Inferring the size of an unobservable population from observable samples

Jack Cao[1*] & Mahzarin R. Banaji[1]

[1]Harvard University, Department of Psychology

*Corresponding author: jackcao@fas.harvard.edu

Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship (DGE 1144152) to J.C.

Open Practices Statement

All data and code are available at [osf.io/7ady5]. All materials are included in Supplemental Materials section at the end of this manuscript.

Abstract

Success in the physical and social worlds often requires knowledge of population size. However, many populations cannot be observed in their entirety, making direct assessment of their size difficult, if not impossible. Nevertheless, an unobservable population size can be inferred from observable samples. We measured people's ability to make such inferences and their confidence in these inferences. Contrary to past work suggesting insensitivity to sample size and failures in statistical reasoning, inferences of populations size were accurate – but only when observable samples indicated a large underlying population. When observable samples indicated a small underlying population, inferences were systematically biased. This error, which cannot be attributed to a heuristics account, was compounded by a metacognitive failure: confidence was highest when accuracy was at its worst. This dissociation between accuracy and confidence was confirmed by a manipulation that shifted the magnitude and variability of people's inferences without impacting their confidence. Together, these results a) highlight the mental acuity and limits of a fundamental human judgment and b) demonstrate an inverse relationship between cognition and metacognition.

Keywords

numerical cognition, population estimates, accuracy, confidence, sampling processes

Introduction

The number of objects in a set has implications for a wide range of human endeavors. The number of goods in a windfall affects whether they can be allocated equitably or efficiently (Blake & McAuliffe, 2011). The number of people in a group predicts judgments of warmth and competence (Cao & Banaji, 2016) and decisions about whether to engage in physical conflict (Pietraszewski & Shaw, 2015). Given the influence of discrete quantity, it is important to assess people's cognitive ability to estimate set size and people's metacognitive ability to know their own limits and flexibilities.

While past work has examined problems where the entire set is visible (LeCorre & Carey, 2007; Libertus, Feigenson, & Halberda, 2011), we focus on problems where it is difficult or impossible to view the entire set. This latter type of problem is ecologically common, for the simple reason that one's visual field is limited but the physical world is vast and includes many ways of rendering sets of objects unobservable: they can be hidden, dispersed, or occluded. Consider, for example, how many people live on your block, how many taxis operate in your city, or how many bicycles are on campus. Each of us has intuitions about these set sizes (i.e., populations) even though only subsets (i.e., samples) have been encountered. How accurate are these intuitions? And to what extent does confidence track these intuitions?

Answering the first question about accuracy requires a normative model against which human judgments can be compared. Johannes Petersen, a 19th century marine biologist, laid the foundations of this model when estimating the number of fish in a fjord. Petersen (1896) accomplished this by taking a random sample of fish at time 1, marking them (e.g., by tagging their fins), and releasing them back into the fjord. At time 2, he took another random sample and counted the number of fish that were resampled.

The intuition behind this method is that the number of resampled fish – the overlap between samples 1 and 2 – is indicative of the total number of fish in the fjord. If the overlap is small, there are likely many fish in the fjord. But if the overlap is large, there are likely few fish in the fjord. The idiom "it's a small world" is commonly expressed when an individual is encountered again; the "small world" references the small population size that explains the reencountering of the same individual.

This intuition is formalized in Bayes' rule, allowing precise inferences of population size N to be made based on the sizes of the two random samples, s_1 and s_2, and the overlap, o, between them (see Supplemental Materials for technical details):

$$P(N \mid s_1, s_2, o) \propto P(\, o \mid N, s_1, s_2) \times P(N)$$

Ecologists have successfully applied this model to studying, for example, the sizes of animal populations (Seber, 1982). In addition to usage by experts, a version of this model may also be used by laypeople who are in situations where the size of a population can be inferred based on the overlap between two observable samples. After all, many populations cannot be directly observed, but samples are frequently observed. Furthermore, recall and recognition memory enables the overlap between samples to be noticed (Tulving, 1999).

But unlike experts, laypeople may be unable to make accurate inferences. Computing $P(N \mid s_1, s_2, o)$ requires sensitivity to sample size, which people appear to lack, as they inadequately weigh sample size when it is presented alongside information such as mean, variance, and qualitative text. This insensitivity has been demonstrated among college students in lab studies (Obrecht, Chapman, & Gelman, 2007), prospective jurors making hypothetical

decisions (Ubel, Jepson, & Baron, 2001), and consumers reading online product reviews (De Langhe, Fernbach, & Lichtenstein, 2016). Given this insensitivity to sample size, it would seem that people would fall short in a task that requires them to make a population size inference based on random samples.

Furthermore, sample size insensitivity is among the many cognitive errors documented by Tversky & Kahneman (1974). Laypeople mistakenly believe that extreme heights are equally likely to be observed in a sample of 1,000 individuals as they are in samples of 100 or even 10 individuals. Similarly, sample size is ignored when people judge that smaller and larger hospitals are equally likely to record an extreme gender imbalance among newborn babies. Both of these cases demonstrate that people do not consider the statistical fact that extreme outcomes are more likely in smaller samples. This failure is reason to suspect that people's ability to use samples to infer a population size is compromised.

In the aforementioned cases, people fall prey to the representativeness heuristic, which undercuts the computation of simpler conditional probabilities like $P(cancer \mid positive\ mammogram)$ (Kahneman & Tversky, 1972). Here, human error has been observed where there are just two hypotheses and one piece of data (i.e., whether someone does or does not have breast cancer given a positive mammogram). By contrast, $P(N \mid s_1, s_2, o)$ is more complex: it involves a theoretically unbounded number of hypotheses and three pieces of data (the size of the first sample, the size of the second sample, the overlap between the two samples). Given these complexities and the comparatively straightforward nature of tasks where people have been shown to fail, it seems unlikely that people can accurately infer the size of an unobservable population from observable samples.

In addition to gauging the accuracy of people's inferences, we also measure people's confidence in their inferences to assess metacognition. Deficiencies in metacognition typically manifest as overconfidence. Physicians are confident in diagnoses that turn out to be incorrect (Christensen-Szalanski & Bushyhead, 1981). Students are confident in exam score predictions that are too high compared to the scores they actually receive (Clayson, 2005). And, as many readers can attest, people are confident that they will finish their work sooner than they actually do (Buehler, Griffin, & Ross, 1994).

Unlike past research where a single, verifiable truth (e.g., the actual diagnosis, exam score, or completion date) was compared to confidence ratings, the current experiments rely on group-level distributions of population estimates because no single estimate is correct per se. Rather, a distribution of estimates represents accuracy (see Fig. 1 for further details). In the forthcoming experiments, the overlap between random samples is parametrically manipulated, resulting in variability in accuracy, as measured by the fit, or lack thereof, between theoretically expected distributions and observed distributions produced by participants. Insofar as confidence is highest when accuracy is likewise highest and lowest when accuracy is lowest, metacognition would be well calibrated. However, if confidence is highest where accuracy is lowest, then people would be overconfident in their ability to infer the size of an unobservable population from observable samples.

Experiment 1

Participants. Data were collected in two independent rounds on Amazon Mechanical Turk. Four hundred twenty four participants were recruited in the first round to demonstrate the effects. Twelve hundred sixty two participants were recruited in the second round to establish replicability and generalizability. Given how similar the results are, data from both rounds are presented together. Across both rounds of data collection, 78 participants did not finish the procedure; 81 participants were excluded for providing population size estimates that were less than the logical minimum (see footnote 1). The final sample consisted of 1,527 participants (M_{age} = 34.73, SD_{age} = 11.21; 819 females, 701 males, 7 unspecified).

Procedure. In the first round of data collection, participants estimated the number of marbles in an urn by using information limited to the sizes two random samples and the overlap between them. The first sample, s_1, was always 10 marbles. The second sample, s_2, was always 5 marbles. The overlap, o, was manipulated between-subjects to be 0 out of 5 marbles (denoted as 0/5) or 4 out of 5 marbles (denoted as 4/5).

Since computing $P(N \mid s_1, s_2, o)$ requires a prior over population size, N, the maximum number of marbles needed to be specified. This value, N_{max}, was manipulated between-subjects to be 50 or 100 marbles. A uniform prior over N was established by informing participants that they could not hear any of the marbles move as samples were taken (see Table 1 for stimuli).

After providing their population estimates, participants expressed how confident they were in their estimates (*1 = Not at all confident … 5 = Extremely confident*). Lastly, participants provided self-perceptions of numeracy by indicating their level of agreement with four

statements (e.g., I feel confident in by ability to solve statistical problems; *1 = Strongly disagree … 5 = Strongly agree*; see Supplemental Materials for all stimuli).

The second round of data collection was the same as the first round except for two differences. First, the type of object was manipulated between-subjects to be marbles (direct replication), spoons (conceptual replication), or bottle caps (conceptual replication). Second, participants did not provide self-perceptions of numeracy.

