

The Effect of Concrete Wording on Truth Judgements: A Preregistered Replication and Extension of Hansen & Wänke (2010)

Emma L. Henderson¹, Frédéric Vallée-Tourangeau², & Daniel J. Simons³

¹ Department of Management, Kingston University, Kingston upon Thames, UK

² Department of Psychology, Kingston University, Kingston upon Thames, UK

³ Department of Psychology, University of Illinois at Urbana-Champaign, USA

When you lack the facts, how do you decide what is true and what is not? In the absence of knowledge, we sometimes rely on non-probative information. For example, participants judge concretely worded trivia items as more likely to be true than abstractly worded ones (the *linguistic truth effect*; Hansen & Wänke, 2010). If minor language differences affect truth judgements, ultimately they could influence more consequential political, legal, health, and interpersonal choices. This Registered Report includes two high-powered replication attempts of Experiment 1 from Hansen and Wänke (2010). Experiment 1a was a dual-site, in-person replication of the linguistic concreteness effect in the original paper-and-pencil format ($n = 253$, $n = 246$ in analyses). Experiment 1b replicated the study with an online sample ($n = 237$, $n = 220$ in analyses). In Experiment 1a, the effect of concreteness on judgements of truth (Cohen's $d_z = 0.08$; 95% CI: [-0.03, 0.18]) was smaller than that of the original study. Similarly, in Experiment 1b the effect (Cohen's $d_z = 0.11$; 95% CI [-0.01, 0.22]) was smaller than that of the original study. Collectively, the pattern of results is inconsistent with that of the original study.

Keywords: replication, truth judgements, truth effect, concreteness, language, Registered Report

The perceived truth of a statement can be influenced by factors other than its probative, informational content (Koriat & Adiv, 2012), including the source of the information, the context in which it is presented, and characteristics of the statement itself (Dechêne, Stahl, Hansen, & Wänke, 2010). This paper examines an effect of the statement wording: Participants judge concretely worded trivia items as more likely to be true than abstractly worded versions of the same content (the linguistic concreteness effect; Hansen & Wänke, 2010). For instance, the statement, “The poet C. Dickens wrote the play *Miss Sara Sampson*,” was judged more likely to be true than the more abstract equivalent, “The play *Miss Sara Sampson* is by the poet C. Dickens.” Across all statements, more concrete versions were judged as more probably true than their abstract equivalents (Cohen's $d_z = .48$).

This manipulation is based on the linguistic category model (Semin & Fiedler, 1988, 1991) which posits that a concrete verb (“wrote”) conjures a vivid, reliable, and easily verifiable image, but an abstract one (“is by”) does not (Semin & Fiedler, 1988). The model was originally designed to assess descriptions of people's behaviour, and it has also been ap-

plied to analyses of persuasion and influence. For example, prosecutors in the Nuremberg trials used concrete language to signpost the responsibility of Nazi generals (Schmid & Fiedler, 1996).

According to the model, descriptive action verbs, such as “wrote” or “punch” require no interpretation; they refer to a single, concrete, behavioural event and convey the perceptual properties of that event (e.g., “A punches B”). All of the concrete statements used by Hansen and Wänke (2010) contained such descriptive action verbs. In contrast, their abstract statements described the same event but required more interpretation (e.g., “A hurts B”). Although their abstract statements were guided by the linguistic category model, they did not fully implement it. Some of their abstract statements contained no state verbs or adjectives, the two categories classified as abstract in the model. Those statements that lacked state verbs or adjectives “map the criteria of the LCM of abstractness (e.g., high stability, low situational dependency)” (J. Hansen, personal communication, January 25, 2018) and rely on characteristics associated with abstract word categories rather than always containing the word categories themselves.

Replication Value

Correspondence concerning this article should be addressed to Emma L. Henderson, Department of Management, Kingston University, Kingston Hill Campus, Kingston Hill, Kingston upon Thames, KT2 7LB. E-mail: e.henderson@kingston.ac.uk

Understanding how and when belief in the truth of a statement is influenced by its superficial characteristics rather than its substance is of great practical and theoretical importance. That a statement's truthiness can influence judgements is well

established (Fazio, Brashier, Payne, & Marsh, 2015; Newman, Garry, Bernstein, Kantner, & Lindsay, 2012; Newman et al., 2015). Most studies examining the factors that influence truth judgements, other than the statement's substance, have focused on the *illusory truth effect* (Hasher, Goldstein, & Toppino, 1977); repeated statements are believed more than new statements. A Google Scholar search for “illusory truth effect” revealed 247 results, compared to 2 for “linguistic concreteness effect” (Google Scholar, August 27, 2018). Yet, Hansen and Wänke's (2010) experiment underlies research on the persuasiveness of concrete language in political communication (Menegatti & Rubini, 2013), voting intentions (Chou & Yeh, 2018), and eyewitness testimony (Kurinec & Weaver III, 2018). Given the ease with which this concrete/abstract manipulation can be applied in practice and the estimated effect from the original study ($d_z = 0.48$), the experimental manipulation merits further investigation and a more precise estimate of the effect size. In practice, manipulation of beliefs via linguistic concreteness might be easier to do and harder for readers to notice. If robust, the effects of linguistic concreteness could potentially be combined with the illusory truth effect (or other such effects) to yield even greater effects on beliefs.

Despite its theoretical and practical implications, the effect of linguistic concreteness on truth judgements has not been independently replicated, either conceptually or closely. In light of this paucity of research, combined with the practical implications of the effect if it proves robust, we designed two high-powered replications using sample sizes substantially bigger than the original study. We undertook this research for three further reasons. First, Hansen and Wänke's (2010) experiment has been heavily cited (102 citations according to Google Scholar, August 27, 2018, approximately 10 times the mean for 2010) and used to motivate research on topics ranging from political persuasion to eyewitness testimony. It has also been discussed in the media as a technique for increasing trustworthiness (e.g., Stott, 2011). Second, the relatively subtle manipulation of concreteness yielded an effect size of Cohen's $d_z = 0.48$, but the sample size ($n=46$) means that the estimate was not precise (95% CI [0.19, 0.78]). A direct replication using the same materials will verify the effect and estimate its size more precisely. Finally our second experiment will directly replicate the original study using the same design, but with a different source of participants (online) to determine whether the effect is equally robust.

