Need for Cognition, Cognitive Load, and Forewarning do not Moderate Anchoring Effects

A Replication Study of Epley & Gilovich (*Journal of Behavioral Decision Making, 2005; Psychological Science, 2006)*

Lukas Röseler[*1,2] Hannah L. Bögler[2] Lisa Koßmann[2]
Sabine M. Krueger[2] Sabrina L. C. Bickenbach[2] Ricarda Bühler[2]
Jasmin della Guardia[2] Lisa-Marie A. Köppel[2] Jarl Möhring[2]
Susanne Ponader[2] Konstantin Roßmaier[2] Jessica Sing[2]

Journal of Comments and Replications in Economics, Volume 3, 2024-8, DOI: 10.18718/81781.38

JEL: D91
Keywords: Anchoring effect, Self-generated, Experimenter-provided, Replication, Anchoring and adjustment

Data Availability: The data and R code to reproduce the results of this replication can be downloaded at JCRE's data archive (DOI: 10.15456/jl.2024270.0728512269).

Please Cite As: Röseler, L. et al. (2024). Need for Cognition, Cognitive Load, and Forewarning do not Moderate Anchoring Effects. A Replication Study of Epley & Gilovich (2005, 2006). *Journal of Comments and Replications in Economics*, Vol 3(2024-8). DOI: 10.18718/81781.38 .

---

[*]Corresponding author. Email: lukas.roeseler@uni-muenster.de
[1] Münster Center for Open Science, University of Münster
[2] University of Bamberg

Declaration: The authors declare that they have no conflicts of interest. The research reported in this paper did not result from a for-pay consulting relationship and our employer has no financial interest in the paper's topic.

Received October 31, 2023; Revised May 15, 2024; Accepted August 26, 2024; Published October 09, 2024.

Abstract

Anchoring, the assimilation of numerical estimates toward previously considered numbers, has generally been separated into anchoring from *self-generated anchors* (e.g., people first thinking of 9 months when asked for the gestation period of an animal) and *experimenter-provided anchors* (e.g., experimenters letting participants spin fortune wheels). For some time, the two types of anchoring were believed to be explained by two different theoretical accounts. However, later research showed crossover between the accounts. What now remains are contradictions between past and recent findings, specifically, which moderators affect which type of anchoring. We conducted three replications ($N_{\text{total}} = 657$) of seminal studies on the distinction between self-generated and experimenter-provided anchoring effects where we investigated the moderators need for cognition, cognitive load, and forewarning. We found no evidence that either type of anchoring is moderated by any of the moderators. In line with recent replication efforts, we found that anchoring effects were robust, but the findings on moderators of anchoring effects should be treated with caution.

1 Introduction

What is the freezing point of vodka? And are there more or fewer than nine African states in the UN? These two seemingly unrelated questions are examples of two different kinds of anchoring questions. That is, both have shown anchoring effects (Epley & Gilovich, 2001; Tversky & Kahneman, 1974), which occur when people's estimates are biased toward previously considered anchors (0°C as the freezing point of water; nine African states), but the two have usually been explained by different mechanisms. Anchoring researchers have hypothesized that most people estimate the freezing point of vodka by memorizing the freezing point of water (self-generated anchor) and adjusting away from it because they know that vodka freezes at colder temperatures. In doing so, they fail to adjust far enough, and their estimates are thereby biased toward the self-generated anchor. This has been referred to as insufficient adjustment model (e.g., Epley & Gilovich, 2001). However, when experimenter-provided anchors are present, people do not adjust away from the anchor but instead engage in hypothesis-consistent testing, that is, they generate reasons in favor of the anchor while simultaneously priming numeric values that are close to the anchor. This process has been referred to as selective accessibility (e.g., Mussweiler & Strack, 1999a).

2 Contradictory Findings from Past Research

The *insufficient adjustment model* and the *selective accessibility model* are the two most prominent anchoring models. Insufficient adjustment was the first explanation for anchoring effects (e.g., Tversky & Kahneman, 1974, p. 1228), whereas selective accessibility was proposed later and made prominent by Mussweiler and Strack (1999a, 1999b). Noting that little attention had been paid to the insufficient adjustment account, Epley and Gilovich (2001, 2004, 2005, 2006, 2010) argued extensively that insufficient adjustment accounts for anchoring effects that are caused by self-generated anchors, and selective accessibility accounts for effects that are caused by experimenter-provided anchors. In their experiments, Epley and Gilovich showed that things such as cognitive load, need for cognition, or forewarnings affect adjustment from *self-generated* anchors but not adjustment from *experimenter-provided* anchors. Notably, the direction of adjustment was known for self-generated anchors but not for experimenter-provided anchors in Epley and Gilovich's experiments (i.e., people were aware that the anchor was too high or too low for the self-generated anchors). By showing that the motivation to be accurate leads to more adjustment away from anchors *only when the direction of adjustment is known* but regardless of the type of anchor, Simmons et al. (2010) made a case that insufficient adjustment accounts for experimenter-provided anchors, too. And later, Chaxel (2014) showed that selective accessibility accounts for self-generated anchors, too. Thus, by 2014, the distinction between the two kinds of anchors that had been fostered by a decade of research had to be dropped again. According to this logic, the moderators that Epley and Gilovich investigated should also moderate experimenter-provided anchoring effects as long as the direction of adjustment is known. Despite their theoretical relevance, the moderators that Epley and Gilovich investigated have received little attention since Simmons et al.'s (2010) findings.

Moreover, recent findings have challenged the view that two different theories are needed to explain the two kinds of anchors: First, Harris et al. (2019) failed to replicate a "signature test for the operation of selective accessibility mechanisms" (Abstract), and Bahník (2021) showed that anchors do not activate information that is consistent with the anchor (see also Frederick & Mochon, 2012 for criticism of the selective accessibility mechanisms). Additionally, in Epley and Gilovich's

(2010) Study 2a, susceptibility to self-generated anchors was correlated with need for cognition (but susceptibility to experimenter-provided anchors was not). However, research testing the reliability of susceptibility to anchoring scores (e.g., Röseler, 2021; Röseler et al., 2019; Schindler et al., 2021) has found extremely low values (i.e., average interitem correlations close to zero). Thereby, despite the findings reported by Epley and Gilovich (2006, Study 2a), variables such as need for cognition cannot be correlated with anchoring.

Table 1: Summary of Findings Regarding the Relationship Between Self-Generated and Experimenter-Provided Anchors

No difference between anchor types	**EP + SA and SG + IA**
- Simmons et al. (2010): Insufficient adjustment is valid for both types of anchors as long as the direction of adjustment is known - Chaxel (2014): Selective accessibility is valid for both types of anchors - Harris et al. (2019) and Bahník (2021): the selective accessibility model is problematic	- Epley and Gilovich (2001, 2004, 2005, 2006): Self-generated anchors are moderated by need for cognition, forewarning, monetary incentives, cognitive load, head movement, arm flexion, and alcohol consumption

Notes: EP = experimenter-provided, SG = self-generated, SA = selective accessibility, IA = insufficient adjustment.

3 *Resolving the Contradictions*

Epley and Gilovich's findings have been fundamental for the development of the insufficient adjustment model of anchoring and are highly cited (e.g., 1328 citations of the 2006 paper according to Google Scholar in April 2024). Yet, they strongly contradict more recent work. The aim of this research is to bring clarity to the contradictions surrounding self-generated and experimenter-provided anchors: How can two mechanisms be responsible for the two types of anchoring if there is no evidence for one of them (i.e., selective accessibility)? Why is there no published evidence that moderators of self-generated anchoring effects also affect experimenter-provided anchoring? We chose to begin by conducting close replications of three seminal studies from Epley and Gilovich's research in which an intervention affected susceptibility to self-generated anchors but not experimenter-provided anchors. We chose studies that could be conducted online as all of them were conducted between June 2020 and March 2022 during the COVID-19 pandemic. We report all studies in the order in which they were conducted. Note that Studies 2 and 3 began at the same time, but the recruitment of participants for Study 3 took longer.

- Study 1 is a replication of Epley and Gilovich (2006, Study 2a) and investigated the moderator *need for cognition*.
- Study 2 is a replication of Epley and Gilovich (2006, Study 2c) and investigated the moderator *cognitive load*.
- Study 3 is a replication of Epley and Gilovich (2005, Study 2) and investigated the moderator *forewarning*.

All available study materials and data sets can be found online (https://osf.io/prwu6). We invite other researchers to reanalyze our data or to conduct studies using our materials and to replicate the results. We report how we determined our sample sizes, all data exclusions (if any), all manipulations, and all measures in the studies (Simmons et al., 2012). In all studies, we used SoSci Survey (Leiner, 2019) to program the studies, and we used R (R Core Team, 2023) and the packages cocor (Diedenhofen & Musch, 2015), data.table (Dowle & Srinivasan, 2024), dplyr (Wickham et al., 2023), ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), lmerTest (Kuznetsova et al., 2017), lubridate (Grolemund & Wickham, 2011), MBESS (Kelley, 2023), metafor (Viechtbauer, 2010), psych (Revelle, 2024), pwr (Champely, 2020), reshape (Wickham, 2007), and xlsx (Dragulescu & Arendt, 2020) to analyze the data. An overview of the main results is provided in Figure 1.

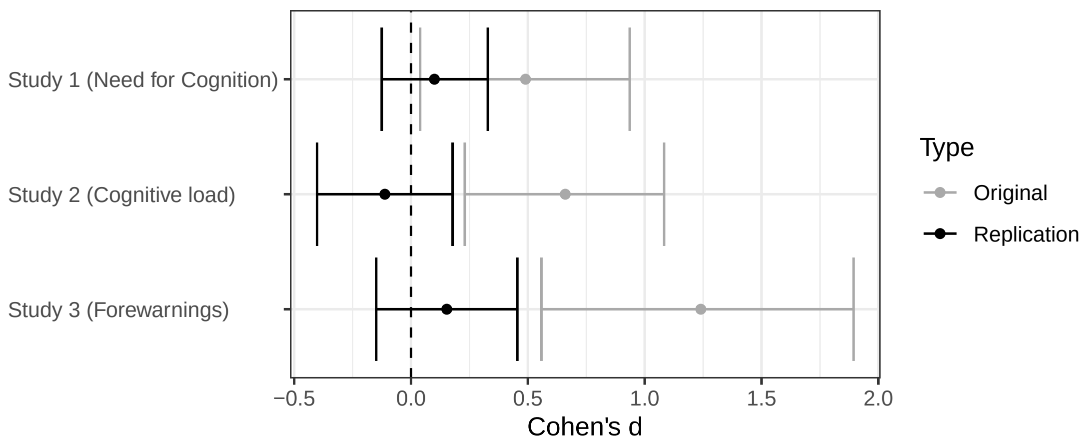

Figure 1: Overview of effects of moderators on adjustment from self-generated anchors in original and replication studies

Notes: Evaluation of results according to LeBel et al.'s (2019) terminology is no signal – inconsistent. All tests for replication studies were preregistered. Code to reproduce this figure: https://osf.io/pe5kn

4 *Study 1: Replication of Epley and Gilovich, 2006, Study 2a (Need for Cognition)*

4.1 *Method*

Need for cognition is defined as "the tendency for an individual to engage and enjoy thinking" (Cacioppo & Petty, 1982, p. 116). In terms of anchoring and adjustment, more thinking corresponds to more adjustment away from the anchor and thus less susceptibility to anchors. If insufficient adjustment occurs for self-generated anchors but not for experimenter-provided anchors, only the former should be correlated with need for cognition. To test whether need for cognition is negatively correlated with susceptibility to self-generated anchors and uncorrelated with susceptibility to experimenter-provided anchors, we conducted a preregistered close replication of Epley and Gilovich (2006, Study 2a).

We used Epley and Gilovich's descriptions to create new materials because the original ones were not available. The entire study (hypothesis, procedure, materials, analysis script) was preregistered (https://osf.io/8k9nt) using the replication recipe (Brandt et al., 2014).

4.1.1 A Priori Sample Size Determination

The original effect size of the difference in susceptibility to self-generated anchors between people who scored high versus people who scored low in need for cognition was $d = 0.49$, CI 95% $[0.039, 0.936]$. The main effect of need for cognition on susceptibility to experimenter-provided anchors was not reported. The effect size for the interaction between need for cognition and anchor type was $\eta^2 = .05$. In the original study, only people with need for cognition scores in the upper and lower quintiles were used. We conducted a simulation to see which correlations this effect would apply to if the data had been collected from people with a normal distribution of need for cognition values instead of the extreme quintiles. We planned to collect data from $N = 240$ participants so that the statistical power for detecting the correlation between self-generated anchors and need for cognition ($r = .16$) would be 80%. We aimed to achieve 80% power only for practical reasons. When self-generated anchoring items are used in past studies, many participants who do not think of the intended self-generated anchor must be excluded (e.g., Epley & Gilovich, 2006, p. 314; more than 10% of all participants, Röseler et al., 2020, p. 8). The simulation and power analyses are available online (https://osf.io/n2gtu).