The values for N_{max} (50 and 100) and overlap (0/5 and 4/5) were adapted from Lee & Wagenmakers (2013, p. 75-76) because these values lead to different theoretical distributions when the overlap is 0/5, but similar distributions when the overlap is 4/5 (Fig. 1). When the overlap is 0/5, the total number of objects can range from 15 ($s_1 + s_2 - o = 10 + 5 - 0 = 15$) to N_{max}.[1] A uniform prior over this range is updated to favor larger population sizes. Thus, the largest possible population size, N_{max}, matters a great deal.

However, when the overlap is 4/5, the total number of objects can range from 11 ($s_1 + s_2 - o = 10 + 5 - 4 = 11$) to N_{max}. A uniform prior over this range is updated to favor smaller population sizes. Because the smallest possible population size is the same irrespective of whether N_{max} is 50 or 100, N_{max} hardly matters at all.

[1] Fifteen ($s_1 + s_2 - o = 10 + 5 - 0 = 15$) is the logical minimum number of objects in the population when s_1 is 10, s_2 is 5 and o is 0/5. Thus, any population size estimates that fall below this value are invalid.

Fig. 1. Uniform prior distributions and theoretical posterior distributions. Theoretical posterior distributions result from 150,000 Markov chain Monte Carlo (MCMC) samples per cell. When the overlap is 0/5, there is a substantial effect of N_{max}, resulting in different posterior distributions (left column). But when the overlap is 4/5, the effect of N_{max} is minimal, resulting in similar posterior distributions (right column).

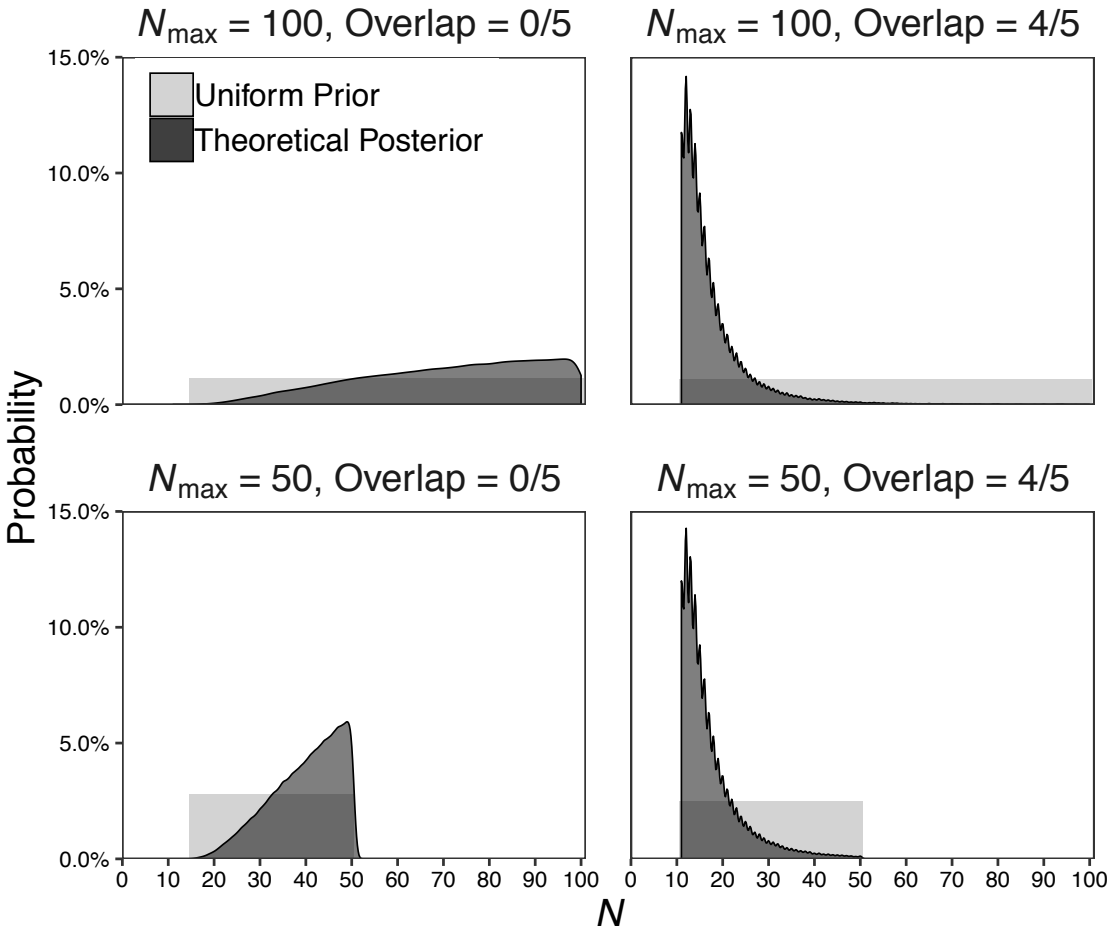


Table 1. Experiment 1, first and second rounds of data collection, marbles condition. Stimuli presented to participants. Highlighted in grey are the between-subjects manipulations. In the second round of data collection, the object type was changed to "bottle caps in a box" or "spoons in a box."

As you read the scenario below, imagine yourself playing the game that's described.

Imagine you're at a state fair where there are many games to play. One game in particular catches your eye. It's called, "*Guess the number of marbles*." You approach the person in charge of the game and ask him how the game works. He shows you an urn and tells you the following information, all of which is true:

- Inside the urn, there are an unknown number of marbles. Nothing else is inside the urn.
- All the marbles are identical and they're all white in color. There are no markings on any of the marbles.
- At most, there are 50 [100] marbles inside the urn.

You can't see through the urn, so aside from picking a random number between 1 and 50 [100], there's no way for you to guess how many marbles there are. You raise this objection, so the person in charge of the game offers you some help. But first, he asks you to put on pair of noise-cancelling headphones so that you can't hear the marbles move inside the urn, which could give you an idea of how many marbles there are. Intrigued you put on the headphones. You watch as the person thoroughly mixes up all the marbles inside the urn. He then randomly pulls out 10 marbles.

The person takes a red permanent marker and draws a large dot on every one of the 10 marbles he pulled out. After the red ink on each marble is completely dry, the person puts the 10 marbles back into the urn. Next, the person thoroughly mixes up all the marbles once again and randomly pulls out more marbles. This time, he pulls out 5 marbles. He shows you these 5 marbles, and you see that none [4] of these 5 marbles have large red dots on them.

At this point, the person asks you to guess how many marbles there are inside the urn.

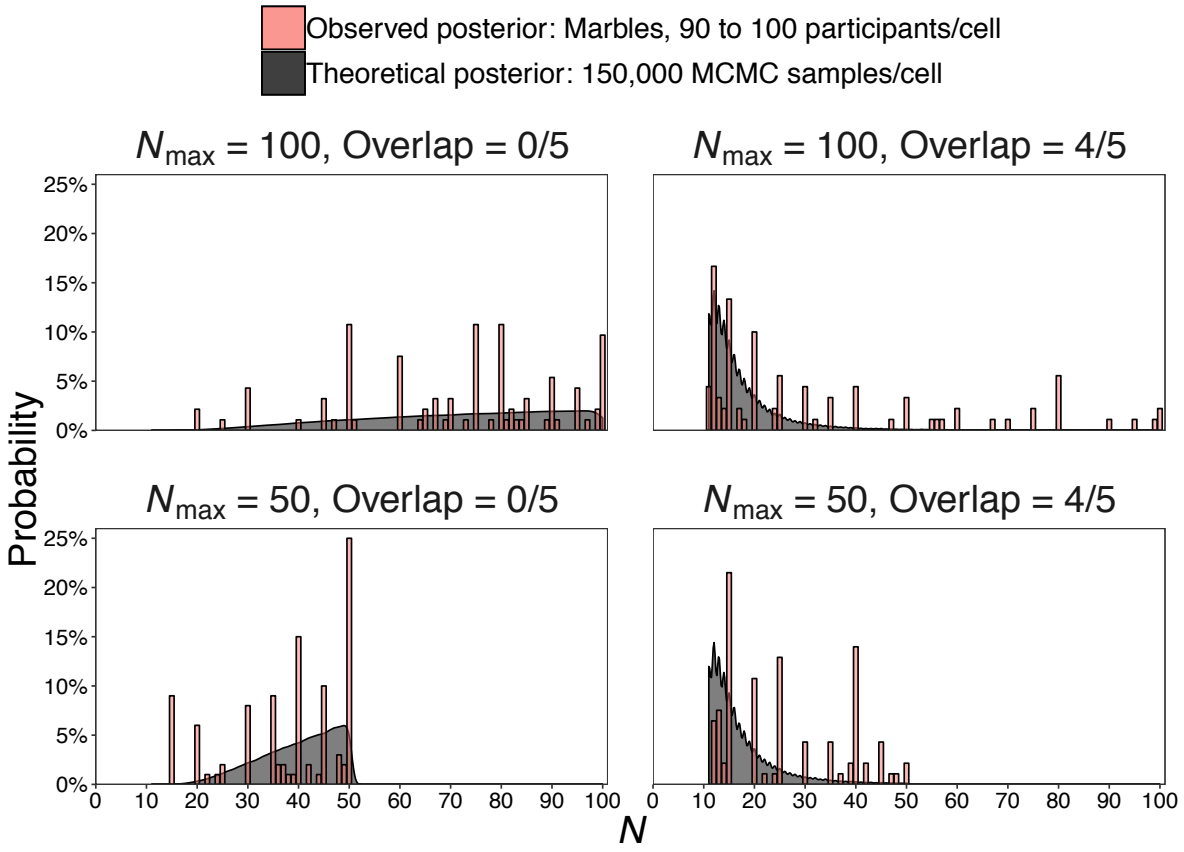
How many marbles do you think are inside the urn? Please type in a number below.

