Judgment and Decision Making, Vol. 13, No. 2, March 2018, pp. 170–184

The category size bias: A mere misunderstanding

Hannah Perfecto* Leif D. Nelson† Don A. Moore‡

Abstract

Redundant or excessive information can sometimes lead people to lean on it unnecessarily. Certain experimental designs can sometimes bias results in the researcher's favor. And, sometimes, interesting effects are too small to be studied, practically, or are simply zero. We believe a confluence of these factors led to a recent paper (Isaac & Brough, 2014, *JCR*). This initial paper proposed a new means by which probability judgments can be led astray: the category size bias, by which an individual event coming from a large category is judged more likely to occur than an event coming from a small one. Our work shows that this effect may be due to instructional and mechanical confounds, rather than interesting psychology. We present eleven studies with over ten times the sample size of the original in support of our conclusion: We replicate three of the five original studies and reduce or eliminate the effect by resolving these methodological issues, even significantly reversing the bias in one case (Study 6). Studies 7–8c suggest the remaining two studies are false positives. We conclude with a discussion of the subtleties of instruction wording, the difficulties of correcting the record, and the importance of replication and open science.

Keywords: subjective probability, judgment, estimation, bias, replication

1 Introduction

Assessments of subjective likelihood are crucial in decision making, since every decision depends on beliefs about its likely consequences. However, because people are imperfect estimators of likelihoods (Edwards, 1968), scholars have long been interested in understanding exactly how people form subjective probability judgments (Kahneman & Tversky, 1972; Gigerenzer & Hoffrage, 1995; Newell, Lagnado & Shanks, 2007). Isaac and Brough (2014) made a contribution to this literature by documenting a new bias in subjective probability judgments: the category size bias.

The category size bias is the phenomenon whereby people judge an individual event to be more likely to occur when it comes from a larger, more likely category. If the features of the category bleed over into the features people ascribe to a particular instance, they may well make this error. The bias is captured in the joke about the farmer whose farm switched, in 1920, from being part of Russia to being part of Poland. When asked what he thought of the change, he responded, "Thank goodness I won't have to endure those harsh Russian winters anymore."

The first author was supported by the University of California, Berkeley Chancellor's Fellowship. The second author was supported by the Barbara and Gerson Bakar Faculty Fellowship. This work was supported in part by the Fetzer-Franklin Fund.

Copyright: © 2018. The authors license this article under the terms of the Creative Commons Attribution 3.0 License.

*Washington University in St. Louis, Olin Business School, Campus Box 1156, 1 Brookings Drive, St. Louis, MO 63130. E-mail: perfecto@wustl.edu.

†University of California, Berkeley.

‡University of California, Berkeley.

The category size bias is not implausible. Prior research has already demonstrated that categorization affects probability judgments (Tversky & Koehler, 1994). Additionally, category membership certainly influences the inferences people draw about its members. In a social domain, this could mean that people judge a criminal defendant as more likely to be guilty when that person comes from an ethnic group stereotypically associated with criminality (Bodenhausen, 1988). Similarly, in a non-social domain, people feel more confident that they will win a raffle when they hold three tickets and seven other people all hold one ticket each than when one other person holds seven tickets (Windschitl & Wells, 1998). It is noteworthy, however, that Windschitl and Wells (1998) found that the effect was isolated to subjective feelings of risk, confidence, or worry. Their participants' judgments of numerical probability were unaffected by their manipulations of category.

A representative paradigm that Isaac and Brough (2014) used to test their category size hypothesis (and a focal paradigm investigated in the present paper) consists of asking participants to consider a 26-sided die, with each face representing a different letter of the alphabet. Participants estimated the probability of rolling either the vowel A (coming from a small category of letters, vowels) or the consonant T (coming from a large category, consonants). Though, of course, each letter would be equally likely to come up, participants judged the A to be less likely to come up than the T.

Isaac and Brough (2014) argued that the category size bias follows the similar intriguing psychological logic as in the above investigations. People infer the characteristics of the category when considering its individual members.

Accordingly, when thinking of the category "consonants," people recognize that the category is large and therefore likely to occur. That "trait," high likelihood, then bleeds over into the assessment of letter T's likelihood. Put another way, the authors ascribe this disparity to participants' "erroneous belief that individual category members inherit the statistical propensities of the parent category" (p. 311).

We suggest a more mundane alternative explanation. It is possible that people are simply confused by the instructions. If the experimental instructions allowed confusion about whether the question is about the instance (the letter T) or the broader category (consonants), unbiased and well-intentioned participants may well respond in ways that appear to reveal bias. Conversational norms dictate that one should only include as much information as is necessary (Grice, 1975). Perhaps, by including redundant information about the category, some participants misunderstood the question and estimated the likelihood of the category: rolling a vowel (23%) or a consonant (77%) rather than the probability of rolling an A or a T. We sought out to understand whether this mundane alternative might explain some or all of the category size bias.

Before collecting any new data it is worth noting that there are some ambiguous indications for confusion in the existing data. In Isaac and Brough's (2014) die task, the correct answer is 3.8% (for both conditions, since the letters A and T have an equal chance of coming up on a roll of the 26-sided die). In the study itself, participants estimated the probability of rolling the vowel A at 11.2% and of rolling the consonant T at 32.2%. Although, *a priori*, we found the authors' hypothesis plausible, we still found these means to be surprisingly, even worryingly, high. Did people really think that a T will come up almost one third of the time?

Our investigation proceeded with the following strategy. First we confirm that the primary finding was replicable (Study 1). Then with that knowledge, we show that the effect can be substantially mitigated by a very mild clarification (Studies 2 to 4) and provide evidence that it really is confusion that drives the bulk of the effect (Study 5). Finally, because the original authors had claimed to address this concern with their final experiment, we gave that study extra scrutiny in Study 6: We found that study to contain a different, but substantial, confound, which when eliminated in an empirical investigation, completely eliminates the reported difference. We conclude with replication attempts for the original paper's remaining two studies.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in each study (Simmons, Nelson and Simonsohn, 2012). All studies, with the exception of Study 2 and Study 4a, were preregistered at the Open Science Framework, and the data for all studies are posted there as well (https://osf.io/wq5n7/). Because the OSF can be tricky to navigate, we provide links to the individual pages for each study, containing materials, data, and preregistrations, as follows: Study 1, Study 2, Study 3, Study 4a, Study 4b, Study 5, Study 6, Study 7, Study 8a, Study 8b, Study 8c.

2 Study 1

Before we could investigate possible boundary conditions, we wanted to replicate Isaac and Brough's (2014) key result. To do this, we conducted a direct replication, using the materials the authors generously included in an appendix of their paper. We focus first on Isaac and Brough's (2014) Study 3. This study manipulates both the category size (small vs. large) and the category's salience (high, medium, low) in a 2x3 between-subjects design. The key hypothesis is that participants should make greater probability estimates for a specific event when it comes from a large category and that category's salience is high.

2.1 Method

We recruited 505 US participants from Amazon Mechanical Turk. We chose this sample size as it is roughly 2.5 times that of the original study (Simonsohn, 2015). All participants considered a 26-sided die, with a different letter of the alphabet on each side. Participants in the small-category conditions estimated the probability of rolling an "A," whereas participants in the large-category conditions estimated the probability of rolling a "T". In the high-salience conditions, participants were first reminded that there are 5 vowels and 21 consonants in the alphabet, and were then asked about rolling "the vowel A" or "the consonant T". Participants in the medium-salience conditions did not receive the initial reminder about letter distribution, but were still asked about "the vowel A" or "the consonant T". Finally, participants in the low-salience conditions also did not receive the initial reminder, and were asked about "the letter A" or "the letter T". As a secondary, exploratory test, we subsequently asked all participants the probability of rolling "the consonant P" and reminded them we were not looking for the category's likelihood.

2.2 Results

We conducted a 2 (category size: large vs. small) x 3 (category salience: high, moderate, low) ANOVA. Neither main effect reached significance, but the predicted interaction did, $F(2, 499) = 3.71$, $p = .025$.[1] Replicating the original study, the category size bias emerged in the high-salience conditions (see Figure 1). These participants gave significantly

[1]Although the data's multi-modal distributions might make non-parametric tests more appropriate, we present results from parametric tests throughout both to facilitate a better comparison with the original paper and because the differences in results are minimal with large sample sizes like ours.