4.1.2 Materials

Need for cognition was measured with the validated German version of the need for cognition scale by Bless et al. (1994). To measure the susceptibility to anchoring, we had to deviate from the original experiment to some degree: There were four self-generated anchoring items and four experimenter-provided anchoring items, the latter of which were taken from Jacowitz and Kahneman (1995). Unfortunately, the original study did not disclose which four of the 15 items (Jacowitz & Kahneman, 1995, p. 1163, Table 1) they used. Thus, we chose items that we believed would work well with a German sample instead of a U.S. sample. Due to the strict exclusion criteria for self-generated anchoring items (i.e., participants must know the self-generated anchor and must indicate that they had thought of it when giving their estimate), we chose eight items, at least four of which had to remain after the exclusion criteria were applied. These items were a combination of items that were translated or adapted from the original study (four items), items that were taken from Röseler et al. (2020; three items), and a newly created item (one item). An overview of all items and their respective source, type, anchor, and true value are available online (https://osf.io/9dnez).

4.1.3 Procedure

Participants were greeted and told that the purpose of the experiment was to test their general knowledge. Due to difficulties in recruiting participants, we added non-monetary incentives for participation for 74% of the final sample (all students were offered course credit but only later participants were offered feedback on the correct values for the general knowledge questions). After we collected demographic data, participants completed the need for cognition scale. Experimenter-provided anchoring items were presented with fixed anchors and comparative questions (e.g., Are there more or fewer than 127 African members in the UN?). For each item, participants answered the comparative question (more or less/fewer), gave their estimate (How many African members are there in the UN?), and—for exploratory purposes—indicated how sure they were about their estimate on a 10-point scale with labeled extremes (not at all, very much). For self-generated anchoring items, participants only gave estimates. The anchoring items were presented on two subsequent pages (one for each type). Then, participants were asked whether they knew the values that they were supposed to use for the self-generated anchoring questions (e.g., the average human body temperature) and whether they thought of the value when giving their estimate (e.g., the lowest temperature measured in a living human). Finally, for exploratory purposes, participants were asked whether they knew or had heard about anchoring effects and whether they had thought about anchoring effects when completing the study. An overview of the procedure is provided in Figure 2.

4.1.4 Design

For the experimenter-provided anchoring items, we manipulated whether participants were given high or low anchors. To minimize programming efforts for the questionnaire and the analyses, participants were given anchors that were either *high-high-low-low* or *low-low-high-high* for the four experimenter-provided anchoring items. All participants gave estimates for all 12 anchoring items. There were no further manipulations. Need for cognition was not manipulated as it is relatively stable over time (Bruinsma & Crutzen, 2018).

4.1.5 Statistical Analyses

As per the preregistered analysis script, susceptibility to anchoring was computed as the absolute difference between anchor and estimate (*absolute adjustment*) for all items. Absolute adjustment was standardized (i.e., mean-centered and scaled) by item and aggregated anchor type. Means were computed by anchor type to operationalize the susceptibility to anchoring.

Participants were excluded if (a) the need for cognition scale had not been answered or (b) they did not know or think of the intended self-generated anchors for at least five out of the eight self-generated anchoring items. Single scores were excluded if (a) they corresponded to the correct value or (b) their absolute standardized (by item) absolute adjustment was $Z > 3$.

4.1.6 Deviations From the Original Study

We decided to deviate from the original study in some respects because of the COVID-19 pandemic, insufficient reporting in the original study, and some other practical reasons. For example, our study was conducted in Germany, requiring us to translate the anchoring questions and think of new ones. For example, few German people know when Washington was elected president, which is an item used by Epley and Gilovich. Moreover, we refrained from using the original anchor of 30 miles/hr

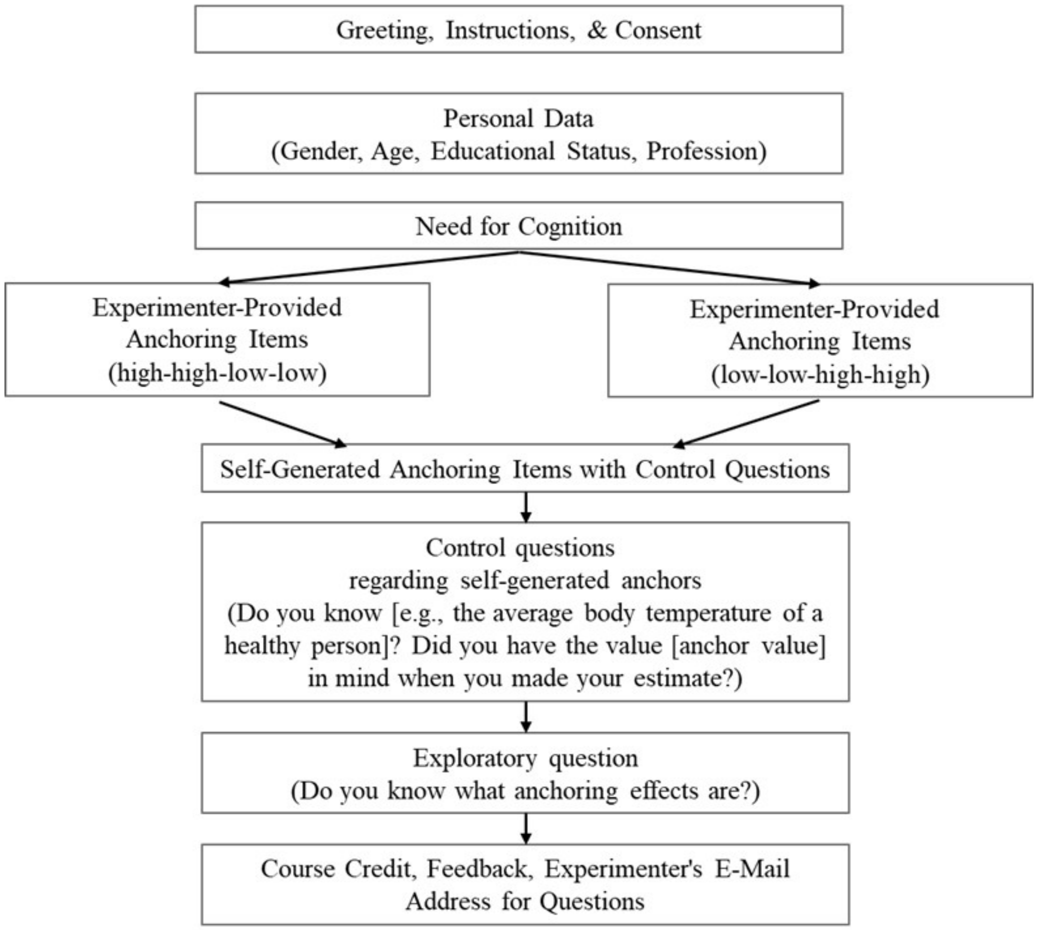

Figure 2: Procedure Used in Study 1

for the maximum speed of a housecat because the actual maximum speed of a house cat is about 30 miles/hr (new anchor: 37.28 miles/hr or 60 km/hr). Another deviation was that we used all the need for cognition scores instead of just the extreme quintiles. This choice was made possible by our larger sample size and more diverse sample. And we used more self-generated anchoring items (eight items for which there had to be at least four estimates instead of only four items) due to strict exclusion criteria. Compensation, type of instructions (e.g., the purpose of the experiment), age, gender, and control questions were not reported in the original study, and thus, we used common standards from the field of anchoring research. We believe that all of these changes were necessary for the study to work. An overview is provided in Table 2.

4.1.7 Deviations From the Preregistration

The original analysis script for the study had been preregistered, but there were errors and code that did not work as intended. Therefore, we decided to deviate from the preregistration in some cases. All differences between the preregistered and final analysis scripts are highlighted. Both R scripts are available online.

(a) The processing of variables had to be refined because the initial codes resulted in the incorrect recognition of some values (e.g., 100.00 was recognized as 100000). (b) We excluded values that were nonsensical (e.g., 0 m for the height of the second highest mountain, fewer than 7 sides for the die with the most sides). (c) One of the self-generated anchoring items was the height of the second highest mountain, which required participants to know the height of Mount Everest (i.e., the highest mountain). At the same time, one experimenter-provided anchoring item was the height of Mount Everest, which required participants not to know the height of Mount Everest. We excluded the latter item from our analyses. (d) Outlier exclusions (i.e., exclusion of mean estimates by type of anchor for values that lay ± 3 SD away from the mean estimate) led to extremely inaccurate values not being excluded (e.g., -73°C for the minimum body temperature of a living human, 89,000 m for the second highest mountain). We changed the code to exclude outliers on an item-by-item basis instead. (e) We changed the test for the need for cognition hypothesis to be one-tailed and thus match the hypothesis written in the preregistration and in the original study.

Table 2: Methodological Differences Between the Original Study and our Replication Study 1

Study feature	**Epley & Gilovich, 2006, 2a**	**This study**	**Reason for change**
Language of questionnaire	English	German	German participants
Type of sample	Only college students	College students and non-students	Heterogeneous sample should increase NFC variance and thereby the effect size
Compensation	Not reported	Course credit and feedback	Facilitate participant recruitment
Type of instructions	Not reported	Test of general knowledge	Original materials were not available
Type of study	Study was conducted on site	Online study	COVID-19, larger sample
Further variables that were collected	Not reported	Age, sex, educational status, profession	Original materials were not available
Need for Cognition	Need for Cognition Scale by Cacioppo and Petty (1982)	Need for Cognition Scale by Bless et al. (1994)	German participants
Preselection of participants	Only participants with extreme NFC scores answered anchoring questions	All participants answered anchoring questions regardless of their NFC score	More efficient; more accurate estimation of effect size because no quintile split was applied
Experimenter-provided anchor	Four items that were not specified any further (there were 15 items in the source; Jacowitz & Kahneman, 1995)	Choice of four experimenter-provided items from the source (Jacowitz & Kahneman, 1995)	Original materials were not available
Self-generated anchor	Four items	Four items from the original study plus four additional items	Participants who did not think of or did not know the anchor would have had to have been excluded
Exclusion criteria	Not reported	Did the participants think of the anchor? Were estimates outliers ($\|Z\| > 3$)?	To prevent distortion by participants who did not think of the anchor or whose estimates were unrealistic
Units of anchor items	Feet, Fahrenheit, mph	Converted to meters, Celsius, km/h	German participants
Exploratory items	Not reported	Asked participants whether they were familiar with anchoring and whether they thought about it while filling out the questionnaire	Exploratory

Notes: NFC = Need for Cognition.

4.2 Results

4.2.1 Sample

Between July 18, 2021 and September 17, 2021, a total of 462 participants clicked on the link to the experiment of which 303 completed the study and were not excluded by our exclusion criteria.

The mean age was 44.49 years (two missing values), and there were 210 female participants and 93 male participants. In order to achieve a large variance in need for cognition, we tried to recruit nonstudents. There were 46 psychology students and 34 students studying other subjects in the sample. As we exceeded our planned sample size, the achieved power for detecting the correlation between self-generated anchors and need for cognition ($r = .16$) was 87.60% and correlations of $r \geq .142$ could be detected with 80% power.

4.2.2 Data Quality Checks

The internal consistency of the need for cognition scale was in line with our expectations, $\alpha = .89$, 16 items, $N = 303$ (Bless et al., 1994, $\alpha = .86$). Values ranged from 20 to 109 ($M = 80.09, SD = 14.86, N = 303$; lowest and highest possible values were 16 and 112, respectively). Also in line with previous research (Röseler et al., 2019; Schindler et al., 2021), the internal consistencies of the experimenter-provided anchoring items and self-generated anchoring items were very low ($\alpha_{\text{experimenter-provided}} = -.25$, three items, $N_{\text{min}} = 303$; $\alpha_{\text{self-generated}} = .10$, eight items, $N_{\text{min}} = 115$). Anchoring effects were significant for all three experimenter-provided anchoring items (all ds ≥ 0.89). To our knowledge, previous research did not test for whether there were actually any anchoring effects for the self-generated anchoring items. We checked for whether the estimate was significantly different from the true value in the direction of the anchor, which was the case for two of the eight items.

4.2.3 Hypothesis Tests

The correlation between susceptibility to self-generated anchors and need for cognition was not significantly different from zero and positive, $r(301) = .050$, 95% CI $[-1, .14]$, $p = .809$ (one-tailed), and the correlation between susceptibility to experimenter-provided anchors and need for cognition was not significantly different from zero, $r(301) = -.011$, $p = .849$ (two-tailed).