The Present Experiments

With guidance from the original authors, we designed a high-powered, pre-registered replication of Experiment 1 from Hansen and Wänke (2010). We aimed to match, as closely as possible, the conditions and methods of the original paper

with an implementation that addressed those factors that the original authors believe are necessary for obtaining the effect. Like the original study, we tested the prediction that participants would judge concretely worded trivia items as more probably true than abstractly worded versions (H1 - confirmatory hypothesis). We also added several enhancements and extensions. First, to test whether the effect would generalise beyond the originally sampled population, we tested participants in the United Kingdom and the United States, both with in-person samples and online. Second, to ensure that our primary hypothesis tests were adequately powered to detect the original effect and to enable a more precise measure of the effect size, we tested approximately five times as many participants as the original experiment. Third, in their fourth study, Hansen and Wänke (2010) inferred that some participants already knew answers to some of the trivia items (i.e., their objective truth value). Consequently, we added a check for prior knowledge of the answers to the trivia questions. Finally, at the suggestion of the first author of the original study, we used an expanded stimulus set to test the exploratory hypothesis that the perceived psychological distance of the statement content would interact with the concreteness of the wording (H2 - exploratory hypothesis).

For both experiments, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures (Simmons, Nelson, & Simonsohn, 2012). Our preregistration, materials, and data are available on the Open Science Framework at <https://osf.io/s2389/>. Our stage 1 manuscript, and a supplement outlining the changes between our stage 1 and stage 2 manuscripts, can also be found there.

Experiment 1a

Experiment 1a was designed to replicate the linguistic concreteness effect (Hansen & Wänke, 2010, Experiment 1). Participants judged the truth of trivia items and we assessed whether, in the absence of self-reported knowledge of the correct answer, their judgements were influenced by the concreteness of the wording.

Method

Our replication follows the procedures of the original paper and uses the original materials provided by the authors (translated from the original German wording). In consultation with J. Hansen (personal communication, January 25, March 01, April 09, and April 16, 2018), we further adapted those materials to our participant populations in order to test the same hypotheses as the original (see below). Differences between this experiment and the original are outlined in the “Known Differences from the Original Study” section below. The experimental procedures were approved by both the Kingston

University Research Ethics Committee and the University of Illinois Institutional Review Board. Participants provided informed consent before participating.

Sampling Plan. There is no clear theoretical lowest effect size of interest for the linguistic concreteness effect that we can use as the basis of a power analysis. As an alternative, we could use the effect size from the original study for power analysis, but that effect size might not reflect the “true” effect due to chance variation, sampling error, and the possibility of publication bias. Consequently, we conducted a sensitivity power analysis using G*Power 3.1.7 (Faul, Erdfelder, Lang, & Buchner, 2007) to determine the smallest effect that we would have high statistical power (95%) to detect given pragmatic constraints on our total sample size. Our preregistered plan was to collect usable data from 210 participants (five-times the original sample size), which would give our study 95% power to detect an effect size of $d_z = 0.228$ at $\alpha = .05$ (one-sided). Hansen and Wänke (2010) reported effect sizes (η_p^2) of .19 and .081 (for Experiments 1 and 2 respectively), which correspond to Cohen’s d_z of 0.477 and 0.292 (Lakens, 2013). Given that both reported effects are larger than $d_z = 0.228$, our planned sample had greater than 95% power to detect the originally reported effects as well (with our sample size, we have greater than 95% power to detect an effect that is 50% the size of the original Experiment 1).

Participants. Undergraduate students (and some masters students in the UK) participated in the study in exchange for course credit or a chance to win one of three £50 prizes. These incentives were used in Hansen and Wänke’s (2010) Experiment 1 and Experiment 2, respectively. Participants were recruited via a dual-site collaboration enabled by StudySwap (Chartier & McCarthy, 2018); approximately half the participants were from Kingston University, UK and half from the University of Illinois at Champaign-Urbana, USA. For recruiting purposes, and in line with the original study, the experiment was described as a “study on truth judgements.” Our final sample was larger than our target sample due to higher signup rates and lower no-show rates than anticipated during scheduling (UK: $n = 130$, $M_{age} = 24.7$; USA: $n = 123$, $M_{age} = 19.3$).

Materials. Two native German speakers translated the original 52 trivia statements from German to English. These items cover a myriad of general knowledge topics including history, geography, and science. Half of the statements are true and half are false. All statements are plausible but describe facts that few participants know. Each trivia item has both an abstract and a concrete version, with concreteness determined using linguistic category model criteria (Semin & Fiedler, 1988, 1991). For example, in the first statement in Table 1, “wrote” is more concrete (i.e., a descriptive action verb) than “is by.” To maximise the chance of observing the linguistic concreteness effect, we took care to ensure that the

English translations complied with the description of the original items. For each statement: 1) the concrete version contains a descriptive action verb; 2) both versions were approximately the same length; and 3) the abstract and concrete versions used equally common language because any unusual words were common to both versions of each statement (e.g., words like “bandoneon” were core to the content of the statement). The translation was checked by the first author of the original paper.

Updated statements.