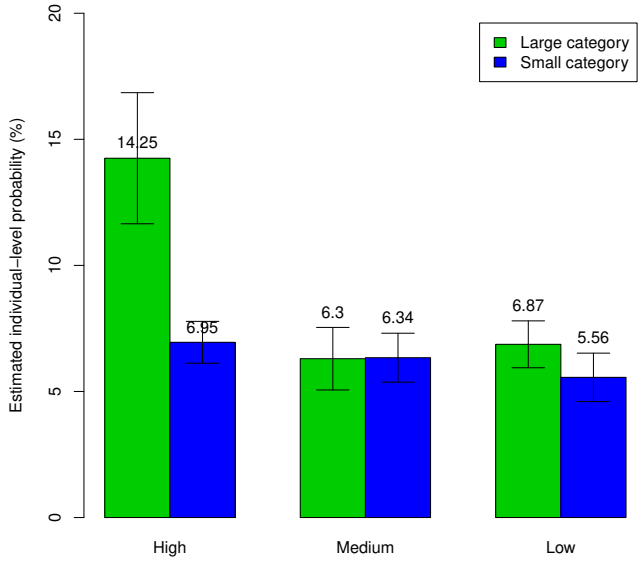

FIGURE 1: Estimated likelihood as a function of category size and category salience from Study 1 (all error bars indicate +/- SE).

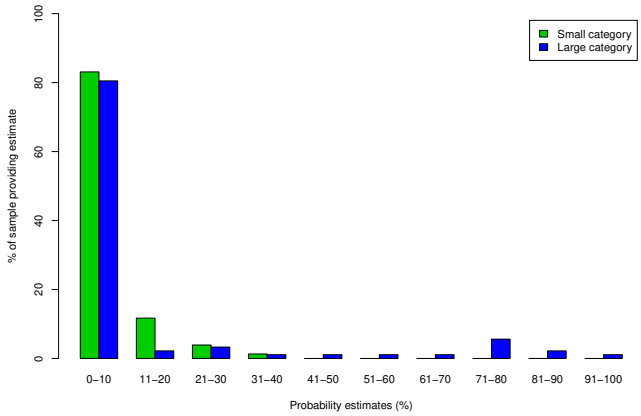

FIGURE 2: The distribution of estimates for estimates from large or small categories in the high-salience conditions in Study 1.

greater estimates of rolling a "T" than of rolling an "A", $t(162) = 2.54$, $p = .012$. When category salience was moderate, $t(164) = 0.03$, $p = .976$, or low, however, the category size bias was not observed, $t(158) = .64$, $p = .523$. Although the original authors found a significant simple effect at medium salience and we did not, we still consider this to be a successful replication.

Probabilities reported by our participants in the high-salience condition (14.3% and 7% for "T" and "A," respectively) were substantially lower than those of the original study. However, considering that the correct answer is 3.8%, they were still considerably higher than what would be expected from an arithmetically industrious participant. We took a closer look at the distribution of responses (see Figure 2).

The pattern is telling. In both conditions the modal response is close to the correct number: 80.5%, guess a number quite close to the correct answer (i.e., a number between 0% and 5%), with a quarter of those participants guessing exactly 3.8%. (Participants in neither the original study nor our replication were explicitly barred from using a calculator if they so desired.)

We looked for exceedingly high and isolated outliers, but what we found was something much more systematic. Aside from the aforementioned cluster of responses at the low end, the distribution in each condition (Figure 2) features its own, smaller cluster at higher levels: around 80% for rolling "consonant T" and around 20% for rolling "vowel A". These three categories — near the correct answer of 3.8%, near 80%, and near 20% — account for 92.8% of responses.[2] We do not believe these secondary clusters to be random: note that 80%, a cluster found only in the "consonant T" condition, is roughly equivalent to the odds of rolling *any* consonant, and 20%, a cluster found primarily in the "vowel A" condition, is roughly equivalent to rolling *any* vowel. This pattern of responses is consistent with the idea that some participants simply misunderstood and answered a different question. Studies 2 and 3 were designed to test this possibility. Specifically, we test whether participants mistake the question of asking for the odds of a specific member of the category to be asking for the odds of any member by clarifying the original phrasing.

3 Study 2

To dispel this confusion, we aimed to clarify the goal of the question for participants. In considering possibilities, we wanted to find a manipulation which was sufficient to clarify, but not so large as to potentially disrupt the hypothesized process. We do so by first asking participants about the category's probability and then asking about the category member's probability: because participants do not expect to be asked the same question twice (Gal & Rucker, 2011), we predict that providing the category's probability first will signal to participants to provide the category member's probability subsequently. In Study 2, we try this alternative phrasing of the probability elicitation to minimize misunderstanding and eliminate the effect in Study 3. If this reminds them of the features of the broader category it could plausibly magnify the category size bias, but we predict that, by clarifying the question, it will reduce or eliminate it.

3.1 Method

We aimed for approximately 100 participants per cell: Given our two-cell design (small vs. large category), we recruited

[2]Moreover, we obtained the original data from the authors and found that they reflected a similar pattern as well.

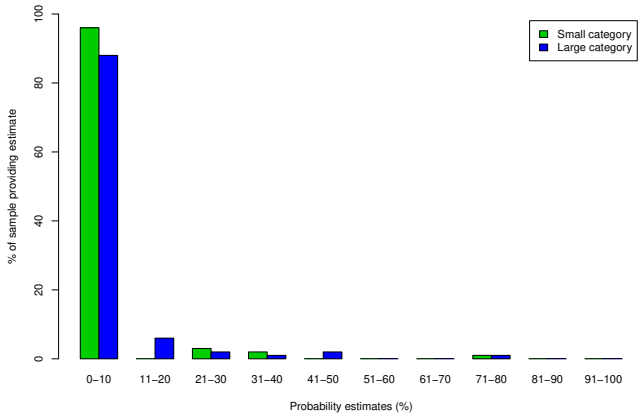

FIGURE 3: Histogram from Study 2 (clarified phrasing).

211 participants from Amazon Mechanical Turk, but excluded 32 participants who reported participating in our previous study (their inclusion does not meaningfully affect the results). Because the critical effect (and, in our case, the only significant one) is the simple effect of large and small categories within high category salience (i.e., likelihood of rolling the consonant T versus the vowel A; the leftmost pair of bars in Figure 1), we focus on these cells for our remaining studies: Participants were asked to consider the 26-sided alphabet die and recall the frequency of vowels and consonants in the alphabet. For the probability elicitation, however, we made two changes. First, participants were asked, out of 100 rolls of the die, how many would be the target letter, as a more intuitive alternative to percent likelihood. Second, and more importantly, we broke down the original phrasing into two questions: first, participants were asked about the probability of a vowel [consonant] and then, underneath, they were asked about the probability of the letter A [T]. We hoped that this clarifying phrasing would better test the existence of a category size bias.[3]

3.2 Results

Using this revised phrasing, the category size bias was reduced to non-significance. Participants gave similar estimates for the likelihood of rolling an "A" (M=5.5%) and "T" (M=7.6%), $t(177) = 1.38$, $p = .168$. Whereas Isaac and Brough (2014) observed a 20% difference between those conditions and our replication in Study 1 observed a 7% difference, importantly, the difference under the clarified phrasing was further reduced to only 2%. Accordingly, if we combine the data from Study 2 with Isaac and Brough's (2014) data, we find significant moderation, $F(1, 231) = 18.04$, $p < .001$. In line with our hypothesis that participants were mistakenly reporting the category probability, we no longer saw the substantial clusters at 80% and 20% (Figure 3). In fact, no participant guessed 20% in the "A" condition, and no participant guessed 80% in the "T" condition. Taken together, these results suggest that the category size bias as reported by Isaac and Brough (2014) may have been largely driven by participants misunderstanding the question being asked.[4]

4 Study 3

In Study 3, we sought to show more directly how the category size bias can be eliminated with this clarified phrasing, by combining Studies 1 and 2 into a single design, rather than comparing across studies.

4.1 Method

We recruited 604 participants from Amazon Mechanical Turk, after excluding three for giving nonsensical answers (e.g., 2,600 rolls out of 100). Because this included a replication, we wanted to have 2.5 times the original sample, but because our key result would be an attenuated interaction, we needed to double that, and recruit five times the original sample (Simonsohn, 2015; see also: http://datacolada.org/2014/03/12/17-no-way-interactions-2/). The study had a 2 (category size: large vs. small) x 2 (phrasing: original vs. clarified) between-subjects design. All participants gave two estimates — the probability of rolling the target letter (as in the original study) and the probability of rolling the category — but we varied the order in which they appeared, and separated them on consecutive pages (Figure 4). The order determined participants' conditions: Participants who gave the target estimate first were in the original-phrasing conditions, and participants who gave the category estimates first were in the clarified-phrasing condition. As in Study 2, for each estimate, participants were asked out of 100 rolls of the die, how many would be [X].

