (2015). *Prior choice recommendations.* https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. doi: 10.18637/jss.v080.i01

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science*. *Nature Human Behaviour*, *2*, 637–644.

Critcher, C. R., & Gilovich, T. (2008). Incidental environmental anchors. *Journal of Behavioral Decision Making*, *21*, 241–251.

Derex, M., Beugin, M.-P., Godelle, B., & Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, *503*, 389–391.

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. *Science*, *337*, 485–487.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PloS ONE*, *11*, e0149794.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., ... Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*. doi: 10.1016/j.joep.2018.10.009

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*, 493–496.

Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. *Organizational Behavior and Human Decision Processes*, *104*, 30–44.

Gneezy, U., Keenan, E. A., & Gneezy, A. (2014). Avoiding Overhead Aversion in Charity. *Science*, *346*, 632–635. doi: 10.1126/science.1253932

Hauser, O. P., Rand, D. G., Peysakhovich, A., & Nowak, M. A. (2014). Cooperating with the future. *Nature*, *511*, 220–223.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

Kidd, D. C., & Castano, E. (2013). Reading Literary Fiction Improves Theory of Mind. *Science*, *342*, 377–380. doi: 10.1126/science.1239918

Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., ... Nosek, B. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69. doi: 10.1111/j.1467-9574.2005.00279.x

Kovacs, A. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press/Elsevier.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lee, S. W. S., & Schwarz, N. (2010). Washing away postdecisional dissonance. *Science*, *328*, 709.

Morewedge, C. K., Huh, Y. E., & Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, *330*, 1530–1533.

Nishi, A., Shirado, H., Rand, D. G., & Christakis, N. A. (2015). Inequality and visibility of wealth in experimental social networks. *Nature*, *526*, 426–429.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*, 293–307.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*, 546–556.

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science*, *338*, 682–685. doi: 10.1126/science.1222426

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*, 776–778.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using informed priors. *Behavior Research Methods*, *51*, 1042-1058.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling (with discussion). *The Annals of Statistics*, *25*, 1371–1470.

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and Categorization*, *1*, 79–98.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian latent–normal inference for the rank sum test, the signed rank test, and Spearman's ρ. *Preprint via* `https://arxiv.org/abs/1712.06941`.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford University Press, USA.

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., ... Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science*, *345*, 75–77.

Zaval, L., Keenan, E. A., Johnson, E. J., & Weber, E. U. (2014). How warm days increase belief in global warming. *Nature Climate Change*, *4*, 143–147.

Zhong, C.-B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science*, *313*, 1451–1452.