Hansen and Wänke (2010) argued that the match between concreteness and psychological/physical distance also influences truth judgements. In their Experiment 4, concretely worded items presented in the foreground of a landscape photograph (i.e., close) were judged to be more true than those presented in the background. Similarly, abstract items presented in the background were judged to be more true than those presented in the foreground. These effects presumably result from a match in the participant’s mindset: both physical proximity and linguistic concreteness activate a more “concrete” mindset which increases judged truth values. A mismatch in those factors reduces truth judgements. In reviewing our replication plan, Hansen suggested that the content of some original items might induce a similar “distance” effect. In the original experiment, some of the statements related to culture and history local to Switzerland, and those statements might be more psychologically distant for a Briton or an American. That distance might interact with the linguistic concreteness effect.¹

The original experiment was conducted at a university in Switzerland. The first author coded each statement as being either spatially close, distant, or neutral from Switzerland, the UK, and the USA, and these judgements were checked by the first author of the original study. We then generated additional trivia items (modelled on the originals) for those deemed close for Swiss participants but far for Britons (8 items) or Americans (18 items).² Thus, participants in the UK judged a total of 60 items and USA participants judged 70 items (see Table 2). The new statements were modified versions of the original items created by swapping words that conveyed spatial distance for our participants for equivalent spatially close words while maintaining the concreteness/abstractness of the original item. For example, we changed “In Hamburg, one can count the largest number of bridges in Europe” to “In London you can count the largest number of surveillance cameras in Europe.” We did not change the actual truth of

¹We analyzed the original data from Hansen and Wänke’s (2010) Experiment 1 and did not observe the predicted effect of distance.

²One statement (about Swiss Cantons) that was likely not understandable for UK and USA participants was amended and remains in the original 52.

Table 1
Examples of Trivia Statements with Abstract and Concrete Phrasings.

Statement	Concrete	Abstract
1	The poet C. Dickens wrote the play Miss Sara Sampson.	The play Miss Sara Sampson is by the poet C. Dickens.
2	The Roe River flows into the Missouri River.	The Roe River is a tributary of the Missouri River.
3	People nicknamed the Cuban composer Esteban Salas y Castro the "Santiago Angel".	The Cuban composer Esteban Salas y Castro was also known as the "Santiago Angel".

Note: Statements 1 and 3 are false: The author of Miss Sara Sampson is Gotthold Ephraim Lessing. Esteban Salas y Castro had no well-known nickname.

the new statements (i.e., if the original statement was true the replacement was also true). The statements, modifications, and plans for confirmatory analyses were discussed with the first author of the original paper. Confirmatory analyses were carried out on both the 52 original statements, and the updated version containing 52 statements in which distant statements have been removed and replaced with close statements. Planned secondary analyses explored whether the linguistic concreteness effect differed for the matched subsets of original and replacement items (8 for Britons and 18 for Americans).

Table 2
Psychological distance of the original items in each country (with numbers of replaced items).

Stimuli	Spatial psychological distance		
	Close	Distant	Neutral
Swiss Original Study			
Original items	20	12	20
UK Replication			
Original items	12	20	20
Changed	8	-8	0
USA Replication			
Original items	3	30	19
Changed	17	-18	1

Note. Original = original 52 statements; Changed = number of added or removed statements needed to match the proportions in original set.

Statement verification.

Before conducting the study, we followed the same procedures used by Hansen and Wänke (2010) to ensure that the concrete versions of the statements were seen as more concrete than were the abstract ones. We combined all trivia items into a single set of 78 (52 original + 18 USA-specific items + 8 Britain-specific items), and then created two sets of 78 items (set A and set B) so that the concrete and abstract version of each item appeared in different sets. Four student raters (2 for set A and 2 for set B), who were blind to the experimental hypothesis and who were briefly trained on the pertinent aspects of the linguistic category model (see <https://osf.io/s2389/> for complete training instructions) then independently coded each

item on a 1 (*most concrete*) to 4 (*most abstract*) scale. For set A, the correlation between raters was $r = .77$. For set B, the correlation between raters was $r = .81$. As in the original experiment, concrete versions were consistently coded as more concrete than their corresponding abstract versions (see Table 3).

In the experiment, the statements were presented in the same two sets (A and B), and in same order as in the original experiment, with the new items randomly interspersed among them (we used <https://www.randomizer.org/> to allocate positions). If a new version of a statement was assigned to a position within five places of the corresponding original statement, it was re-randomised. In each set, half of the statements are actually true and half are false. Each trivia item appears only once in each set, in either its abstract or concrete form; statements presented as concrete in set A were presented as abstract in set B, and vice versa. The concreteness and actual truth of the statements were fully crossed.³

In the original study the statements were presented across four pages, with the following number of statements on each page: 15 (including instructions), 17, 17, 3. We standardised the number of statements presented across the paper-and-pencil (Experiment 1a) and online formats (Experiment 1b). The first page presented the instructions and four statements; each page thereafter contained six statements (except that the last page in the UK set contained two statements). The UK set consisted of 11 pages and the USA set consisted of 12 pages.⁴

Procedure. The experiment followed the procedure used by Hansen and Wänke (2010), including directly translated

³Note that in the original study, sets A and B had unequal numbers of concrete and abstract versions of the items (the design was not fully crossed between truth value and concreteness). To fully cross the factors in the replication, we swapped the concrete and abstract versions of two of the original items between set A and set B.

⁴Due to a copy/paste error in creating the printed packets for the USA versions, the item numbering was out of sequence (... 38, 39, 40, 44, 45, 46, 41, 42, 43, 47, 48 ...). We noticed the error after testing had already started, so we did not change it for the remaining participants. The sequence was correct for the USA online version and for both laboratory and online UK versions.

Table 3

Coder concreteness ratings for concrete and abstract statements, their difference, and the confidence interval around that difference

Statements	n	Concrete M(SD)	Abstract M(SD)	Diff	95% CI
Overall	78	1.67 (0.69)	3.27 (0.59)	-1.60	[-1.81, -1.40]
SetA	78	1.75 (0.75)	3.28 (0.60)	-1.53	[-1.85, -1.22]
SetB	78	1.59 (0.62)	3.26 (0.60)	-1.67	[-1.96, -1.39]
True	39	1.77 (0.81)	3.36 (0.57)	-1.59	[-1.91, -1.28]
False	39	1.57 (0.53)	3.19 (0.61)	-1.61	[-1.89, -1.34]
Old	52	1.78 (0.72)	3.27 (0.62)	-1.49	[-1.74, -1.23]
New	26	1.45 (0.57)	3.28 (0.56)	-1.83	[-2.19, -1.48]

instructions. It was administered as a paper-and-pencil questionnaire study to students enrolled in introductory psychology and other undergraduate and masters psychology classes. The experiment was conducted in classrooms. Participants were given one of the two versions of the questionnaire (set A or set B) containing 60 (UK) or 70 (USA) statements in a fixed random order. Questionnaire packs were distributed to participants in each sample in alternating order to ensure that approximately equal numbers of participants received each set. In each set, half the statements were actually true and half were false, and for each actual truth value, half the statements were abstract and half concrete. Items that were concrete in Set A were abstract in Set B, and vice versa. Participants were asked to judge the truth of each statement on a scale ranging from 1 (*definitely false*) to 6 (*definitely true*; Hansen & Wänke, 2010, p. 1579). In short, English-speaking participants at each testing site were randomly assigned to a 2 (concreteness of statements: concrete vs. abstract) x 2 (actual truth: true vs. false) x 2 (statement set: set A vs. set B) mixed design with the first two factors varied within participants and the last factor varied between participants.