4.3 Discussion

We conducted a close replication of Epley and Gilovich (2006, Study 2a) to test whether need for cognition was correlated with susceptibility to self-generated anchoring items but not with susceptibility to experimenter-provided anchoring items. Despite the fact that we made changes that we believed were necessary for the study to work, there were no correlations between need for cognition and susceptibility to anchoring.

A notable deviation is the difference in items. Epley and Gilovich chose four out of the 15 items reported by Jacowitz and Kahneman (1995) but did not report the exact items. If the hypothesis might not have been true for some items, we believe that they should have justified their choice. If anything, we believe that the choice of items may have increased the chances to replicate the effect as we translated the items to German language and culture. For example, we took the "Mount Everest" or the "Telephone" items from Jacowitz and Kahneman (1995). These items should work for German participants like they work for North American participants (see https://osf.io/9dnez) for the complete table of items. Figure 8 also provides an overview of the item-specific associations with all moderators. In fact, the largest positive and negative effects are for new items (election

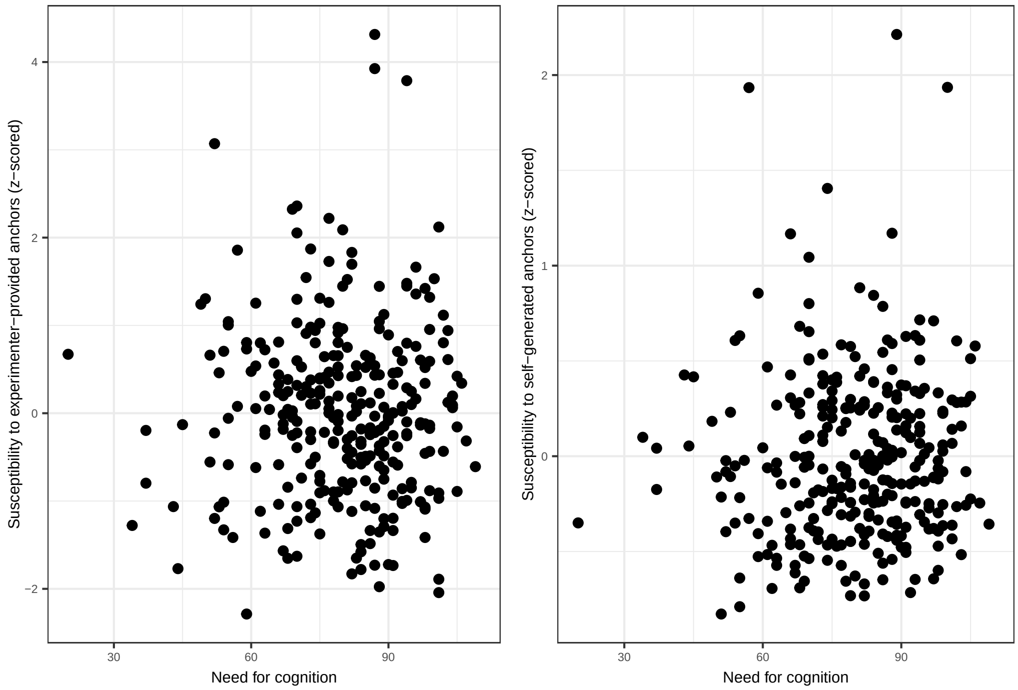

Figure 3: Scatterplots for the Relationship Between Need for Cognition and Susceptibility to the Two Types of Anchors

Notes: For the sake of transparency, we included outliers in these plots. Code to reproduce figure: https://osf.io/wzhmyhttps://osf.io/pe5kn

year of the German chancellor Gerhard Schröder and Number of Bundesländer in the BRD before the reunion with the DDR).

As the original correlation between susceptibility to self-generated anchoring items and need for cognition had a very wide confidence interval (95% CI $[0.039, 0.936]$), a very large sample size would be necessary to show that the nonsignificant correlation was smaller than the original correlation. However, as our sample size was about 3.7 times the original sample size ($N_{\text{original}} = 81, N_{\text{replication}} = 303$), we do not think that there was enough evidence in favor of a correlation. Note that we did not consider only extreme values in need for cognition, which necessarily led to a smaller expected correlation ($r = .16$), but we did account for this fact in our a priori determination of the necessary sample size. Importantly, anchoring effects in our study were relatively small for experimenter-provided anchoring items as only two out of four items displayed anchoring effects.

We also analyzed the openly available data from Yoon et al. (2021) where (among others) original items such as "the year the Boston Tea Party occurred", were used. We found that even for an American sample, this original item did not work as most people did not think of the correct self-generated anchor (declaration of independence) and 75% of all values were beyond the interval of 1776 and 1773 (https://osf.io/y2mr9, slide 13).

The self-generated anchoring items were even less effective, although we do not know of previous studies that had tested for anchoring effects with this type of anchor. We do not think that significant anchoring effects are necessary to test the hypothesis. On the contrary, very large effect sizes might even attenuate the role of individual differences (Hedge et al., 2018).

5 Study 2: Replication of Epley and Gilovich, 2006, Study 2c (Cognitive Load)

5.1 Method

To test whether susceptibility to self-generated anchors would be found to be positively correlated with cognitive load, we conducted a preregistered (https://osf.io/b7nt8) close replication of Epley and Gilovich (2006, Study 2c). We hypothesized that cognitive load leads to smaller absolute adjustment from self-generated anchors but has no effect on absolute adjustment from experimenter-provided anchors.

5.1.1 A Priori Sample Size Determination

We sought to replicate the original effect of $d = 0.66$, 95% CI $[0.230, 1.083]$ reported by Epley and Gilovich (2006). The original study had a sample size of $N = 94$ participants, 46 of whom belonged to the experimental group. We determined our sample size using the small telescopes approach (Simonsohn, 2015). Following this approach, "if the true effect is zero, a replication needs 2.5 times as many observations as the original study to have about 80% power to reject $d_{33\%}$" (Simonsohn, 2015, p. 565). Accordingly, our target sample size was 235 participants.

5.1.2 Materials

For the experimenter-provided anchors, Epley and Gilovich used nine not further specified items from the 15 items reported by Jacowitz and Kahneman (1995, p. 1163, Table 1). Five of these items required upward adjustment from the anchor, and four required downward adjustment. For our study, we selected nine items from the original source (Table 5, Items 13 to 22). We chose items that were as neutral as possible, meaning that they were not US-specific and therefore more likely to work with German participants. We made small adjustments to Items 13, 15, 16, 19, and 20, to facilitate estimations for our German participants (e.g., "Election year of German chancellor Schröder" instead of "Number of Lincoln's presidency"; see Table 5). As in the original study, we also chose five items that required upward adjustment and four items that required downward adjustment.

For the self-generated anchors, we used four items from the original study (see Table 5, Items 23, 24, 26, 28), adapted three of the original items (Table 5, Items 22, 29, 30), and added two new items (Table 5, Items 25, 27) so that the items would work with a German sample instead of a US-American one. An overview of all items is available online (https://osf.io/ktjc4).

5.1.3 Procedure

Participants were asked to participate in an online survey that would test their common knowledge and were instructed to answer the questions that appeared on their screen. After they had given consent to participate, they were asked to provide basic socioeconomic data. In the experimental group, we induced cognitive load by asking participants to memorize an eight-letter string before asking them to answer one of the items from our lists of experimenter-provided anchoring questions and self-generated anchoring questions (see Table 5, Items 13 to 30). Then, they were asked to type in the eight-letter string they had to memorize. This procedure was repeated for all items. The order of the anchoring items and letter strings was not randomized. Participants in the control group were also asked to memorize the eight-letter strings, but the retrieval question immediately followed the memorization so that there was no additional cognitive load when they answered the anchoring questions.

In both groups, each anchoring question was followed by a control question ("When answering this question, did you think of XX as a reference value?") to ensure that participants were thinking of the intended anchor. At the end of the survey, each participant was asked if they had tried to memorize the eight-letter strings or if they had used another method of retrieval. The wording of the question ensured that the participants understood that there would be no repercussions if they answered honestly ("The following question is solely used to interpret your answers in the scientific context of the study, so please answer honestly. There will be no repercussions, no matter what your answer is. Did you answer the questions about the eight-letter strings without any help/aid and did you rely solely on your own memory? Yes. No."). Finally, the participants were able to request course credit. An overview of the procedure is provided in Figure 4.

5.1.4 Statistical Analyses

To test our hypotheses on cognitive load, we followed the analysis plan from the original study and conducted a 2×2 (anchor type: self-generated vs. experimenter-provided) $\times$ 2 (cognitive load: yes vs. no) ANOVA and used independent-samples t tests to compare the z-transformed means. To determine whether the t test showed a significant group difference, we used the standard $p < .05$ criterion.

We computed the absolute amount of adjustment by determining the absolute difference between the anchor value and a participant's generated estimate. We then computed the mean adjustment and mean z-scores for all anchor items.
Additionally, we tested for anchoring effects by testing whether the mean estimates (per question) were significantly different from the anchor in the direction of the true value.

We excluded participants' responses if the participant gave the exact answer to the anchoring question, if their answer was off by more than three standard deviations in any direction, or if the participant answered "no" to the control question about their knowledge of the anchor (see Procedure).

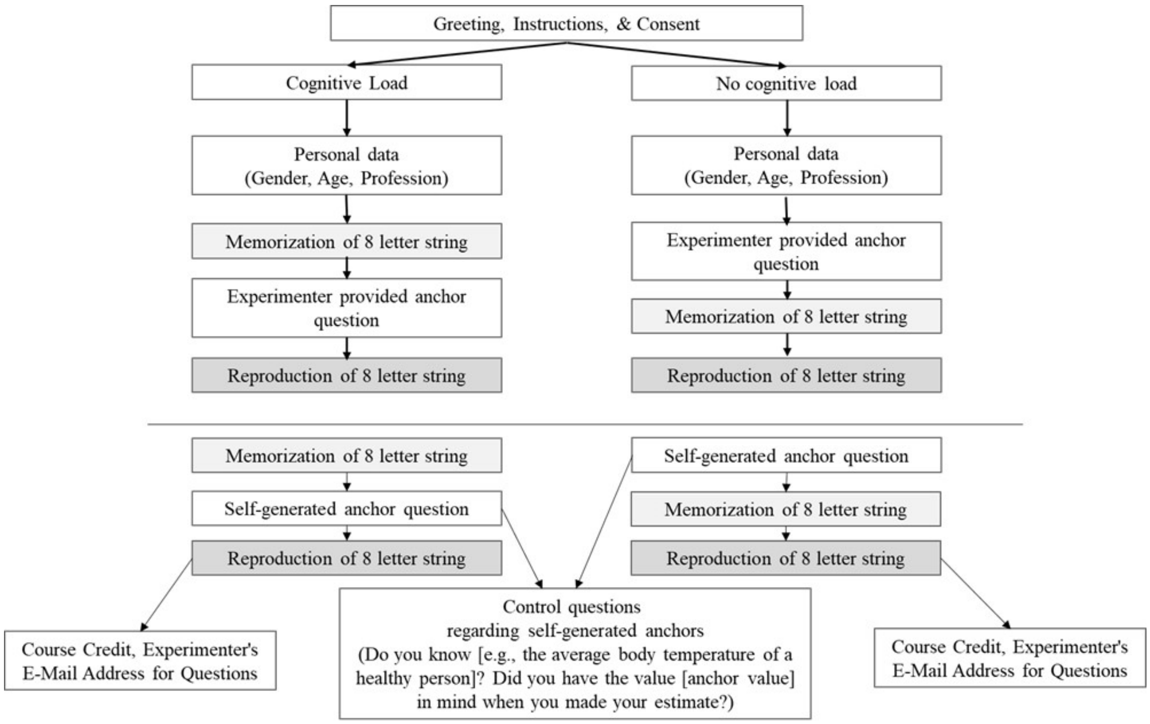

Figure 4: Procedure Used in Study 2

5.1.5 *Deviations From the Original Study*

Due to the COVID-19 pandemic, this replication was conducted online. As participants were German, we had to change some of the items (e.g., "Election year of German chancellor Schröder" instead of "Number of Lincoln's presidency") or create new ones. A comprehensive list of deviations from the original study is presented in Table 3.