[Participant types in estimate here]

We're interested in your intuitions. So don't make any complicated calculations or think too hard. Just put down when you think!

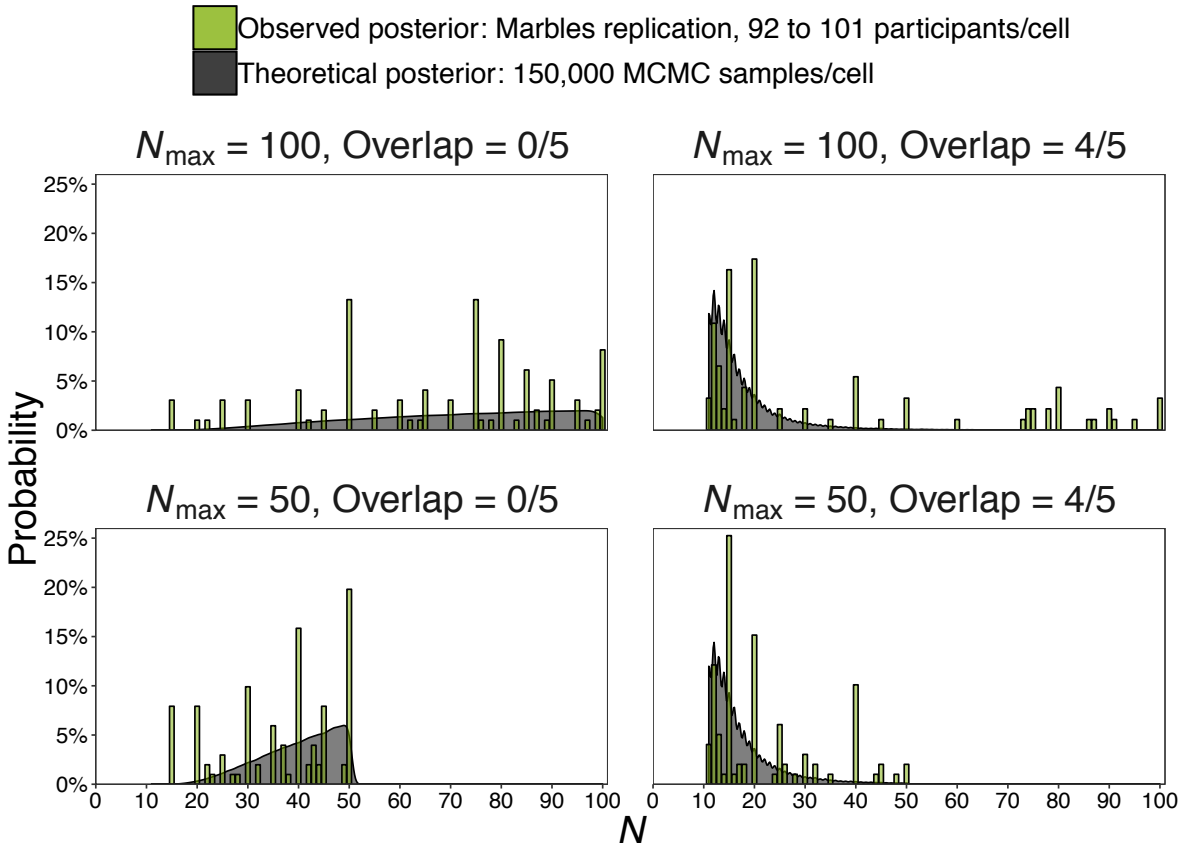
Results. To assess the accuracy of people's inferences, we compared theoretical and observed distributions. Across all conditions, including replications, there was a qualitative match between the two distributions (see probability density functions, PDFs, in Figs. S1-S4).[2] When the overlap was 0/5, estimates were left-skewed, as participants tended to provide higher estimates closer to N_{max}. When the overlap was 4/5, estimates were right-skewed, as participants tended to provide lower estimates irrespective of N_{max}.

Fig. S1. Experiment 1, first round of data collection, marbles condition. Observed vs. theoretical distributions.



[2] Supplemental figures, denoted with an *S* preceding the number, are included in the main text for convenience.

212 **Fig. S2.** Experiment 1, second round of data collection, replication of marbles condition.
213 Observed vs. theoretical distributions.



214

215

216

217

218

219

220

221

222

223

Fig. S3. Experiment 1, second round of data collection, spoons condition. Observed vs. theoretical distributions.

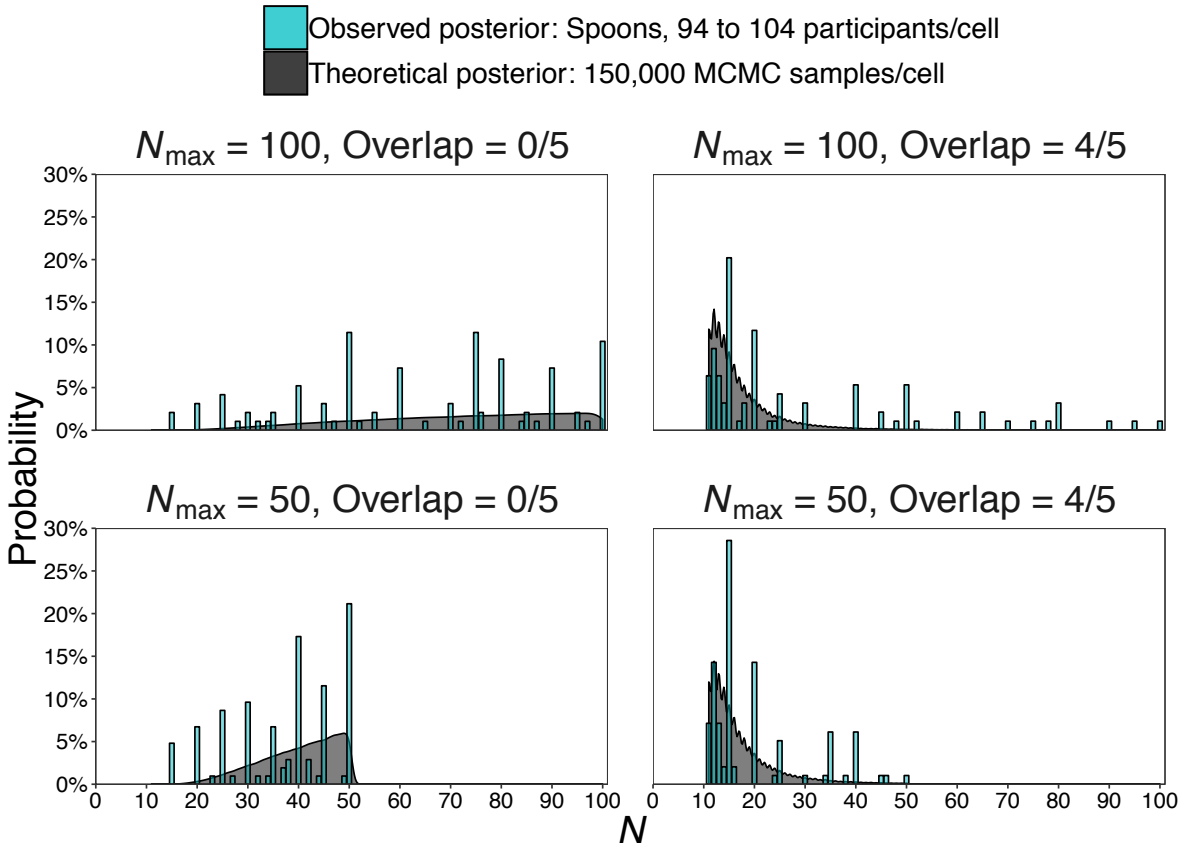
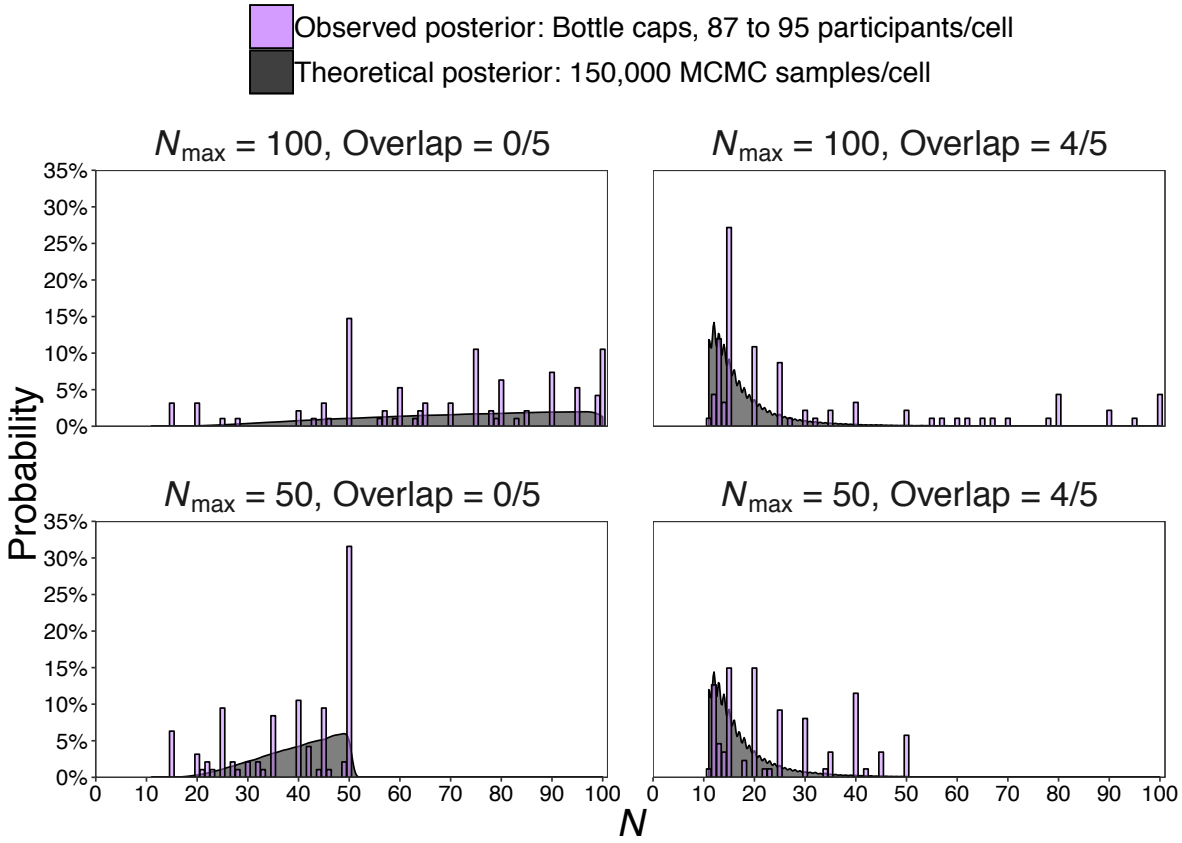


