Failures to replicate a key result of the selective accessibility theory of anchoring

Adam J. L. Harris, Fi B. N. Blower, Sophie A. Rodgers, Sandra Lagator, Elise Page, Adam Burton, Diana Urlichich, and Maarten Speekenbrink

University College London

Author Note

Adam J. L. Harris, Fi B. N. Blower, Sophie A. Rodgers Sandra Lagator, Elise Page, Adam Burton, Diana Urlichich and Maarten Speekenbrink, Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, United Kingdom.

We thank Thomas Mussweiler for helpful email correspondence and Jenni Rodd for discussions.

AJLH and MS conceived the project, and were involved in the design of all experiments. Additionally, FBNB and SAR ran and designed Experiments 1 and 2, SL and EP ran and designed Experiments 3 and 4, AB and DU ran and designed Experiment 5. MS programmed the studies and analysed the data. AJLH and MS wrote the manuscript.

All materials, data, and pre-registrations associated with this project can be found at https://osf.io/jtx34/. Please note that experiment numbers do not necessarily match those in

the manuscript. This is clarified both in footnotes within the manuscript, and within a 'read me' document in the zip file containing all the experimental tasks and data at the OSF link.

The research described here was previously presented as a poster at the 2018 meeting of the Society for Judgment and Decision Making in New Orleans (November, 2018).

Correspondence concerning this article should be addressed to Adam J. L. Harris (email: adam.harris@ucl.ac.uk)

Word count (main text including abstract, 'context of the research', and Footnotes, excluding Figures, Tables and their captions): 10,848.

Abstract

Numerical anchoring effects describe the assimilative effect of a previously presented number on subsequent numerical estimates. Such effects are robust and consequential. A number of different accounts have been proposed to explain these effects. What is currently unclear is under which situations different mechanisms play more or less critical roles. An extant test from the literature is proposed as a 'signature test' for the operation of selective accessibility mechanisms. Four experiments were conducted to ascertain the evidence for selective accessibility with this test, tests that subsequently failed. A fifth experiment employed a different methodology, and again failed to show evidence for selective accessibility. Subsequent discussion suggests that the robustness of anchoring effects is remarkable, but the theoretical basis for some previous tests of the selective accessibility account of anchoring is shaky, and we advise against its use in this capacity.

Failures to replicate a key result of the selective accessibility theory of anchoring

Tversky and Kahneman (1974) famously asked their participants to estimate the percentage of African countries in the United Nations. Before providing their estimate, participants were asked whether the percentage was larger or smaller than a number that was randomly produced by a wheel of fortune. Participants for whom the wheel produced a larger number estimated a higher percentage of African countries in the United Nations than did those for whom the wheel produced a smaller number. Tversky and Kahneman referred to this as an anchoring effect, which in this instance is clearly a bias, for a random number produced by a wheel of fortune should not, rationally, influence one's estimates. Since Tversky and Kahneman's seminal work, anchoring effects have been observed in many domains, across a myriad of areas of psychology. In the applied arena, these include the pricing of real estate by estate agents (Northcraft & Neale, 1987), sentencing decisions of judges (for a review see Englich, 2006), students' evaluations of course instructors (Thorsteinson, Breier, Atwell, Hamilton, & Privette, 2008), negotiations (Galinsky & Mussweiler, 2001), supermarket purchase decisions (Wansink, Kent, & Hoch, 1998), and the payment of credit card bills (Navarro-Martinez, Salisbury, Lemon, Stewart, Matthews, & Harris, 2011; Stewart, 2009). Theoretically, anchoring has been proposed as a potential mechanism underlying numerous phenomena, including: hindsight bias (Hawkins & Hastie, 1990), overconfidence (Block & Harper, 1991), preference reversals in choice (e.g., Lichtenstein & Slovic, 1971), and has even been proposed as an explanation for probability weighting functions elicited in decision-making experiments (Hogarth & Einhorn, 1990).

Numerous accounts have been put forward to explain the anchoring effect, including anchoring-and-adjustment (Tversky & Kahneman, 1994), numeric priming (Wilson, Houston, Etling, & Brekke, 1996; Wong & Kwong, 2000), magnitude priming (Oppenheimer,

LeBoeuf, & Brewer, 2008; see Sleeth-Keppler, 2013, for a related account), and scale distortion (Frederick & Mochon, 2012; Mochon & Frederick, 2013).

Selective Accessibility

Arguably the dominant account of standard anchoring effects (where judgments are assimilated to an externally provided standard), however, is that of selective accessibility (Mussweiler & Strack, 1999, 2000a, 2000b, 2001; Strack & Mussweiler, 1997; see also, Chapman & Johnson, 1994, 1999). On this account, when answering the initial comparative question (e.g., "Does a giraffe weigh less than 100 lbs."), participants engage in hypothesis-consistent search. For this particular question, this means that participants will initially recruit information consistent with a giraffe weighing less than 100 lbs. As a consequence of this hypothesis-consistent search, when answering the subsequent question asking for an estimate of the weight of a giraffe, information consistent with the giraffe weighing less than 100 lbs. will be more accessible in memory and hence estimates will be lower than if asked about a higher (e.g., 5,000 lbs.) anchor value. In the standard anchoring paradigm, participants are asked whether the target (e.g., weight of a giraffe) is *greater or less* than the anchor value (e.g., 100 lbs.). Mussweiler and Strack propose that the hypothesis that is tested in this case is that a giraffe's weight is *equal* to 100 lbs. The observation that anchoring effects are no different when either of these two comparative questions ('greater or less' versus 'equal to') are asked (Mussweiler & Strack, 1999) is supportive of this hypothesis. From our perspective, we see the close link between selective accessibility and fundamental cognitive processes (confirmatory hypothesis testing and semantic priming, Mussweiler & Strack, 1999; Strack & Mussweiler, 1997) as a very desirable property in its favour (see also Newell & Shanks, 2014). This does not, however, mean that selective accessibility underlies all anchoring effects.

Pluralism and the Present Aim

Discourse in the anchoring literature now tends to recognise the likelihood that multiple processes underlie anchoring effects (see e.g., Bahník, Englich, & Strack, 2017; Frederick & Mochon, 2012; Simmons, LeBoeuf, & Nelson, 2010). Whilst some suggest the simultaneous operation of multiple mechanisms (Chaxel, 2015; Simmons et al., 2010), others prefer the potential for different processes underlying different anchoring effects (Bahník et al., 2017; Frederick & Mochon, 2012). Whilst all accounts have been supported by data from experiments carefully designed specifically to test theoretical predictions, there is an open question of how to determine whether a particular process has generated any given anchoring effect. Alternatively, where an experiment supports, for example, scale distortion processes (Frederick & Mochon, 2012), how can one determine whether selective accessibility processes might *also* be playing a role? The question addressed in the present paper is therefore whether there is a signature test for selective accessibility that can be straightforwardly added to any extant demonstration of anchoring, and which does not require a complete experimental redesign.

Such a methodological test is important both for the development of effective debiasing interventions where desirable (targeting the appropriate anchoring mechanism), as well as theory development. The identification of situations in which different theories of anchoring do operate in parallel (Chaxel, 2015; Simmons et al., 2010), or accurate identification of situations that do or do not facilitate certain mechanisms, will extend current theoretical understanding of these effects, enabling the development of more complete models. We agree with Turner and Schley (2016, p. 2) that:

> "having several non-mutually exclusive theories is acceptable…when there are multiple unique cognitive processes involved in the decision and it is clear under which conditions each theoretical mechanism plays a more versus less critical role in

the decision process. We suggest that the anchoring literature lacks on this latter point…"

An example of the potential applicability of an appropriate 'signature test' comes from our own work. Harris and Speekenbrink (2016) provided a demonstration of anchoring occurring across two different response scales. For example, participants asked to estimate the weight of an elephant in tons, and then asked whether a giraffe weighed more or less than an elephant, subsequently provided higher estimates of a giraffe's weight in lbs. than they did in a control condition. Selective accessibility predicts such an effect, on the basis that information consistent with a giraffe being heavy is subsequently more accessible, which will influence estimates regardless of the scale used. Additionally, however, requesting participants to compare the weights of the elephant and the giraffe might prompt participants to employ an approximate conversion between the two response scales, triggering scale distortion processes. Whilst Harris and Speekenbrink undertook three further experiments, they still acknowledged the potential for multiple explanations for these results.

Inappropriate 'signature' tests for selective accessibility.

Support for selective accessibility stems from a variety of methodologies. The majority of them, however, require specific experimental designs and cannot, therefore, necessarily determine the operation of selective accessibility processes for any ambiguous anchor effect. Demonstrations that priming or prompting the consideration of anchor inconsistent knowledge reduces the size of the anchoring effect (Chapman & Johnson, 1999; Chaxel, 2015; Mussweiler, 2002; Mussweiler, Strack, & Pfeiffer, 2000) requires an amendment to the *set-up* of the anchoring experiment, thus potentially altering the original effect. In addition, whilst predicted by selective accessibility accounts, such a manipulation might additionally have an influence on an individual's subjective certainty in the direction of adjustment (see Simmons et al., 2010), or on the diffusion of the prior distribution of one's

estimate (see Turner & Schley, 2016). It is therefore not clear that selective accessibility is the only account that predicts such effects.

Early support for selective accessibility demonstrated the importance of the applicability of the anchor to the target. Strack and Mussweiler (1997) demonstrated that an anchor pertaining to the height of the Brandenburg Gate did not influence subsequent estimates of the width of the Brandenburg Gate (but see Frederick & Mochon, 2012, who did report an anchoring effect in such an instance). Whilst supportive, such a result cannot be used to demonstrate where selective accessibility effects *are* in operation. Note, also, that there is no a priori specification of what constitutes 'applicability' (between the anchor and the target). Harris and Speekenbrink (2016), for example, argued that an anchoring effect of an animal's weight on estimates of its height are consistent with selective accessibility, due to the strong correlation between height and weight.

Finally, support for selective accessibility has been obtained from response latencies for the provision of comparative and absolute estimates in light of plausible versus implausible anchors (Mussweiler and Strack, 1999, 2000b; Strack & Mussweiler, 1997). Critically, support for selective accessibility processes contributing to anchoring effects is obtained from the comparison of results from different types of anchors. Once again, therefore, this does not satisfy our requirements for a simple, direct signature test, which does not require altering the standard elements of the demonstration.

Selective accessibility of anchor consistent information.

The methodology we identify as the most likely candidate for a 'signature test' of selective accessibility specifically addresses the question of whether information consistent with the anchor value is *selectively* accessible following the presentation of the anchor. Support for this proposition has been obtained from a variety of methods, and has recently been labelled the most direct support for selective accessibility by some selective

accessibility proponents (Bahník et al., 2017, p. 233). Bahník and Strack (2016), for example, tested the prediction that an anchor would have no influence on estimates of a target which would already be predicted to recruit anchor consistent information, even in the absence of the anchor. Specifically, estimates of the mean summer temperature in New York City were not influenced by an anchor asking whether the *annual* temperature in New York City is greater or less than 102°F (a high anchor). The assumption is that participants recruited information consistent with the summer temperature of New York City when answering the comparative question, but they would have recruited this information to answer a question about the summer temperature of the city in any case[1]. By contrast, an anchoring effect was observed when the comparison question referred to the summer temperature in New York City (presumably now the anchor brings to mind particularly hot summer days).

The explanations for the results in Bahník and Strack (2016) contain a number of uses of the term 'presumably.' How can direct evidence for these explanations be obtained? Mussweiler and Strack (2000a) and Mussweiler and Strack (1999) obtained qualitative reports from their participants following a comparative question. The content of these reports was seen to assimilate towards the provided anchor values. For example, participants who compared the River Elbe with a large anchor, in a free-report task reported features coming to mind that were more consistent with long extensions of the River Elbe, than did those in a low anchor condition (Mussweiler & Strack, 1999). Whilst such methodologies provide support for the selective accessibility of information consistent with the anchor, they have not been widely used. One reason might be because of the subjectivity associated with coding

[1] Note that an anchoring effect was observed when the winter temperature was the target judgment. On the basis that the high anchor for the annual temperature is predicted to bring to mind summer-consistent information, such a result might be considered surprising. This relates to our query about the definition of applicability, raised in the previous section.

qualitative responses. In addition, the data provided in the focal methodology of this article is continuous in such a way as might provide subsequent stronger evidence for the selective accessibility hypothesis – a point to which we will return presently.

Assessing selective accessibility through lexical decisions (a potential 'signature test') – results to date.

Our target 'signature test' addresses the same question as those in the preceding section, namely, does an anchor selectively increase the accessibility of anchor-consistent knowledge? This is a key claim of the selective accessibility account, and a number of studies have used our proposed 'signature test' in support of this claim (Englich, Mussweiler, & Strack, 2006, Study 4; Mussweiler & Englich, 2005, Study 3; Mussweiler & Strack, 2000a, Studies 1 & 2). The key dependent variable in these studies is response latency in a categorisation task that is completed following the comparative judgment.[2] The more accessible information is, the faster it is predicted to be recognised (Mussweiler & Strack, 2001a). Thus, responses to anchor-relevant information are predicted to be quicker in conditions where the anchor is consistent with that information than when it is inconsistent.