4.2 Results

We conducted a 2x2 ANOVA on target estimates. The results reveal a significant main effect of category size, $F(1, 599) = 3.93$, $p = .048$, as well as of phrasing, $F(1, 599) = 6.24$, $p = .013$. Overall, the interaction shows that our additional factor of phrasing weakly attenuated the original effect, $F(1, 599) = 3.08$, $p = .080$. (An alternative approach could be to combine the data from Study 1 [high salience conditions only] and

[3]This was the earliest study we conducted, and did so purely out of curiosity. Our lab had also not yet instituted a pre-registration policy at the time. For these reasons alone, we did not pre-register Study 2.

[4]One of us (D.A.M.), in parallel to Study 2, tried a similar approach. He recruited 86 participants from MTurk and employed a within-subjects version of Study 2's design, including the two-step revised phrasing. As in Study 2, this clarified wording rendered the original effect non-significant with participants who were asked about the vowel A (M=6.6%) giving statistically identical estimates to participants who were asked about the consonant T (M=7.7%), $t(85)=0.60$, $p = .549$.

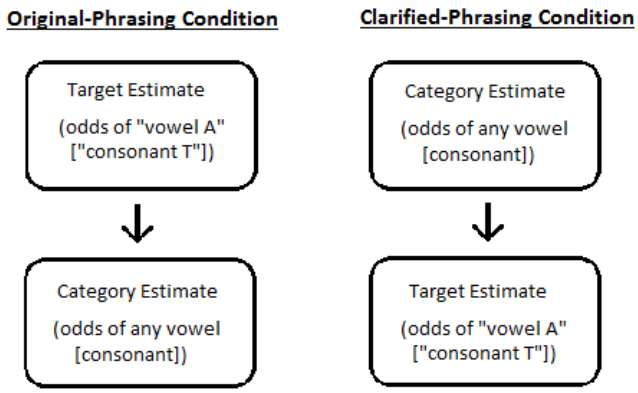

FIGURE 4: Procedure for the two conditions in Study 3:

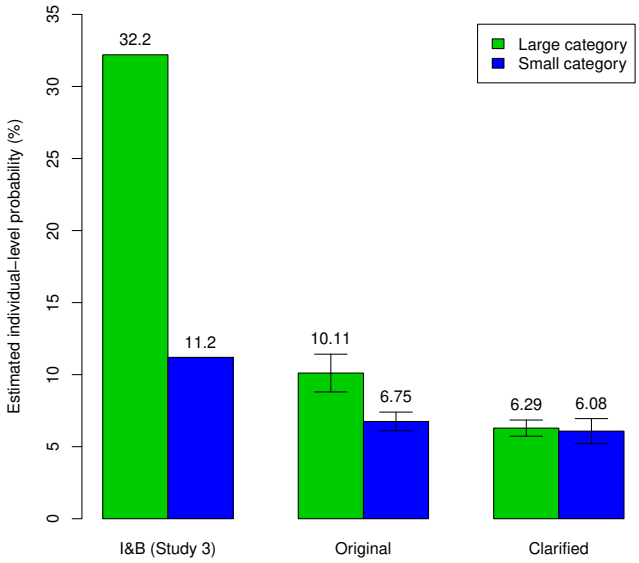

FIGURE 5: Means from Isaac and Brough (2014) Study 3 (left), and means from our replication (middle) and clarified phrasing (right), showing marginal moderation of our replication of the effect.

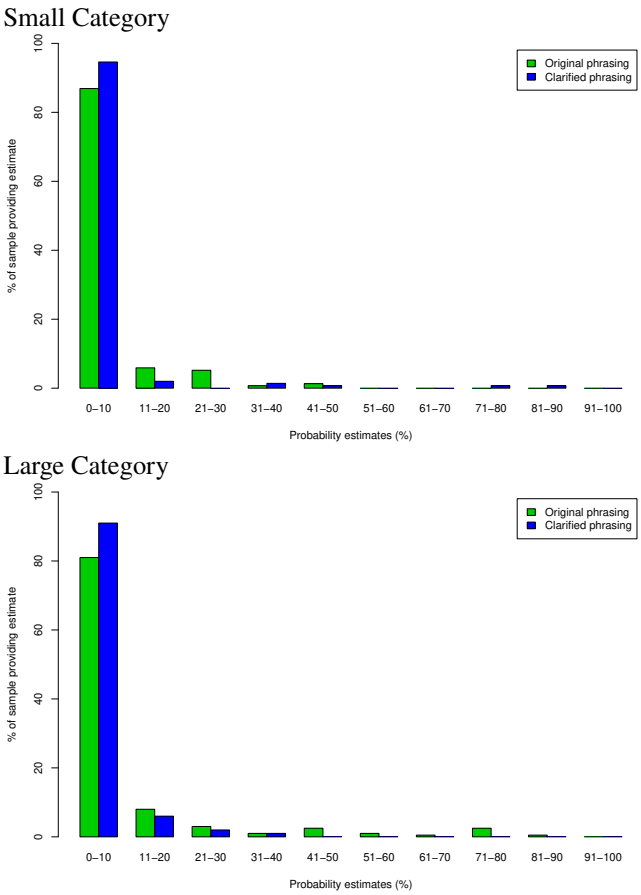

FIGURE 6: Histograms from Study 3, comparing original to clarified phrasing for the small (top panel) and large (bottom panel) category size.

Study 2. Doing so reveals another weak attenuation of the category size effect by phrasing, $F(1, 371) = 3.60$, $p = .058$.) All of these effects were driven by the substantially higher mean from the original/large-category condition (see Figures 5 and 6 for distributions). As in Study 1, we replicated the original effect, $t(303) = 2.30$, $p = .022$.[5] Consistent with our hypothesis, the clarified phrasing, even on separate pages, again reduced this effect to non-significance, $t(296) = 0.20$, $p = .844$.

With this subtle, minor manipulation in Study 3, we eliminated the category size bias. Merely by answering an additional question before the original, unchanged target question, participants were now largely able provide correct answers. Note that, although we are far from the first to show participants giving the right answer to the wrong question (Kahneman & Frederick, 2002), this is not an example of attribute substitution: Providing the probability of rolling a "T" should be no harder or less accessible than providing the probability of rolling any consonant, yet a small percentage of participants reliably and incorrectly report the latter. Moreover, this is not moderation in the theoretical sense — this clarification manipulation should not have eliminated the original effect if the theory behind it were correct — instead, it is more evidence for our confusion hypothesis.

5 Study 4a

Thus far, we have focused exclusively on Isaac and Brough's (2014) Study 3. However, this is not the only study in the paper; nor is it the only one susceptible to misunderstanding. In our fourth study, we apply the same debiasing approach as we did in Study 3, but to Isaac and Brough's (2014) Study 1.

[5]The astute reader may note a drop in effect size from that of Study 1. We attribute this decline primarily to eliciting probability in the more intuitive terms of die rolls out of 100, as opposed to as a percentage.

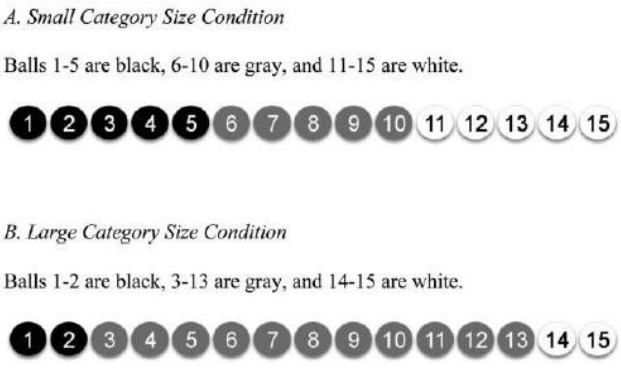

FIGURE 7: Stimuli from Isaac and Brough's (2014) Study 1, used in our Studies 4a and 4b.

5.1 Method

We recruited 515 participants from Amazon Mechanical Turk. Because the original study appeared to be adequately powered, with the largest sample size in the paper (N = 223), we were not able to quintuple the sample size on MTurk for this study, as we had done previously. Instead, we tried to give both the replication and debiased conditions (description to follow) approximately the same number of participants as in the original. No participants were excluded.