Fig. S4. Experiment 1, second round of data collection, bottle caps condition. Observed vs. theoretical distributions.



There are two challenges to quantifying the difference between theoretical vs. observed distributions. The first is the disparity in samples sizes: 150,000 Markov chain Monte Carlo (MCMC) samples for each theoretical distribution vs. an average of 95 participants in each observed distribution. The second is that participants tended to provide round numbers as estimates.

To overcome these challenges, the following steps were taken. First, each observed PDF was converted to a cumulative density function (CDF). Then, samples of size 95 – the average number of participants in each condition – were randomly drawn from the theoretical

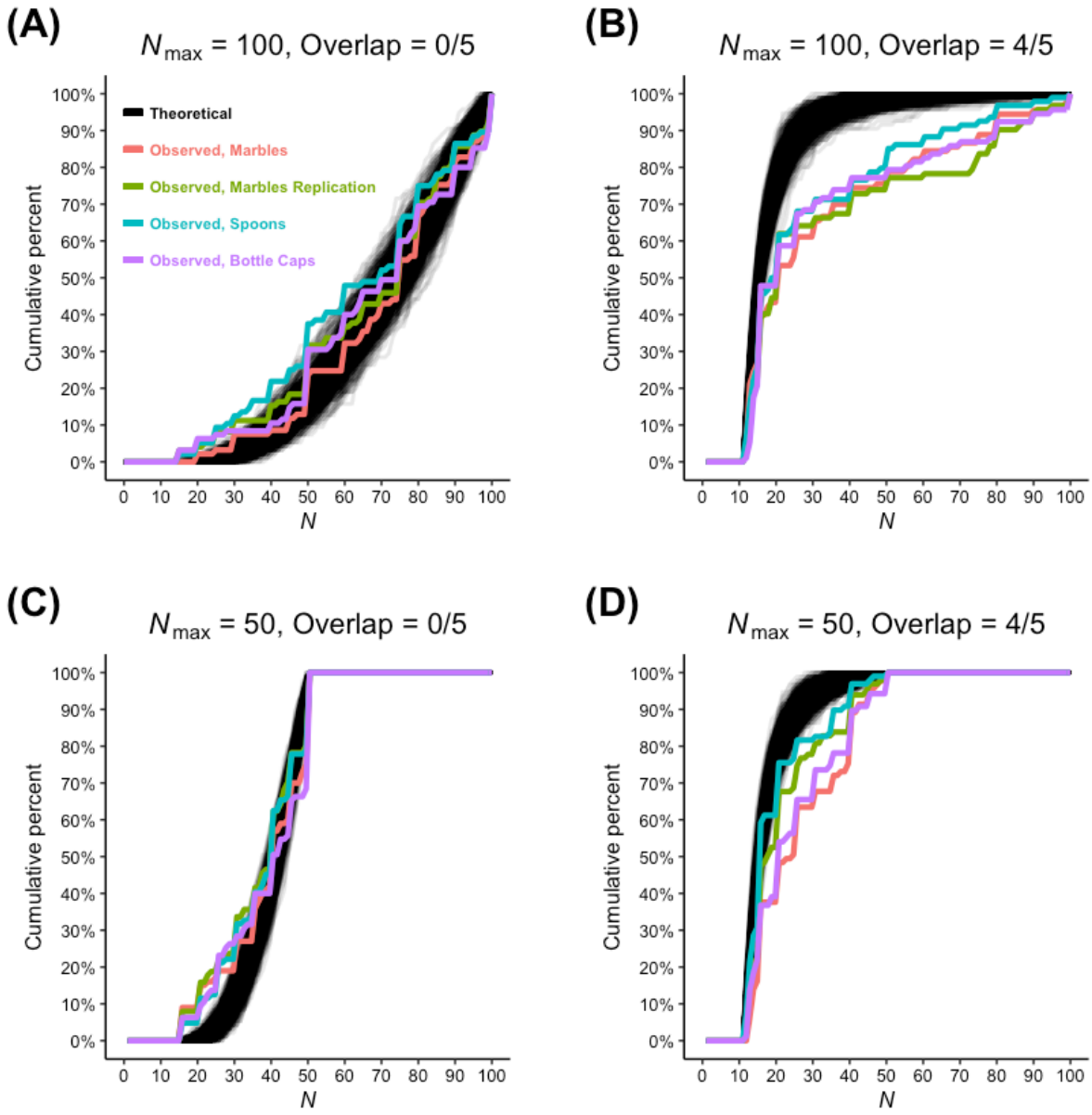
distributions. One thousand of these distributions were drawn in each condition and converted to CDFs to form de facto null hypotheses. Insofar as observed CDFs fall within the bootstrapped theoretical CDFs, people's estimates would be accurate. Calculating the absolute difference in area under the curve (AUC) between the observed CDF and each theoretical CDF enables precise quantification. The average of these absolute differences, *M*ΔAUC, indexes the degree to which each observed distribution differs from the theoretical distribution, with zero indicating no difference and higher values indicating greater deviation.

When the overlap between samples was 0/5, indicating a large population, inferences were quite accurate. Observed CDFs resembled the corresponding theoretical CDFs, as shown by small average AUC differences, both when N_{max} was 100 [Fig. 2A; *M*ΔAUC ranged from 1.62 to 7.12] and when N_{max} was 50 [Fig. 2C; *M*ΔAUC ranged from 1.59 to 3.49].

However, when the overlap between samples was 4/5, indicating a small population, participant's inferences erred in the direction of overestimation. Observed CDFs deviated from corresponding theoretical CDFs, as shown by high average AUC differences, both when N_{max} was 100 [Fig. 2B; *M*ΔAUC ranged from 11.42 to 16.76] and when N_{max} was 50 [Fig. 2D; *M*ΔAUC ranged from 2.71 to 8.64]. These deviations indicate that a higher than expected proportion of participants gave high population size estimates, a result that is also visually apparent in the thicker right tails of the corresponding PDFs in Figs. S1-S4.