Englich et al. (2006) did not use a Lexical Decision Task, but their method was conceptually similar enough to the one utilised in the current experiments that we discuss their results in this section. Englich et al. (2006) investigated legal experts' susceptibility to anchoring effects in criminal sentencing judgments. Consequently, their categorisation task was whether a statement pertaining to a shoplifting case they had already studied corresponded to an incriminating or an exculpatory argument. Receipt of a low sentence anchor was predicted to speed recognition, and therefore responses to, exculpatory statements

[2] In Mussweiler and Strack (2000a, Study 1) the categorisation task came after the absolute question, creating the confound that participants' final answers, rather than the anchor value may have been causing the increased accessibility of anchor-relevant words.

(based on enhanced cognitive accessibility following consideration of the low sentence anchor), whilst a high anchor should speed up recognition of incriminating statements. With a total sample of 57 legal experts, incriminating arguments were categorised faster following presentation of a high anchor than they were in the low anchor condition, $t(55) = 2.03$, $p = .047$, with no effect observed for exculpatory statements.[3]

Mussweiler and Englich (2005, Study 3) investigated the influence of anchor values presented subliminally. Participants were instructed to think about the average price of a new midsize car, whilst focussing on a 'flickering' nonsense letter string. The flickering was due to the subliminal presentation of either a high (40,000) or low (20,000) anchor value.[4] Subsequent to this, participants completed a Lexical Decision Task (LDT) in which they were presented with non-words, neutral (non car-related) words, words associated with expensive cars, and words associated with inexpensive cars. Mechanisms of selective accessibility afford the prediction that information consistent with expensive cars will be more accessible following the presentation of a high anchor, whilst information consistent with inexpensive cars will be more accessible following the presentation of a low anchor. Subsequently, LDT responses to expensive cars are predicted to be faster following a high anchor and responses to inexpensive cars are predicted to be faster following a low anchor. Directionally, the results matched these predictions and the reliability of the result was shown with a significant interaction, $F(1, 35) = 4.28$, $p = .046$ (N = 37).[5] Simple effects were not reported.

[3] Precise *p*-values for *t*-tests are calculated from the information provided in the original manuscript using the calculator at http://www.socscistatistics.com/pvalues/tdistribution.aspx.

[4] This study was run in Germany, before the introduction of the Euro currency, so that the currency in which participants were presumed to be thinking was German Marks.

[5] Precise *p*-values for ANOVAs are calculated from the information provided in the original manuscript using the calculator at http://www.danielsoper.com/statcalc3/calc.aspx?id=7

Mussweiler and Strack (2000a) conducted two studies using the same rationale as in Mussweiler and Englich (2005), but using a standard supraliminal anchoring task prior to the LDT. In Study 1, the LDT was presented following the final absolute estimate, thus allowing the possibility that the observed selective accessibility was a consequence of participants' absolute estimates rather than the anchor value. In Study 2, this confound was avoided by presenting the LDT immediately after the comparative judgment. Study 1 asked participants about the annual mean temperature in Germany (high anchor = 20°C; low anchor = 5°C). The LDT subsequently included words associated with summer (predicted to be facilitated by the high anchor), words associated with winter (predicted facilitation by the low anchor), neutral words, and non-words. Study 2 was the supraliminal version of the car price question used in Mussweiler and Englich (2005) (although it is unclear why 'slow' and 'fast' were categorised as neutral words in Mussweiler & Strack, but as associated with inexpensive and expensive cars respectively in Mussweiler & Englich). Study 1 revealed the predicted interaction, $F(1, 26) = 4.53$, $p = .043$ (N = 28), but with two-tailed *t*-tests[6] the individual results were not significant for either the winter words, $t(26) = 1.68$, $p = .105$, or the summer words, $t(26) = 0.63$, $p = .534$. Study 2 again revealed the predicted interaction, $F(1, 28) = 6.57$, $p = .016$ (N = 30), but two-tailed tests were not significant for either the expensive car words, $t(28) = 1.49$, $p = .147$, or the inexpensive car words, $t(28) = 1.50$, $p = .145$.

The results of all the studies reviewed above are typically presented as evidence for the selective accessibility hypothesis of anchoring. What the review makes clear, however, is that the degree of support is somewhat underwhelming with a lack of predicted simple effects, and inconsistencies between the results for high and low anchors. The sample sizes

[6] Although Mussweiler and Strack (2000a) report one-tailed *t*-tests, we report two-tailed results to maintain consistency within our manuscript.

for these reaction time studies are also relatively low. Given the potential for such methodologies as 'signature tests' of selective accessibility, establishing their suitability in this capacity is critical. This is a primary aim of the present paper.

If the effects reported are replicated they demonstrate that the anchor value increases the accessibility of consistent information. The continuous nature of the data, however, enables an additional test, if the overall effect is observed, to determine whether this selective accessibility contributes to the anchoring effect. Namely, across participants, a correlation should be observed between participants' absolute judgments and the differential speed of responses to anchor-consistent versus anchor-inconsistent words. [7]

Overview of the Present Paper

Experiments 1 and 2 are large-sample (at least relative to the original sample sizes) replications of Studies 1 and 2 from Mussweiler and Strack (2000a), which are the studies using the proposed 'signature test' that are closest to prototypical anchoring studies. Upon failing to replicate the original results, Experiments 3 and 4 adjusted the anchor values in an (ultimately unsuccessful) attempt to strengthen the effect. Experiment 5 subsequently explored an alternative, Continuous IDentification (CID), method, which has been shown to be a more sensitive measure than LDT for measuring differences in perceptual fluency (Yang, Huang, & Shanks, 2017). Upon failing to (conceptually) replicate the results described above in any of these experiments, we discuss the methodological implications of the results, as well as the implications for the selective accessibility account of anchoring.

Methodological Note

All experiments reported in this manuscript were pre-registered on the Open Science Framework (https://osf.io/jtx34/). Experiments 1 and 2 were pre-registered together, as were

[7] Such an analysis would have been exploratory in the current experiments, since we did not pre-register it.

Experiments 3 and 4. Experiment 5 and the experiments reported in the Supplementary Materials were pre-registered individually. All were pre-registered and run in the chronological order they are reported in the manuscript[8]. All materials and data are available at https://osf.io/jtx34/. All experiments were conducted in line with the ethical guidelines of the British Psychological Society, and received ethics committee approval from the Ethics Chair of the Department of Speech, Hearing and Phonetic Sciences, UCL (SHaPS-2015-AH-017).

There were various reasons why direct replications were not possible (expanded upon in the individual methods sections), but in updating the methodology we aimed to maximise the quality of the experiments whilst remaining as close as possible to the original methods.

Experiments 1 and 2

Experiments 1 and 2 are essentially the same experiment using different materials. They are therefore described together here. We first outline the methodological changes that were necessary to replicate studies conducted in Germany prior to the year 2000 (Mussweiler & Strack, 2000a), in London in late 2015 / early 2016.

Anchoring Task Changes

Firstly, it was necessary to ensure the suitability of our materials for present day UK participants. The anchor values for the two experiments were the 5th and 95th percentile estimates of a calibration study in which 100 participants - approached on the streets of London and Cambridge, UK (there were no differences between the estimates from the two

[8] Although the original pre-registration of Experiments 1 and 2 recognised Supplementary Experiment 1 as a necessary experiment in the event of a failed replication in Experiments 1 and 2, Experiments 1 and 2, we saw Experiments 3 and 4 as more logical next experiments. Supplementary Experiment 1 then followed them.

sampling locations) - were asked to estimate the “annual average temperature in the UK” (in degrees Celsius) and “the average price of a new car in the UK” (in GBP).

Lexical Decision Task (LDT) Changes

It was necessary to use English rather than German words in these experiments. In line with potential temporal-cultural differences in the implications of different words, we chose to follow the same pre-testing rules as Mussweiler and Strack in choosing the words for our LDTs, rather than necessarily choosing the English translation of their words. In their Study 1, Mussweiler and Strack used seven summer words, seven winter words, 34 neutral words and 12 non-words. They used fewer words in Study 2. In order to maximise the reliability and power of our experiment, we aimed to use the larger number in both our experiments. As documented below, our pre-testing made that possible, with enough words satisfying the criteria outlined in Mussweiler and Strack (2000a). Because of our use of English-speaking participants, we also used different non-words in the task, with the words taken from the ARC non-word database (http://www.cogsci.mq.edu.au/research/resources/nwdb/nwdb.html).

The critical car-related words in Mussweiler and Strack’s Study 2 were makes and models of cars. It was felt that determining whether such proper nouns were words or non-words would be potentially confusing to participants (you can’t, for example, use proper nouns in the board game ‘Scrabble’) in the absence of additional instructions. The acronym, BMW, and abbreviation, VW, used by Mussweiler and Strack were thought to be particularly troublesome and these seeming non-words were not included in the pre-testing. In order to address the ambiguity problem of proper nouns, we amended the instructions for the LDT, to read: “Do the following collection of letters have meaning for an English speaking person?” At the very start of the experiment, before the anchoring task, participants were provided with

examples of what is meant by this through the use of well-known proper nouns and brand names for products other than cars.

Procedural Changes

In Mussweiler and Strack's Study 1, the LDT was administered after participants made an absolute judgment. Study 2 administered the LDT before the absolute judgment in order to prevent the possibility that it was the absolute judgment rather than the comparative judgment that gave rise to the selective accessibility effects observed in the LDT. We consequently used this presentation order in both our experiments. In this procedure, Mussweiler and Strack (Study 2) included seven 'practice experiments'. To reduce fatigue, and increase power, we used two 'practice experiments' (see 'procedure' section for details), which we felt was sufficient to ensure participants were comfortable with the task.

Method

Participants.

A critical factor in a replication project is the statistical power of the experiments. Assuming that a reasonable estimate of the population standard deviation is the mean of the standard deviations of the two conditions, the effect sizes (*d*) of the four *t*-tests reported in Mussweiler and Strack (2000a, Study 1, Study 2) from the LDTs were .29, .76, .56 and .55. These represent the difference in the log-transformed response times between the high and low anchor conditions for summer words, winter words, expensive car words and inexpensive car words respectively. A reasonable (and conservative) estimate of the predicted effect size is .5. To ensure that the power of the individual *t*-tests to detect a true effect was at least 80%, 63 participants were required in each experimental condition, for a total of 126 participants in each experiment (from Howell, 1997). In order to fulfil the counterbalancing requirements

(see below), this number was increased to 128. This is a considerable increase over the 28 and 30 participants recruited in Mussweiler and Strack's Study 1 and 2 respectively, and ensured power of greater than 98% for the interaction term in the ANOVA[9]. Native English speakers (40 male, 87 female, 1 other; aged 18-53, median = 20) were recruited from the Division of Psychology and Language Sciences participant panel at University College London and through personal contacts and advertising of the experimenters. Each participant participated in both experiments, with the order of experiments counterbalanced between participants. Moreover, the combination of anchor conditions was counterbalanced (see Appendix 1 for the assignment of participants to conditions). Two experimenters (FBNB & SAR) tested 64 participants each, and they ran the same number of each combination of conditions and experiment orders (see Appendix 1).

Design and materials.

The two experiments each employed a 2 (anchor – high/low) x 2 (word type) mixed design, with the former factor manipulated between-participants and the latter manipulated within-participants. Absolute estimates were analysed as a function of the first factor, whilst log-transformed reaction times were analysed as a function of the full 2x2 design. The difference between the two experiments was that the topic was 'temperature' in Experiment 1 and car prices ('cars') in Experiment 2.

The anchor-values were the 5th and 95th percentile estimates of a calibration experiment asking participants about both mean temperature and car prices (N = 100; order of

[9] The exact figure depends on the assumed correlation between the within-subject measures, and was calculated using Gpower (Faul, Erdfelder, Lang, & Buchner, 2007), assuming a medium effect size (eta_p^2 = .06), and a non-negative correlation between within-participant measures. Published effect sizes have been argued to be likely to be overestimates of the true size of an effect (e.g., Greenwald, 1975). We note that this interaction term has >80% power to detect sample sizes as small as eta_p^2 = .028.

questions counterbalanced). For Experiment 1, the resulting low and high anchors were 8 and 20 degrees Celsius (the median of the calibration group's estimates was 14.5 degrees). For Experiment 2, they were 6,000 and 30,000 GBP (median = 13,500 GBP).

The words for the LDTs were pre-tested in an online survey hosted on qualtrics.com, with 20 participants recruited from experimenters' (FBNB & SAR) social networks. All pilot participants pre-tested words both for Experiment 1 and Experiment 2 (the order was counterbalanced). To pre-test the words for use in Experiment 1, participants were asked to indicate on a 9-point scale (-4 = strongly associated with winter; +4 strongly associated with summer) the degree to which 94 words related to the concept of winter or summer. The seven words with a mean rating closest to -4 were chosen as 'winter words' (three of these overlapped with the English translations from Mussweiler & Strack, 2000a), the seven with a rating closest to +4 were chosen as 'summer words' (two overlapped with Mussweiler & Strack), and the 34 words with ratings closest to zero were chosen as neutral words. The list of words is shown in Table 1, along with their mean ratings (note that all the words used by Mussweiler & Strack were included in this pre-test). Note that, as in Mussweiler and Strack (2000a), all target words had absolute mean ratings greater than 2, and none of the neutral words did. The 12 non-words were taken from the ARC non-word database (http://www.cogsci.mq.edu.au/research/resources/nwdb/nwdb.html), with the search specifying that words could only be returned that had legal bigrams, and be between three and ten letters long.