With the help of the appendix in the original paper and correspondence with the authors, we recreated the original paradigm to the best of our ability. We presented participants with an image of a set of balls that will be placed in an urn, from which one was to be drawn. All participants saw a row of gray, black, and white balls, numbered 1–15 (Figure 7), and estimated the probability of drawing ball number eight. In the small-category condition, balls 1–5 were black, 6–10 were gray, and 11–15 were white. In the large-category condition, balls 1–2 were black, 3–13 were gray, and 13–15 were white. Therefore, the size of the category "gray," of which ball number eight was always a part, varied between conditions. In every condition, however, the correct answer was 6.67%.

As in our Study 3, we also implemented a mild, secondary manipulation to test the possibility that confusion was driving the results. As in Study 3, all participants gave two estimates — the probability of drawing ball number 8 and the probability of drawing a gray ball — but we varied the order in which these questions appeared, and separated them on consecutive pages. Therefore, participants who first estimated the probability of drawing ball 8 were in the original-phrasing conditions. Participants who first estimated the probability of drawing a gray ball were in the clarified-phrasing condition, resulting in a 2 (category size: small vs. large) x 2 (phrasing: original vs. clarified) between-subjects design. (Because this study was also run before a lab-wide policy for preregistration, only the materials and data are posted.)

5.2 Results

We conducted a 2x2 ANOVA on target estimates. However, we found no significant effects: not of category size, $F(1, 511) = 0.15$, $p = .696$ nor phrasing, $F(1, 515) = 3.17$, $p = .076$. The interaction was also not significant, $F(1, 515) = 0.63$, $p = .428$. Participants in the replication conditions (correctly) estimated that ball 8 was equally likely to be drawn when it was from a small (M = 8.84%, SD = 10.55) vs. large (M = 9.87%, SD = 12.04) category, $t(242) = 0.72$, $p = .475$. Participants in the debiased conditions were similarly indifferent to category size ($M_{small} = 7.96\%$, $SD_{small} = 7.06$; $M_{large} = 7.62\%$, $SD_{large} = 9.54$), $t(269) = 0.34$, $p = .732$.

6 Study 4b

The results from Study 4a casts doubt on the reliability of the category size bias. However, because we did not (yet) believe the original effect could be a false positive, we conducted a direct replication in a much larger sample: Study 4a, with its factorial design, had matched only the original sample size, rather than increasing by the conventional 2.5 times (Simonsohn, 2015). In addition, Study 4b gave us the opportunity to pre-register our materials and analysis plan.

6.1 Method

Participants were recruited as part of larger project that utilized a private, online survey panel company. Participants completed an unrelated survey online and were then redirected to the present study upon completion. The pre-registered sample for the first survey was 1,500 complete, attentive responses. However, to make up for attention-check failures in the main survey, more than 1,500 participants were recruited (although not all chose to participate in Study 4b). In the end, Study 4b had 2,564 participants. This large sample size was useful to enhance statistical power and thereby prevent one possible cause of Study 4a's failure to replicate.

As in our Study 3 and per our pre-registration, we excluded 11 participants for nonsensical answers (e.g., likelihoods greater than 100%), and we also excluded 307 participants for failing an attention check at the end of the study. The attention check asked participants to consider the same image of the set of balls (still numbered 1–15), and asked for the probability that ball #27 will be drawn (answer: 0%). Otherwise, the materials and analysis plan were identical to those of Study 4a.

6.2 Results

We conducted a 2x2 ANOVA on target estimates. We found a significant main effect of category size, $F(1, 2244) = 31.47$, $p < .001$, as well as of phrasing, $F(1, 2244) = 56.42$, $p < .001$. More importantly, consistent with our account, these main

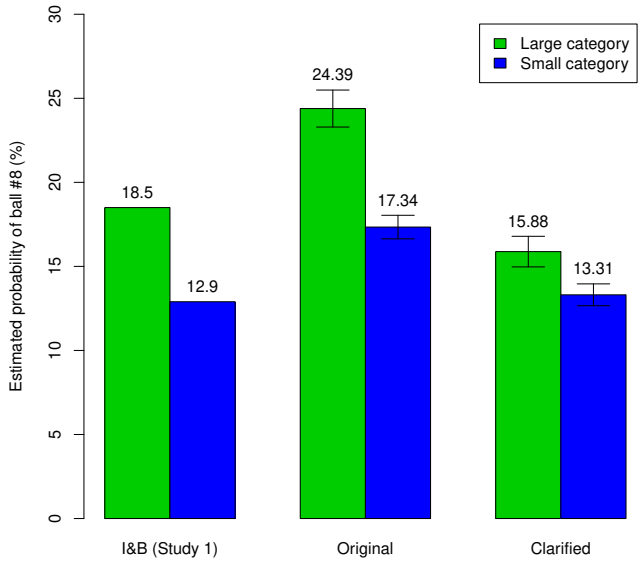

FIGURE 8: Means from Study 4b, showing moderation of the original effect.

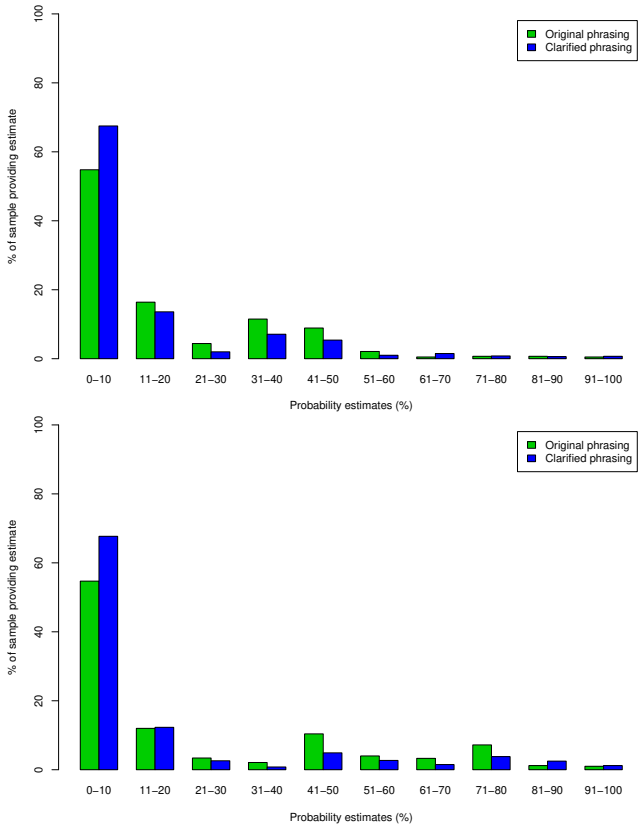

FIGURE 9: Histograms from Study 4b, comparing original to clarified phrasing for the small (top panel) and large (bottom panel) category size.

effects were qualified by a significant interaction, $F(1, 2244) = 6.80$, $p = .009$. (When we include the 307 participants who gave a nonzero probability of drawing a nonexistent ball, this interaction is not quite significant: $F(1, 2548) = 3.74$, $p = .053$. This apparent change is likely due to these additional participants not reading our crucial task instructions either.) As in Study 3, these effects were largely driven by the higher mean from the original/large-category condition (see Figure 8 for mean probability estimates; see Figure 9 for their distributions). We successfully replicated the original effect, $t(1144) = 5.44$, $p < .001$. Although the effect did not completely disappear in the clarified-phrasing conditions, it was significantly smaller, $t(1100) = 2.31$, $p = .021$.

With the previous studies, we have demonstrated the replicability of the category size bias, as Isaac and Brough (2014) tested it. However, these studies suggest that their proposed psychology may not be what caused the original results. Instead, participants may have misunderstood the instructions in such a way as to artificially create results the category size bias would predict: Participants frequently mistakenly provide probability estimates for any member of the category, rather than a specific one, which leads participants to report large-category members as more likely, on average. When we clarify our question and reduce confusion by asking participants for both the category and category-member probabilities, the category size bias disappears (Studies 2–3) or shrinks considerably (Study 4b, if it appears at all). Both the original effect and our moderation replicate across two different samples, probability elicitations, and study designs.