In Experiment 4 of Hansen and Wänke (2010), which used a subset of these statements, the authors inferred from the pattern of responses that a few participants knew the answers to some items. We added a check for prior knowledge to ensure that ratings were of items with unknown truth value. After completing all truth judgements, participants viewed the list of items again, and indicated next to each item if they knew the answer to that item. After completing the trivia items and the knowledge check, participants reported their age, gender, nationality, the number of years they had lived in the UK/USA, and whether they had used any sources to find out answers to any of the items. Finally, participants were thanked and debriefed. The experimental tasks were self-paced and took approximately 10-20 minutes to complete. The experimenter remained in the room for the duration of the experiment. Given that successful recruiting from the subject pool in the USA required a longer testing session (approximately 40 minutes was needed to receive a full credit), most participants in the USA completed an additional packet

of questionnaires following completion of the tasks for this study (see online supplement for more information).

The experimental data were entered into spreadsheets. The UK data files were verified by re-entering all numbers and cross-checking discrepancies. The USA data files were verified by reading aloud the entered numbers from the spreadsheet while an assistant verified that they matched the responses in the packets. Any entirely ambiguous responses (e.g., two numbers marked) were coded as missing. These verified data files are stored on OSF along with the data from Experiment 1b.

Results

Analysis scripts were generated from pilot data that was created by having the first author repeatedly complete each survey herself (varying her responses to questions to allow tests of various exclusion rules). All analyses were written using R (R Core Team, 2018) and the following packages: tidyverse (Wickham, 2017), janitor (Firke, 2018), datatable (Dowle & Srinivasan, 2018), varhandle (Mahmoudian, 2018), ez (Lawrence, 2016), BayesFactor (Morey & Rouder, 2018), summarytools (Comtois, 2018), and bootES (Gerlanc & Kirby, 2015). This manuscript was written in RMarkdown (Allaire et al., 2018) and formatted using papaja (Aust & Barth, 2018), knitr (Xie, 2015), kableExtra (Zhu, 2018), and xtable (Dahl, 2016). The RMarkdown file includes the full analysis script and results are analyzed and inserted into the manuscript without human intervention. The scripts, data, and RMarkdown files are available at <https://osf.io/s2389/>. Unless explicitly noted otherwise, all exclusion rules and analyses followed the pre-registered plan specified in our stage 1 manuscript.

For the primary analyses, data were pooled across country (UK and USA) and across set (A and B). The original study excluded no participants. We excluded responses to any items that were already known by a participant (as indicated by checking the box next to that item in the knowledge check), regardless of whether their actual answer was correct or incorrect. We excluded data from any participant who elected

to end their participation prior to completing the study ($n=4$), who self-reported using technological aids to answer questions ($n=2$), or who responded uniformly (e.g., always answer 1) to all statements in either the original 52 items or the new set of 52 items ($n=0$).

In addition to the preregistered exclusion criteria, we excluded participants who reported knowing 59 or more items ($n=1$) because they could not be included in the primary analyses after excluding “known” items (see Table 4). Finally, we did not enter data from one additional USA participant who the experimenter observed marking responses in a pattern (1-2-3-4-5-6-5-4-3-2-1, etc.) without reading the items.

For both Experiment 1a and 1b, as in the original paper, our primary, confirmatory analyses examined the effect of concreteness of language on the perceived truth of trivia statements, with the six-point Likert ratings as the dependent measure. The linguistic concreteness effect predicts that Likert scores should be higher for concretely worded statements than for more abstractly worded statements. We separately computed each participant’s mean rating across items falling into each combination of the truth of the statement (true/false) and the concreteness of the statement (concrete/abstract). Our confirmatory hypothesis tests were based on the data after exclusions and after removing any items that participants reported having known previously, and the online supplement presents further exploratory analyses including items that participants reported knowing already.

Primary confirmatory analyses. The original study used a mixed-design ANOVA to analyse the effects of concreteness, actual truth, and set. Given that we had no a-priori hypotheses about actual truth or set, we did not use an ANOVA for our confirmatory hypothesis test. For completeness, we report the results of a comparable ANOVA (adding country as a factor) in the online supplementary materials at <https://osf.io/s2389/>.

As a test of the linguistic concreteness effect, we directly compared the average responses to concrete and abstract statements in a paired, one-sided t-test for the original 52 items (H1). Average ratings for concrete items ($M = 3.57$, $SD = 0.41$) were about the same as those for abstract items ($M = 3.54$, $SD = 0.41$), $t(245) = 1.21$, $p = .115$, $BF_{10} = 0.29$ (The Bayes Factor used $r_{scale} = 0.336$, the d_r effect size for the original study, as an informed alternative hypothesis.) The Bayes Factor shows that our observed difference is 3.45 times more consistent with the null hypothesis of no difference or a negative effect than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen’s $d_z=0.08$; 95% CI: [-0.03, 0.18]) to the criterion value from our sensitivity power analysis (Cohen’s $d_z = 0.228$) to determine whether the observed ef-

Table 4
Participants recruited, excluded, and analyzed, separated by country, set, and gender for Experiment 1a

Group	N recruited	N excluded	N analyzed
UK Set A			
Male	8	0	8
Female	55	2	53
Gender Variant	0	0	0
Not Reported	1	1	0
UK Set B			
Male	11	1	10
Female	52	0	52
Gender Variant	1	0	1
Not Reported	2	2	0
USA Set A			
Male	17	1	16
Female	43	0	43
Gender Variant	1	0	1
Not Reported	1	0	1
USA Set B			
Male	30	0	30
Female	30	0	30
Gender Variant	0	0	0
Not Reported	1	0	1
Total	253	7	246

Note. Recruited includes all participants who started the study, even if they did not complete it.

fect was “inferior” to that planned minimum effect (Lakens, Scheel, & Isager, 2018). Because the upper bound of the confidence interval was smaller than 0.228, the observed difference between truth ratings for the concrete and abstract statements was statistically inferior to a positive effect of Cohen’s $d_z = 0.228$.