Table 3: Methodological Differences Between the Original Study and our Replication Study 2

Study feature	**Epley & Gilovich, 2006, 2c**	**This study**	**Reason for change**
Language of questionnaire	English	German	German participants
Type of sample	Only college students	College students and non-students	Heterogeneous sample should increase the effect size
Type of study	Study was conducted on site	Online study	COVID-19, larger sample
Compensation	Not reported	Course credit or none	Facilitate participant recruitment
Type of instructions	Not reported	Test of general knowledge	Original materials were not available
Additional variables that were collected	Not reported	Age, sex, educational status, profession	Original materials were not available
Letter strings/Cognitive Load	Eight-letter strings	Memorization of experimenter-provided eight-letter strings	Eight-letter strings from the original study were unknown
Order of presentation	The order of presentation of the two types of questions was counterbalanced	The order of presentation was the same for each participant	Presenting the questions in a counterbalanced order would be too complex
Experimenter-provided anchor	Nine items, which were not specified any further (there were 15 items in the source; Jacowitz & Kahneman, 1995);	Choice of nine experimenter-provided items from the source (Jacowitz & Kahneman, 1995)	Original materials were not available
Self-generated anchor	Nine items	Four items from the original study plus five additional items (two completely new items, three adapted from the original items)	German participants have different common knowledge than American participants; Participants who did not think of the anchor might have to be excluded
Exclusion criteria	Not reported	Participants had to think of the anchor; outliers were removed	Otherwise, there would be distortion from participants who did not think of the anchor; outliers would have distorted effect size estimates

5.1.6 Deviations From the Preregistration

We had planned to recruit 235 participants for our replication, but at the end of testing on March 22nd, 2022, only 183 people had participated. We did not deviate from the preregistration in any other regard. Despite the smaller-than-planned sample size, the statistical power for the original effect size ($d = 0.66$) was 99.75% ($d_{\min} > 0.369$), and the power for the interaction effect ($\eta^2 = .03$) was 81.16%.

5.2 Results

5.2.1 Sample

Our sample consisted of 183 participants (123 women, 58 men, 2 other) of which 127 were students. Their ages ranged from 18 years to 65 years, with a median age of 24 years.

5.2.2 Data Quality Checks

In a one-sided t test, anchoring effects were present for seven out of nine experimenter-provided anchoring items (see Table 5, Items 13, 14, 15, 17, 18, 20, 21) and six out of nine self-generated anchoring items (see Table 5, Items 23, 24, 26-29). Note that the test in Table 5 is more sophisticated than the preregistered one-sided t test because the test we used also required that the mean adjustment values were not in the direction opposite the anchor.

5.2.3 Hypothesis Tests

The first hypothesis was that cognitive load does not have an effect on experimenter-provided anchor questions. A Welch two-sample t test revealed that cognitive load did not have a significant effect on experimenter-provided anchors, $t(180.79) = 0.75, p = .451, d = -0.112$, 95% CI $[-0.402, 0.178]$, and participants' adjustment scores were similar in the two conditions ($M_{\text{no cognitive load}} = 0.05, SD_{\text{no cognitive load}} = 0.44, N_{\text{no cognitive load}} = 90, M_{\text{cognitive load}} = 0.00, SD_{\text{cognitive load}} = 0.47, N_{\text{no cognitive load}} = 93$; see also Figure 5). The second hypothesis predicted that cognitive load would lead to less adjustment from the anchor for self-generated anchoring items. This was not the case, $t(176.58) =$ -1.34, $p = .908$, $d = 0.198$, 95% CI $[-0.093, 0.488]$. The interaction between forewarning condition and anchor type in the 2 × 2 repeated-measures ANOVA was also not significant, $F(1, 362) = 2.18$, $p = .141$, $\eta^2 = .006$, 90% CI $[.000, .026]$.

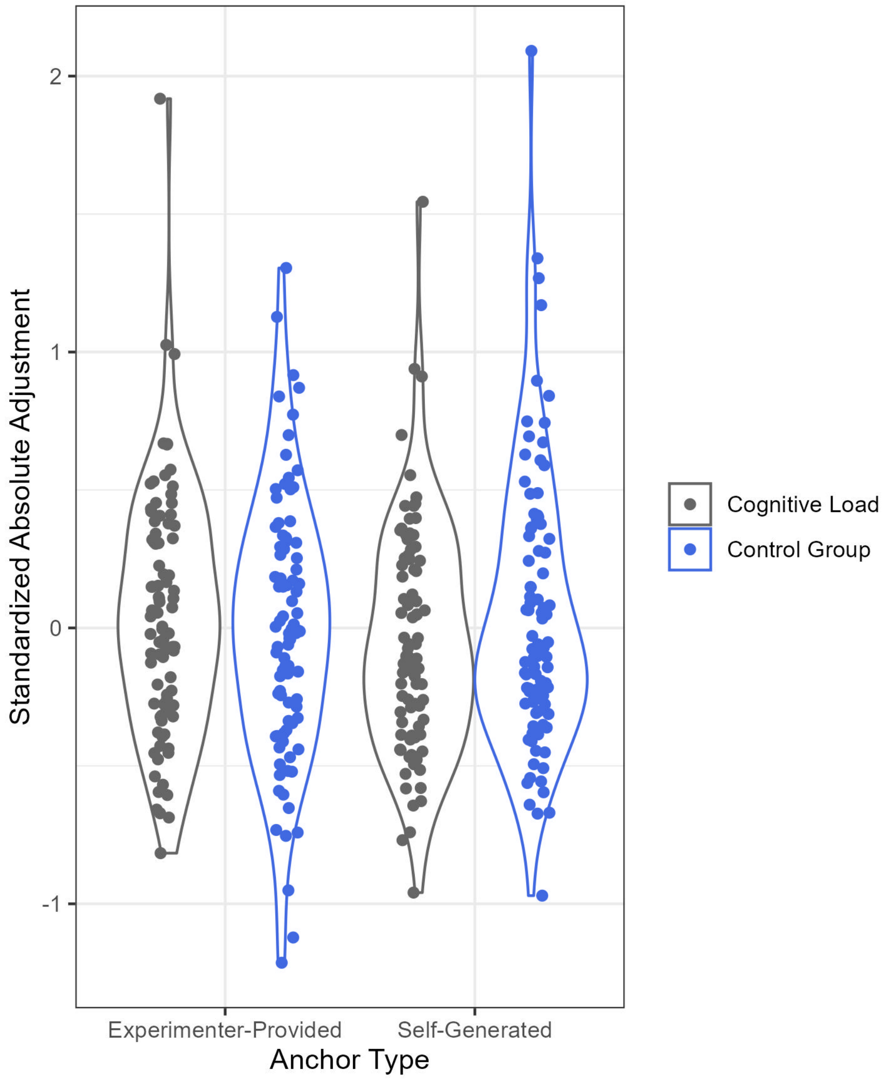

Figure 5: Effects of Cognitive Load on Adjustment from Experimenter-Provided and Self-Generated Anchors

Notes: Code to reproduce this figure: https://osf.io/g29w8

5.3 Discussion

In our replication of Epley and Gilovich (2006, Study 2c), we tested whether adjustment from self-generated but not experimenter-provided anchors decreased when participants experienced cognitive load. We deviated from the original study in that we had to create new items for the

German (instead of US-American) participants. Difficulties in recruiting participants resulted in a final sample size of $N_{\text{replication}} = 183$ ($N_{\text{target}} = 235$, $N_{\text{original}} study = 94$), but statistical power was still >99% for the original effect of cognitive load on adjustment from self-generated anchors. According to our simple test of anchoring, anchoring occurred for 13 out of 18 items. Most importantly, we could not replicate the original finding that cognitive load affected adjustment from self-generated anchors but not from experimenter-provided anchors.

6 Study 3: Replication of Epley and Gilovich, 2005, Study 2 (Forewarning)

6.1 Method

In their Study 2, Epley and Gilovich (2005) found that warning participants that their adjustments from anchors were insufficient led to increases in adjustments from *self-generated* anchors but not from experimenter-provided anchors. We conducted a preregistered (https://osf.io/f5sj8) replication of this study.

6.1.1 A Priori Sample Size Determination

As effect sizes were not reported in the paper, we calculated them on the basis of the reported results. The t test for the self-generated anchors yielded Cohen's $d = 1.24$, CI 95% $[0.558, 1.894]$, whereas the t test for the experimenter-provided anchors yielded $d = 0$ (we assumed a null effect because the t test was reported as nonsignificant with t < 1; Epley & Gilovich, 2005, p. 207). The interaction effect was $\eta^2 = .17$, 90% CI $[.033, .319]$. To determine the necessary sample size, we used the small telescopes approach (Simonsohn, 2015). When applied to the present replication study, the approach indicated that we needed 120 participants (2.5 multiplied by the original 48 participants). The final sample size exceeded $N = 120$, so we tested for whether the results differed when they were based on the first 120 participants versus the entire sample.

6.1.2 Materials

The anchoring items (six self-generated and six experimenter-provided anchoring items) for measuring susceptibility to anchor values were presented in the replication study. Most of these items differed from the items used in the original study. Two items, one self-generated and one experimenter-provided, were taken from the original study. The reasons for the deviations were that we adapted the items to the German culture and language and the fact that some of the items did not show significant anchoring effects in our Study 1 reported above. Items that could not be adapted were replaced by newly created ones. Details about the items used in the original and replication studies and their respective source, type, anchor, and true value are listed in a separate table available online (https://osf.io/dwm82, file name: Anchoring_items; see also Table 5).

6.1.3 Procedure

The study was conducted online. The introductory text of the online study and the recruitment text both stated that general knowledge and estimation questions were the subjects of the study. We recruited participants by contacting other universities in Germany (Bavaria), social media, and private acquaintances. The introductory text in the questionnaire explained that the study would take about 10 min. Course credit was offered as a non-monetary incentive, and participants had the option to receive feedback from the study. After participants provided informed consent, they were randomly assigned to the condition with or without forewarning. In the forewarning condition, participants were informed that research had demonstrated that judgments are strongly influenced by the information that first comes to mind. The example from the original study was retained: For example, real estate agents' estimates of a house's value are influenced by the value of the previously inspected house. Participants were told that it is suspected that individuals begin from the value that first occurs to them and subsequently fail to sufficiently adjust away from that value. The instructions, similar to the ones used in the original study were: In the following, you will be asked some questions. Either certain values will be given or you will have a certain value in your mind. Please try not to be influenced by these numbers. Participants were asked to avoid using any auxiliary sources, and they were reassured that it did not matter if they were not sure about the answer. All instructions and questions were translated into German.

The experimenter-provided anchoring items were asked first, followed by the self-generated anchoring items. For all experimenter-provided anchoring items, the first question was whether the true value was above or below the given anchoring value (e.g., whether the Rhine is shorter or longer than 2,000 km). Participants could choose between "more/greater/longer" or "less/smaller /fewer/shorter." Subsequently, they were asked to enter their estimated value in an open field.

The self-generated anchoring items were presented afterwards. After all items has been presented, the participants were asked whether they knew the expected anchor value (e.g., whether they knew the average body temperature of a human being) and whether they had this value in mind when they gave their estimate (e.g., Did you think of approximately 36/37 degrees Celsius when you made your estimate?).

At the end of the questionnaire, questions about demographic variables regarding gender, age, and occupation were asked. Finally, as an exploratory question, the participants were asked whether they knew what anchoring effects were. An email address for questions was given, and participants had the option to provide their email address for participation in future studies. An overview of the procedure is provided in Figure 6.

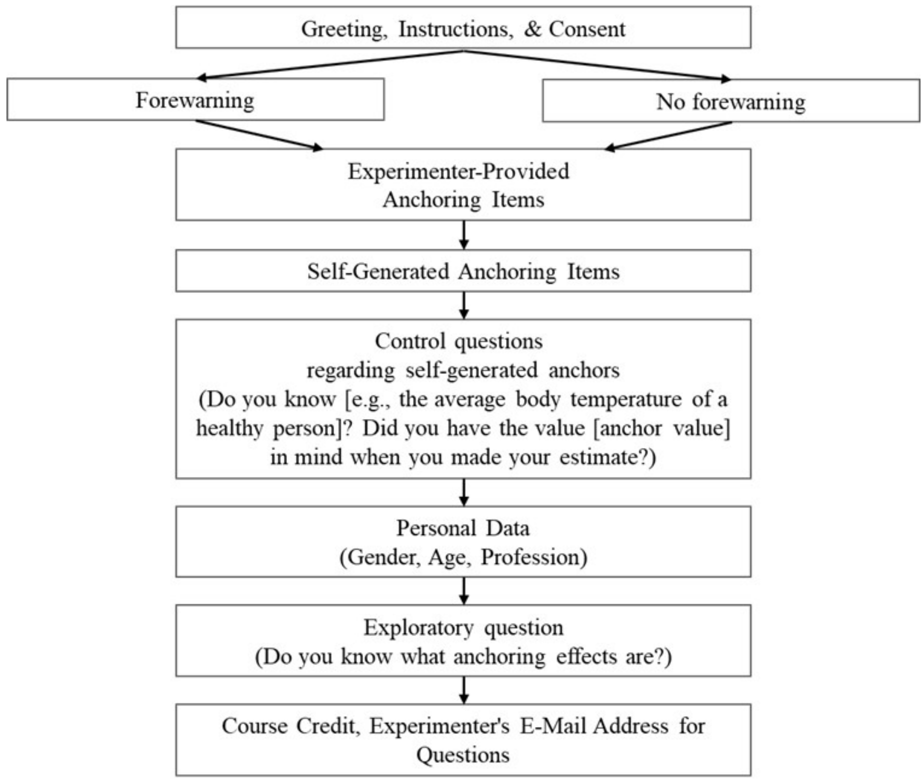

Figure 6: Procedure Used in Study 3

6.1.4 Design

The present study used a 2 (anchor type: self-generated vs. experimenter-provided) × 2 (forewarning: yes vs. no) design. Unlike in the original study, participants were not all given the same anchor values for the experimenter-provided anchoring items but were randomly assigned to different versions of the questionnaire with high and low anchor values. To simplify the programming, four versions of the questionnaire were created: In the first version, high anchor values were specified for the first three experimenter-provided anchoring items and low anchor values for the following three experimenter-provided anchoring items (high-high-high-low-low-low). This version was available once with and once without forewarning. For the other two versions of the questionnaire, the first three experimenter-provided anchoring items were presented with low anchor values, and the following three with high anchor values (low-low-low-high-high-high). A distinction was also made between items with and without forewarning for this order of anchoring values. Thus, there were four versions of the questionnaire in total. The self-generated anchoring items were identical for all participants. The order of the 12 items was not randomized.