7 Study 5

So far, we have claimed that the category size bias is driven by confused participants, by supposedly alleviating this confusion and eliminating the effect. However, we have not yet directly tested whether this manipulation is reducing confusion, or even if confusion is what is causing these odd distributions (e.g., the long tails in Figure 9). The goal of Study 5, then, was to replicate our attenuation of the category size bias with our clarification manipulation, but include a measure of participant confusion. We do so here by asking participants to recall the event of which they predicted the likelihood. Although we are relying on any confused participants to self-identify with their mistake (and this is therefore an imperfect measure), we nevertheless predict that, because our clarification manipulation reduces confusion, its results should look like original-phrasing's results with the confused participants removed.

7.1 Method

We recruited 817 (60.1% female, $M_{age} = 33.8$) participants from Amazon Mechanical Turk. The procedure for Study 5 was identical to that of Study 3 with one addition: After

making each of the two likelihood estimates (for both the category and for a specific member of that category to obtain), participants were asked to report back the question they had just answered. Specifically, participants were reminded that on the previous page, they had reported the likelihood of an event occurring to be [X]%. They were then presented with seven options and asked which event has a likelihood of [X]%? The options were (with order randomized for each participant): a consonant, a vowel, a "T", an "A", a "5", "I just wrote down a random number", and Other; they were the same for both questions (category and target) as well as for both conditions (small and large category). At the conclusion of the study, in addition to reporting age and gender as in Study 3, participants also reported if they had taken a similar study before.

7.2 Results

Four participants (0.5%) reported having taken a similar study before and were excluded from analyses. (Their inclusion does not meaningfully affect the results.) We then conducted a 2x2 ANOVA on target estimates, which revealed significant effects of both category size, $F(1, 809) = 15.82$, $p < .001$ and question order (our means of reducing confusion), $F(1, 809) = 17.25$, $p < .001$. Crucially, and replicating our Study 3, these effects were qualified by a significant interaction, $F(1, 809) = 4.45$, $p = .035$. Participants in the original-phrasing conditions participants believed rolling the consonant "T" (M = 15.83%, SD = 25.68%) was more likely than rolling the vowel "A" (M = 8.76%, SD = 10.82%), $t(396) = 3.64$, $p < .001$, in line with the category size bias. However, in line with our confusion account, this effect shrank to insignificance in the clarified-phrasing conditions: Participants believed the consonant "T" (M = 8.56%, SD = 15.33%) was just as likely to be rolled next as the vowel "A" (M = 6.38%, SD = 10.64%), $t(413) = 1.66$, $p = .097$.

We turn next to the recall question participants answered after making this target estimate. Overall, most participants recalled the question correctly, with 70.9% in the large-category/consonant conditions and 68.8% in the small-category/vowel conditions. However, of the remaining participants who gave an incorrect answer, the modal response was that they were asked about any consonant [vowel] (13.5% of total responses in the consonant conditions, 9.3% in the vowel conditions). Although these are not large percentages, given that the real answer is so low (3.8%) and these incorrect answers are so much higher (approximately 20% or 80% depending on condition), only a small minority need to be confused in this way to sway the entire study's results.

Therefore, we have two ways of "debiasing" the data: looking at the data of participants in our clarified conditions and excluding participants who self-identified as needing clarification and being confused. These two approaches should both yield significant evidence that appears to support the

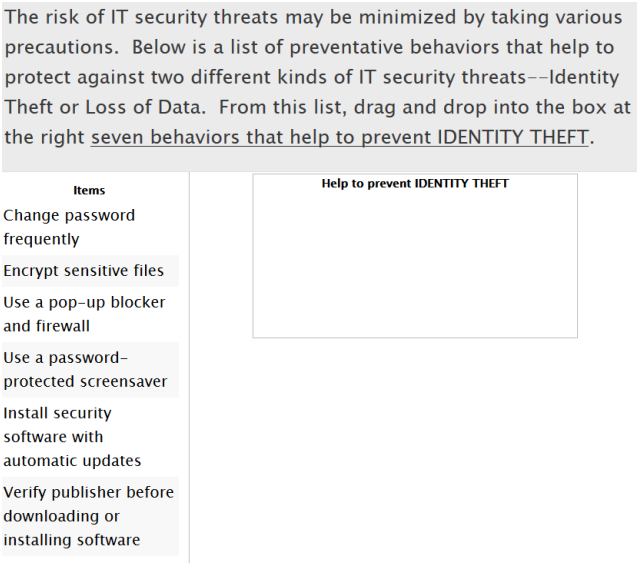

FIGURE 10: Screenshot from the original paper's Study 5. Participants dragged seven items on the left (abridged here) to the box on the right to create the large category.

category size bias in the original condition or form, and non-significant effects with understanding participants. We saw previously that our clarifying manipulation successfully attenuated the category size bias (interaction: $F(1, 809) = 4.45$, $p = .035$). When we exclude participants who reported misunderstanding the question, this attenuation is reduced to insignificance, $F(1, 580) = 2.12$, $p = .146$, suggesting that now the original-phrasing conditions pattern with the clarified-phrasing conditions (however, our study was underpowered to detect this three-way interaction: $F(1, 809) = 0.91$, $p = .342$). The original-phrasing conditions also no longer show a significant category size bias, ($M_{large} = 8.30$, $SD_{large} = 14.51$; $M_{small} = 5.62$, $SD_{small} = 5.39$), $t(235) = 1.90$, $p = .059$. Looking more closely at the distribution of responses, it appears that this small effect is driven largely by a very small minority of five participants (2.5%) in the large-category condition with a category-sized estimate (e.g., 70%–85%; note also the substantial difference in variance between the two groups). We believe that these participants may have been confused too, but did not want to admit it in their response.

8 Study 6

In their paper's final study, possibly anticipating this specter of confusion, Isaac and Brough (2014) designed a study to "show that the category size bias does not stem from a simple misunderstanding of statistical principles or experimental instructions" (p. 318). We agree with their assessment; our debiasing method from the previous studies would not

be appropriate here. However, we were still able to implement a debiasing procedure, which will become clear after describing the study's paradigm.

Instead of using pre-existing categories (e.g., consonants and vowels), they asked participants to sort targets into categories based on their opinions. Specifically, participants saw a list of nine behaviors that guard against computer threats (e.g., use a firewall, avoid opening unknown attachments) and grouped them into two categories: behaviors that guard against data loss or against identity theft. The behaviors were chosen so that two would fall in one category and seven in the other; and participants were told to assign seven to the corresponding larger category (Figure 10). Participants then estimated the likelihood of each of the nine activities. Isaac and Brough (2014) hypothesized that if a threat has several protective behaviors, it should be perceived as more risky, so the behaviors listed under it should be more important to do (and likely to be done).

In concept, this paradigm should avoid the confusion confound as intended. In practice, however, we believe a new confound was built into the research design: Instead of having participants select seven behaviors for the large category and two behaviors for the small category, due to the limitations of the Qualtrics survey software, Isaac and Brough (2014) had participants choose behaviors *only* for the large category, calling the unchosen behaviors the small category (Figure 10). If participants were selecting behaviors they believed were important (for the large category) and leaving behind (for the small category) behaviors they believed were unimportant, then they would report likelihoods consistent with the category size bias, but for an entirely different, mundane reason: judged importance.

Hence, in Study 6, we had three goals: First, as in previous studies, we wanted to replicate the original design (which we were able to do with the original materials, thanks to correspondence with the original authors), the *choose-7* condition. Second, we wanted to test our hypothesis that participants were selecting behaviors that they believed were important. To do this, we added a second condition in which the larger category was now made up of leftovers and participants were asked to select behaviors for the smaller category, the *choose-2* condition. The category size bias would still predict greater reported likelihoods for behaviors in the large category — the bias depends on the categorizations; it is indifferent to how the categories came about — but if we are correct that participants believe their selected behaviors are more important than their unselected behaviors, we predicted this choose-2 condition should fully reverse the original effect. Finally, we created a third condition that best represents Isaac and Brough's (2014) original intention: participants select all nine behaviors and categorize all of them into the two groups without leftovers, the *two-groups* condition (Figure 11). Because all behaviors are selected here, and thus considered at least somewhat important, we predicted a null effect of category size.