To pre-test the words for use in Experiment 2, participants were asked to indicate, on a 9-point scale (-4 = strongly associated with inexpensive cars; +4 = strongly associated with expensive cars), the extent to which 81 words relate to the concept of inexpensive cars or expensive cars. Six of the words used in Mussweiler and Strack were *not* included in the pre-test (BMW, VW, Golf, Fiesta, slow, fast). The former two because it was unclear to us that

they are words, and the latter two due to the inconsistency in their coding between Mussweiler and Englich (2005 – as inexpensive and expensive car words) and Mussweiler and Strack (2000a – as neutral words).[10] Participants were also instructed to use a response option labelled "unrelated to cars" for words they thought were unrelated to cars. This replaced a question in Mussweiler and Strack, which required participants to rate how ambiguously the word was related to cars. We took the absence of any 'unrelated to cars' responses for any of the target words as indication that they were unambiguously associated with cars. The inexpensive car words were then those target words with the lowest mean ratings. As in Mussweiler and Strack, all ratings were less than zero (see Table 1). The expensive car words all had mean ratings higher than +3 (as in Mussweiler & Strack). The neutral words were those with the highest number of 'unrelated to cars' responses (see Table 1). There were six words for which 80% of respondents responded thus, and 'portrait' was chosen as it had the mean rating closest to zero from the other respondents. The words were selected on the basis of their meaning rather than being closely matched on psychophysical properties, as the critical tests are within-item tests across the two anchor conditions.

The LDT was programmed using the PsychoPy software (version 1.82.01, http://www.psychopy.org/; the full code has been uploaded to https://osf.io/jtx34/) and run on desktop computers operating Windows 7 and displays with a resolution of 1600 x 900 pixels. Following the procedure of Mussweiler and Strack, each trial began by displaying a fixation cross on the centre of the screen for 400ms. Immediately following, the letter string (word or non-word) was displayed until the participant provided a response. Responses were given via

[10] The middle two were excluded because we had considered solving the problem of ambiguity in whether proper nouns are words or not, by replacing the LDT with a categorisation task ('Is the word a brand of car?'). We subsequently decided that this was too great a departure from the original methodology. As our pre-test provided us with words that matched the criteria used in Mussweiler and Strack (2000a), we deemed it unnecessary to repeat it.

the Q and P keys on the computer keyboard (marked with blue and yellow stickers respectively). Assignment of the keys to the response options was counterbalanced, with Q (P) representing a meaningful string for half the participants and a non-meaningful string for the other half (participants were presented with the appropriate version of a crib sheet that stated, "Does this word have meaning? Blue = Yes / No, Yellow = Yes / No" for the duration of the experiment). Once a response was given, the letter string disappeared from the screen and, after a pause of 3 seconds, the next string was displayed. The order in which the letter strings were presented was randomised for each participant in the experimental program.

Table 1. *Letter strings chosen for the LDTs in the two experiments. Their mean ratings are included in parentheses.*

EXPERIMENT 1		EXPERIMENT 2	
Word type	Letter String (mean rating in pre-test)	Word type	Letter String
Summer words	summer (+4); suncream (+3.55); august (+3.15); barbeque (+3.05); sandcastle (+2.85); sandals (+2.85); hot (+2.85)	Expensive car words	limousine (+3.40); maserati (+3.20); bentley (+3.35); lamborghini (+3.70); porsche (+3.50); rolls royce (+3.90); ferrari (+3.90)
Winter words	winter (-4); snow (-3.7); icicle (-3.7); freezing (-3.45); hibernate (-3.45); january (-3.45); frost (-3.4);	Inexpensive car words	volkswagen (-0.25); citroen (-0.60); fiat (-0.95); hyundai (-0.50); kia (-2.00); nissan (-0.95); peugeot (-0.85)
Neutral words	sobering (-0.4); shoe (-0.35); eat (-0.25); equipment (-0.25); predator (-0.2); house (-0.2); boardroom (-0.15); pencil (-0.1); dexterity (-0.1); analyse (-0.05); land (0); circling (0); horn (0); sector (0.05); boldly (0.05); cat (0.05); devil	Neutral words	portrait (80%); ordinate (85%); observer (85%); liberal (85%); cotton (90%); appetite (90%); forest (90%); locust (90%); time (90%); always (90%); fracture (90%); dove (90%); piece (90%); democracy (90%); bluster (90%); diary (95%); circus (95%);

	(0.1); voice (0.1); clean (0.1); buffalo (0.1); table (0.1); hair (0.2); citizen (0.2); dog (0.2); walk (0.25); basin (0.3); agility (0.35); arm (0.35); trunk (0.4); potential (0.5); tiger (0.55); cow (0.55); water (0.6); horse (0.6)		drone (95%); bait (95%); birthplace (95%); bottle (95%); background (95%); paper (100%); write (100%); candle (100%); edit (100%); capture (100%); crop (100%); day (100%); zoo (100%); costume (100%); soup (100%); clay (100%); stapler (100%)
Non-words	sckood gluphs sproped sckrulled glebb frusk splooged moarph broge fluilts floaphts geigs	Non-words	frirmths durpths skronnth thourmb jouche kulced dwoacsed twimed vakes kirmbed doarged sowntse
Practice words (non-words first)	woned jenths sought absorb glasgow france dentist chase festival cabinet	Practice words (non-words first)	thonz shrinths involve scrape portsmouth germany dancer click carnival carrier

Note: The ratings are from the -4 to +4 scales for all categories except for the neutral words in Experiment 2. These are the percentage of responders who indicated that the word was unrelated to cars.

Procedure.

Participants were tested in individual cubicles in the Dept. of Experimental Psychology, UCL. Participants consented to take part in what was ostensibly described as a pre-test for the construction of a general-knowledge questionnaire, in which “variations on traditional

methods that use general-knowledge questions will be compared with modern methods that analyse how quickly and accurately people respond to words." (as in Mussweiler & Strack, 2000a). Also following Mussweiler and Strack, these instructions additionally stated "Some of the questions require comparison with a given number. These numbers were chosen randomly, with a mechanism like a 'wheel of fortune.' This is to minimise any influence they might have on your answers and so we can assess the impact of different question formats." Following these 'standard' instructions, participants received the amended instructions for the LDT. Specifically, they were informed that one of the tasks would be to indicate whether letter strings have meaning for an English speaking person. As an example, they were told that STEAVES does not mean something to an English speaking person, whilst AMAZING does, as it is a word. In addition, although they are proper nouns, LONDON, COLGATE, ALDI, IKEA and KIT-KAT also mean something to an English speaking person.

Participants subsequently completed two practice 'experiments', which were not analysed (in line with our pre-registration protocol, https://osf.io/jtx34/). These practice experiments (see Appendix 2 for full details) followed exactly the same procedure as the critical experiments (comparative question, LDT [70 letter strings, with matching proportions of words to non-words as in Table 1], absolute question).[11] This is in line with the procedure of Mussweiler and Strack (2000a, Study 2), who gave their participants four practice 'experiments'. As our participants are completing two critical experiments, and to reduce fatigue, we used two. Participants completed the two critical experiments in a pre-determined order (see Appendix 1). For each experiment, participants were first presented

[11] Although no information was available as to the subject used in the practice trials mentioned in Mussweiler and Strack (2000a), we chose two domains seemingly unrelated to the critical tasks (length of the M25 motorway and height of Big Ben).

with the comparative judgment task: “Is the annual average temperature in the UK higher or lower than [anchor]°C” (Experiment 1) or “Is the average price of a new car in the UK higher or lower than £[anchor]?” (Experiment 2). Participants were then presented with the LDT. As in Mussweiler & Strack (Study 1), 10 practice letter strings – 2 non-words and 8 neutral words, matching the proportion of words to non-words in the critical trials - preceded the 60 critical strings (see Table 1). Following the LDT, participants provided their exact estimate for the average temperature, or new car price. They then proceeded to the next experiment.

A short filler task was included between each experiment (i.e., after Practice Experiment 1, Practice Experiment 2 and the first critical experiment). This task was a simple forward letter span task. After a fixation cross (1000 ms), participants were presented with a sequence of consonants (drawn randomly without replacement from the set of all consonants), each displayed for 800 ms, followed by a 200 ms blank screen. After presentation of the final letter in the string, participants were prompted to type in the sequence in the order of presentation. There was no time limit for their response and no feedback was provided about the correctness of their response. Sequences increased from 4, 5, 6, to 7 letters. The filler task took between 0.57 and 2.10 minutes. Results of this task were not analysed (see https://osf.io/jtx34/).

After completing the whole experimental session, participants were thanked and debriefed.

Results

Exclusion criteria.

Although Mussweiler and Strack (2000a) do not indicate that any data were excluded for any of their analyses, with absolute estimates made on an unbounded scale there is the potential for outliers to distort the data. In line with our pre-registration (https://osf.io/jtx34/),

we would have excluded any extreme responses (e.g., above 100°C or £1,000,000), but there were none in the data. Following such exclusions, within each condition, responses further than three standard deviations from the mean were to be excluded. As a result, we removed one response from the low anchor condition in Experiment 1, one response from the low anchor condition, and two responses from the high anchor condition in Experiment 2. Similarly, for the reaction time data, following log transformations (as in Mussweiler & Strack), any trials with reaction times more than three standard deviations either side of the mean were excluded from analysis. This resulted in excluding 60 responses from the low anchor and 68 responses from the high anchor condition in Experiment 1, and excluding 59 responses from the low anchor and 40 responses from the high anchor condition in Experiment 2.

Planned analyses (*https://osf.io/jtx34/*).

Absolute estimates.

Estimates of the average UK temperature (Experiment 1) and the average price of a new car (Experiment 2) were both higher following a high anchor than a low anchor (Experiment 1: $t(125) = 4.84$, $p < .001$; Experiment 2: $t(123) = 5.89$, $p < .001$; see Table 2 for full descriptive statistics).

Table 2. *Descriptives of the absolute estimates for all experiments*

	Low anchor condition			High anchor condition		
Experiment	n	mean	sd	n	mean	sd
1	63	12.68	2.14	64	14.70	2.55
2	63	10754.14	5061.05	62	19741.92	10974.91
3	62	10.71	4.05	64	15.88	3.40
4	63	13817.27	6366.24	63	61452.29	79174.26
5	63	11.03	4.49	64	16.16	3.09

Lexical decisions.

Mussweiler and Strack (2000a) tested the effect of the anchors on lexical decisions using "facilitation scores" in which, for each participant, the average log RT to neutral and non-words was subtracted from the average log RT for the two types of target words (i.e., words associated with summer and winter words in Experiment 1, and words associated with inexpensive and expensive cars in Experiment 2). Employing this analysis, there was no evidence of an interaction between anchor and word type (summer, winter) in Experiment 1, $F(1,126) = 0.07$, $p = .785$, $\eta_G^2 < .001$, or between word type (expensive, inexpensive) and anchor in Experiment 2, $F(1,126) = 1.71$, $p = .193$, $\eta_G^2 = .004$. The main effects of word type and anchor were not significant in either Experiment ($Fs < 1$).

We also pre-registered a standard 2 (anchor: high vs low) x 4 (word type: summer/expensive, winter/inexpensive, neutral, non-word) ANOVA, in which 'raw' log RTs were analysed (i.e., without subtracting the average log RT to neutral and non-words). The results of this analysis matched those of the previous one, with no interaction between word type and anchor in Experiment 1, $F(2.45,308.26) = 0.31$, $p = .776$, $\eta_G^2 < .001$, or Experiment 2, $F(2.65,333.48) = 0.86$, $p = .452$, $\eta_G^2 = .001$ (note a Greenhouse-Geisser correction of the degrees of freedom was applied due to violation of sphericity). For both experiments, there was a significant main effect of word type: for Experiment 1, $F(2.45,308.26) = 261.69$, $p < .001$, $\eta_G^2 = .236$, and for Experiment 2, $F(2.65,333.48) = 128.48$, $p < .001$, $\eta_G^2 = .171$. The main effect of anchor was not significant for either experiment ($Fs < 1$; see Figure 1 for descriptives).