The same analysis conducted on the revised set of 52 items – replacing items that were close for the Swiss participants in the original study with new items that were close for the UK or USA participants (H1) – revealed a pattern that was similar to that for the original 52 items: Average ratings for concrete items ($M = 3.58$, $SD = 0.40$) were again about the same as those for abstract items ($M = 3.55$, $SD = 0.40$), $t(245) = 1.60$, $p = .056$, BF_{10} (with $r_{scale} = 0.336$) = 0.49. The Bayes Factor shows that our observed difference is 2.06 times more consistent with the null hypothesis of no difference or a negative effect than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen’s $d_z=0.10$; 95% CI: [0.00, 0.20]) to the criterion value from our sensitivity power analysis (Cohen’s $d_z = 0.228$) to determine whether the observed effect was “inferior” to that planned minimum effect. Because

the upper bound of the confidence interval was smaller than 0.228, the observed difference between truth ratings for the concrete and abstract statements in the revised set of items was statistically inferior to a positive effect of Cohen's $d_z = 0.228$.

Secondary exploratory analyses. Hansen and Wänke (2010) found that physical distance moderated the linguistic concreteness effect (Experiment 4). In their study, items were displayed against a photographic background so that they appeared either near or far. Concrete items were judged to be more true when they were close and abstract items were judged to be more true when they were far. In consulting with Hansen about the design of our replication, he suggested a conceptual replication of that effect based on the geographic proximity of the item contents to our participants. That suggestion motivated the addition of the new items, but it also permits a conceptual replication of the proximity effect. We compared truth ratings for the original “distant” versions of statements (those judged to be geographically “close” for Swiss participants but remote for participants in the UK or USA) with the new replacements for those items (8 original and updated items for the UK, and 17 for the USA; in the USA, one additional close item was replaced by a neutral item to ensure a fully crossed design with a total of 18 new items) that were intended to be “close” for our participants (see Table 5). For close items, the difference between concrete and abstract should be positive, because of the conceptual “match” between concrete and close and the mismatch between abstract and close. In contrast, for distant items, the difference between concrete and abstract should be negative, because of the conceptual “mismatch” between concrete and distant and the match between abstract and distant. Consequently, we compared difference scores (*Concrete – Abstract*) between the original (distant) and replacement (close) items with a one-sided t-test (H2).

Partially consistent with the prediction that a match between proximity and concreteness would increase truth judgements, the difference between concrete and abstract was positive for the close items ($M=0.06$), but it was also positive for the distant items ($M=0.02$), and near zero in both cases, $t(245) = 0.67$, $p = .253$.

Experiment 1b

The research reported in Hansen and Wänke (2010) tested undergraduate participants in person using paper-and-pencil materials. This extension attempted to replicate the linguistic concreteness effect using the same materials as Experiment 1a but in an online setting.

A growing literature suggests that people process online material more superficially, relying on heuristics to judge message

credibility (Metzger & Flanagin, 2013; Sundar, Knobloch-Westerwick, & Hastall, 2007) and believability (Sungur, Hartmann, & Koningsbruggen, 2016). If so, we might expect to observe a larger linguistic concreteness effect online. Conversely, a recent meta-analysis of studies of the illusory truth effect (Dechêne et al., 2010) showed a reduction in effect size online; when judgements of a set of repeated statements were compared to judgements of new statements (between-items), the effect size was reduced from $d = .59$ using paper-and-pencil to $d = .30$ on the computer. The reasons for this reduction are unclear, but the authors suggested it might be due to differences in presentation time (i.e., constrained intervals or participant paced) or presentation appearance (i.e., how many statements are presented at once). Given that Experiment 1b samples from a different population using a different medium, differences in absolute performance levels and the size of the concreteness effect could differ between Experiments 1a and 1b for many reasons. Hence, rather than directly comparing the effect sizes in the two studies, we report whether the linguistic concreteness effect emerges in each study relative to the same standard set by our sensitivity analysis.

Method

Participants. As for Experiment 1a, our plan was to continue recruiting participants until we had usable data from 210 participants, with approximately half from the USA and half from the UK. Participants were recruited and tested online using the Prolific platform and Qualtrics. We used Prolific's pre-screening to ensure that participants were between 18 and 65 years of age, listed English as their first language, and had a “participation on Prolific” approval rating of 98% or higher (Final sample: UK: $n = 120$, $M_{age} = 34.3$; USA: $n = 117$, $M_{age} = 33.2$). The experimental procedure was approved by the Kingston University Research Ethics Committee, and participants provided informed consent before completing the study. Each participant was randomly assigned to set A or set B, and as in Experiment 1a, they completed equal numbers of items in each cell of a design that fully crossed concrete/abstract and true/false. Upon completion of the experiment, participants received £2.18 as compensation.

Materials and procedure. Except as noted, the materials and procedure matched those used in Experiment 1a. To ensure that the formatting, font size, and number of statements on each page were the same between Experiments 1a and 1b, we created the Qualtrics survey used in Experiment 1b first and produced the paper-and-pencil version from that version. To promote consistency in the appearance of the items, we constrained the study to allow participation only via a desktop or laptop computer (rather than a handheld device). At the end of the experiment, participants reported the type of device they used to complete the survey and whether or not they used

Table 5
Means and SDs for new items (close) and distant items (replaced original) for Experiments 1a and 1b.