271 **Fig. 2.** Experiment 1. Theoretical vs. observed cumulative density functions in each condition.
272

273



274

275

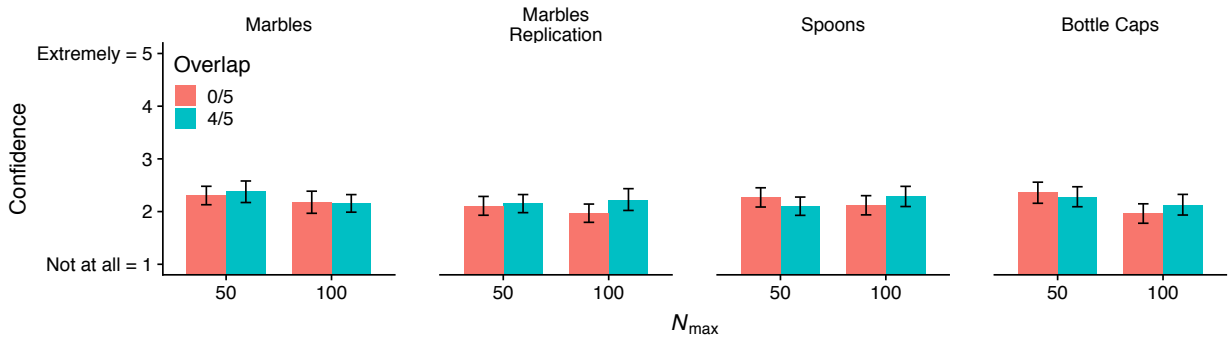
276

277

278

After inferring the population size, participants rated how confident they were in their inferences. If confidence ratings tracked accuracy, then participants would have been more confident when the overlap was 0/5 than when the overlap was 4/5. However, no main effect of overlap was observed [$F(1, 1511) = 1.24$, $P = 0.27$, $\eta_p^2 = 0.0008$]. Despite differences in accuracy that depended on the overlap between observable samples, participants expressed similar confidence ratings across all conditions (Fig. S5). In Experiment 2, the overlap between the two samples was parametrically manipulated to assess accuracy and confidence across the full range of possible overlaps between samples (i.e., from 0/5 through 5/5).

Fig. S5. Experiment 1. Average confidence ratings in population size estimates. Error bars are 95% confidence intervals.



Experiment 2

Having shown that inferences were more accurate when the overlap indicated a large population and less accurate when the overlap indicated a small population, we sought to test the full range of overlap values. On the one hand, inferences might only be accurate when there is no overlap between samples and erroneous otherwise. On the other hand, inferences may be accurate across small and moderate overlap values and err when the overlap is high. Testing the complete range of overlap conditions would provide a fuller picture of people's cognitive and metacognitive abilities in this domain.

Participants. Five hundred forty seven participants were recruited from Amazon Mechanical Turk. Two hundred ninety one participants were excluded for failing attention checks (see Supplemental Materials for these checks). [3] Another seventy five participants were excluded for providing one or more population size estimates below the logical minimum. The final sample consisted of 181 participants (M_{age} = 39.26, SD_{age} = 11.49; 95 females, 85 males, 1 unspecified).

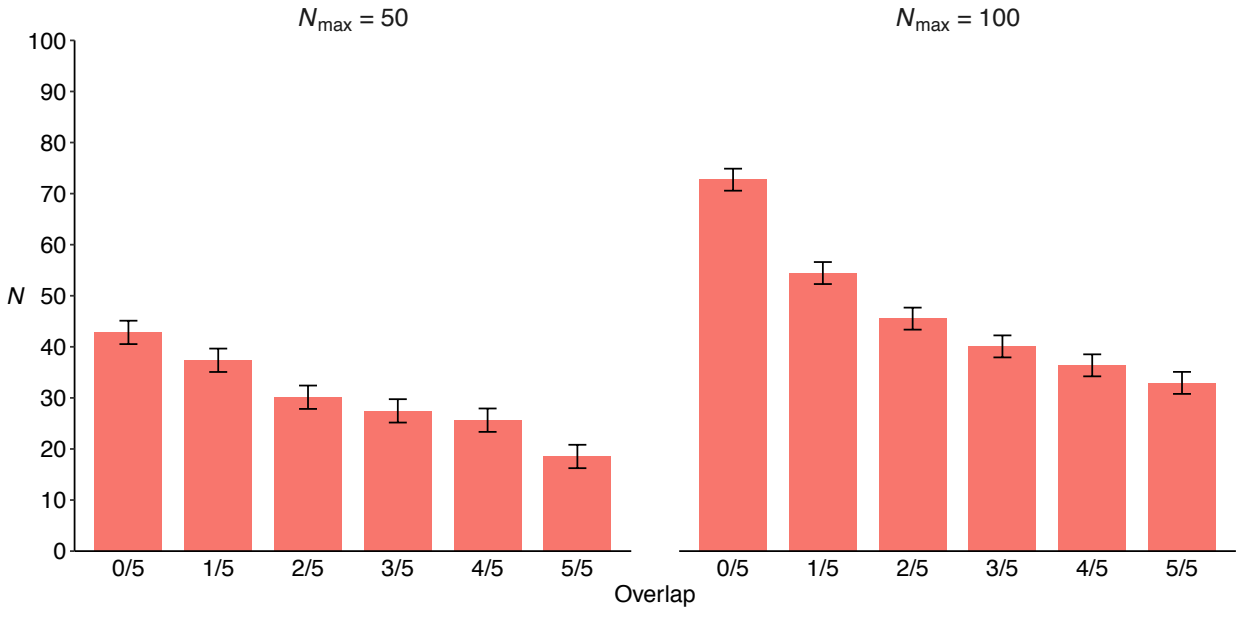
Procedure. As in Experiment 1, participants estimated the number of marbles in an urn based on the sizes of two random samples (s_1 = 10; s_2 = 5) and the overlap between them. N_{max} was manipulated between-subjects to be 50 or 100. The overlap was manipulated within-subjects to range from 0/5 to 5/5. For each overlap value, participants estimated the total population size and gave a confidence rating in their estimate on a scale from 0 (*Not at all confident*) to 100

[3] Experiments 2 and 3 were conducted after summer 2018 when researchers observed a drop in data quality from Amazon Mechanical Turk (Bai, 2018). To guard against this concern, far more participants than necessary were recruited and stringent manipulation checks were included, resulting in the exclusion of many data points from the analysis. Although Experiments 1 and 4 do not contain these manipulation checks, this is not a concern since data quality was assessed in accordance with Bai (2018) and the results are robust and replicate.

(*Extremely confident*). For each participant, the order in which the different overlap values were presented was randomized.

Results. Average estimates of population size decreased as the overlap increased from 0/5 to 5/5, both when N_{max} was 50 and 100 (Fig. S6). This result indicates that people were able to intuit the negative relationship between population size and the overlap between random samples. However, considerable variability in accuracy emerged when theoretical vs. observed distributions were compared. Specifically, people were more accurate when the overlap was small or moderate than when the overlap was large, in which case the tendency was to overestimate the population size.

Fig. S6. Experiment 2. Average estimates of population size. Error bars are 95% confidence intervals.



When N_{max} was 50, observed distributions matched with theoretical distributions for overlap values of 0/5, 1/5, 2/5, and 3/5 [MΔAUC values were relatively small, ranging from 1.03 to 3.14]. However, observed distributions deviated from theoretical distributions for overlap values of 4/5 [Fig. 3; MΔAUC = 8.58] and 5/5 [MΔAUC = 6.57]. Similar results emerged when N_{max} was 100 (Fig. 4).

Fig. 3. Experiment 2. Theoretical vs. observed cumulative density functions when N_{max} was 50.