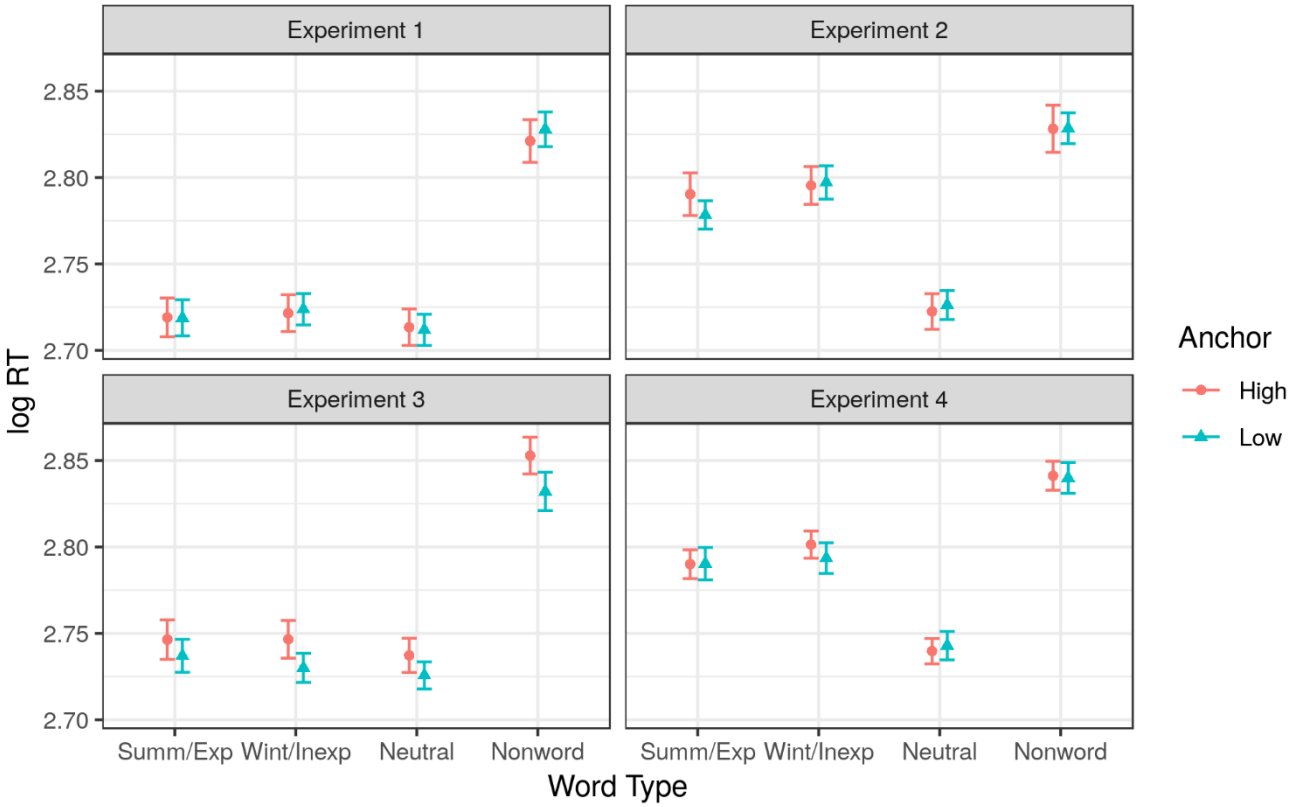

Figure 1. Mean log reaction times in Experiments 1–4. Error bars are +/- 1 S.E.

Combining Experiments 1 and 2.

As this potentially offers more power, we also pre-registered a linear mixed-effect model analysis, combining the data from Experiment 1 and Experiment 2. The analysis was conducted with the lme4 (Bates, Mächler, Bolker, & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017) packages for the R statistical computing environment. These models include orthogonal contrasts for all factors (Word type, Anchor, and Experiment). For word type, contrast C_1 compares non-words to words, contrast C_2 compares neutral words to target words, and contrast C_3 compares summer/expensive words to winter/inexpensive words. We estimated two models which had fixed effects for word type, anchor, and experiment, as well as all two- and three-way interactions, but differed in their random-effects structure. Model 1 included random participant-specific intercepts

crossed with random word-specific intercepts. Model 2 included, in addition, random participant-specific slopes for word-type, task and anchor. Model 2 fitted significantly better than Model 1, $\chi^2(5) = 432.93$, $p = .001$, hence we report the results for that model. The main interest is in the two-way interaction between anchor and contrast C_3, which was not significant, $t(577.01) = -1.12$, $p = .26$ and the three-way interaction between Experiment, anchor, and contrast C_3, which was also not significant, $t(604.01) = -0.77$, $p = .44$.

Experiments 3 and 4

Experiments 1 and 2 demonstrated a standard anchoring effect in absolute estimates, which persisted despite an interleaved LDT of 70 items. There was, however, no evidence of enhanced accessibility of anchor-related constructs observed in the LDT. Closer inspection of the materials is suggestive of why this might be the case. The high anchor in the car scenario was £20,000. It appears unlikely that this would be high enough to prime high-end cars such as Maserati and Rolls Royce – as used in the current experiment. Moreover, the temperate nature of the UK climate might make it unlikely that 10 degrees and 18 degrees Celsius are distinct enough to prime concepts of winter and summer respectively. Experiments 3 and 4 replicated Experiments 1 and 2, but used different anchor values. Rather than asking pre-test participants what the average temperature is in the UK / what the average price of a new car is, pre-test participants were directly asked what temperature is associated with the critical words in the LDT.

Method

Participants.

128 native English speakers participated. Due to an administrative error, demographic data was only collected from 105 participants (76 female, 29 male; aged 18-50 years, median = 20).

Design, materials and procedure.

Forty-eight, additional, pre-test participants completed the 'High anchor' pre-test, and 48 completed the 'Low anchor' pre-test. Participants were approached around the UCL campus, and asked: "If you think of a *day in the UK and these words* / *a car and these brands*:…What *temperature* / *price* comes to mind?" (the italics are added here to show the words that were different for the different questions). The "…" indicates that the seven critical words from either the high or low anchor condition were then provided in bold text. Participants provided an estimate for both the car and temperature words in a single anchor condition. The median response was used as the anchor in the main experiment. Thus, in Experiment 3 the anchor values were 25ºC and 2ºC. In Experiment 4, they were £200,000 and £13,000.

An additional change from Experiments 1 and 2 was that "shroomed" was not included as a non-word in the practice experiment. Participants in Experiments 1 and 2 reported uncertainty as to whether it was a real word or not. It was therefore replaced with "skreets" in Experiments 3 and 4, which was obtained from the ARC database, includes the same number of letters and syllables, as well as legal bigrams.

The final change from Experiments 1 and 2 was that Experiments 3 and 4 were run by SL and EP, following the same counterbalancing procedure (see Appendix 1).

All other aspects of the method were the same as Experiments 1 and 2.

Results

Exclusions.

Following the same criteria as Experiments 1 and 2, two participants were excluded (from the low anchor condition of Experiment 3) for providing estimates greater than 100ºC. One response from both the low and high anchor conditions of Experiment 4 was excluded for being more than three standard deviations from the mean. Similar exclusions of extreme log

reaction times resulted in excluding 64 responses from the low anchor and 59 responses from the high anchor condition in Experiment 3, and excluding 71 responses from the low anchor and 50 responses from the high anchor condition in Experiment 4.

Planned analyses.

Absolute estimates.

Once again, estimates of the average UK temperature (Experiment 3) and the average price of a new car (Experiment 4) were both higher following a high anchor than a low anchor (Experiment 3: $t(124) = 7.76$, $p < .001$; Experiment 4: $t(124) = 4.76$, $p < .001$; see Table 2).

Lexical decisions.

Analyses of the LDT revealed the same pattern of results as in Experiments 1 and 2. In the analysis of facilitation scores, there was no evidence of an interaction between anchor and word type (summer, winter) in Experiment 3, $F(1,126) = 0.40$, $p = .529$, $\eta_G^2 = .001$, or between word type (expensive, inexpensive) and anchor in Experiment 4, $F(1,126) = 1.71$, $p = .193$, $\eta_G^2 = .004$. The main effects of word type and anchor were not significant in either experiment ($Fs < 1$).

The standard 2 (anchor: high vs low) x 4 (word type: summer/expensive, winter/inexpensive, neutral, non-word) ANOVA also did not show a significant interaction between word type and anchor in Experiment 3, $F(2.41,303.92) = 0.44$, $p = .683$, $\eta_G^2 = .001$, or Experiment 4, $F(2.78,350.73) = 0.46$, $p = .697$, $\eta_G^2 = .001$ (Greenhouse-Geisser correction due to violation of sphericity). For both experiments, there was a significant main effect of word type (Experiment 3: $F[2.41,303.92] = 187.18$, $p < .001$, $\eta_G^2 = .247$; Experiment 4: $F[2.78,350.73] = 140.71$, $p < .001$, $\eta_G^2 = .217$). The main effect of anchor

was not significant for either experiment ($Fs(1,126) < 1.35$; see Figure 1 for descriptives).

Combining Experiments 3 and 4.

The same linear mixed-effect models were estimated as for Experiments 1 and 2. As previously, the model with participant-specific slopes for word-type, task and anchor fitted significantly better than the model with only participant-specific intercepts crossed with random word-specific intercepts, $\chi^2(5) = 432.93$, $p = .001$, hence we report the results for that model. Neither the two-way interaction between anchor and contrast C_3, $t(577.01) = -1.12$, $p = .26$, nor the three-way interaction between experiment, anchor and contrast C_3 was significant, $t(604.01) = 0.77$, $p = .44$.

Interim Discussion (Experiments 1-4)

All four experiments reported thus far have observed significant anchoring effects, but no evidence has been found for a selective increase in the accessibility of anchor-related concepts, as measured by an LDT (thus failing to replicate the results reported in Mussweiler & Strack, 2000a). To determine the significance of this result for the selective accessibility theory of anchoring, we ran a validation experiment (for full details see Supplementary Experiments 1 & 2). The purpose of this experiment was to determine whether our LDT was sensitive to a general priming effect. Participants were primed with the concept of summer or winter (Supplementary Experiment 1) and expensive or inexpensive cars (Supplementary Experiment 2) with pictures, before completing the respective LDTs. Once again, there was no difference in participants' reaction times across the experimental conditions. Consequently, the null results observed in Experiments 1-4 are somewhat uninformative as regards the mechanism underlying the anchoring effect (if direct priming does not affect responses in the LDT, a failure to observe an effect in the LDT cannot be taken as evidence that priming is not occurring in the anchoring question).

Whilst our results are not especially informative as to the status of the selective accessibility theory of anchoring, they do demonstrate the difficulty associated with the use of the LDT for testing the mechanisms underlying anchoring effects (a point to which we return in the General Discussion). Experiment 5 sought to propose a solution to this difficulty by utilising a novel method, with a similar logic to the LDT, the Continuous IDentification (CID) task. The CID has been used in memory research, especially in the area of repetition priming (e.g., Stark & McClelland, 2000; Ward, Berry, & Shanks, 2013), and has explicitly been shown to be a more sensitive measure than LDT for measuring differences in perceptual fluency (Yang, Huang, & Shanks, 2017). If support for selective accessibility were obtained through faster (correct) responses to anchor consistent words using the CID task, these results, coupled with the results of Experiments 1-4, would suggest it as a more sensitive measure for assessing selective accessibility, and therefore a more suitable 'signature test' for detecting the operation of selective accessibility processes.

Experiment 5[12]

Experiment 5 used a CID task as a test of the selective accessibility of anchor-related concepts. The experiment focussed on the anchoring of temperature estimates (as in Experiments 1 & 3). We added an additional analysis in Experiment 5. The first three words in the CID task were fixed. After two practice words, the third word was always 'summer.' Analyses of responses to this single word (predicted to be quicker in the high anchor condition than the low anchor condition) allow a test of the central hypothesis with reduced interference from other items. Additional changes were made to the methodology (reducing

[12] This was pre-registered as Experiment 7. The original Experiment 5 is Supplementary Experiment 1.

the number of practice trials and attempting to better standardise the words used) in an attempt to maximise the power and reliability of the experiment.

Method

Participants.

88 female and 40 male participants (aged 18 to 33; median = 20) were recruited for Experiment 5 through the UCL Psychology and Language Sciences research subject pool and through approaching students around the UCL campus. All participants had a proficient level of English, but 38% reported a language other than English as their native tongue. AB and DU tested 64 participants each (see Appendix 3 for counterbalancing protocol).

Design.

The design follows Experiments 1 and 3, with a 2(anchor) x 2(summer/winter word) mixed design.

Materials and procedure.

We did not use the 'M25' practice task in this experiment. Participants answered the 'Big Ben' comparative question, before completing a CID task consisting of 44 words unrelated to temperature (or summer or winter; see Table 3), in random order. Following the CID task, participants provided an absolute estimate for the height of Big Ben. As in Experiments 1-4, participants next completed the digit-span task, before being presented with the temperature anchoring question (from Experiment 3). Following this, participants completed the critical CID task. The first two words were "cotton" and "dancer" (randomised order) with "summer" always the third word presented. The remainder of the words (see Table 3) were presented in a randomised order. Following the CID task, participants provided their absolute estimate of the annual average temperature in the UK. At the end of the experiment, participants were

presented with a set of measures relevant to climate change perceptions. These were part of a separate project, investigating the influence of anchoring on such perceptions (and behavioral intentions), and are not considered further here.

Table 3. *Words used in the CID task (Experiment 5).*

Summer words	Winter words	Neutral words	Words in practice (Big Ben) experiment
summer	winter	cotton	pencil
warm	cold	dancer	forest
sand	snow	horn	shoe
beach	frost	piece	small
pool	coat	write	paper
sandal	gloves	voice	table
sunny	chilly	palace	always
		clay	hair
		devil	house
		land	lake
		walk	time
		horse	clean
		monkey	bottle
		lion	slow
		trunk	ghost
		tiger	basin
		errand	advice
		cube	cure
		feet	bush
		mercy	flock
		kick	hold
		insect	moment
		scrape	circus
		lamp	noun
		note	path
		which	where
		part	pony
		chase	essay
		click	graze
		coyote	sector
			boldly
			drone

	locust
	diary
	bait
	edit
	crop
	absorb
	sought
	soup
	dove
	pond
	turn
	hood

The CID task.

The nature of the masks used in the CID task meant that it was important to balance the length of the words used. All words were therefore chosen to be between four and six letters in length, with the critical (summer/winter) words balanced (as far as possible) for length. This was facilitated with the original pre-test data, as well as an additional pre-test (raw results available at https://osf.io/6txdr/). “Sunny” was not pre-tested, but was assumed to be associated with summer, following the association of “sun”, which was too short for inclusion in this experiment. Moreover, to balance the length of the words, the singular, “sandal,” was included in place of the pre-tested “sandals.”