6.1.5 Statistical Analyses

As for the previous replication studies, we first calculated the absolute difference between anchoring value and estimate (*absolute adjustment*). Absolute adjustment was standardized (i.e., mean-centered and scaled) by anchor type. Means were computed by anchor type to operationalize the susceptibility to anchoring. Estimates were excluded item wise (a) if participants did not know the intended self-generated anchor or did not have the value in mind during their estimation, (b) if they specified the true value, or (c) if their standardized absolute adjustment was $Z > 3$. Participants were included in the analyses if they answered at least one experimenter-provided anchoring item and one self-generated anchoring item.

6.1.6 Deviations From the Original Study

Due to the COVID-19 pandemic, insufficient reporting in the original study, and practical reasons, the replication study differed from the original one in some respects. Just like the other replication studies described earlier, the study was conducted in Germany. This is why instructions, anchoring items, and other questions were translated into German and adapted to the German culture and general knowledge. Likewise, units were converted to the German standard (e.g., miles to kilometers). The original study was carried out at a Boston train station, but due to the pandemic, we chose to conduct the study online. Other information, such as type of sample, compensation, and type of demographic data collected, was not reported in the original study. Additional details about the deviations are given in Table 4.

Table 4: Methodological Differences Between the Original Study and Replication Study 3

Study feature	**Epley & Gilovich, 2005, Study 2**	**This study**	**Reason for change**
Language of questionnaire	English	German	German participants
Type of sample	Unknown	College students and non-students	Sufficient participants
Compensation	Some candy	Course credit	Sufficient participants
Type of study	Study was conducted in a Boston train station	Online study	COVID-19, larger sample
Type of instructions	Spoken	Written form, similar instructions	Online study did not allow for spoken words
Personal data	Not reported	Three individual-related pieces of data (age, sex, profession) were collected	We did not know which kinds of individual-related data were collected in the original study
Experimenter-provided anchoring items	6 items (from Jacowitz & Kahneman, 1995)	1 item from original study, 5 different items (see extra file) - As neutral as possible, not US-specific - Adapted to German culture	- German students have little US-specific knowledge - Items that were not significant in our previous replication study were changed
Experimenter-provided anchors: Variation of anchor values	Every participant received the same anchor values	Participants were randomly given high or low anchor values	To test for whether anchoring effects occurred: comparison of adjustment from two different directions
Self-generated anchors	6 items: 2 items from previous research from Epley and Gilovich; Epley and Gilovich (2001, 2004), 4 other items (self-generated)	1 item from original study used, 5 different items (see extra file) - As neutral as possible, not US-specific - Adapted to German culture	- German students have little US-specific knowledge - Items that were not significant in our previous replication study were changed
Order of self-generated and experimenter-provided anchor questions	Order was counterbalanced	Order was fixed (first experimenter-provided, then self-generated)	The original study found no influence of the order of anchor types on any of the reported results
Exploratory question	None	Question about whether the phenomenon of "anchoring effects" was known	Exploratory
Exclusion criteria	No details about outlier exclusion	Item-based exclusion if participants deviated more than 3 SD from the average item estimate	Other distortion of the results by extreme answers
Data analysis	Not specified	Linear mixed-effects model (lme4 R-package)	

6.1.7 Deviations From the Preregistration

Instead of the planned 120 participants, our sample size comprised 171 participants. We included all data in the analysis to increase the statistical power, but we tested whether including only the first 120 participants would have yielded different results. Concerning the analysis script, we made small adjustments. In the sample description part, we added one function to determine the actual sample size. Moreover, we conducted a power analysis for the original effect size and a sensitivity analysis for the given sample size. In the hypothesis testing section, we added one line each for the self-generated and experimenter-provided anchors to provide descriptive statistics (*M*, *SD*) regarding anchor susceptibility grouped by forewarning condition. We made another small change in this section so that the Cohen's *d* output would be presented to 3 or more decimal places. Another addition consisted of an effect size calculation with CIs for the ANOVA. Coding of temperature estimates was corrected as the preregistered functions did not work due to the participants using a mix of commas and points as decimal symbols. Finally, we carried out exploratory analyses. We tested whether the hypotheses would have been confirmed if the items that did not show significant anchoring effects had been excluded. Another issue was that some participants were already aware of anchoring effects before the study, as the exploratory question we asked at the end of the study had revealed. We therefore examined whether adjustment increased when these individuals were included in the group with forewarning. A final supplement was a graph showing the days of the study on the x-axis and the number of participants on the y-axis.

6.2 Results

6.2.1 Sample

The study took place between January 19, 2022 and February 3, 2022. During this period, 220 participants completed the online questionnaire, of which 49 had to be removed due to our exclusion criteria. Our final sample comprised 171 participants, 94 of whom were forewarned about anchoring effects, and 77 who were not. The mean age was 26.98 years (one missing value), and of all participants, 114 were women, 56 were men, and one was diverse. Regarding participants' professions, the largest part of the sample ($N = 105$) reported that they were university students, 52 were employed, five were in vocational training, one was a high school student, one was unemployed, and the other seven selected the category "other."

As we exceeded our planned sample size of 120 participants, the power we achieved for the t test of the self-generated anchoring items was > 99.99%, and mean differences of $d > 0.384$ could be detected with 80% power. For the interaction between forewarning condition and anchoring type, we also achieved a power of > 99.99%.

6.2.2 Data Quality Checks

All six experimenter-provided anchoring items produced significant anchoring effects (all $d \geq 0.60$), that is, there was a significant difference between the estimates given after considering high anchors and those given after considering low anchors. As in Study 2, we checked for the self-generated anchoring items if the absolute level of adjustment was smaller than the adjustment necessary to estimate the true value. We found significant anchoring effects only for three out of the six self-generated anchoring items (Table 5, Items 37, 41, and 42). In total, 75% of all anchoring items showed significant anchoring effects, a finding that was lower than the expected 80%.

6.2.3 Hypothesis Tests

One-tailed independent-samples t tests revealed that adjustment in the forewarning condition was not significantly greater than in the control condition for the experimenter-provided ($M_{\text{no forewarning}} =$

$-0.04, SD_{\text{no forewarning}} = 0.48, N_{\text{no forewarning}} = 77, M_{\text{forewarning}} = 0.04, SD_{\text{forewarning}} = 0.51, N_{\text{no forewarning}} = 94)$, $t(165.71) = -0.996, p = .160, d = 0.153$, 95% CI $[-0.149, 0.455]$, or for the self-generated anchoring items $(M_{\text{no-forewarning}} = -0.06, SD_{\text{no-forewarning}} = 0.51, N_{\text{forewarning}} = 77, M_{\text{forewarning}} = 0.01, SD_{\text{forewarning}} = 0.53, N_{\text{forewarning}} = 94)$, $t(165.39) = -0.78, p = .219, d = 0.120$, 95% CI $[-0.182, 0.421]$. The interaction between forewarning condition and anchor type in the 2 × 2 repeated-measures ANOVA was also not significant, $F(1, 169) = 0.016$, $p = .899$, $\eta^2 < .001$, 90% CI $[.000, .008]$ (see also Figure 7).

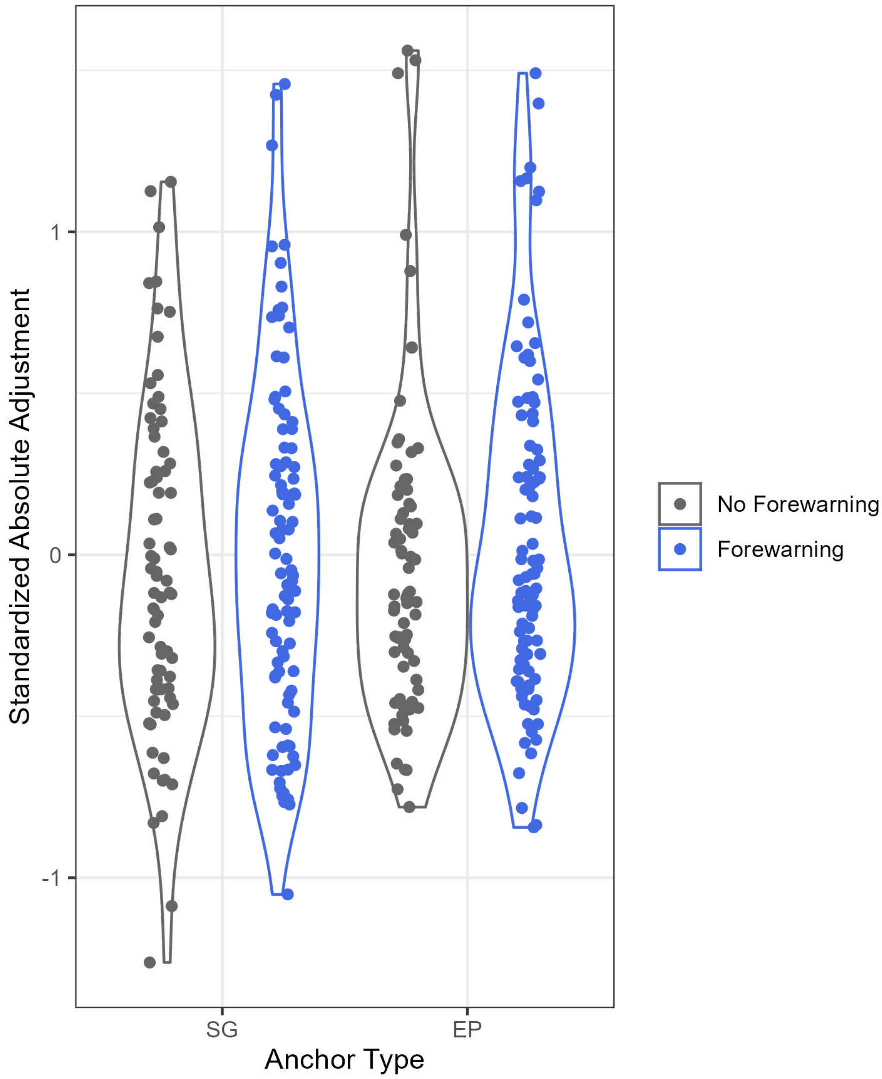

Figure 7: Effects of Forewarnings on Adjustment from Experimenter-Provided and Self-Generated Anchors

Notes: Code to reproduce this figure: https://osf.io/qvud5

6.3 Discussion

In our replication of Epley and Gilovich (2005, Study 2), we tested whether adjustment from self-generated but not experimenter-provided anchors increased when participants were forewarned about misleading anchors. We deviated from the original study in that we used both high and low anchors for experimenter-provided anchoring questions so that we could check whether anchoring actually occurred with classical tests, in that our study was conducted online, and the forewarning was written and not spoken.

Estimates for high anchors were significantly higher than estimates for low anchors for all six items (all d > 0.60). Adjustment from self-generated anchors was insufficient for only three out of the six items. Most importantly, neither absolute adjustment from self-generated anchors nor absolute adjustment from experimenter-provided anchors was moderated by forewarning. Using only the items that displayed anchoring effects did not change this. Note that having a written instead of a verbal forewarning may have contributed to the manipulation not working. Thus, we deem the replication an informative failure: There is evidence against the hypothesis that forewarning moderates adjustment from self-generated anchors.