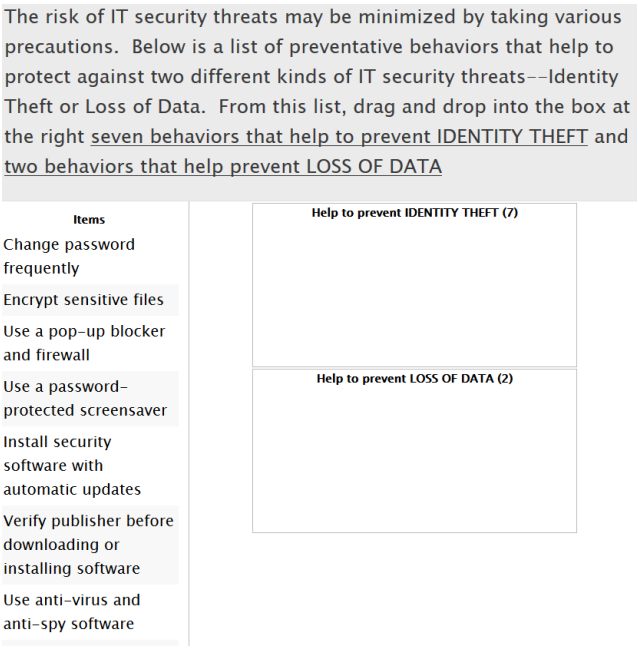

FIGURE 11: The two-groups condition from Study 6, modifying the original paper's Study 5 by providing one box for each category.

8.1 Method

We recruited participants as part of larger project that utilized the same private, online survey-panel company as Study 4b. Again, participants first completed another online survey and were then redirected to the present study upon completion. The first survey determined the sample size we could obtain. That survey had a pre-registered sample size of 1500 complete, attentive responses. However, to make up for attention-check failures the sample size is larger than 1,500. In the end, Study 6 had 1,551 participants. We selected this study to receive this large sample because of its unusual design and importance to the present argument.

Using the methods section and the appendix from the original paper and correspondence with the original authors, we recreated the original paradigm to the best of our ability. In the original study, participants read nine behaviors that could be used to protect against computer threats, and had to categorize each behavior according to which of two threats (identity theft or data loss) the behaviors better protected against. One threat requires seven behaviors; the other requires two (with threat counterbalanced across conditions). Hence, for each behavior, participants decided for themselves whether it was in the small or large category.

The instructions then reminded all participants of their categorization and asked them to report the likelihood of

carrying out each behavior in the next three months (1 = Very Unlikely; 7 = Very Likely). All participants also reported if they had done any of the nine behaviors in the last six months and completed a manipulation check.

We had two versions of each condition, to counterbalance the order of the two groups; however, we collapse across these groups in our analyses, for three conditions in total. We suspected that, if simply selecting behaviors increased their importance and likelihood ratings, the choose-2 condition should show the opposite of the original effect: higher likelihood ratings for the behaviors categorized in the *smaller* group. However, choosing all nine would give equal importance to all, hence we predicted a null effect in the two-groups condition.

8.2 Results

Isaac and Brough (2014) did not exclude participants based on their manipulation check, hence neither do we. The choose-7 and choose-2 conditions had validation in place that ensured participants selected the correct number of behaviors before moving on (seven and two, respectively). However no such validation was possible in the two-groups condition, which may be why Isaac and Brough (2014) opted for an alternative design. Most participants (63.8%) created the categories as instructed. Of those that did not follow directions exactly, most (50.8%) put more behaviors in the large category than the small category. We exclude the 17.8% of participants who treated the large category as small, and vice versa, to avoid wrongly influencing our results.

Replicating the original study, when participants were choosing behaviors for the large category (choose-7), they reported higher likelihood of carrying out behaviors from the large category (M = 5.78) than of carrying out behaviors from the small category (M = 4.82), $t(529) = 14.87$, $p < .001$. This result is similar to the original in Isaac and Brough (2014), and was highly significant, suggesting another successful replication.

Moving to our two new conditions, we first looked at the choose-2 condition. Consistent with our hypothesis, the pattern we found previously reversed in the choose-2 condition: participants reported *lower* likelihoods of carrying out behaviors from the large category (M = 5.43) than of carrying out behaviors from the small category (M = 5.84), $t(520) = 7.51$, $p < .001$. Again, the original theory is about the size of the category and not about the order in which the items were chosen, so there should be no difference from the choose-7 condition. Instead the results were in the opposite direction and highly significant. That pattern is highly inconsistent with the category size bias but entirely in keeping with the prediction that participants were selecting the behaviors that they believed were important, creating a category of important behaviors and a category of unimportant behaviors.

Finally, turning our attention to the design that best captured the original plan of Isaac and Brough (2014), in the two-groups condition, we found participants reported similar likelihoods of carrying out behaviors from the large category (M = 5.29) than of carrying out behaviors from the small category (M = 5.18), $t(408) = 1.60$, $p = .110$. (Recall this excludes participants who categorized fewer behaviors in the large category than in the small category. Including them does not change this null result: $M_{large} = 5.27$, $M_{small} = 5.20$, $t(495) = 1.11$, $p = .267$).

Accordingly, the condition X category size interaction was significant, $F(2, 1457) = 7.08$, $p = .001$. This result, while again not consistent with the hypothesis based on the category size bias, is consistent with ours. Overall, the results of Study 6 suggest that, as in the previous studies, a quirk in the original design unrelated to the hypothesized psychological mechanism may have caused the original result.

Thus far we have demonstrated that the category size bias, as put forth by Isaac and Brough (2014), can be attributed to one of two confounds, in instruction or design. When we rectify these confounds, the category size bias reduces or even reverses. However, our critiques address only their Studies 1, 3, and 5. Their Studies 2 and 4 remain untested. Because these paradigms do not as easily accommodate our clarification manipulation from our Studies 1–4 and their statistical evidence was weaker, we first made only replication attempts — should they be successful, we would then consider alternative clarification methods. However, the industrious reader will find that our concerns about the replicability of these two studies were warranted.

9 Study 7

In Study 7, we attempt to replicate the second study from the original paper. This paradigm is similar to that of the original Study 1 (balls in an urn; our Studies 4a and 4b) and Study 3 (alphabet die; our Studies 1–3), but with conveniently less confusing instructions. As a result, although our clarification manipulation used in previous studies cannot easily be applied here, we instead attempt to measure confusion by asking participants. If the original effect replicates, we would expect that excluding any confused participants would eliminate the effect.

9.1 Method

We recruited 206 US participants (53.9% female; M_{age} = 32.8) from Amazon Mechanical Turk. We chose this sample size as it is roughly 2.5 times that of the original study (Simonsohn, 2015). All participants saw a photograph of a glass urn filled with blue and yellow tickets. As in the original, the urn contained 81 blue tickets and 9 yellow and was pre-tested such that the median participant estimate was

approximately accurate. (We again thank the original authors for additional information about the study's procedure, and their suggestions of ways our urn could more closely match the one from their study.) Participants were told that this urn contained all the tickets of those who had already participated in an upcoming lottery.

On the next page, they were shown a ticket of their own, and asked to imagine they had the opportunity to buy this very last ticket to the lottery. Participants were randomly assigned to see either a blue (large category) or yellow (small category) ticket, which served as the category-size manipulation. Participants were then invited to "fill out" their ticket, by providing their worker ID, maximum willingness to pay, and percent likelihood that they would win. Participants were informed that, hypothetically, the winning bid would pay out tenfold; however, they could bid no more than $10.

To test for possible participant confusion, we subsequently asked all participants to identify how they believe the lottery could be won, from four options: (1) If the [blue/yellow] ticket with my Worker ID is drawn. (2) If any [blue/yellow] ticket is drawn. (3) If my [blue/yellow] ticket is NOT drawn. (4) If ANY ticket that is not [blue/yellow] is drawn. (The option order was randomized for each participant.) Finally, participants reported their age, gender, the color of their ticket (blue or yellow), and how many tickets in total they believed were in the urn, as a confirmation of our pretest.

9.2 Results

Contrary to what the category size bias would predict, participants with a yellow ticket were willing to pay statistically similar amounts (M = $6.53, SD = $3.34) to participate in the lottery as participants with a blue ticket (M = $6.48, SD = $3.29). $t(204) = 0.09$, $p = .926$. A similar equivalence emerged in participants' estimated likelihood of winning: $M_{yellow} = 9.73\%$, $SD_{yellow} = 20.34\%$; $M_{blue} = 12.65\%$, $SD_{blue} = 23.97\%$; $t(204) = 0.94$, $p = .351$.

On average, participants were close to the true number of tickets in the urn: the correct answer was 91; the average guess was 92.6 and the modal guess was 100. Participants were also largely able to correctly recall the color of their ticket (97.3%) and how to win the lottery (95.1%). Excluding the few participants who answered either of these checks incorrectly does not meaningfully change the results. Therefore, it does not appear that confusion influenced the (absence of) results, and the original may have indeed been a false positive result.