The CID methodology was based on that in Yang et al. (2017). A fixation cross was shown for 500ms. For each word, there were 15 cycles. At the first cycle, the word was presented for 17ms, followed by a mask for 233ms - adding up to a 250ms cycle. Then for each cycle, the length of time that the word is shown increases by 17ms, and the mask decreases by 17ms. The length of the mask is always the same size as the words (e.g.., "####" for "horn" and "######" for summer). The mask was presented in a slightly larger font size (120%).

Participants were instructed to press the space bar as soon as they thought they recognised the word and were then required to type it in. Whilst participants were instructed to answer all questions as quickly and accurately as possible, to ensure that participants were recognising the words, incorrect responses (or responses not made before the end of the 'cycles') were 'penalised' with a 10 second delay between entering the response and proceeding to the next trial. This ensured that there was no time advantage to participants for not responding appropriately to the stimuli.

Results

Exclusions.

There were no responses above 100°C in Experiment 5, but in the low anchor condition one absolute estimate was removed because it was further than three standard deviations from the mean. Similar exclusions of extreme log reaction times resulted in excluding 11 responses from the low anchor and 17 responses from the high anchor condition.

Planned analyses.

Absolute estimates.

Once again, estimates of the average UK temperature were higher following a high anchor than a low anchor, $t(125) = 7.50$, $p < .001$ (see Table 2).

CID results.

Analyses of the CID revealed the same pattern of results as in Experiments 1-4. In the analysis on facilitation scores, there was no evidence of an interaction between anchor and word type (summer, winter), $F(1,126) = 0.41$, $p = .522$, $\eta_G^2 = .001$. Neither the main effect of word type, $F(1,126) = 2.77$, $p = .099$, $\eta_G^2 = .008$, nor the main effect of anchor, were significant, $F(1,126) < 0.01$, $p = .983$, $\eta_G^2 < .001$.

Focusing only on the word "summer", which was the first target word to appear, did not show a significant effect of anchor condition, $\Delta M = -0.01$, 95% CI , 0.03, $t(121) = -0.65$, $p = .517$.

The standard 2 (anchor: high vs low) by 3 (word type: summer, winter, neutral) ANOVA also did not show a significant interaction between word type and anchor, $F(1.84,231.32) = 0.26$, $p = .756$, $\eta_G^2 < .001$ (Greenhouse-Geisser correction applied due to violation of sphericity). There was a significant main effect of word type, $F(1.84,231.32) = 15.60$, $p < .001$, $\eta_G^2 = .008$, while the main effect of anchor was not significant, $F(1,126) < 0.01$, $p = .963$, $\eta_G^2 < .001$ (see Figure 2 for descriptives).

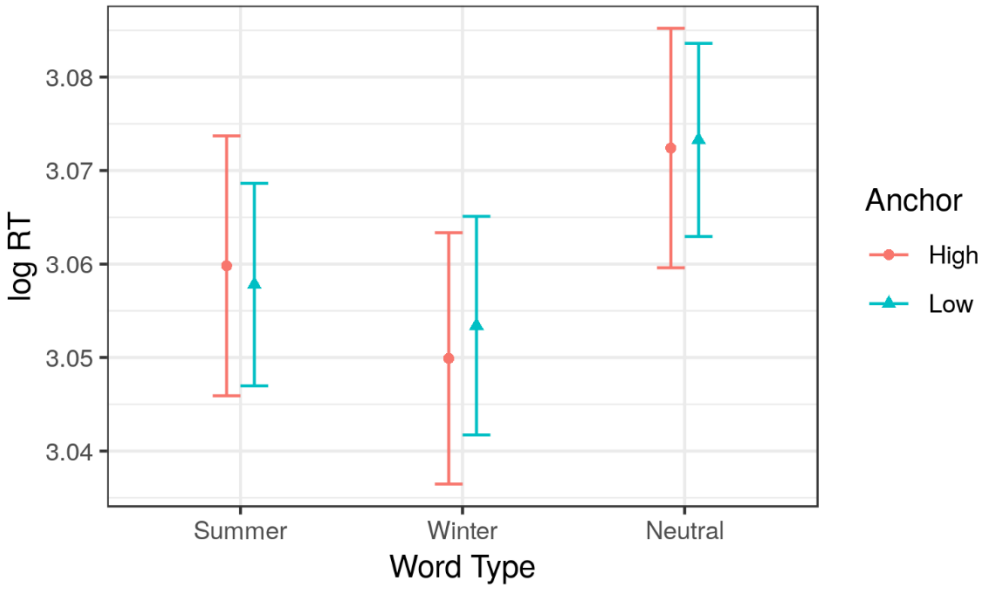

Figure 2. Mean log reaction times in Experiment 5. Error bars are +/- 1 S.E.

Linear mixed-effect analysis.

A comparison between a model with crossed random intercepts for participants and stimuli did not fit significantly worse than a model with additional participant-specific slopes for the two contrasts comparing neutral words to target words (C_1) and comparing summer

words to winter words (C_2), $\chi^2(2) = 1.82$, $p = .402$. Hence, we report results for the first, simpler model. This analysis did not show evidence of the crucial interaction between anchor condition and C_2, $t(5008.22) = -0.67$, $p = .51$. In fact, none of the fixed effects were significant (all $|t| < 1.4$).

Discussion

Experiment 5 tested a new 'signature test' for the operation of selective accessibility processes. As with Experiments 1-4, no evidence was obtained that anchor-consistent concepts were recognised faster than anchor-inconsistent concepts. As with Experiments 1-4, it is important to demonstrate the sensitivity of this new methodology to detecting priming effects. Supplementary Experiment 3[13] was therefore run using the same images from Supplementary Experiment 1, with the CID task. Once again, there was no evidence that our task was sensitive to a priming influence of seasonal pictures. It is worth noting that none of the supplementary experiments were run under optimal test conditions (see Supplementary Materials for details). A positive result in these experiments would, however, have demonstrated the suitability of the current methodologies for investigating selective accessibility processes. Whilst we could have followed these experiments up with more controlled validation experiments, the pattern of results in the current experiments is clear: signatures of selective accessibility processes are hard to detect using lexical tests of accessibility. Consequently, regardless of whether effects *can* be observed, these are not good signature tests. Our subsequent review of the literature around LDTs and semantic priming (presented in the General Discussion) supports this conclusion.

[13] Pre-registered as Experiment 8.

Exploratory Bayesian meta-analysis

Across five experiments, we found no evidence for the enhanced accessibility of anchor-congruent information following an initial comparison question. A Bayesian meta-analysis[14] (Rouder & Morey, 2011) on the critical interaction (anchor x word type (summer/expensive vs winter/inexpensive) suggested the data were 21.10 times more likely under the null hypothesis than the alternative, thus providing 'strong' evidence in favour of the null (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009, p. 228).

General Discussion

Five experiments followed the logic of Mussweiler and Strack (2000a), testing whether anchor-consistent concepts were selectively more accessible following responses to a comparative question in a standard anchoring task. Experiments 1-4 employed LDTs (following Mussweiler & Strack, 2000a). Experiment 5 employed a CID. In each experiment, we tested the hypothesis that words associated with the high anchor would be recognised faster following a high anchor, and words associated with the low anchor would be recognised faster following a low anchor. We observed no support for this hypothesis in any individual experiment, or in a subsequent meta-analysis across the experiments.

Two 'validation experiments' (Supplementary Experiments 1-3) were run in addition to the five reported above. These experiments tested for the power of the LDT (Supplementary Experiments 1 & 2) or the CID (Supplementary Experiment 3) to detect direct priming effects. Such evidence was not observed in either experiment. Without this

[14] For each experiment, we computed a *t*-statistic for the critical interaction between word type (restricted to summer/expensive vs winter/inexpensive) and anchor (high vs low). The resulting five *t*-values were then entered in a Bayesian meta-analysis (using the meta.ttestBF function in the BayesFactor package (Morey & Rouder, 2018). The null-hypothesis assumes the true standardized effect size underlying each t-statistic is 0, while the alternative hypothesis uses a Cauchy prior (mean 0 and scale factor 0.7071) on the standardized effect size (see Rouder & Morey, 2011) for further details.

evidence, the consequences of the current results for the status of the selective accessibility theory of anchoring are unclear. If the method does not reveal direct priming effects, the failure to observe priming effects following an anchoring task does not imply that the anchor is not priming anchor-consistent concepts. On the basis of these findings, and current evidence in the psychology of reading, we do, however, argue that the current methodology is not fit for purpose as a signature test of selective accessibility.

These methodologies are not a suitable 'signature test' for selective accessibility

The main analysis in the current experiments (following Mussweiler & Strack, 2000a) concerns average reaction times across seven critical words for each of two categories (e.g., summer and winter). These words are interspersed in a list of 56 additional items (26 additional items in Experiment 5, which required no non-words). The logic of the current methodology requires that the semantic priming effect of the anchor extends throughout the LDT/CID task. Upon closer reflection, this seems unlikely to occur. Semantic priming effects in LDTs are predominantly observed within 2000 ms of the target stimulus (e.g., Jones & Estes, 2012), and are typically not observed with more than a single interleaving trial (e.g., Becker, Moscovitch, Behrmann, & Joordens, 1997; Dannenbring & Briand, 1982; Henderson, Wallis, & Knight, 1984; McNamara, 1992; Monsell, 1985; Ratcliff, Hockley, & McKoon, 1985; Zeelenberg & Percher, 2002).

Mussweiler and Strack (2000a) cite Neely (1991) in support of using LDTs, stating that "it is a basic finding that a target word is recognized faster if an associatively or semantically related word was presented beforehand" (Mussweiler & Strack, 2000a, p. 1041). This is indeed a basic finding in cognitive psychology. However, as we have already hinted at, the difficulty in the instantiation of this in the current rationale lies exactly with the low-level nature of this result. Recently presented words will affect response times to subsequent related words, thus crowding out any influence of the anchor. Indeed, Neely (1991)

concentrates his discussion of LDTs on demonstrating trial-to-trial priming effects, where the prime is the immediately preceding word. The SOA (Stimulus Onset Asynchrony) between the prime and target is understood to be critical in observing semantic priming effects with LDTs. In the context of the Neely review, it is noteworthy that the longest SOA explicitly mentioned is 2000 ms., indicating the degree to which the current methodology is removed from standard instantiations of LDTs. One major take-home message from Neely (1991) would appear to be the fragility of semantic priming effects (as measured by LDTs)[15]. More specifically, however, the observations reviewed above would all argue against the use of LDTs as they have been employed in the anchoring literature (Englich et al., 2006; Mussweiler & Englich, 2005; Mussweiler & Strack, 2000a), where there is a considerable temporal gap between the 'prime' (the anchor) and the critical LDT trials, which are interleaved with a large number of related and unrelated words.

In light of these insights, can the present methodology be adapted to increase its utility? Experiment 5 introduced a relatively novel task that has been shown to be a more powerful test of priming effects, to no avail. The critiques advanced above apply to Experiment 5 too, especially with the presence of the interleaving items. The 'one-shot' analysis of 'SUMMER' in that experiment likely provides better test conditions than the aggregate analysis. One potential solution, therefore, might be to include a number of one-shot tests using the CID methodology, and a variety of different anchoring questions. The difficulty associated with such an approach is evidenced by the quote from Mussweiler and Strack (2000a) above. Whilst Mussweiler and Strack reported that their experiments provided evidence of *selective* accessibility, that hypothesis is somewhat difficult to falsify with the

[15] This fragility is further evidenced by studies demonstrating that a mere categorical relationship between words does not reliably produce priming effects – rather, an associative relationship is also required (e.g., Abad, Noguera, & Ortells, 2003).

LDT (or CID) methodology: “a target word is recognized faster if an associatively or semantically related word was presented beforehand.” In other words, effects in LDTs are not highly discriminatory. Thoughts of summer might prime the associatively related concept of winter. Potentially, there is scope for thorough piloting in trial-to-trial LDTs, ensuring that choices of ‘low anchor words’ are not primed by ‘high anchor words’, in order to generate a set of ‘one-shot’ stimuli. Given our experience with this project, we are, however, sceptical about the likely success of such an approach (as a suitable method).