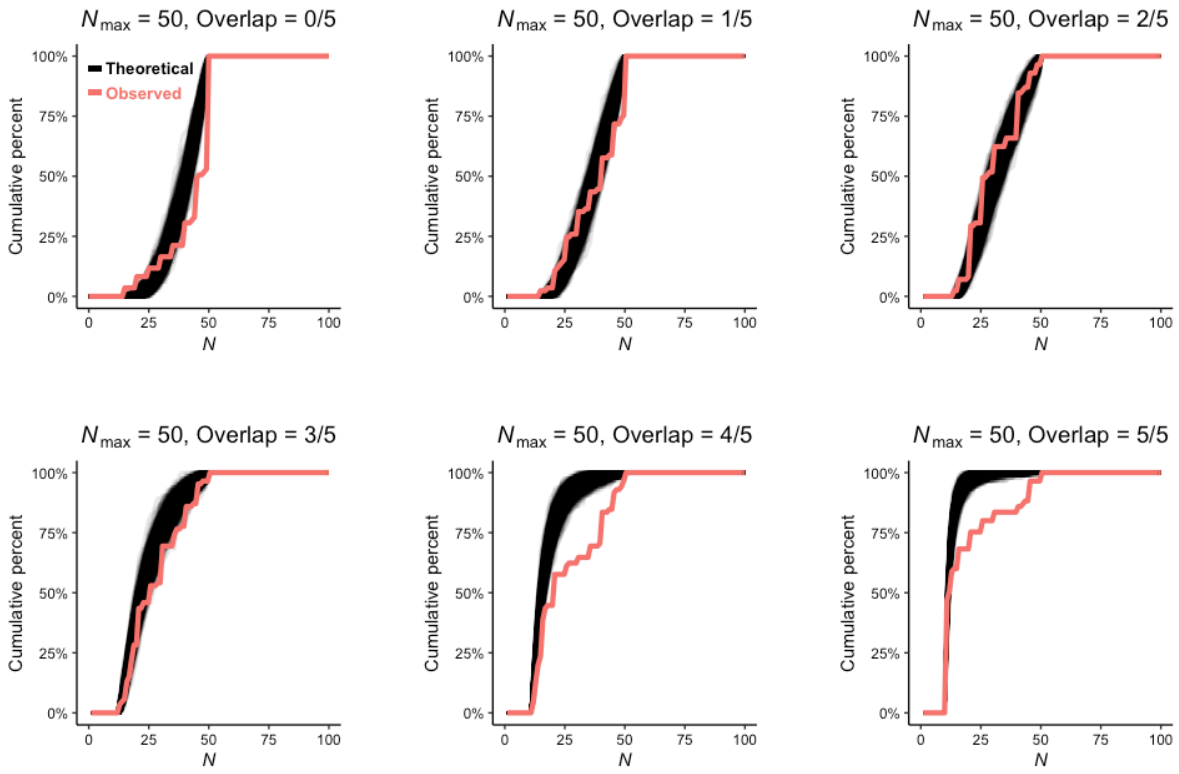
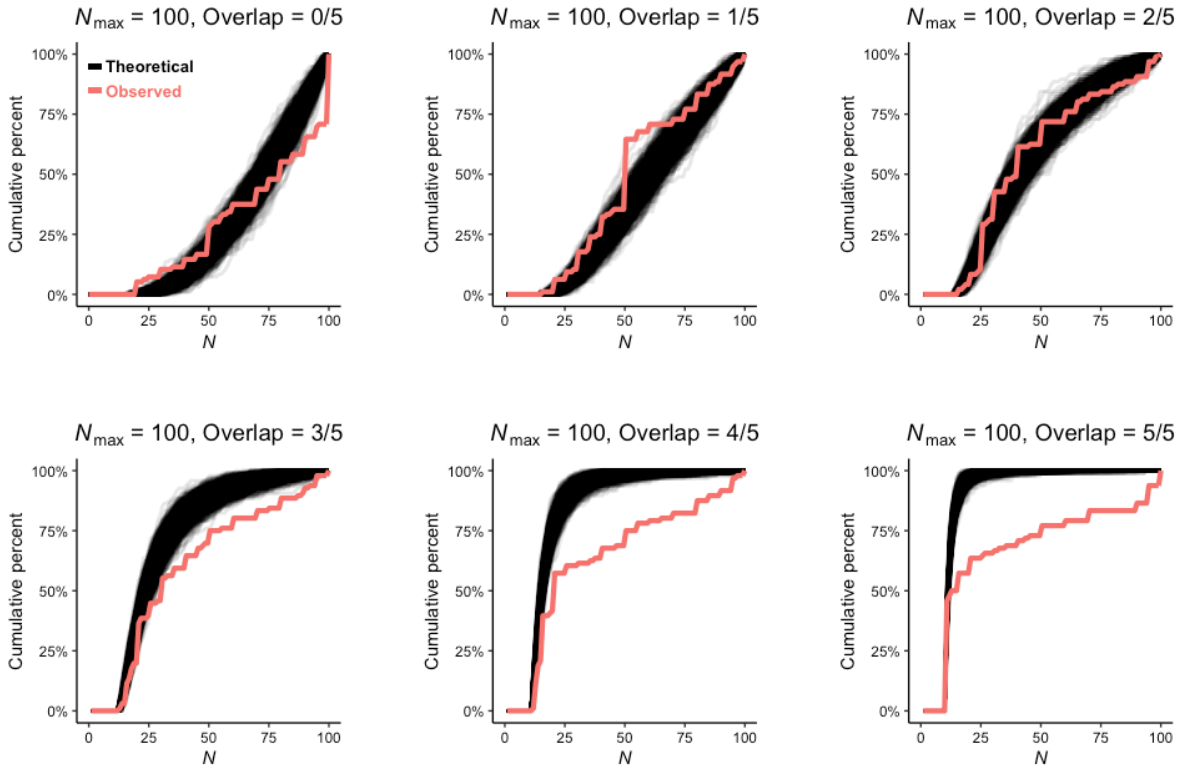
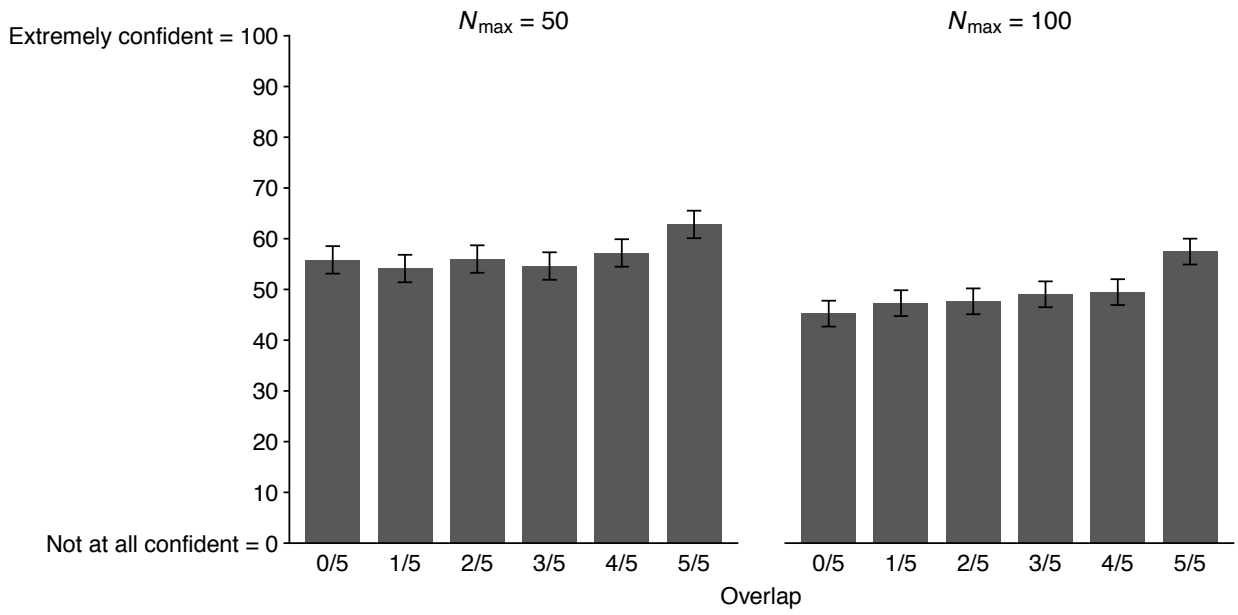


Fig. 4. Experiment 2. Theoretical vs. observed cumulative density functions when N_{max} was 100.



If confidence tracked with accuracy, then participants would have been more confident in their inferences when the overlap was small to moderate and less confident when the overlap was large. But to the contrary, participants expressed the greatest confidence when the overlap was 5/5 – when accuracy was at or near its worst – and similar, lower levels of confidence when the overlap ranged from 0/5 to 4/5, which was when inferences were more accurate (Fig. 5). This dissociation between accuracy and confidence emerged both when N_{max} was 50 [bs < -5.62, SEs < 1.99, ts(895) < -2.83, Ps < 0.005, rs > 0.09] and when N_{max} was 100 [bs < -7.98, SEs < 1.87, ts(895) < -4.27, Ps < 0.001, rs > 0.14].

Fig. 5. Experiment 2.Average confidence in population size inferences. Error bars are 95% confidence intervals.



Experiment 3

The results so far raise the question of why inferences are more accurate when observable samples indicate a large population and less accurate when observable samples indicate a small population. One possible explanation is anchoring (Tversky & Kahneman, 1974; Epley & Gilovich, 2006). According to this heuristics account, people anchored their estimates on N_{max} (50 or 100, depending on condition). As a salient and explicitly mentioned possible population size, N_{max} would produce the observed effects if people used it as an anchor. Recall that small overlaps (e.g., 0/5) indicate a large population, so anchoring on N_{max} would enable accurate inferences, which were observed in the previous experiments. Further recall that large overlaps (e.g., 4/5) indicate a small population, so anchoring on N_{max} would reduce accuracy via overestimation, which was also observed in the previous experiments. Experiment 3 tested this heuristics account of the observed effects.

Participants. One thousand seven hundred four participants were recruited from Amazon Mechanical Turk. Five hundred forty seven participants were excluded for failing attention checks (see Supplemental Materials for these checks). The final sample consisted of 1,157 participants (M_{age} = 35.59, SD_{age} = 11.75; 643 females, 503 males, 11 unspecified).

Procedure. The same 2 (N_{max}: 50 vs. 100) x 2 (overlap: 0/5 vs. 4/5) between-subjects design from Experiment 1 was adapted to include an additional between-subjects factor, N_{min} absent vs. N_{min} present. The N_{min} absent conditions were identical to those in Experiment 1: the small possible population size was not explicitly mentioned, although it was computable ($s_1 + s_2$

– o). By contrast, the N_{min} present conditions included an additional sentence stating what the smallest possible population size could be (see Supplemental Materials for stimuli).