Item type	Concrete M(SD)	Abstract M(SD)	Diff	r
Experiment 1a				
Close	3.64 (0.62)	3.58 (0.56)	0.06 (0.72)	0.26
Distant	3.57 (0.56)	3.55 (0.54)	0.02 (0.64)	0.32
Experiment 1b				
Close	3.59 (0.59)	3.55 (0.62)	0.05 (0.79)	0.16
Distant	3.60 (0.60)	3.54 (0.57)	0.06 (0.61)	0.46

Table 6
Participants recruited, excluded, and analyzed, separated by country, set, and gender for Experiment 1b

Group	N recruited	N excluded	N analyzed
UK Set A			
Male	22	0	22
Female	37	1	36
Gender Variant	1	0	1
Not Reported	0	0	0
UK Set B			
Male	23	1	22
Female	37	5	32
Gender Variant	0	0	0
Not Reported	0	0	0
USA Set A			
Male	19	4	15
Female	37	2	35
Gender Variant	1	0	1
Not Reported	0	0	0
USA Set B			
Male	33	2	31
Female	27	2	25
Gender Variant	0	0	0
Not Reported	0	0	0
Total	237	17	220

Note. Recruited includes all participants who started the study, even if they did not complete it.

any technology to aid their responses. The UK survey can be viewed at <https://bit.ly/2NrUKmc>, and the USA survey can be viewed at <https://bit.ly/2PLgrPF>.

Results

The planned data analysis and exclusion rules were identical to those of Experiment 1a, with an added criterion to account for overly fast or slow completion of the study in the absence of an experimenter observing data collection in person. We set the “maximum time allowed” to 45 minutes within the Prolific settings, and we also excluded participants who completed the study in less than 3 minutes.⁵ We excluded data

from any participant who elected to end their participation prior to completing the study ($n=0$), who self-reported using technological aids to answer questions ($n=9$), who responded uniformly (e.g., always answer 1) to all statements in either the original 52 items or the new set of 52 items ($n=1$), or who reported knowing 59 or more items ($n=1$) because they could not be included in the primary analyses after excluding “known” items.

Given that online participants could cheat by looking up the answers, and that we could not identify overly long response times to individual questions using Qualtrics, we used the data from Experiment 1a to establish a plausible accuracy level (because participants in Experiment 1a could not easily cheat in answering questions). We calculated the mean number of questions that each participant correctly answered in Experiment 1a, where we operationally defined a correct answer as a response of 1 (*definitely false*) when the statement was false and 6 (*definitely true*) when the statement was true. We excluded any participant in Experiment 1b whose percentage correct according to that same standard was more than three standard deviations above the mean from Study 1a (Experiment 1a $M = 0.08$, $3SD$ cutoff = 0.34; total excluded $n=9$; note, though, that 3 of those participants had already been excluded for self-reported use of technological aids).

Given that we anticipated needing to replace some excluded participants, we initially collected data from 240 participants, with the plan to test additional batches of 20 participants as needed until we achieved final sample with usable data from at least 210 participants (see Table 6).

Primary confirmatory analyses. As in Experiment 1a, we compared the average responses to concrete and abstract statements in a paired, one-sided t-test for the original 52 items (H1). Average ratings for concrete items ($M = 3.66$, $SD = 0.42$) were about the same as those for abstract items ($M = 3.63$, $SD = 0.38$), $t(219) = 1.61$, $p = .055$, BF_{10} (with

⁵The first author was able to complete the survey in approximately 2 minutes when responding randomly to all items and neglecting to read the instructions. In pilot testing of the online version of the study (Experiment 1b), no participant completed the study in less than 5 minutes.

$r_{scale} = 0.336) = 0.51$. The Bayes Factor shows that our observed difference is roughly equally consistent with the null hypothesis of no difference as with a distribution centred at the original effect size; it does not favour either hypothesis over the other by more than a 2:1 ratio (although it is 1.95 times more consistent with the null than the alternative).

Given that the t-test was not statistically significant, we compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.11$; 95% CI: [-0.01, 0.22]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.228$) to determine whether the observed effect was "inferior" to that planned minimum effect (Lakens et al., 2018). Because the upper bound of the confidence interval was smaller than 0.228, the observed difference between truth ratings for the concrete and abstract statements was statistically inferior to a positive effect of Cohen's $d_z = 0.228$.

The same analysis conducted on the revised set of 52 items – replacing items that were close for Swiss participants with new items that were close for the UK or USA participants (H1) — revealed a pattern that was similar to that for the original 52 items: Average ratings for concrete items ($M = 3.65$, $SD = 0.42$) were again about the same as those for abstract items ($M = 3.63$, $SD = 0.39$), $t(219) = 0.95$, $p = .170$, BF_{10} (with $r_{scale} = 0.336) = 0.24$. The Bayes Factor shows that our observed difference is 4.23 times more consistent with the null hypothesis of no difference than with a distribution centred at the original effect size.

Given that the t-test was not statistically significant, we again compared the upper confidence bound around the observed effect (observed effect: Cohen's $d_z = 0.06$; 95% CI: [-0.05, 0.17]) to the criterion value from our sensitivity power analysis (Cohen's $d_z = 0.228$) to determine whether the observed effect was "inferior" to that planned minimum effect. Because the upper bound of the confidence interval was smaller than 0.228, the observed difference between truth ratings for the concrete and abstract statements in the revised set of items was statistically inferior to a positive effect of Cohen's $d_z = 0.228$.

Secondary exploratory analyses. As in Experiment 1a, we tested whether a match between proximity and concreteness increased truth ratings by comparing difference scores (*Concrete – Abstract*) between the original (distant) and replacement (close) items (H2). Partially consistent with the prediction that a match between proximity and concreteness would increase truth judgements, the difference between concrete and abstract was positive for the close items ($M = 0.05$), but it was also positive for the distant items ($M = 0.06$), and near zero in both cases, $t(219) = -0.26$, $p = .603$.