Durable anchoring effects

The previous section has stressed the fragility and brevity of semantic priming effects. The consistent positive result in our experiments, however, was a significant effect of the anchor, even though 70 LDT trials (40 CID trials) separated the comparative question and the absolute question (for other instances of long term anchoring effects, see e.g., Mussweiler, 2001). How does such an effect persist across these intervening items, where more than one typically wipes out semantic priming effects in LDTs (e.g., Becker et al., 1997; Dannenbring & Briand, 1982; Henderson, Wallis, & Knight, 1984; McNamara, 1992; Monsell, 1985; Ratcliff, Hockley, & McKoon, 1985; Zeelenberg & Percher, 2002)? The first, and very plausible, possibility is that participants approximate an absolute answer when answering the comparative question, which is then utilised in the generation of the absolute answer. The anchoring effect observed after 70 LDTs would therefore rely more on recall of this previous answer, which may result from (e.g.) selective accessibility, anchoring-and-adjustment, scale distortion… It is also, however, plausible that this is an instance of semantic priming. Whilst long-term semantic priming (spanning more than a single interleaving item) has been elusive in standard LDT studies, there is increasing evidence that it can be observed at longer time lags, across multiple interleaving trials, when the participant’s response relies on sufficient processing of semantics (Becker, Moscovitch, Behrmann, & Joordens, 1997; Joordens &

Becker, 1997; Ray & Bly, 2007; Rodd, Lopez Cutrin, Kirsch, Millar, Davis, 2013; Tse & Neely, 2007; Woltz, 2010; Woltz, Sorensen, Indahl, & Splinter, 2010). Becker et al. (1997), for example, observed no evidence for long term semantic priming in an LDT, but priming after more than 20 interleaved items when participants' task was to determine whether a word was a living thing or not (thus requiring semantic processing). They argued that no effect is observed in LDTs, as orthographic processing is sufficient to make a word/non-word judgment, before semantic processing is required. On this logic, anchoring effects such as those observed here might, in themselves, be considered evidence for a semantic priming effect underlying anchoring. However, the obvious circularity in this argument is clear.

Conclusion

Our experiments failed to replicate previous results supporting the predictions of the selective accessibility theory of anchoring. Given problems associated with the methodology (related to the documented fragility and [non-]selectivity of semantic priming effects, as measured by LDT tasks), we do not believe the current results necessarily provide additional evidence for or against that theory, which we still consider a good explanation for judgmental anchoring. We do, however, caution against the use of such methodologies for investigating the underlying processes in anchoring effects, and hope that the present experiments are enlightening for researchers designing tests to discriminate theories of anchoring. Given the issues identified with these methods, and the null results reported here, we also suggest that supportive evidence for selective accessibility obtained in the past using this method is not a strong part of the evidence base for that theory.

Context of the Research

In an early draft of their manuscript, Harris and Speekenbrink (2016) challenged the necessity of Frederick and Mochon's (2012) Scale Distortion theory of anchoring to account for extant results in the literature (including those presented in Frederick & Mochon, 2012). Whilst we

argued that selective accessibility processes *could* explain the results we obtained, we were under pressure to demonstrate that they *did.* The one candidate test that we could identify in the literature to provide such a 'signature test' was the one described in Mussweiler and Strack (2000a, Studies 1 & 2), and evaluated in the present manuscript. Given some inconsistencies with the use of this test (as described in the introductory section 'Assessing selective accessibility through lexical decisions [a potential 'signature test'] – results to date), we were hesitant to use it. Subsequently, in light of this hesitancy, and our perception of the utility of such a test, we decided it would be a fruitful endeavour to replicate the original studies. We continue to agree with Turner and Schley (2017) in their evaluation that the pluralism in our understanding of the anchoring effect is problematic, given our limited understanding of the scope of each mechanism, and the potential interactions between them.

References

Abad, M. J. F., Noguera, C., & Ortells, J. J. (2003). Influence of prime-target relationship on semantic priming effects from words in a lexical-decision task. *Acta Psychologica, 113,* 283-295.

Adaval, R., & Wyer, R. S. Jr. (2011). Conscious and nonconscious comparisons with price anchors: Effect on willingness to pay for related and unrelated products. *Journal of Marketing Research, 48,* 355-365.

Bahník, Š., Englich, B., & Strack, F. (2017). Anchoring effect. In R. F. Pohl (Ed.). *Cognitive Illusions: Intriguing Phenomena in Thinking, Judgment, and Memory (2nd ed.)* (pp. 223-241). Hove, UK: Psychology Press.

Bahník, Š., & Strack, F. (2016). Overlap of accessible information undermines the anchoring effect. *Judgment and Decision Making, 11*, 92-98.

Bates, D., Mächler, M, Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48.

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1059-1082.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49,* 188-207.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: a review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19,* 7-42.

Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7,* 223-242.

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes, 79,* 115-153.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgment of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.*), Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). Cambridge, England: Cambridge University Press.

Chaxel, A-S. (2014). The impact of procedural priming of selective accessibility on self-generated and experimenter-provided anchors. *Journal of Experimental Social Psychology, 50,* 45-51.

Dannenbring, G. L., & Briand, G. (1982). Semantic priming and the word repetition effect in

a lexical decision task. *Canadian Journal of Psychology, 36,* 435-444.

Englich, B. (2006). Blind or biased? Justitia's susceptibility to anchoring effects in the courtroom based on given numerical representation. *Law & Policy*, *28*, 497-514.

Englich, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*, 188-200.

Epley, N. (2004). A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment. In D. J. Koehler, & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 240-257). Oxford, UK: Blackwell.

Epley, N,. & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12,* 391-396.

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin, 30,* 447-460.

Epley, N., & Gilovich, T. (2005). When effortful thinking influences judgmental anchoring: Differential effects of forewarning and incentives on self-generated and externally provided anchors. *Journal of Behavioral Decision Making, 18,* 199-212.

Epley, N., & Gilovich, T. (2006). The anchoring and adjustment heuristic: Why adjustments are insufficient. *Psychological Science, 17,* 311-318.

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*, 175-191.

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141,* 124-133.

Galinksy, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology, 81,* 657-669.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20.

Harris, A. J. L., & Speekenbrink, M. (2016). Semantic cross-scale numerical anchoring. *Judgment and Decision Making, 11,* 572-581.

Hawkins, S. A., & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin, 107,* 311-327.

Henderson, L., Wallis, J., & Knight, K. (1984). Morphemic structure and lexical access. In H. Bouma & D. Bouwhuis (Eds.), *Attention and Performance X: Control of Language Processes* (pp. 211-224). Hillsdale, NJ: Erlbaum.

Hogarth, R. M., & Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management Science, 36,* 1161-1166.

Jones, L. L., & Estes, Z. (2012). Lexical priming: associative, semantic, and thematic influences on word recognition. In In J. S. Adelman (Ed.), *Current Issues in the Psychology of Language. Visual Word Recognition: Meaning and context, individuals and development* (pp. 44-72). New York: Psychology Press.

Joordens, S., & Becker, S. (1997). The long and short of semantic priming effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1083-1105.

Kuznetsova, A., Brockhoff P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1-26.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46-55.

McNamara, T. P. (1992). Theories of Priming: I. Associative distance and lag. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1173-1190.

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes, 122,* 69-79.

Monsell, S. (1985). Repetition and the lexicon. In A. W. Ellis (Ed.), *Progress in the Psychology of Language, Vol. 2* (pp. 147-195). London: Lawrence Erlbaum Associates.

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Mussweiler, T. (2001). The durability of anchoring effects. *European Journal of Social Psychology, 31,* 431-442.

Mussweiler, T. (2002). The malleability of anchoring effects. *Experimental Psychology, 49,* 67-72.

Mussweiler, T., & Englich, B. (2005). Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organizational Behavior and Human Decision Processes, 98,* 133-143.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35*, 136-164.

Mussweiler, T., & Strack, F. (2000a). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology, 78*, 1038-1052.

Mussweiler, T., & Strack, F. (2000b). Numeric judgments under uncertainty: The role of knowledge in anchoring. *Journal of Experimental Social Psychology, 36*, 495-518.

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes, 86*, 234-255.

Navarro-Martinez, D., Salisbury, L. C., Lemon, K. N., Stewart, N., Matthews, W. J., & Harris, A. J. L. (2011). Minimum required payment and supplemental information disclosure effects on consumer debt repayment decisions. *Journal of Marketing Research, 48,* S60-S77.

Newell, B. R., & Shanks, D. R. (2014). Prime numbers: anchoring and its implications for theories of behavior priming. *Social Cognition, 32,* 88-108.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84-97.

Ratcliff, R., Hockley, W., & McKoon, G. (1985). Components of activation: Repetition and priming effects in lexical decision and recognition. *Journal of Experimental Psychology: General, 114,* 435-450.

Ray, S., & Bly, B. M. (2007). Investigating long-term semantic priming of middle- and low-familiarity category exemplars. *The Journal of General Psychology, 134,* 453-466.

Rodd J. M., Lopez Cutrin B., Kirsch H., Millar A., & Davis M. H. (2013) Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language*, *68*, 180–198.

Rouder, J. N. & Morey, R. D. (2011). A Bayes Factor Meta-Analysis of Bem's ESP Claim. *Psychonomic Bulletin & Review, 18,* 682-689.

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

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors? *Journal of Personality and Social Psychology, 99,* 917-932.

Stewart, N. (2009), The cost of anchoring on credit-card minimum repayments. *Psychological Science*, *20*, 39–41.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73*, 437-446.

Thorsteinson, T., Breier, J., Atwell, A., Hamilton, C., & Privette, M. (2008). Anchoring effects on performance judgments. *Organizational Behavior and Human Decision Processes, 107,* 29-40.

Tse, C-S., & Neely, J. H. (2007). Semantic and repetition priming effects for Deese/Roediger – McDermott (DRM) critical items and associates produced by DRM and unrelated study lists. *Memory & Cognition, 35,* 1047-1066.

Turner, B. M., & Schley, D. R. (2016). The anchor integration model: A descriptive model of anchoring effects. *Cognitive Psychology, 90,* 1-47.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Wansink, B., Kent, R. J., & Hoch, S. J. (1998). An anchoring and adjustment model of purchase quantity decisions. *Journal of Marketing Research, 35,* 71-81.

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125,* 387-402.

Woltz, D. J. (2010). Long-term semantic priming of word meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1510-1528.

Woltz, D. J., Sorensen, L. J., Indahl, T. C., & Splinter, A. F. (2015). Long-term semantic priming of propositions representing general knowledge. *Journal of Memory and Language, 79-80,* 30-52.

Wright, W. F., & Anderson, U. (1989). Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organizational Behavior and Human Decision Processes, 44,* 68-82.

Zeelenberg, R., & Pecher, D. (2002). False memories and lexical decision: even twelve primes do not cause long-term semantic priming. *Acta Psychologica, 109,* 269-284.

List of Appendices

Appendix 1

Participant number	Which Expt. first	Expt. 1 anchor (1 = low; 2 = high)	Expt. 2 anchor (1 = low; 2 = high)	EXPERIMENTER
1	1	1	2	1
2	1	2	1	1
3	1	1	1	1
4	1	2	2	1
5	2	1	2	1
6	2	2	1	1
7	2	1	1	1
8	2	2	2	1
9	1	1	2	1
10	1	2	1	1
11	1	1	1	1
12	1	2	2	1
13	2	1	2	1
14	2	2	1	1
15	2	1	1	1
16	2	2	2	1
17	1	1	2	1
18	1	2	1	1
19	1	1	1	1
20	1	2	2	1
21	2	1	2	1
22	2	2	1	1
23	2	1	1	1
24	2	2	2	1
25	1	1	2	1
26	1	2	1	1
27	1	1	1	1
28	1	2	2	1
29	2	1	2	1
30	2	2	1	1
31	2	1	1	1
32	2	2	2	1
33	1	1	2	1
34	1	2	1	1
35	1	1	1	1
36	1	2	2	1
37	2	1	2	1
38	2	2	1	1

39	2	1	1	1
40	2	2	2	1
41	1	1	2	1
42	1	2	1	1
43	1	1	1	1
44	1	2	2	1
45	2	1	2	1
46	2	2	1	1
47	2	1	1	1
48	2	2	2	1
49	1	1	2	1
50	1	2	1	1
51	1	1	1	1
52	1	2	2	1
53	2	1	2	1
54	2	2	1	1
55	2	1	1	1
56	2	2	2	1
57	1	1	2	1
58	1	2	1	1
59	1	1	1	1
60	1	2	2	1
61	2	1	2	1
62	2	2	1	1
63	2	1	1	1
64	2	2	2	1
65	1	1	2	2
66	1	2	1	2
67	1	1	1	2
68	1	2	2	2
69	2	1	2	2
70	2	2	1	2
71	2	1	1	2
72	2	2	2	2
73	1	1	2	2
74	1	2	1	2
75	1	1	1	2
76	1	2	2	2
77	2	1	2	2
78	2	2	1	2
79	2	1	1	2
80	2	2	2	2
81	1	1	2	2
82	1	2	1	2
83	1	1	1	2

84	1	2	2	2
85	2	1	2	2
86	2	2	1	2
87	2	1	1	2
88	2	2	2	2
89	1	1	2	2
90	1	2	1	2
91	1	1	1	2
92	1	2	2	2
93	2	1	2	2
94	2	2	1	2
95	2	1	1	2
96	2	2	2	2
97	1	1	2	2
98	1	2	1	2
99	1	1	1	2
100	1	2	2	2
101	2	1	2	2
102	2	2	1	2
103	2	1	1	2
104	2	2	2	2
105	1	1	2	2
106	1	2	1	2
107	1	1	1	2
108	1	2	2	2
109	2	1	2	2
110	2	2	1	2
111	2	1	1	2
112	2	2	2	2
113	1	1	2	2
114	1	2	1	2
115	1	1	1	2
116	1	2	2	2
117	2	1	2	2
118	2	2	1	2
119	2	1	1	2
120	2	2	2	2
121	1	1	2	2
122	1	2	1	2
123	1	1	1	2
124	1	2	2	2
125	2	1	2	2
126	2	2	1	2
127	2	1	1	2
128	2	2	2	2

Note: For Experiments 3 and 4, the experiments are represented by 1 and 2 respectively here.

Appendix 2

The questions in the critical experiments concern prices and temperature. The questions in the practice experiments concern height and length.