7 Mini Meta-Analysis

To assess the difference between self-generated and experimenter-provided anchoring items across all three replication studies, we conducted a mini meta-analysis (e.g., Goh et al., 2016) using a random-effects multilevel model for all correlations between the respective moderators and absolute adjustment. Correlations were nested in studies. We tested for moderator effects of the moderator/study and anchor type (self-generated vs. experimenter-provided) and their interaction. The overall effect size was $r = -.024$, 95% CI $[-.074, .025]$, $p = .333$, $k = 42$, $N_{\text{total}}[402, 649]$. There was no effect of any of the moderators, $Q(5) = 4.57, p = .471$, and no residual heterogeneity, $Q(36) = 32.33, p = .644, I^2 = -.113$. An overview of all correlations between absolute adjustment and 0-1 scores is provided in Figure 8 and Table 5. Note that correlations of absolute adjustment scores and moderators were similar to those between 0-1 scores and moderators, $r(40) = .816, p < .001$.

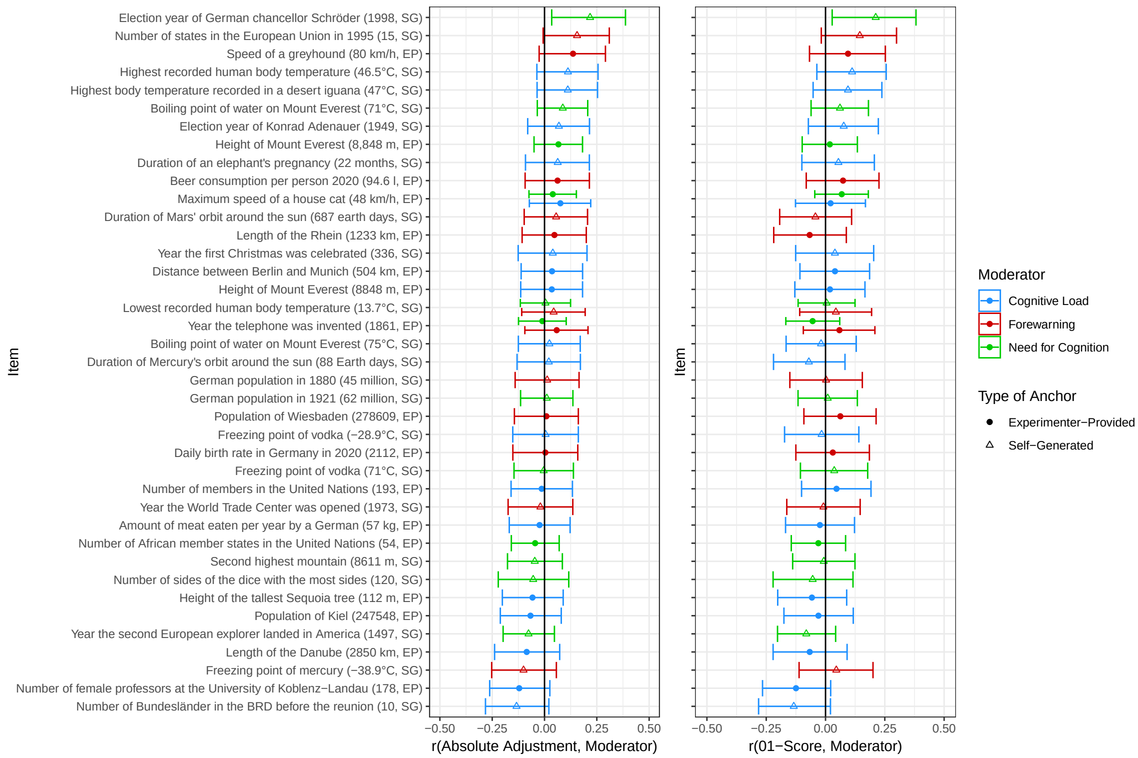

Figure 8: Correlations Between Absolute Adjustment (left, preregistered) and 01-Scores (right, exploratory) With Moderators for All Items

Notes: EP = experimenter-provided anchor (grey background), SG = self-generated anchor (blue background). Absolute adjustment represents the absolute difference between anchor and estimate. 0-1 scores are 0 if the true value was estimated and 1 if the anchor was estimated. Scores that do not include 0 and are between 0 and 1 suggest that anchoring occurred and are bold. Scores above 1 indicate overadjustment. Scores below 0 indicate adjustment in the wrong direction. Numbers in parentheses represent the correct answer. Error bars represent 95% confidence intervals. Higher correlations indicate more adjustment and less anchoring given high values for need for cognition, the presence of cognitive load, or the presence of a forewarning. Outliers ±3 SD were excluded. Code to reproduce this figure: https://osf.io/6vhmu

Table 5: 01-Scores and Correlations With Moderators for all Anchoring Items

No.	Study	Item	n	01-score	r (moderator, absolute adjustment)	r (moderator, 01-score)
1	Need for Cognition	Height of Mount Everest (8,848 m, EP)	284	**0.925 [0.888, 0.962]**	0.067 [-0.05, 0.182]	0.018 [-0.098, 0.134]
2	Need for Cognition	Number of African member states in the United Nations (54, EP)	293	**0.612 [0.547, 0.677]**	-0.045 [-0.159, 0.07]	-0.03 [-0.144, 0.085]
3	Need for Cognition	Year the telephone was invented (1861, EP)	296	1.376 [1.166, 1.586]	-0.01 [-0.124, 0.104]	-0.054 [-0.167, 0.06]
4	Need for Cognition	Maximum speed of a house cat (48 km/h, EP)	299	**0.995 [0.843, 1.148]**	0.04 [-0.074, 0.152]	0.068 [-0.045, 0.18]
5	Need for Cognition	Lowest recorded human body temperature (13.7°C, SG)	267	**0.363 [0.334, 0.392]**	0.005 [-0.116, 0.124]	0.005 [-0.116, 0.124]
6	Need for Cognition	Second highest mountain (8611 m, SG)	223	3.296 [2.561, 4.031]	-0.047 [-0.177, 0.085]	-0.007 [-0.139, 0.124]
7	Need for Cognition	Number of sides of the die with the most sides (120, SG)	135	**0.244 [0.202, 0.286]**	-0.054 [-0.221, 0.116]	-0.054 [-0.221, 0.116]
8	Need for Cognition	German population in 1921 (62 million, SG)	246	1.669 [1.552, 1.785]	0.011 [-0.114, 0.136]	0.009 [-0.116, 0.134]
9	Need for Cognition	Election year of German chancellor Schröder (1998, SG)	113	**0.962 [0.861, 1.064]**	0.218 [0.034, 0.387]	0.212 [0.028, 0.381]
10	Need for Cognition	Year the second European explorer landed in America (1497, SG)	253	15.102 [12.185, 18.019]	-0.076 [-0.198, 0.047]	-0.081 [-0.202, 0.043]
11	Need for Cognition	Boiling point of water on Mount Everest (71°C, SG)	261	**0.47 [0.396, 0.544]**	0.087 [-0.034, 0.207]	0.061 [-0.061, 0.181]
12	Need for Cognition	Freezing point of vodka (71°C, SG)	191	**0.874 [0.736, 1.012]**	-0.004 [-0.146, 0.138]	0.037 [-0.106, 0.178]
13	Cognitive Load	Length of the Danube (2850 km, EP)	157	1.263 [1.164, 1.362]	-0.085 [-0.239, 0.072]	-0.067 [-0.221, 0.091]
14	Cognitive Load	Height of Mount Everest (8,848 m, EP)	176	**0.682 [0.627, 0.738]**	0.035 [-0.114, 0.182]	0.019 [-0.129, 0.166]

15	Cognitive Load	Amount of meat eaten per year by a German (57 kg, EP)	182	**0.733 [0.681, 0.784]**	-0.024 [-0.168, 0.122]	-0.024 [-0.168, 0.122]
16	Cognitive Load	Distance between Berlin and Munich (504 km, EP)	178	1.204 [1.118, 1.29]	0.036 [-0.112, 0.182]	0.04 [-0.108, 0.186]
17	Cognitive Load	Height of the tallest Sequoia tree (112 m, EP)	181	**0.741 [0.682, 0.799]**	-0.057 [-0.202, 0.089]	-0.057 [-0.202, 0.089]
18	Cognitive Load	Number of members in the United Nations (193, EP)	179	**0.506 [0.44, 0.573]**	-0.014 [-0.16, 0.133]	0.046 [-0.101, 0.192]
19	Cognitive Load	Number of female professors at the University of Koblenz-Landau (178, EP)	181	1.046 [0.971, 1.12]	-0.121 [-0.262, 0.025]	-0.125 [-0.266, 0.022]
20	Cognitive Load	Population of Kiel (247548, EP)	180	**0.311 [0.174, 0.448]**	-0.067 [-0.211, 0.08]	-0.03 [-0.175, 0.117]
21	Cognitive Load	Maximum speed of a house cat (48 km/h, EP)	177	**0.380 [0.339, 0.422]**	0.076 [-0.073, 0.221]	0.021 [-0.127, 0.168]
22	Cognitive Load	Election year of Konrad Adenauer (1949, SG)	175	2.064 [1.58, 2.549]	0.069 [-0.08, 0.215]	0.077 [-0.072, 0.223]
23	Cognitive Load	Highest recorded human body temperature (46.5°C, SG)	176	**0.720 [0.685, 0.756]**	0.112 [-0.037, 0.256]	0.112 [-0.037, 0.256]
24	Cognitive Load	Boiling point of water on Mount Everest (75°C, SG)	176	**0.695 [0.540, 0.850]**	0.023 [-0.125, 0.171]	-0.019 [-0.166, 0.129]
25	Cognitive Load	Highest body temperature recorded in a desert iguana (47°C, SG)	180	1.251 [1.039, 1.463]	0.111 [-0.036, 0.253]	0.095 [-0.052, 0.238]
26	Cognitive Load	Freezing point of vodka (-28.9°C, SG)	157	**0.503 [0.424, 0.581]**	0.005 [-0.152, 0.162]	-0.017 [-0.173, 0.14]
27	Cognitive Load	Number of Bundesländer in the BRD before the reunion (10, SG)	162	**0.825 [0.771, 0.88]**	-0.134 [-0.282, 0.021]	-0.134 [-0.282, 0.021]
28	Cognitive Load	Duration of an elephant's pregnancy (22 months, SG)	164	**0.404 [0.337, 0.472]**	0.063 [-0.091, 0.214]	0.054 [-0.1, 0.206]

29	Cognitive Load	Duration of Mercury's orbit around the sun (88 Earth days, SG)	168	**0.534 [0.424, 0.645]**	0.021 [-0.131, 0.171]	-0.07 [-0.219, 0.082]
30	Cognitive Load	Year the first Christmas was celebrated (336, SG)	142	2.501 [2.125, 2.877]	0.04 [-0.126, 0.203]	0.04 [-0.126, 0.203]
31	Forewarning	Population of Wiesbaden (278,609, EP)	165	**0.73 [0.645, 0.815]**	0.009 [-0.144, 0.162]	0.062 [-0.091, 0.213]
32	Forewarning	Speed of a greyhound (80 km/h, EP)	148	**0.783 [0.63, 0.936]**	0.136 [-0.026, 0.291]	0.095 [-0.068, 0.252]
33	Forewarning	Length of the Rhein (1233 km, EP)	163	**0.449 [0.342, 0.557]**	0.047 [-0.107, 0.2]	-0.067 [-0.218, 0.088]
34	Forewarning	Daily birth rate in Germany in 2020 (2,112, EP)	160	**0.355 [0.223, 0.487]**	0.004 [-0.151, 0.159]	0.031 [-0.125, 0.185]
35	Forewarning	Beer consumption per person 2020 (94.6 l, EP)	162	**0.465 [0.123, 0.807]**	0.062 [-0.093, 0.214]	0.074 [-0.081, 0.225]
36	Forewarning	Year the telephone was invented (1861, EP)	168	1.376 [1.077, 1.675]	0.058 [-0.094, 0.208]	0.059 [-0.094, 0.208]
37	Forewarning	Lowest recorded human body temperature (13.7°C, SG)	167	**0.386 [0.348, 0.424]**	0.044 [-0.109, 0.194]	0.044 [-0.109, 0.194]
38	Forewarning	Year the World Trade Center was opened (1973, SG)	161	**0.939 [0.828, 1.051]**	-0.02 [-0.174, 0.135]	-0.009 [-0.163, 0.146]
39	Forewarning	German population in 1880 (45 million, SG)	165	1.29 [1.203, 1.377]	0.013 [-0.14, 0.165]	0.002 [-0.15, 0.155]
40	Forewarning	Freezing point of mercury (-38.9°C, SG)	158	0.41 [0.22, 0.599]	-0.1 [-0.253, 0.057]	0.045 [-0.111, 0.2]
41	Forewarning	Duration of Mars' orbit around the sun (687 earth days, SG)	167	0.085 [-0.049, 0.218]	0.056 [-0.097, 0.206]	-0.043 [-0.193, 0.11]
42	Forewarning	Number of states in the European Union in 1995 (15, SG)	147	**0.868 [0.779, 0.958]**	0.156 [-0.007, 0.31]	0.145 [-0.018, 0.299]

Notes: EP = experimenter-provided anchor, SG = self-generated anchor. Values in parentheses represent correct answers. Values in brackets represent 95% confidence intervals. 0-1 scores are 0 if the true value was estimated and 1 if the anchor was estimated. Scores that do not include 0 and are between 0 and 1 suggest that anchoring occurred and are bold. Scores above 1 indicate overadjustment. Scores below 0 indicate adjustment in the wrong direction. Higher correlations indicate more adjustment and less anchoring given high values for need for cognition, the presence of cognitive load, or the presence of a forewarning. Outliers ±3 SD were excluded.