Known Differences from the Original Study

The instructions, measures, and procedures were adapted directly from those of the original study. The original study was conducted in German at the University of Basel in Switzerland, whereas our study was conducted in English at universities in the UK and USA. The first author of the original study reviewed the translated statements and agreed that the procedures should work with our populations. Upon realising that truth value and concreteness were not fully crossed in the original study design, we exchanged the concrete and abstract versions of two items across sets A and B to ensure that each set had equal number of items for each combination of true/false and concrete/abstract. Our primary analysis combined across sets, and there is no theoretical reason to expect this change to affect the outcome. Participants in the original study were all undergraduate psychology students who received course credit. Our sample in the USA also consisted of undergraduate psychology students who received course credit or extra credit for their participation. Our sample in the UK was composed of undergraduates from psychology and also included some masters students. For the UK sample, participants had a chance to win one of three £50 prizes rather than receiving course credit. This compensation was commensurate with that used by the original authors in their Experiment 2 which tested the same hypothesis and used the same materials as Experiment 1 (Hansen & Wänke, 2010, p. 1580). We added a check to ensure that participants did not actually know the answers to any questions (see Procedure section).

We included additional, culturally-aligned trivia items to the study (see Materials section). Our participants therefore completed 60 (UK) or 70 (USA) statements rather than 52 in the original study. The Qualtrics platform constrained the presentation format of the statements resulting in more white space between statements than in the original questionnaire. The number of statements presented on each page was identical for our paper-and-pencil and online formats, and differed from the original study (see Materials section). We discussed these changes in advance with the first author of the original paper, and neither we nor they expected these changes to affect the outcome.

In experiment 1b, data collection occurred online rather than using the paper-and-pencil format of the original study.

Discussion

In Experiment 1a we attempted to replicate the linguistic concreteness effect from Experiment 1 of Hansen and Wänke (2010) in which participants judged concretely worded trivia items as more probably true than abstractly worded versions (H1). Concrete items were not rated as significantly truer

than abstract items for either the original items or the revised set of items, which is inconsistent with the original study. The Bayes Factor for the original set favoured the null - a distribution centred at no effect — over a distribution centred at the original effect size by a 3.45:1 ratio. For the revised set, it favoured the null by a ratio of 2.06:1. For the original items, the upper bound of the confidence interval around the effect was smaller than our smallest effect of interest, and therefore also smaller than the original effect size, meaning that the data were inconsistent with the original finding. Similarly, for the revised items, the data were inconsistent with the original finding. Collectively, these results do not provide evidence for a linguistic concreteness effect on truth judgements.

In Experiment 1b, we extended our test of the linguistic concreteness effect to an online sample. Inconsistent with the original study, concrete items were not rated as significantly truer than abstract items for either the original items or the revised set of items. The Bayes Factor for the original set favoured the null over a distribution centred at the original effect size by a 1.95:1 ratio. For the revised set, it favoured the null by a ratio of 4.23:1. For the original items, the upper bound of the confidence interval around the effect was smaller than our smallest effect of interest, and therefore also smaller than the original effect size, meaning that the data were inconsistent with the original finding. Similarly, for the revised items, the data were inconsistent with the original finding. Collectively, these results do not provide evidence for a linguistic concreteness effect on truth judgments.

In designing these replications, we consulted the first author of the original study to ensure that our replication matched the procedures necessary to test the original hypothesis and to verify that any changes were consistent with the original conceptualization of the hypothesis. Still, by necessity, some aspects of the design differed between the original study and our replication attempt, and those differences might contribute to the different outcome.

First, our study used English rather than German materials. Although the change in language might contribute, neither we nor Hansen suggested theoretical reasons why translated materials would be ineffective in producing the effect. Indeed Hansen and Wänke's (2010) manipulation was based on the linguistic category model (Semin & Fiedler, 1988, 1991) which was developed based on experiments with English-speaking participants.

Second, in developing the protocol, Hansen suggested the possibility that perceived psychological distance might interact with the experimental manipulation (H2). Consequently, we added additional trivia items intended to match the "distances" of those items for our participants to the distance of the items for the original Swiss participants. Our study showed no effect of this distance manipulation; the close-distant effect

was close to zero and numerically in the opposite direction to the prediction. Although it is possible that adding more trivia items to the original set of 52 might dampen the effect, we saw no difference in the pattern of results for the UK participants (60 items) and USA participants (70 items) in either study. If testing language, perceived proximity, or number of items explain the different patterns of results between our studies and the original study, then the effect might be specific to theoretically uninteresting aspects of the testing context.

Our use of a 6-item response scale maximized the chances of observing an effect because it lacked a neutral mid-point; participants were forced to lean toward true or false for each statement. Consequently, even a small linguistic concreteness effect should nudge participants to make the appropriate directional response, leading to a measurable difference. Using a scale with a neutral midpoint (e.g., 4 on a 1-7 scale) would allow participants to ignore a slight sense of truth or falsity.⁶ Future research could consider using a scale with a neutral midpoint. Future studies would also need a substantially larger sample size in order to have adequate sensitivity to measure a much smaller effect.

The aim of the present studies was to accumulate evidence for the reliability of the linguistic concreteness effect and provide a robust estimate of its size for use in subsequent studies. Our experimental design and analyses were planned to optimise the chances of observing the effect: In Experiment 1a we collected data in a setting comparable to that of the original study and used paper/pencil materials matched as closely as possible to the original study. Experiment 1b adopted those materials for online testing with a broader population using Prolific. Each study had greater than 95% power to detect an effect half the size of the original, and each produced evidence more consistent with the absence of an effect than with the original effect. Across these two studies, our analyzed sample (466) was approximately ten times the size of the original study (n=46). Although no single study is definitive about the existence of an effect, our studies raise doubt about the reliability of using concrete/abstract language as a way to manipulate the judged truth of trivia statements.