If people anchor their estimates on N_{max}, then the presence of N_{min} should diminish this effect by rendering N_{max} less salient. In fact, people may instead anchor on N_{min} due to its intentional placement at the end of the vignette, right before the dependent measure was taken. This heuristics account would therefore predict a main effect such that estimates of population size are systematically lowered when N_{min} is present compared to when it is absent.

This systematic lowering, however, would have different implications for an overlap of 0/5 vs. 4/5. Recall that in Experiments 1 and 2 where N_{min} was absent, estimates were largely accurate when the overlap was 0/5. A systematic lowering would result in underestimation. Also recall that in Experiments 1 and 2 where N_{min} was absent, estimates were too high when the overlap was 4/5. A systematic lowering would result in accuracy, or, at the very least, attenuated overestimation. Insofar as these predictions are supported, anchoring would be a parsimonious account of how people infer the size of an unobservable population from observable samples.

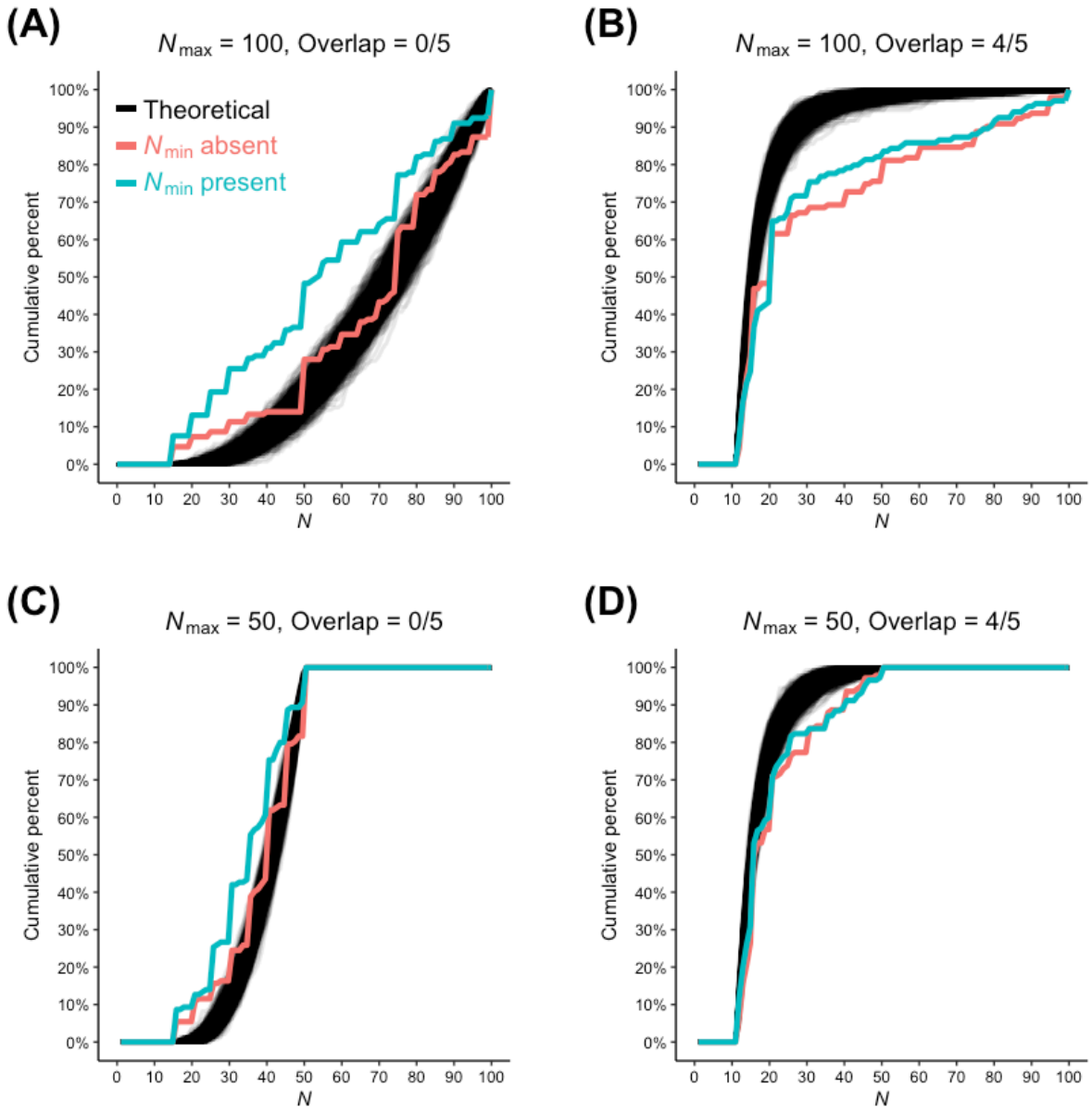
Results. The N_{min} absent conditions replicated previous findings, further underscoring the robustness of these effects. When the overlap between samples was 0/5, participants were largely accurate in their population size inferences. But when the overlap was 4/5, participants tended to overestimate the population size.

The N_{min} present conditions partially support an anchoring account. For an overlap of 0/5, participants underestimated the population size when N_{min} was present compared to when it was absent. This underestimation occurred both when N_{max} was 100 [Fig. 6A; b = -12.37, SE = 2.24, $t(1149) = -5.52$, $P < 0.0001$, $r = 0.16$] and when N_{max} was 50, though this latter effect was

smaller [Fig. 6C; b = -3.70, SE = 2.23, t(1149) = -1.66, P = 0.10, r = 0.05]. Observed CDFs in the N_{min} present conditions were shifted to the left relative to observed CDFs in the N_{min} absent conditions, leading to inaccurate inferences that fell outside and to the left of the bounds of the bootstrapped theoretical CDFs.

Although the results from the 0/5 overlap conditions support the anchoring account, the results from the 4/5 overlap conditions did not. Irrespective of whether N_{min} was absent or present, participants tended to overestimate the population size – both when N_{max} was 100 [Fig. 6B; b = -2.43, SE = 2.31, t(1149) = -1.05, P = 0.29, r = 0.03] and when N_{max} was 50 [Fig. 6D; b = -0.52, SE = 2.27, t(1149) = -0.23, P = 0.82, r = 0.007]. Together, these findings indicate that anchoring can explain greater accuracy when the overlap is small and indicative of a large population. However, anchoring cannot explain the tendency for people to overestimate when the overlap is large and indicative of a smaller population.

Fig. 6. Experiment 3. Theoretical vs. observed cumulative density functions in each condition.



Experiment 4

Previously, people expressed the most confidence in population size inferences that were among the least accurate. Experiment 4 further probed this dissociation between cognition and metacognition through a manipulation that would affect one construct but not the other. Inspired by implicit social cognition studies that establish a lack of association by showing effects on explicit but not implicit measures (e.g., Gregg, Seibt, & Banaji, 2006), Experiment 4 induced priors over the population size to be high or low. If cognition and metacognition are indeed dissociated, this manipulation of priors would affect the magnitude and variability of people's inferences, but not people's confidence in these inferences.

The rationale is as follows. If the prior over population size is low, meaning smaller estimates are initially more likely, then average estimates should be lower compared to when the prior is high, meaning larger estimates are initially more likely. These different priors should also affect the variability of participants' estimates: estimate variability should be lower when the prior and data are consistent relative to when the prior and data are inconsistent. The prior and data are consistent when a) the prior is low and the overlap is large (e.g., 4/5) because both components suggest a small population, or b) when the prior is high and the overlap is low (e.g., 0/5) because both components suggest a large population. In these cases, uncertainty is reduced, which should result in lower variability. Conversely, the prior and data are inconsistent when a) the prior is low and the overlap is low, or b) when the prior is high and the overlap is large. In these cases, uncertainty is exacerbated, which should result in higher estimate variability.