10 Study 8a

In Studies 8a-c, we attempt to replicate the fourth study from the original paper. This paradigm is similar to that of their Study 6 (sorting behaviors into two categories; our Study 6), in that participants are asked to categorize items. However, to the best of our knowledge, this paradigm does not suffer from the confound detailed in Study 6. As in Study 7 then, given that none of our previous clarification manipulations could be applied, we simply attempted to replicate the original result.

10.1 Method

We recruited 404 US participants (57.4% female, M_{age} = 34.4) from Amazon Mechanical Turk. We chose this sample as it is about 4.5 times the original sample size, to accommodate the additional power needed for a replication and testing an interaction effect. No participants were excluded. Participants were shown eight mascot logos for teams competing in the NCAA men's basketball tournament. Six of the logos showed animals; two showed humans. Six of the logos showed only the mascot's head; two showed the whole body. Participants were randomly assigned to categorize the eight mascots into two groups, based on one of the two aforementioned dimensions (i.e., animal vs. human, head vs. body). This way, the same team could be categorized differently between the conditions.

On the next page, participants were asked the odds against a particular team winning the tournament, and were provided a definition of odds against. Some participants estimated the odds against the Florida State Seminoles (a human head); others estimated the odds against the Wisconsin Badgers (a full-bodied animal). Participants next estimated the odds against the Ohio State Buckeyes (no mascot logo provided, to act as a control). Per our communication with the original authors, participants' estimates were capped at 100. Finally, participants provided their age and gender.

10.2 Results

We conducted a 2 (Team: FSU vs. Wisconsin) x 2 (Categorization: Face/Body vs. Animal/Human) ANOVA on participants odds against estimates. We found no effect of team ($F(1, 400) = 0.23$, $p = .636$) or categorization ($F(1,400) = 0.80$, $p = .372$). However, a significant, albeit small, interaction effect did emerge: $F(1, 400) = 4.10$, $p = .044$. Unpacking this interaction, we found that, consistent with the category size bias, participants categorizing the Wisconsin mascot into the larger category estimated lower odds against (M = 32.54 : 1, SD = 28.41) than did participants categorizing it into the smaller category (M = 40.92, SD = 29.13), $t(198) = 2.05$, $p = .042$. However, this effect did not emerge for participants categorizing the Florida State mascot (M_{large} = 36.47 : 1, $SD_{large} = 27.02$; $M_{small} = 39.71$, $SD_{small} = 30.34$), $t(202) = 0.81$, $p = .422$.

11 Study 8b

Because of the ambiguous results from Study 8a, we decided to run this replication a second time, with a greater sample size.

11.1 Method

We recruited participants as part of larger project that utilized the same private, online survey-panel company as Study 4b. Again, participants first completed another online survey and were then redirected to the present study upon completion. The first survey determined the sample size we could obtain. That survey had a pre-registered sample size of 1500 complete, attentive responses. However, to make up for attention-check failures the sample size is larger than 1,500. In the end, Study 8b had 1,819 participants (53.4% female; M_{age} = 45.7). No participants were excluded

The procedure for Study 8b was identical to that of Study 8a, with one exception: Some participants were asked to make their estimate in the form of odds against, as in the original study; others, though, were asked to estimate it in the form of a percentage likelihood. The original authors used odds against as a clever means of avoiding the large-answer/large-category association from previous studies, thereby ruling out alternative explanations like anchoring. However, this response format should not be required for the original effect to obtain, given that the rest of their studies retain this large-answer-with-large-category association. We include this response format factor purely because we were concerned that the unintuitive format of odds against may have muddled the original effect and played a role in the inconclusive results of Study 8b. Participants should be much more familiar with estimating likelihoods in percentages, eliminating this concern.

11.2 Results

We conducted two 2 (Team: FSU vs. Wisconsin) x 2 (Categorization: Face/Body vs. Animal/Human) ANOVAs, on participants' odds against and percentage likelihood estimates. However, in both analyses, we found no significant effects (F's < 1.18). Looking first to the original odds against measure, participants provided similar responses when they categorized the target mascot in a large category ($M_{Wisconsin}$ = 32.32 : 1, $SD_{Wisconsin}$ = 29.32; M_{FSU} = 32.16 : 1, SD_{FSU} = 29.91) as when they categorized the target mascot in a small category ($M_{Wisconsin}$ = 33.45 : 1, $SD_{Wisconsin}$ = 30.32; M_{FSU} = 35.25 : 1, SD_{FSU} = 30.33), $t_{Wisconsin}(471) = 1.12$, $p = .264$; $t_{FSU}(467) = 0.41$, $p = .680$.

Looking next to our percentage likelihood measure, participants provided similar responses when they categorized the target mascot in a large category ($M_{Wisconsin}$ = 30.96%, $SD_{Wisconsin}$ = 23.21%; M_{FSU} = 31.96%, SD_{FSU} = 25.92%) as when they categorized the target mascot in a small category ($M_{Wisconsin}$ = 33.20%, $SD_{Wisconsin}$ = 24.58%; M_{FSU} = 33.19%, SD_{FSU} = 25.84%), $t_{Wisconsin}(411) = 0.95$, $p = .343$; $t_{FSU}(436) = 0.50$, $p = .616$.

12 Study 8c

Normally, unambiguously null results as we found in Study 8b would suggest that the original finding and the partially, barely significant replication were false positives. However, the uniformity between measures that should lead to diverging means (i.e., percent likelihood, where larger numbers signal higher likelihood, and odds against, where lower numbers signal higher likelihood), presented a cause for concern. Did these participants misunderstand odds against? To ensure that there were no sample-specific issues with replicating the original effect, we ran a direct replication of Study 8b, with a sample from Amazon Mechanical Turk.

12.1 Method

We recruited 808 US participants (53.3% female, M_{age} = 35.2) from Amazon Mechanical Turk. We chose this sample size as it is approximately double that of Study 8a, of which half of this study is a direct replication. The procedure was identical to that of Study 8b — replication of the original effect, varying response format — with one addition: After providing their estimates, participants were asked to identify the sport discussed in the survey (basketball) from four options, as an attention check.

12.2 Results

Thirty-seven (4.6%) of participants failed the attention check. Per our pre-registration, we exclude these participants for our main analyses (although their inclusion does not meaningfully change the results). As in Study 8b, we conducted two 2 (Team: FSU vs. Wisconsin) x 2 (Categorization: Face/Body vs. Animal/Human) ANOVAs, on participants' odds against and percentage likelihood estimates. Also as in Study 8b, we found no significant effects whatsoever (F's < 1.75). Looking first to the original odds against measure, participants provided similar responses when they categorized the target mascot in a large category ($M_{Wisconsin}$ = 37.79 : 1, $SD_{Wisconsin}$ = 31.23; M_{FSU} = 42.99 : 1, SD_{FSU} = 32.40) as when they categorized the target mascot in a small category ($M_{Wisconsin}$ = 41.28 : 1, $SD_{Wisconsin}$ = 32.26; M_{FSU} = 38.04 : 1, SD_{FSU} = 29.68), $t_{Wisconsin}(190) = 0.76$, $p = .450$; $t_{FSU}(193) = 1.11$, $p = .267$.