8 *General Discussion*

We conducted preregistered replications of three seminal findings on moderators of self-generated and experimenter-provided anchoring. In cases where anchors were varied between participants (e.g., one low and one high anchor was used per anchoring item), anchoring effects were large. Self-generated anchoring items are characterized by the same value consistently coming to participants' minds, and anchors can thereby not be manipulated. We tested whether adjustment from self-generated anchors toward correct values was insufficient and found that 15/21 anchoring items displayed what we would consider anchoring effects. In the other cases, people adjusted in the "wrong" direction (i.e., away from the true value instead of toward the true value), adjusted until they arrived at the true value, or adjusted too far). Evidence of openly available datasets using original items from Epley and Gilovich suggests that their items should not be used to investigate adjustment from self-generated anchors (e.g., most Americans do not think of the declaration of independence when asked about the year the Boston Tea Party occurred; https://osf.io/y2mr9). Most importantly, none of the moderators (need for cognition, cognitive load, forewarning) were associated with adjustment from anchors.

A prominent deviation of our studies is that all replication studies were conducted online whereas the original studies were not. There is currently no way to test whether this has affected the quality of the data and we believe that it did not: First, anchoring effects were as large and heterogeneous as usual, which they may not have been if participants acted differently from the original settings. Generally, anchoring effects do not differ between on-site and online studies (e.g., Röseler & Schütz, 2022, Table 3, p. 22). Second, it is unlikely that all moderators would have been affected equally. For example, need for cognition is a personality trait and measured with a validated scale und thus unlikely to be affected by this type of variation. Moreover, the general level of cognitive load could differ between online and laboratory settings (in both possible ways), but we do not see a reason why forewarning should work better in a laboratory than in an online study. Third, to actually compare data between the original and replication studies, the original study's datasets would be needed. As the original studies have been conducted approximately 20 years ago, no data is available.

Historically, the distinction between the two types of anchors led to a debate spanning 13 years: Epley and Gilovich reconciled the insufficient adjustment model and the selective accessibility model of anchoring in 2001 by suggesting that one had to be applied to self-generated anchoring items and the other to experimenter-provided anchoring items. Afterward, Simmons et al. (2010) and Chaxel (2014) revoked these findings, suggesting that both accounts were valid for both types of anchors. Note that recent findings have suggested that the selective accessibility model is invalid (Bahník, 2021; Harris et al., 2019). Our results support this idea as none of our three replication studies revealed any evidence that there is a difference between adjustment from experimenter-provided and self-generated anchors. Both types of anchors provoked anchoring effects, both adjustment scores were unreliable, and neither score was correlated with any of the moderators we investigated (i.e., need for cognition, cognitive load, forewarning).

Note that our results do not suggest that *none* of Epley and Gilovich's findings on moderators can be replicated (e.g., p-curve analysis indicated overall high power; https://osf.io/c6x4q). We chose to replicate the findings on three specific moderators (i.e., need for cognition, cognitive load, and forewarning) because all of them could easily be measured or manipulated, and the global COVID-19 pandemic prevented us from using manipulations that required participants to be observed in the lab. Additional moderators consist of nodding versus shaking one's head, arm movement (flexion vs. tension), financial incentives, and alcohol consumption. Although we cannot say whether our null findings can be generalized to these moderators, we believe that given our studies' higher degree of transparency and an overall larger sample size, reports of differences between moderators of experimenter-provided and self-generated anchors should generally be taken with a grain of salt. We also do not expect moderators to be associated with adjustment from experimenter-provided anchors if the direction of adjustment is known as there is overwhelming evidence against the hypothesis that these scores are more reliable (Röseler et al., 2022). To further clarify the role of potential moderators of anchoring effects, we think that Simmons et al.'s (2010; monetary incentives for accuracy) and Chaxel's (2014; priming selective accessibility) findings on moderators should be replicated, too. For example, the former add to other findings that monetary incentives decrease anchoring (e.g., Epley & Gilovich, 2005; LeBoeuf & Shafir, 2009; Meub et al., 2013) but are inconsistent with null findings (e.g., Enke et al., 2021; Li et al., 2021; Wilson et al., 1996). Thereby, we encourage adherence to transparently described experimental procedures such as ours or that of other researchers (e.g., Cheek & Norem, 2022; Mayer & Rebholz, 2024) to minimize variation and to beware of compatibility of incentives (Hertwig & Ortmann, 2001). Overall, there is hardly any doubt that anchoring effects are large and robust. But apart from this general finding, anchoring research is no exception to research that has problems with replicability (e.g., Bahník, 2021; Harris et al., 2019; Röseler et al., 2020; Röseler et al., 2021; Röseler et al., 2019; Shanks et al., 2020). We recommend that anchoring researchers put greater emphasis on the replicability of previous findings as this could have saved us 13 years of researching and debating.

CRediT Author Statement

- Lukas Röseler: *Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing, review and editing*
- Hannah L. Bögler, Lisa Koßmann, Sabine M. Krueger: *Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing*
- Sabrina L. C. Bickenbach, Ricarda Bühler, Jasmin della Guardia, Lisa-Marie A. Köppel, Jarl Möhring, Susanne Ponader, Konstantin Roßmaier, Jessica Sing: *Formal analysis, Investigation, Methodology, Software, Writing – review and editing*

Acknowledgments

We thank Leonie Beuerle, Lena Eckert, Afra Fischer, Maxi Görnitz, Alexander Jiranek, Rebekka Keller, Greta Kick, Clara Köth, Rebekka Kraus, Katharina M. Kroworsch, Lisa Lederer, Johanna Popp, Emilia Schramm, and Annkathrin Zorbach for their help in designing and conducting the replication studies. We thank Jane Zagorski for language editing.

References

Auguie, B., Antonov, A. & Auguie, M. B. (2017). “Package ‘gridExtra’”. *Miscellaneous Functions for “Grid” Graphics*, 9. PDF: https://cran.r-project.org/web/packages/gridExtra/gridExtra.pdf.

Bahník, Š. (2021). “Anchoring Does Not Activate Examples Associated with the Anchor Value”. DOI: 10.31234/osf.io/4j5wb.

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M. & Hocking, T. (2024). "data.table: Extension of ‘data.frame". URL: https://cran.r-project.org/web/packages/data.table/index.html.

Bless, H., Fellhauer, R. F., Bohner, G. & Schwarz, N. (1991). "Need for Cognition: eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben", volume 1991/06 of *ZUMA-Arbeitsbericht*. Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. URL: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-68892.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R. & van ’t Veer, A. (2014). “The Replication Recipe: What Makes for a Convincing Replication?” *Journal of Experimental Social Psychology*, 50: 217–224. DOI: 10.1016/j.jesp.2013.10.005.

Bruinsma, J. & Crutzen, R. (2018). “A Longitudinal Study on the Stability of the Need for Cognition”. *Personality and Individual Differences*, 127: 151–161. DOI: 10.1016/j.paid.2018.02.001.

Cacioppo, J. T. & Petty, R. E. (1982). “The Need for Cognition”. *Journal of Personality and Social Psychology*, 42(1): 116–131. DOI: 10.1037/0022-3514.42.1.116.

Champely, S. (2020). "pwr: Basic Functions for Power Analysis". URL: https://CRAN.R-project.org/package=pwr.

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. DOI: 10.1016/j.jesp.2013.09.005.

Cheek, N. N. & Norem, J. K. (2022). “Individual Differences in Anchoring Susceptibility: Verbal Reasoning, Autistic Tendencies, and Narcissism”. *Personality and Individual Differences*, 184(184): 111212. DOI: 10.1016/j.paid.2021.111212.

Diedenhofen, B. & Musch, J. (2015). “Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations”. *PLoS One*, 10(4): e0121945. DOI: 10.1371/journal.pone.0121945.

Dragulescu, A. & Arendt, C. (2020). "xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files". URL: https://CRAN.R-project.org/package=xlsx.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T. & van de Ven, J. (2021). "Cognitive Biases: Mistakes or Missing Stakes?" *The Review of Economics and Statistics*, 105(4). DOI: $10.1162/rest_{a0}1093$.

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(5): 391–396. DOI: 10.1111/1467-9280.0037.

Epley, N. & Gilovich, T. (2004). "Are Adjustments Insufficient?" *Pers Soc Psychol Bull*, 30(4): 447–460. DOI: 10.1177/0146167203261889.

Epley, N. & Gilovich, T. (2005). "When Efortful Thinking Influences Judgmental Anchoring: Differential Effects of Forewarning and Incentives on Self-generated and Externally Provided Anchors". *Journal of Behavioral Decision Making*, 18(3): 199–212. DOI: 10.1002/bdm.495.

Epley, N. & Gilovich, T. (2006). "The Anchoring-and-adjustment Heuristic: Why the Adjustments are Insufficient". *Psychological Science*, 17(4): 311–318. DOI: 10.1111/j.1467-9280.2006.01704.x.

Epley, N. & Gilovich, T. (2010). "Anchoring Unbound". *Journal of Consumer Psychology*, 20(1): 20–24. DOI: 10.1016/j.jcps.2009.12.005.

Frederick, S. W. & Mochon, D. (2012). "A Scale Distortion Theory of Anchoring". *Journal of Experimental Psychology: General*, 141(1): 124–133. DOI: 10.1037/a0024006.

Goh, J. X., Hall, J. A. & Rosenthal, R. (2016). "Mini Meta-analysis of Your Own Studies: Some Arguments on Why and a Primer on How". *Social and Personality Psychology Compass*, 10(10): 535–549. DOI: 10.1111/spc3.12267.

Grolemund, G. & Wickham, H. (2011). "Dates and Times Made Easy with Lubridate". *Journal of statistical software*, 40: 1–25. DOI: 10.18637/jss.v040.i03.

Harris, A. J. L., Blower, F. B. N., Rodgers, S. A., Lagator, S., Page, E., Burton, A., Urlichich, D. & Speekenbrink, M. (2019). "Failures to Replicate a Key Result of the Selective Accessibility Theory of Anchoring". *Journal of Experimental Psychology: General*, 148(9): e30–e50. DOI: 10.1037/xge0000644.

Hedge, C., Powell, G. & Sumner, P. (2018). "The Reliability Paradox: Why Robust Cognitive Tasks Do Not Produce Reliable Individual Differences". *Behavior Research Methods*, 50(3): 1166–1186. DOI: 10.3758/s13428-017-0935-1.

Hertwig, R. & Ortmann, A. (2001). "Experimental Practices in Economics: a Methodological Challenge for Psychologists?" *Behavioral and Brain Sciences*, 24(3): 383–403. DOI: 10.1017/S0140525X01004149.

Jacowitz, K. E. & Kahneman, D. (1995). "Measures of Anchoring in Estimation Tasks". *Personality and Social Psychology Bulletin*, 21(11): 1161–1166. DOI: 10.1177/01461672952111004.

Kelley, K. (2023). MBESS: The MBESS R Package. URL `https://CRAN.R-project.org/package=MBESS`. R package version 4.9.3.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). "LmerTest Package: Tests in Linear Mixed Effects Models". *Journal of Statistical Software*, 82(13). DOI: 10.18637/jss.v082.i13.

LeBel, E. P., Vanpaemel, W., Cheung, I. & Campbell, L. (2019). "A Brief Guide to Evaluate Replications". *Meta-Psychology*, 3. DOI: 10.15626/MP.2018.843.

LeBoeuf, R. A. & Shafir, E. (2009). "Anchoring on the "Here" and "Now" in Time and Distance Judgments". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1): 81–93. DOI: 10.1037/a0013665.

Leiner, D. J. (2019). "SoSci Survey (version 3.1. 06)[computer software]". URL: https://www.soscisurvey.de/.

Li, L., Maniadis, Z. & Sedikides, C. (2021). "Anchoring in Economics: A Meta-analysis of Studies on Willingness-to-pay and Willingness-to-accept". *Journal of Behavioral and Experimental Economics*, 90(101629). DOI: 10.1016/j.socec.2020.101629.

Mayer, M. & Rebholz, T. R. (2024). "Navigating Anchor Relevance Skillfully: Expertise Reduces Susceptibility to Anchoring Effects". URL: https://osf.io/preprints/psyarxiv/69jwr.