⁶Hansen and Wänke (2010) reported no difference in average ratings for true and false items. Across our studies and conditions, a post-hoc analysis showed that true statements were rated slightly higher than false statements (less than 0.20 rating points on average), regardless of whether or not we excluded items that participants claimed to have known). This small difference is difficult to interpret, but it is consistent with a slight bias to respond on the larger end of the scale (toward true) coupled with some limited sense about the truth or falsity of items even when participants did not know the answer.

References

- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Chang, W. (2018). *Rmarkdown: Dynamic documents for r*. Retrieved from <https://CRAN.R-project.org/package=rmarkdown>
- Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
- Chartier, C. R., & McCarthy, R. J. (2018). StudySwap tutorial.
- Chou, H.-Y., & Yeh, M.-H. (2018). Minor language variations in campaign advertisements: The effects of pronoun use and message orientation on voter responses. *Electoral Studies, 51*, 58–71.
- Comtois, D. (2018). *Summarytools: Tools to quickly and neatly summarize data*. Retrieved from <https://CRAN.R-project.org/package=summarytools>
- Dahl, D. B. (2016). *Xtable: Export tables to latex or html*. Retrieved from <https://CRAN.R-project.org/package=xtable>
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review, 14*(2), 238–257.
- Dowle, M., & Srinivasan, A. (2018). *Data.table: Extension of 'data.frame'*. Retrieved from <https://CRAN.R-project.org/package=data.table>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144*(5), 993.
- Firke, S. (2018). *Janitor: Simple tools for examining and cleaning dirty data*. Retrieved from <https://CRAN.R-project.org/package=janitor>
- Gerlanc, D., & Kirby, K. (2015). *BootES: Bootstrap effect sizes*. Retrieved from <https://CRAN.R-project.org/package=bootES>
- Hansen, J., & Wänke, M. (2010). Truth from language and truth from fit: The impact of linguistic concreteness and level of construal on subjective truth. *Personality and Social Psychology Bulletin, 36*(11), 1576–1588.
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*(1), 107–112.
- Koriat, A., & Adiv, S. (2012). Confidence in one's social beliefs: Implications for belief justification. *Consciousness and Cognition, 21*(4), 1599–1616.
- Kurinec, C. A., & Weaver III, C. A. (2018). Do memory-focused jury instructions moderate the influence of eyewitness word choice? *Applied Psychology in Criminal Justice, 14*(1), 55–69.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. *Frontiers in Psychology, 4*, 863.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 2515245918770963*.
- Lawrence, M. A. (2016). *Ez: Easy analysis and visualization of factorial experiments*. Retrieved from <https://CRAN.R-project.org/package=ez>
- Mahmoudian, M. (2018). *Varhandle: Functions for robust variable handling*. Retrieved from <https://CRAN.R-project.org/package=varhandle>
- Menegatti, M., & Rubini, M. (2013). Convincing similar and dissimilar others: The power of language abstraction in political communication. *Personality and Social Psychology Bulletin, 39*(5), 596–607.
- Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220.
- Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>
- Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review, 19*(5), 969–974.
- Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D. S., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, 41(5), 1337.

- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Schmid, J., & Fiedler, K. (1996). Language and implicit attributions in the nuremberg trials analyzing prosecutors' and defense attorneys' closing speeches. *Human Communication Research*, 22(3), 371–398.
- Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54(4), 558.
- Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European Review of Social Psychology*, 2(1), 1–30.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN*. Retrieved from <https://ssrn.com/abstract=2160588>
- Stott, P. (2011). Don't sound like a liar at work. *Newsday*. Retrieved from <https://www.newsday.com/news/new-york/don-t-sound-like-a-liar-at-work-1.3031372>
- Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology*, 58(3), 366–378.
- Sungur, H., Hartmann, T., & Koningsbruggen, G. M. van. (2016). Abstract mindsets increase believability of spatially distant online messages. *Frontiers in Psychology*, 7, 1056.
- Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from <https://CRAN.R-project.org/package=tidyverse>
- Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from <https://yihui.name/knitr/>
- Zhu, H. (2018). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from <https://CRAN.R-project.org/package=kableExtra>

NA

Appendix
Disclosures

Contributions

- Contributed to conception and design: ELH, FVT, DJS
- Contributed to acquisition of data: ELH, DJS
- Contributed to analysis and interpretation of data: ELH, DJS
- Drafted and/or revised the article: ELH, FVT, DJS
- Approved the submitted version for publication: ELH, FVT, DJS

Acknowledgements

We thank Jochim Hansen for timely and insightful help and support in developing the protocol, Chris Chambers for providing valuable comments on the study design and on an earlier draft of this report, Stefanie Schmeer and Inge Kimbell for help translating the materials, and Katherine Wood and Nick Fox for programming assistance. Thanks to Giles Henderson and Jordan Simons for assistance with verifying manually entered data from the lab versions of the study. Thanks also to Richard Morey for help with effect size estimates for the Bayes Factor calculations.

Funding

This work was supported by a Prolific Junior Researcher Grant and a PsyPAG Research Grant awarded to ELH.

Competing interests

The authors declare that they have no competing interests.

Additional files submitted with manuscript:

- Appendix S1: List of UK and USA statements and statement characteristics
- Supplemental analyses

Additional files at <https://osf.io/s2389/>:

- Qualtrics scripts for the online experiment (also linked within the manuscript)
- Experiment 1a information sheet and all four packets
- Full analysis R script
- Simulated data files used in developing the analysis script (and to produce the placeholder values in the Stage 1 Registered Report manuscript)
- Sample output files based on the simulated data illustrating the format that the actual output files will take after running the script on the real data.
- R Markdown file used to generate the manuscript (both Stage 1 and Stage 2)
- Provisionally accepted Stage 1 manuscript and preregistration
- Public data files (with potentially identifying demographics for Experiment 1a masked)
- Password protected versions of Experiment 1a data files with all demographics included
- Lab log with time-of-testing notes from Experiment 1a
- Supplemental analyses
- Additional tasks used in testing of the USA sample (after completion of the primary tasks)
- File documenting the changes between the stage 1 and stage 2 manuscript

Data accessibility statement

All the materials, scripts, data, and RMarkdown files are available at <https://osf.io/s2389/>.