Comparative judgments: “Is Big Ben taller or shorter than 15 meters high?”

“Is the M25 longer or shorter than 600 miles?”

Absolute judgments: “How tall is Big Ben?”____ metres

“How long is the M25?”____miles

Lexical Decision Tasks:

The number of total letter strings (70) and the ratio of words to non-words (4:1) is the same as in the critical experiments. No words that, at face value, are related to cars or seasons were included, but some proper nouns were included (as in Experiment 2). Which LDT was associated with which General Knowledge question was randomised.

	LDT 1	LDT 2
Words	hoover snickers oxford asda ducati carlsberg aldi kleenex nescafe manchester abandon absolute accurate	cambridge twix tesco waitrose heineken andrex ikea kenco smarties liverpool ability actual against

	advice	always
	butcher	butler
	butterfly	brutality
	brim	bush
	cabbage	candle
	capacity	cathedral
	detective	dependable
	cure	cube
	desire	describe
	doorway	dreamer
	eternal	evidence
	expression	exhaustion
	errand	essay
	feet	fell
	favourite	farmyard
	fantasy	fidgety
	flock	flood
	graduate	gorilla
	gravity	gratitude
	gracious	graze
	haddock	harbour
	hold	hood
	horizontal	hostility
	infant	insect
	inferno	incident
	jaw	jog
	kick	kid
	land	lamp
	lantern	laundry
	mechanic	medicine
	mercy	merry
	moment	modern
	noun	note
	ornament	orchestra
	path	part
	pathetic	patient
	robot	raspberry
	turnip	tarnish
	twin	turn
	whisky	water
	where	which
	pony	pond
	advantage	adventurer
Non-words	thrauphth	ghroil
	flewd	phlourche
	lawpth	semmed
	spleace	skrife

	furge quoud swawte wrarve brule shroomed duntz bleafs vourque phroarlte	garphed drufe squarl skups thalck scruilds gnuicked bloage tomps teigged

Appendix 3

Participant number	Anchor (1 = low; 2 = high)	Climate change consensus condition (1 - present; 2 - absent)	EXPERIMENTER
1	1	1	1
2	2	1	1
3	1	2	1
4	2	2	1
5	1	1	1
6	2	1	1
7	1	2	1
8	2	2	1
9	1	1	1
10	2	1	1
11	1	2	1
12	2	2	1
13	1	1	1
14	2	1	1
15	1	2	1
16	2	2	1
17	1	1	1
18	2	1	1
19	1	2	1
20	2	2	1
21	1	1	1
22	2	1	1
23	1	2	1
24	2	2	1
25	1	1	1
26	2	1	1
27	1	2	1
28	2	2	1
29	1	1	1
30	2	1	1
31	1	2	1
32	2	2	1
33	1	1	1
34	2	1	1
35	1	2	1
36	2	2	1
37	1	1	1
38	2	1	1

39	1	2	1
40	2	2	1
41	1	1	1
42	2	1	1
43	1	2	1
44	2	2	1
45	1	1	1
46	2	1	1
47	1	2	1
48	2	2	1
49	1	1	1
50	2	1	1
51	1	2	1
52	2	2	1
53	1	1	1
54	2	1	1
55	1	2	1
56	2	2	1
57	1	1	1
58	2	1	1
59	1	2	1
60	2	2	1
61	1	1	1
62	2	1	1
63	1	2	1
64	2	2	1
65	1	1	2
66	2	1	2
67	1	2	2
68	2	2	2
69	1	1	2
70	2	1	2
71	1	2	2
72	2	2	2
73	1	1	2
74	2	1	2
75	1	2	2
76	2	2	2
77	1	1	2
78	2	1	2
79	1	2	2
80	2	2	2
81	1	1	2
82	2	1	2
83	1	2	2

84	2	2	2
85	1	1	2
86	2	1	2
87	1	2	2
88	2	2	2
89	1	1	2
90	2	1	2
91	1	2	2
92	2	2	2
93	1	1	2
94	2	1	2
95	1	2	2
96	2	2	2
97	1	1	2
98	2	1	2
99	1	2	2
100	2	2	2
101	1	1	2
102	2	1	2
103	1	2	2
104	2	2	2
105	1	1	2
106	2	1	2
107	1	2	2
108	2	2	2
109	1	1	2
110	2	1	2
111	1	2	2
112	2	2	2
113	1	1	2
114	2	1	2
115	1	2	2
116	2	2	2
117	1	1	2
118	2	1	2
119	1	2	2
120	2	2	2
121	1	1	2
122	2	1	2
123	1	2	2
124	2	2	2
125	1	1	2
126	2	1	2
127	1	2	2
128	2	2	2

Note: ‘Climate change consensus condition’ is unrelated to the current project.

Supplementary Experiments 1 & 2[16]

In our original pre-registration document, we outlined a "potential validation study" which would:

"…only be run in the result of a failure to replicate the key results of Mussweiler and Strack (the interaction term). In this instance, the failure to replicate might lie in failings with the LDT. Consequently, a standard priming task would be employed in which participants will be shown pictures of summer or winter scenes before completing the 'summer/winter' LDT, or pictures of expensive or inexpensive cars before completing the 'cars' LDT. The same interaction term as tested for in Experiment 1 will be tested, but the 'high/low' anchor variable will be replaced with 'summer/winter' pictures or 'expensive/inexpensive' cars."

Method

As with Experiments 1 & 2 and Experiments 3 & 4, the two proposed experiments are essentially the same experiment using different materials. They are therefore described together here. For complete detail of the method, please refer to the program which has also been uploaded to the Open Science Framework website.

Participants and procedure.

106, predominantly female, first year UCL psychology students were recruited as part of a practical demonstration. In Supplementary Experiment 1 (S Experiment 1), there were 54 participants in the summer condition, and 52 in the winter condition. In Supplementary

[16] Pre-registered as Experiments 5 & 6.

Experiment 2 (S Experiment 2), there were 53 participants in the expensive condition, and 53 in the inexpensive condition.

Participants attended in two groups, who attended sessions at different times over the course of the same afternoon (2:00 and 3:00 pm, GMT, Thursday 9th February, 2017). The majority of participants completed the study in a communal computer room, whilst a minority completed it in adjoining experimental cubicles, which were used as the number of students attending the first session outnumbered the number of computers in the communal room.

In Experiments 1-4, the experimenter remained with participants whilst they read through the instructions for the LDT to clarify that they understood. This procedure was amended for the current set-up. The computer room in which the experiment was run is a teaching room. The experimenter (AJLH) introduced the general theme of the experiment to the whole class and took participants through the 'does it have meaning' instructions. The PowerPoint slides for this introduction are uploaded on the Open Science Framework database. Prior to the class, participants were informed that the 'lab' would be on 'risk perception.' The experiment is ostensibly a general knowledge test, so participants were informed how anchoring can relate to risk perception in the class feedback session (two weeks later), when they were thoroughly debriefed.

Participants were given the opportunity to ask questions, before being provided with a login and password to enable them to access an experimental desktop on the computers, which included a link to the experiment. They entered their unique participant ID (see materials section), gender and age and proceeded through the experiment in silence.

Participants first completed the two practice experiments (exactly as in Studies 1-4). They then completed the two critical experimental tasks (in counterbalanced order), before finally completing an additional conceptual replication of the Mussweiler and Strack (2000) method, which used dog weights as stimuli. Participants completed this final task primarily as

a pedagogical exercise, as the Mussweiler and Strack (2000) paradigm formed the basis of the class feedback session two weeks later. The reason why this is predominantly pedagogical is that the anchor values and chosen words were not pre-tested. The setup of the experiment is described briefly at the bottom of this document (it can be found in full in the program code, https://osf.io/6txdr/).

Once all participants completed the experimental session, they were dismissed.

We attempted to ensure motivation of participants by reminding them that they would have to write a report on the study themselves, thus it was in their own best interests to concentrate fully on the experimental tasks and to keep their eyes on their own computers. Two weeks later, the experimenter (AJLH) provided participants with the results of the final conceptual replication to write-up for course credit.

Design and materials.

The two experiments were both based on a 2 (prime) x 2 (word type) mixed design, with the former factor manipulated between-participants and the latter manipulated within-participants. The difference between the two experiments is that the primes will be pictures of summer/winter scenes in S Experiment 1 and expensive/inexpensive cars in S Experiment 2. The experiment was counterbalanced as in previous experiments (see Appendix 1), although the same experimenter (AJLH) tested all participants. Likewise, the same forward letter span was used as a filler task, as in Experiments 1-4.

For the critical experimental trials, participants were informed:

“You will now be shown a number of images. Please look carefully at them and keep them in mind as you will be asked questions about them later."

They were then presented with the condition-specific images (each displayed for 1.5 seconds, with an ITI of 0 seconds) in a randomised order. Following that, they completed the LDT. After completion of the LDT, they were asked three questions.

For S Experiment 1:

“Thinking about the pictures you saw, which season describes these best? (type 1 for winter, 2 for spring, 3 for summer, 4 for autumn)"

“Thinking about the pictures you saw, on a day like those shown, what temperature is it? (in degrees Celcius [sic])"

“thinking about the pictures you saw, describe in a short sentence (e.g. 6 words) what a person might do on a day like those."

For S Experiment 2:

“Thinking about the pictures you saw, what type of car describes these best? (type 1 for luxury, 2 for economy, 3 for people carrier, 4 for commercial)"

"Thinking about the pictures you saw, how much would a car like those shown cost? (in pounds)"

"Thinking about the pictures you saw, describe in a short sentence (e.g. 6 words) who would drive a car like those."

The second question of each set served as a manipulation check. The additional questions were designed to encourage concentration on the global properties of the pictures, to ensure appropriate attention in the second experiment participants complete as well as the first.

All images had no licensing restrictions for their usage, and are included with the pre-registered experimental program files for these experiments.

There were two additional minor changes from previous experiments:

1) It was not possible to place colored stickers on the computer keyboards in this communal teaching facility. Thus, references to 'blue' and 'yellow' keys were changed to '*P*' and *'Q*' keys. Each workstation had a piece of paper in front of it, reminding participants of the meaning of the keys (as in Experiments 1-4). In addition, a unique participant ID was hand-written on the top-left corner of this piece of paper. At the beginning of the experimental session, the experimenter (AJLH) made participants aware of this, and instructed them to enter the number when requested at the start of the experiment. The ID was, of course, appropriately linked to the meanings of *P* and *Q* displayed on the reminder paper, as outlined in the original Appendix 1.

2) Participants entered demographic information at the start of the experiment, rather than this being recorded separately. In addition to age and gender, they indicated whether their native language was English or 'other.' The final question was required as it is inappropriate to exclude non-native English speakers from a class exercise. All participants are necessarily fluent in English to be undergraduate students at UCL.

Results

Exclusion criteria.

For the reaction time data, following log transformations, any trials with reaction times more than 3 standard deviations either side of the mean were excluded from analysis. This resulted in excluding 57 responses from the winter prime, and 51 responses from the summer prime condition in S Experiment 1, and excluding 57 responses from the inexpensive, and 55 responses from the expensive prime condition in S Experiment 2.

Pre-registered manipulation check.

After the LDT task, participants in the "temperature" experiment decided which season the images represented, and the likely temperature. 49 out of 54 participants primed with summer images chose "summer"" as the representative season, with the remainder choosing "spring". All participants primed with winter images chose "winter" as the representative season. Participants primed with summer images rated the temperature higher (mean = 28.61, SD = 4.48) than those primed with winter images (mean = 0.25, SD = 0.76), $t(104) = 44.99$, $p < .001$. In the "cars" experiment, participants judged the type of car most representative of the images, and the price of the car. All participants primed with images of expensive cars chose "luxury" as the most representative type of car. 39 out of 53 participants primed with images of inexpensive cars chose "economy", while 6 chose "carrier", 7 chose "commercial", whilst only one chose "luxury". After excluding nine extreme responses ($\leq £100$ and $\geq £1{,}000{,}000$), those participants primed with expensive cars rated the price as higher (mean = 324,824, SD = 737,199) than those primed with inexpensive cars (mean = 7,350, SD = 14,545), $t(100) = 3.07$, $p = .003$.

Planned analyses.

Lexical decisions.

In the analysis on facilitation scores, there was no evidence of an interaction between prime and word type (summer, winter) in S Experiment 1, $F(1,104) = 0.11$, $MSE = 0.00$, $p = .741$, $\hat{\eta}_G^2 = .001$, or between word type (expensive, inexpensive) and prime in S Experiment 2, $F(1,104) = 1.46$, $MSE = 0.00$, $p = .229$, $\hat{\eta}_G^2 = .004$. The main effects of word type and prime were not significant in S Experiment 1 (all $Fs < 1$). In S Experiment 2, there was a significant main effect of word type, $F(1,104) = 16.00$, $MSE = 0.00$, $p < .001$, $\hat{\eta}_G^2 = .039$, but not of prime ($F < 1$).