Meub, L., Proeger, T. & Bizer, K. (2013). "Anchoring: A Valid Explanation for Biased Forecasts when Rational Predictions are Easily Accessible and Well Incentivized?" *cege Discussion Papers No 166*. URL: https://www.econstor.eu/bitstream/10419/78599/1/756663717.pdf.

Mussweiler, T. & Strack, F. (1999a). "Comparing is Believing: A Selective Accessibility Model of Judgmental Anchoring". *European Review of Social Psychology*, 10(1): 135–167. DOI: 10.1080/14792779943000044.

Mussweiler, T. & Strack, F. (1999b). "Contamination & Correction of Social Judgements: Strategies of Correction Revisited: Theory-based Adjustment versus Recomputation".

Pearson, K. & Filon, L. N. G. (1898). "VII. Mathematical Contributions to the Theory of Evolution.— IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation". *Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character*, 191(0): 229–311. DOI: 10.1098/rsta.1898.0007.

Revelle, W. (2024). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. URL: https://cran.r-project.org/web/packages/psych/index.html.

Röseler, L. (2021). Anchoring Effects: Resolving the Contradictions of Personality Moderator Research. Ph.D. thesis, Hochschule Harz. PDF: https://fis.uni-bamberg.de/server/api/core/bitstreams/6b8bc0b9-31e8-41e1-a052-4982dc18ea66/content.

Röseler, L. & Schütz, A. (2022). "Hanging the Anchor Off a New Ship: A Meta-Analysis of Anchoring Effects". URL: https://osf.io/preprints/psyarxiv/wf2tn.

Röseler, L., Schütz, A., Baumeister, R. F. & Starker, U. (2020). "Does Ego Depletion Reduce Judgment Adjustment for Both Internally and Externally Generated Anchors?" *Journal of Experimental Social Psychology*, 87(103942). DOI: 10.1016/j.jesp.2019.103942.

Röseler, L., Schütz, A. & Starker, U. (2019). "Cognitive Ability Does Not and Cannot Correlate with Susceptibility to Anchoring Effects". URL: https://osf.io/preprints/psyarxiv/bnsx2.

Röseler, L., Weber, L., Helgerth, K., Stich, E., Günther, M., Wagner, F.-S. & Schütz, A. (2022). "Measurements of Susceptibility to Anchoring are Unreliable: Meta-analytic evidence from more than 50,000 anchored estimates". URL: https://osf.io/preprints/psyarxiv/b6t35.

Röseler, L., Schütz, A., Blank, P. A., Dück, M., Fels, S., Kupfer, J., Scheelje, L. & Seida, C. (2021). "Evidence against Subliminal Anchoring: Two close, Highly Powered, Preregistered, and Failed Replication Attempts". *Journal of Experimental Social Psychology*, 92(104066). DOI: 10.1016/j.jesp.2020.104066.

Schindler, S., Querengässer, J., Bruchmann, M., Bögemann, N. J., Moeck, R. & Straube, T. (2021). "Bayes Factors Show Evidence Against Systematic Relationships Between the Anchoring Effect and the Big Five Personality Traits". *Scientific Reports*, 11(1): 7021. DOI: 10.1038/s41598-021-86429-2.

Shanks, D. R., Barbieri-Hermitte, P. & Vadillo, M. A. (2020). "Do Incidental Environmental Anchors Bias Consumers' Price Estimations?" *Collabra Psychology*, 6(1): 19. DOI: 10.1525/collabra.310.

Simmons, J., Nelson, L. & Simonsohn, U. (2012). "A 21 Word Solution". *SSRN Electron Journal*. DOI: 10.2139/ssrn.2160588.

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(6): 917–932. DOI: 10.1037/a0021540.

Simonsohn, U. (2015). "Small telescopes: Detectability and the Evaluation of Replication Results". *Psychological Science*, 26(5): 559–569. DOI: 10.1177/0956797614567341.

Team, R. C. & others" (2023). R Core Team: R Foundation for Statistical Computing. URL:

Tversky, A. & Kahneman, D. (1974). "Judgment under Uncertainty: Heuristics and Biases". *Science*, 185(4157): 1124–1131. DOI: 10.1126/science.185.4157.1124.

Viechtbauer, W. (2010). "Conducting Meta-Analyses in R with the metafor Package". *Journal of Statistical Software*, 36(3). DOI: 10.18637/jss.v036.i03.

Wickham, H. (2007). "Reshaping Data with the reshape Package". *Journal of Statistical Software*, 21(12). DOI: 10.18637/jss.v021.i12.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL: https://ggplot2-book.org/.

Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. (2023). dplyr: A Grammar of Data Manipulation. URL: https://CRAN.R-project.org/package=dplyr.

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(4): 387–402. DOI: 10.1037/0096-3445.125.4.387.

Yoon, H., Scopelliti, I. & Morewedge, C. K. (**2021**). "Decision Making Can Be Improved Through Observational Learning". *Organizational Behavior and Human Decision Processes*, 162: 155–188. DOI: 10.1016/j.obhdp.2020.10.011.

Appendix

Study 1: Exploratory Tests

To allow for comparisons between adjustments away from different anchors for different items, we computed standardized adjustment scores by dividing the difference between estimate and anchor by the difference between true value and anchor. Scores for all questions and their correlations with cognitive load along with correlations of absolute adjustment and cognitive load are displayed in Figure 8 and Table 5.

Moreover, we tested whether offering the inclusion of participants who were recruited with non-monetary incentives to the participants affected the results. Due to having too few participants, we provided later volunteers with feedback about their susceptibility to anchoring and the accuracy of their estimates (non-monetary incentive). The need for cognition scores in the sample that received incentives was significantly higher, $t(153.42) = 2.27$, $p = .024$ (two-tailed; $M_{\text{no incentive}} = 77.03$, $SD_{\text{no incentive}} = 13.66$, $N_{\text{no incentive}} = 80$; $M_{\text{incentive}} = 81.19$, $SD_{\text{incentive}} = 15.14$, $N_{\text{incentive}} = 223$). Susceptibility to anchoring did not differ between the samples (both $p > .201$). As both of the susceptibility to anchoring scores were unreliable, their correlation was very low, too, $r(301) = .060$, $p = .298$ (two-tailed). The correlations between susceptibility to the experimenter-provided anchoring items and need for cognition and susceptibility to the self-generated-anchoring items and need for cognition did not differ, $z = 0.78$, $p = .435$ (two-tailed; Pearson & Filon, 1898). Furthermore, neither type of anchoring was related to whether people knew or thought about anchoring effects during the experiment (both $p > .274$). Detailed results can be obtained via the analysis script (https://osf.io/wzhmy/).

As no exclusion criteria were reported in the original study, we re-ran our analyses without excluding participants. Still, despite leverage points, correlations between need for cognition and adjustment from self-generated anchors ($r[301] = -.100$, $p = .042$, one-tailed) and adjustment from experimenter-provided anchors ($r[301] = -.011$, $p = .846$) were close to zero. Note that excluding the middle quintiles obviously did not render the effects significant, either (see analysis script for exact results, https://osf.io/wzhmy/).

Due to the gender imbalance, we tested whether gender affected the relationship between NFC and anchoring. Correlations between adjustment from self-generated anchors were $r_{\text{female}}(208) = .047$, 95% CI $[-.089, .181]$ and $r_{\text{male}}(91) = .056$, 95% CI $[-.149, .257]$. This is in line with meta-analytical findings (e.g., Röseler & Schütz, 2022, Table 2, p. 21).

Finally, we divided the difference between anchor and estimate by the difference between anchor and true value per person and per item. This additional procedure leads to an already standardized 0–1 score instead of an absolute adjustment score that has to be z-transformed before aggregation (e.g., Yoon et al., 2021, p. 14). 0–1 scores are 0 if the true value was estimated and 1 if the anchor was estimated. The results this procedure yielded did not differ from those that we obtained with the preregistered tests.

Study 2: Exploratory Tests

For exploratory purposes, we included a question about whether participants prevented themselves from experiencing cognitive load on the letters task, for example, by writing down the letter strings instead of memorizing them. A total of 11 participants indicated that they had done so. Excluding these participants did not affect the results, as the interaction was still not significant, $F(1, 296) = 1.82$, $p = .179$.

To assess the cognitive load manipulation, we analyzed the memory task. Participants correctly reproduced 43% of all letter strings ($SD = 0.30\%$, $N = 93$). Performance declined slightly over the 18 items, $r(16) = -.484$, $p = .042$, and was reliable, $\alpha = .91$. Participants' adjustment scores were not correlated with the number of memorized letter strings ($r_{\text{sg}}[91] = -.176$, $r_{\text{ep}}[91] = .067$).

To test if the effects were dependent on our exclusion criteria, we tested them again but without applying our exclusion criteria. This led to the inclusion of 3 more participants and a total sample size of $N = 186$. Cognitive load was still not associated with adjustment from experimenter-provided anchors, $d = -.105$, 95% CI $[-0.393, 0.183]$, or adjustment from self-generated anchors, $d = 0.060$, 95% CI $[-0.228, 0.347]$.

As in Study 1, we computed adjustment scores by dividing the difference between the estimate and anchor by the difference between the true value and anchor. These 0–1 scores for all questions and their correlations with cognitive load along with the correlations between absolute adjustment and cognitive load are presented in Figure 8 and Table 5.

Study 3: Exploratory Tests

To test whether our analyses were distorted by poorly functioning anchoring items, we repeated our tests using only the items with significant anchoring effects. Being forewarned still did not significantly increase adjustment for the self-generated anchoring items ($M_{\text{no-forewarning}} = -0.07$, $M_{\text{forewarning}} = 0.03$, $SD_{\text{no-forewarning}} = 0.74$, $SD_{\text{forewarning}} = 0.65$, $t(151.13) = -0.98$, $p = .164$, $d = 0.155$, 95% CI $[-0.150, 0.459]$).

Some participants already knew about anchoring effects, which might have had a similar effect as the forewarning introduction in our study. Therefore, we combined the people who were forewarned or already knew about anchoring effects and compared them with the other participants. There was still no significant difference in the adjustments for the experimenter-provided ($M_{\text{no-forewarning}} = -0.07$, $M_{\text{forewarning}} = 0.04$, $SD_{\text{no-forewarning}} = 0.42$, $SD_{\text{forewarning}} = 0.53$, $N_{\text{no-forewarning}} = 54$, $N_{\text{forewarning}} = 117$, $t(126.16) = -1.409$, $p = .081$, $d = 0.153$, 95% CI $[-0.149, 0.455]$) or for the self-generated anchoring items ($M_{\text{no-forewarning}} = -0.04$, $M_{\text{forewarning}} = -0.01$, $SD_{\text{no-forewarning}} = 0.50$, $SD_{\text{forewarning}} = 0.53$, $N_{\text{no-forewarning}} = 54$, $N_{\text{forewarning}} = 117$, $t(108.28) = -0.280$, $p = .390$, $d = -0.120$, 95% CI $[-0.182, 0.421]$).

Because our sample size exceeded our targeted sample of 120 participants even after we removed participants on the basis of the exclusion criteria, we reran the statistical analyses with only the first 120 participants for comparison. The smaller sample yielded no remarkable differences in results. Therefore, we used the total sample to achieve higher power.

As in Studies 1 and 2, we computed standardized 0–1 scores by dividing the difference between estimate and anchor by the difference between true value and anchor. Scores for all questions and their correlations with forewarning along with correlations between absolute adjustment and forewarning are presented in Figure 8 and Table 5.

Transparency and Openness Statement

This report is an exhaustive report on all data available from research projects related to the replication of studies with self-generated and experimenter-provided anchors of which LR was the principal investigator. This includes not only null findings or unexpected findings but also studies that are considered to have failed, with careful explanation of the circumstances of the failure (e.g., experimental error, failed manipulation check). We carefully explain the contexts in which these data were collected and whether the data were connected to published studies (e.g., dropped experiments). As suggested by Simmons et al. (2012), we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All materials are available online (https://osf.io/prwu6).

- Instructions to reproduce the reported results: https://osf.io/qmkn6
- Study 1
  - Preregistration: https://osf.io/8k9nt
  - Materials: https://osf.io/9dnez
  - Raw Data: https://osf.io/4d8vj
  - Processed Data: https://osf.io/gvp9n
  - Differences between the original study and our replication: https://osf.io/xayzt
- Study 2
  - Preregistration: https://osf.io/b7nt8
  - Materials: https://osf.io/ktjc4
  - Raw Data: https://osf.io/egtm2
  - Processed Data: https://osf.io/g29w8
  - Differences between the orginal study and our replication: https://osf.io/hb57n
- Study 3
  - Preregistration: https://osf.io/f5sj8
  - Materials: https://osf.io/dwm82
  - Raw Data: https://osf.io/xezqv
  - Processed Data: https://osf.io/xdqn8
  - Differences between the original study and our replication: https://osf.io/rmusg