Looking next to our percentage likelihood measure, participants provided similar responses when they categorized the target mascot in a large category ($M_{Wisconsin}$ = 23.57%%, $SD_{Wisconsin}$ = 19.89%; M_{FSU} = 27.00%, SD_{FSU} = 22.45%)

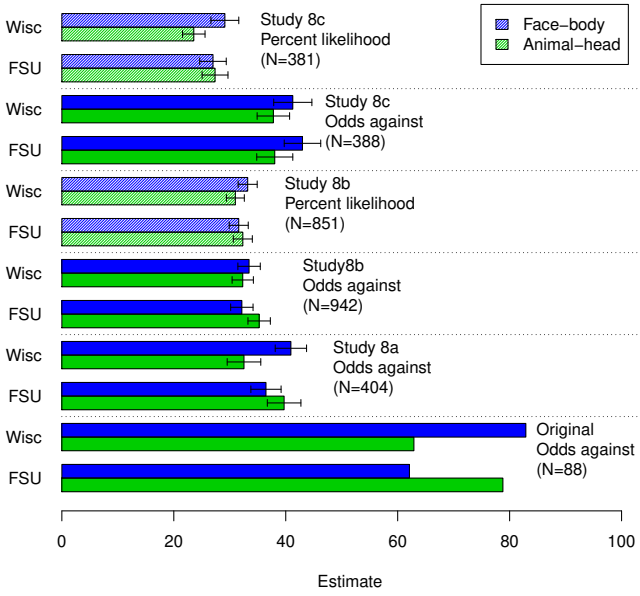

FIGURE 12: Means from the original paper's Study 4 (bottom) and our three replication attempts (Studies 8a–8c). Dotted pairs of bars represent the alternative response format, percent likelihood, which should show the opposite pattern of the original. All effects from all three replications were non-significant, with the exception of one simple effect in the predicted direction (Study 8a, Wisconsin) and one marginal simple effect in the opposite direction (Study 8c, percent likelihood, Wisconsin).

as when they categorized the target mascot in a small category ($M_{Wisconsin}$ = 29.12%, $SD_{Wisconsin}$ = 23.62%; M_{FSU} = 27.37%, SD_{FSU} = 23.55%), $t_{Wisconsin}(185) = 1.74$, $p = .083$; $t_{FSU}(192) = 0.11$, $p = .977$. (Note that the marginally significant simple effect for Wisconsin is in the opposite direction the bias would predict.) Although we found no evidence of the category size bias, we did find overall higher estimates when participants provided percent likelihoods (M = 39.94, SD = 31.31) than when they provided odds against (M = 26.73, SD = 22.43), $F(1, 768) = 45.11$, $p < .001$, relieving our concerns that confusion about the response formats could be masking the original effect. See Figure 12 for a visual summary of our three failed replications.

13 General discussion

We believe that the data we present suggest an alternative understanding of the category size bias in probability judgments as detailed in Isaac and Brough (2014). Their key results, while robust, appear to be largely attributable to confusion or artifacts of research design. The data suggest that a systematic misunderstanding of the question may have led participants mistakenly estimate the size of the category of events, rather than the likelihood of one of its constituent events. Our first five studies showed that, for two of the experiments in Isaac and Brough (2014) instructional confusion offers a parsimonious account of the results. Study 6 showed that, in the study designed by the original authors to avert confusion, a different procedural artifact underlies the reported effect. Moreover, altering that artifact led to a complete reversal of the original effect, despite remaining a context in which the category size bias should have emerged. In our final set of studies, we examine findings that may not have been driven by confusion, but also do not offer reliable support for the original effect. See Table 1 for a summary of our results.

It is important to acknowledge the limitations of our evidence, however. First, though we replicated and moderated three of the five studies from the original, we were not able to replicate — and therefore not able to moderate — the two remaining studies. Additionally, it is worth noting that across Studies 2–4 we frequently found that there was still some non-significant whisper of an effect in our "debiased" conditions (with one exception, a significant effect in Study 4b). We have largely interpreted that result as a function of the modesty of our clarification manipulation — had we pushed harder to ensure that people correctly understood the instructions then the effect would be reduced completely, supported by the Study 5 finding that it is participants who report being confused that are behind the bulk of the effect. Additionally, the subtlety of our clarification measure still requires some level of attention in participants; those who rush through the survey are both likely to make an error and not realize they have done so. Nevertheless, an alternative could be that there is a hint of the category size bias which still creeps into estimates, even after the instructions have been clarified. We still believe this to be unlikely, though, given the failed replications of two of the original studies, and the significant reversal of the effect in a context ripe for the bias to emerge (our Study 6, *two-groups* condition).

It bears mentioning that the original authors did not propose or test any mechanism that could lead participants to adopt this belief of inherited statistical traits. This is not a criticism of Isaac and Brough (2014) — all three of us prefer writing and reading more effects-driven papers, too — however, it does muddle the present discussion of our alternative mechanism. Were the effect more reliable without the aforementioned confusion and confounds, future research on its mechanism would be beneficial.

An additional defense of Isaac and Brough (2014) could focus on ecological validity. It is entirely plausible that the world presents people with ambiguous decision problems for which it is easy to confuse the target with the broader category from which it comes. For example, people might judge the likelihood of winning a specific Bingo tournament to be more likely when they think of the tournament as coming from a large category (all Bingo tournaments) than when coming from a small category (Bingo tournaments played

Table 1: An overview of the original authors' and our own sets of studies.

Isaac and Brough's (2014) studies	Our interpretation	Our studies
Study 1 (numbered balls in an urn)	Participant confusion	Study 4b
Study 2 (colored tickets in an urn)	False positive	Study 7
Study 3 (alphabet sdie)	Participant confusion	Studies 1, 2, 3, 5
Study 4 (categorizing mascots)	False positive	Studies 8a, 8b, 8c
Study 5 (categorizing IT behaviors)	Mechanical confound	Study 6

this weekend). Consumers confusing target and category will make predictable errors in this kind of situation.

The broader significance of our results is twofold: theoretical and practical. First, this paper underscores the importance of understanding the perspective and mind-set of individuals participating in our studies. Some of what appears to be bias can result from sensible judgments, given the imperfect information understood by research participants (Dawes & Mulford, 1996; Juslin, Winman & Olsson, 2000). We would never claim that all biases are simply the result of confusing instructions, but it is undeniably the case that small changes in the wording of experimental instructions can sometimes have profound effects on how participants think about the problem (Cheng & Holyoak, 1985; Kotovsky & Simon, 1990). Understanding these subtleties is essential to understanding when a particular effect will occur and when it will evaporate.

Second, this paper highlights the difficulty of correcting, or even beginning a dialogue about possibly correcting, the scientific record. The original version of this manuscript was submitted, reviewed, revised, reviewed, and ultimately rejected, by a different journal, one that is more relevant to the original paper. (However, the editors of said journal requested that it remain unnamed.) All readers and decision makers were thorough and detailed, and were attuned to many shortcomings in this paper, many of which are undoubtedly still present and detected by other readers. On the other hand, the journal also makes clear that replications or corrections are held to a higher standard than original work.

For example, in rejecting this manuscript, the editor noted the ways in which our results do replicate some of the original Isaac and Brough (2014) results, thanks to participants' confusion. The editor wrote to us that, "It seems likely that the category size bias is multiply determined and there are no doubt moderators that make it more or less likely to emerge. Your chosen path with this revision however did not attempt to uncover moderators of the effect, so much as try to negate the effect altogether." If an editor believes in the truth of an effect, then uncovering moderators is a logical and sensible request. To researchers convinced that the original effect is an artifact or a false-positive, such a request sounds not so different from asking a revision to successfully forecast trends in coin flip outcomes. Unfortunately, as the editor informed us in response to a follow-up message, "Studies that fail to replicate prior results without a focused examination of the (theoretical) conditions under which the results hold/do not hold have historically not tended to do well at [this journal]" (parentheses theirs).

Finally, we would like to conclude with a note on transparency and open science. Only through Isaac and Brough posting their materials in the paper's appendix were we able to start thinking about this problem, and only through generous openness with their data and procedures in correspondence could we properly investigate it. Although transparency is generally discussed in the context of false positives, it also facilitates the investigation of these more traditional meta-scientific questions, which are equally important. We commend Isaac and Brough for their data practices and hope more researchers follow suit or exceed them in the near future. In fact, were their original data made available to reviewers, this error might have been spotted in the review process and saved us from recruiting these 10,090 participants to learn more about it. It is our sincere hope that these practices will facilitate a more cumulative science in which scholars follow up on each other's published work, testing their theories and applying them in novel ways.

References

Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology, 55*(5), 726-737.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*(4), 391–416.

Dawes, R. M., & Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? Organizational Behavior and Human *Decision Processes, 65*(3), 201–211.

Gal, D., & Rucker, D. D. (2011). Answering the unasked question: Response substitution in consumer surveys. *Journal of Marketing Research*, *48*(1), 185–195.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. J. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts*. New York: Academic Press.

Isaac, M. S., & Brough, A. R. (2014). Judging a part by the size of its whole: The category size bias in probability judgments. *Journal of Consumer Research, 41*(2), 310–325.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107*(2), 384–396.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*, pp. 49–81. New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454.

Kotovsky, K., & Simon, H. A. (1990). What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology, 22*(2), 143–183.

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). *Straight choices: The psychology of decision making*. New York: Routledge.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Available at SSRN 2160588.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547-567.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology, 75*(6), 1411–1423.