A 2 (prime: summer/expensive vs winter/inexpensive) x 4 (word type: summer/expensive, winter/inexpensive, neutral, non-word) ANOVA also did not show a significant interaction between word type and prime in S Experiment 1, $F(2.42,251.43) = 0.98$, $MSE = 0.00$, $p = .387$, $\hat{\eta}_G^2 = .002$, or S Experiment 2, $F(2.64,274.34) = 0.62$, $MSE = 0.00$, $p = .585$, $\hat{\eta}_G^2 = .001$ (Greenhouse-Geisser correction due to a violation of sphericity). For both experiments, there was a significant main effect of word type: for S Experiment 1, $F(2.42,251.43) = 30.25$, $MSE = 0.00$, $p < .001$, $\hat{\eta}_G^2 = .055$, and for S Experiment 2, $F(2.64,274.34) = 82.04$, $MSE = 0.00$, $p < .001$, $\hat{\eta}_G^2 = .139$. The main effect of prime was not significant for either experiment (both $F < 1$; see Figure S1 for descriptives).

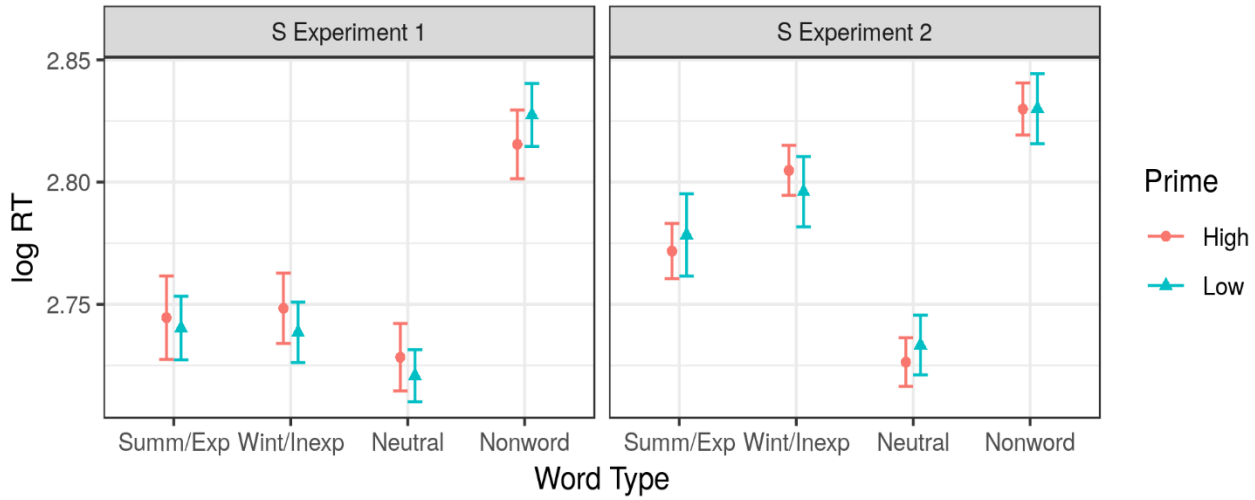

Figure S1. Mean log reaction times in Supplementary Experiments 1 and 2. The 'High' prime is summer in S Experiment 1 and expensive cars in S Experiment 2. The 'Low' prime is winter in S Experiment 1 and inexpensive cars in S Experiment 2. Error bars are +/- 1 S.E.

Combining S Experiments 1 and 2.

As before, we estimated two models which had fixed effects for word type, prime, and experiment, as well as all two- and three-way interactions, but differed in their random-effects structure. Model 1 included random participant-specific intercepts crossed with random word-specific intercepts. Model 2 included, in addition, random participant-specific slopes for word-type, experiment and prime. The model with participant-specific slopes for word-type, experiment and prime fitted significantly better than the model with only participant-specific intercepts crossed with random word-specific intercepts, $\chi^2(5) = 232.70$, $p =< .001$, hence we will report the results for that model. Neither the two-way interaction between prime and contrast C_3, $t(560.93) = 1.32$, $p = .19$, nor the three-way interaction between experiment, prime and contrast C_3 was significant, $t(543.39) = 0.7$, $p = .49$.

Supplementary Experiment 3[17]

Method

For complete detail of the method, please refer to the program which has also been uploaded to the Open Science Framework website.

Participants and procedure.

125, predominantly female, first year UCL psychology students were recruited as part of a practical demonstration. There were 55 participants in the summer condition, and 70 in the winter condition.

Participants attended in three groups, who attended sessions at different times over the course of the same afternoon (2:00, 3:00, and 4:00 pm, GMT, Thursday 8th February, 2018). The majority of participants completed the study in a communal computer room, with a minority completing it in adjoining experimental cubicles, which were used as the number of students attending some sessions outnumbered the number of computers in the communal room.

The experimenter (AJLH) provided generic instructions to participants before instructing them to start the experimental program on their individual computers. Predominantly, participants were asked to read the instructions carefully, and reminded that they would be writing up the results of this experiment, so it was in their interests to pay attention and complete the experiment with due diligence. In addition (as with Supplementary Experiments 1 & 2), prior to the class, participants were informed that the 'lab' would be on 'risk perception.' The experiment is ostensibly a general knowledge test, so participants were again informed how anchoring can relate to risk perception in the class feedback session (two weeks later), when they were thoroughly debriefed. Specific instructions were presented in the experimental program, which has been uploaded to OSF (https://osf.io/jtx34/).

[17] Experiment 8 in the pre-registration

Participants were provided with a login and password to enable them to access an experimental desktop on the computers, which included a link to the experiment. They entered their unique participant ID (see materials section), gender, age, and indicated whether English is their native language, before proceeding through the experiment in silence.

Participants first completed the 'Big Ben' practice experiment (exactly as in Experiment 5). They next completed the critical experimental task, before finally completing a conceptual replication (using the CID) of the Mussweiler and Strack (2000) method. They completed this final task solely as a pedagogical exercise, as the Mussweiler and Strack (2000) paradigm will be the basis of the class feedback session two weeks later. The reason why this is solely pedagogical is that we made no effort to control or pre-test the words used in this CID task. Because this task occurs after the critical task, we do not expand on it here. Full details of it can, however, be found in full in the program code.

Once all participants have completed the experimental session, they were dismissed. Participants were fully debriefed as to the purpose of the experiment in a class feedback session two weeks following the experimental session.

We aimed to ensure motivation of participants by reminding them that they would have to write a report on the study themselves, thus it was in their own best interests to concentrate fully on the experimental tasks and to keep their eyes on their own computers. Two weeks later, the experimenter (AJLH) provided participants with the results of the final conceptual replication to write-up for course credit.

Design and materials.

The analysis of the CID section of the critical task was analysed in the same way as Experiment 5, with 'prime [summer/winter]' replacing 'anchor [high/low]' as the between-participants factor.

For the critical experimental trials, participants were informed:

“You will now be shown a number of images. Please look carefully at them and keep them in mind. You will later be asked questions relating to the season depicted" (note the slight change in wording from Supplementary Experiment 1 – designed to increase the power of the priming manipulation).

They were then presented with the condition-specific images (each displayed for 1.5 seconds, with an ITI of 0 seconds) in a randomised order. In a change from the order of Supplementary Experiment 1, so as to increase the strength of the prime, participants were next asked the three questions below, before continuing to the CID task (which was identical to that used in Experiment 5):

“Thinking about the pictures you saw, which season describes these best? (type 1 for winter, 2 for spring, 3 for summer, 4 for autumn)"

“Thinking about the pictures you saw, on a day like those shown, what temperature is it? (in degrees Celcius [sic])"

“Thinking about the pictures you saw, describe what you might do on a day like those." (note that this sentence was slightly reworded from Experiment 5, encouraging participants to potentially write a little more, and also to take a first person perspective in relation to the priming manipulation).

Results

Exclusion criteria.

Exclusions of extreme log reaction times resulted in excluding 18 responses from the low anchor and 10 responses from the high anchor condition.

Pre-registered manipulation check.

After the CID task, participants decided which season the images represented, and the likely temperature. 53 out of 55 participants primed with summer images chose “summer”" as the representative season, while one participant chose “spring” and another “autumn”. All participants primed with winter images chose “winter” as the representative season. Participants primed with summer images rated the temperature higher (mean = 28.49, SD = 5.5) than those primed with winter images (mean = 0.25, SD = 0.76), $t(123) = 41.21$, $p < .001$.

Planned analyses.

Lexical decisions.

In the analysis on facilitation scores, there was no evidence of an interaction between prime and word type (summer, winter), $F(1,123) = 0.84$, $MSE = 0.00$, $p = .361$, $\hat{\eta}_G^2 = .002$. Neither the main effect of word type, $F(1,123) = 0.39$, $MSE = 0.00$, $p = .536$, nor the main effect of prime were significant, $F(1,123) = 0.50$, $MSE = 0.00$, $p = .481$.

Focussing only on the word “summer”, which was the first target word to appear, did show a significant effect of prime, $\Delta M = 0.06$, 95% CI [0.00, 0.11], $t(121) = 2.01$, $p = .046$, but in the opposite direction to that predicted. Responses to the word summer were faster following winter primes (mean log RT = 2.98, SD = 0.14) than summer primes (mean log RT = 3.04, SD = 0.18).

A 2 (prime: summer / winter) x 3 (word type: summer / winter / neutral) ANOVA also did not show a significant interaction between word type and prime, $F(1.86,229.16) = 0.71$, $MSE = 0.00$, $p = .482$, $\hat{\eta}_G^2 = .000$ (Greenhouse-Geisser correction applied due to a violation of sphericity). There was a significant main effect of word type, $F(1.86,229.16) = 24.78$, $MSE = 0.00$, $p < .001$, $\hat{\eta}_G^2 = .011$, while the main effect of prime was not significant, $F(1,123) = 0.96$, $MSE = 0.04$, $p = .328$, $\hat{\eta}_G^2 = .007$ (see Figure S2 for descriptives).

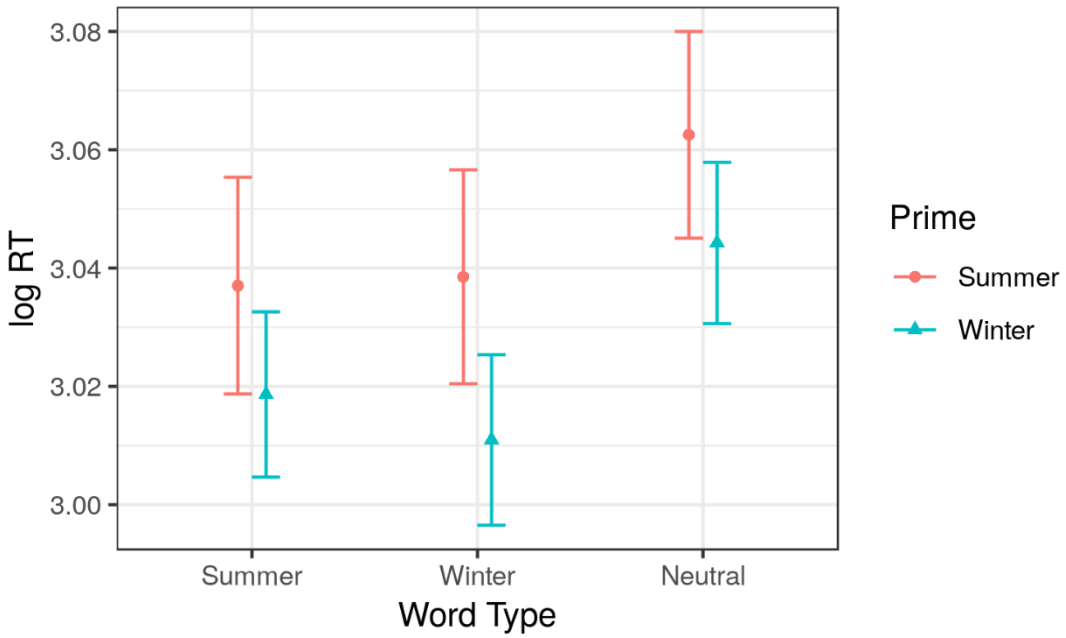

Figure S2. Mean log reaction times in Supplementary Experiment 3. Error bars are +/- 1 S.E.

Linear mixed-effect analysis.

A comparison between a model with crossed random intercepts for participants and stimuli did not fit significantly worse than a model with additional participant-specific slopes for the two contrasts comparing neutral words to target words (C_1) and comparing summer words to

winter words (C_2), $\chi^2(2) = 1.54$, $p = .402$. Hence, we report results for the first, simpler model. This analysis did not show evidence of the crucial interaction between prime and C_2, $t(4896.11) = 0.95$, $p = .34$. The only significant effect was the main effect of contrast C_1, $t(38.97) = 2.05$, $p = .047$. None of the other fixed effects were significant (all $|t| < 1$).

Exploratory analyses.

When running S Experiment 3, we noticed that some computers appeared to have lower refresh rates, which resulted in a slower presentation rate of the CID, making the task appear easier. Subsequent investigation revealed that this was related to a different graphics driver in some computers (57 computers had a 'normal' driver, and 68 had a 'slow' driver). Including a factor for Driver in the linear mixed effects model showed a significant effect of Driver, $F(1, 123.37) = 9.29$, $p = .003$, whereby participants' responses were indeed faster on 'slower' computers. Consequently, we repeated all the analyses above, excluding data from the 'slow' computers. The significance patterns revealed in the main analysis were all replicated in this analysis (all F values from the critical interaction in the ANOVAs < 0.2; in the linear mixed-effect analysis, the interaction between prime and C_2: $t(2199) = 0.55$, $p = 0.58$; for the word 'summer', responses were faster following the winter prime [mean log RT = 3.00] than the summer prime [mean log RT = 3.09], $t[54] = 2.97$, $p = .004$).