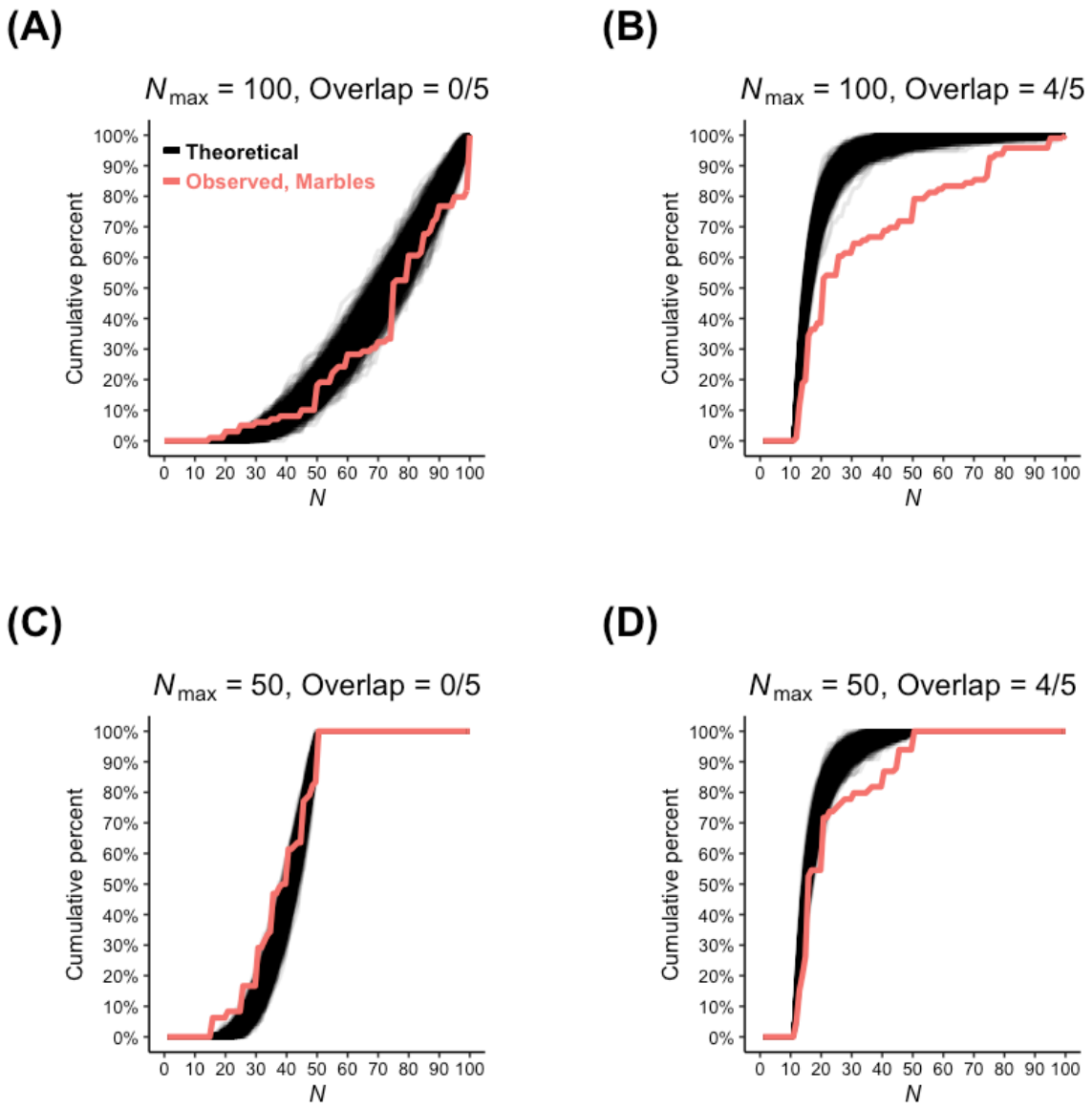
While the above results, if they emerge, would show an impact on cognition, dissociated metacognition would be supported by participants expressing the same level of confidence regardless of how much uncertainty is in the task, which is a direct function of the consistency,

or lack thereof, between the prior and data. That is, if participants make population size inferences that are in line with manipulated priors but express the same confidence irrespective of whether the priors and data are consistent, then dissociation would be further supported.

Participants. One thousand three hundred six participants were recruited from Amazon Mechanical Turk. Ninety-two participants did not begin the procedure, 6 participants began the procedure but did not finish, and 57 participants were excluded for providing estimates below the logical minimum. The final sample consisted of 1,151 participants (M_{age} = 36.26, *SD* = 12.01; 683 females, 465 males, 3 unspecified).

Procedure. The same 2 (N_{max}: 50 vs. 100) x 2 (overlap: 0/5 vs. 4/5) between-subjects design from Experiment 1 was adapted to include an additional between-subjects factor, low vs. uniform vs. high prior. The uniform prior conditions were identical to those in Experiment 1 and replicate previous results (Fig. S7). In the low and high prior conditions, the prior over population size was induced by telling participants that when the marbles were randomly sampled, it sounded like there were few or many marbles inside, respectively (see Supplemental Materials for stimuli). After estimating the population size, participants rated how confident they were in their estimates (1 = Not at all confident … 5 = Extremely confident). Lastly, participants completed a measure of probabilistic reasoning (delMas , Garfield, Ooms, & Chance, 2007).

Fig. S7. Experiment 4. Replication of previous results.



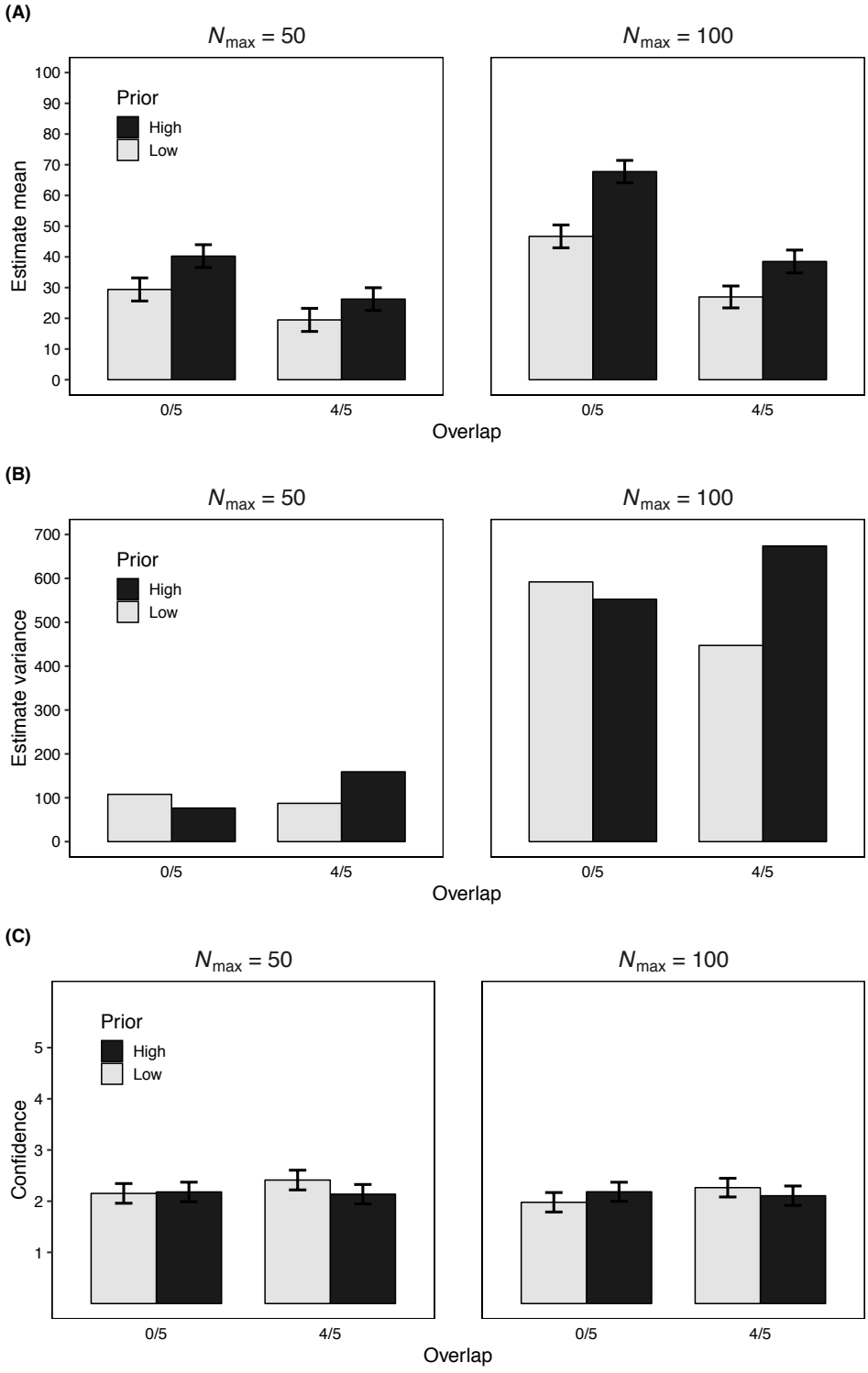
Results. As expected, average estimates were lower when the prior was low and higher when the prior was high (Fig. 7A). This pattern emerged when N_{max} was 50 for an overlap of 0/5 [$M_{\text{Low Prior}}$ = 29.39 vs. $M_{\text{High Prior}}$ = 40.24; b = -10.85, $t(1139)$ = -4.03, P = 0.0001, r = 0.12] and for an overlap of 4/5 [$M_{\text{Low Prior}}$ = 19.48 vs. $M_{\text{High Prior}}$ = 26.25; b = -6.77, $t(1139)$ = -2.52, P = 0.01, r = 0.07]. The same pattern also emerged when N_{max} was 100 for an overlap of 0/5 [$M_{\text{Low Prior}}$ = 46.67 vs. $M_{\text{High Prior}}$ = 67.77; b = -21.10, $t(1139)$ = -7.95, $P < 0.0001$, r = 0.23] and an overlap of 4/5 [$M_{\text{Low Prior}}$ = 26.95 vs. $M_{\text{High Prior}}$ = 38.50; b = -11.55, $t(1139)$ = -4.39, $P < 0.0001$, r = 0.13].

Also as expected, estimate variability was lower when the prior and overlap were consistent compared to when the prior and overlap were inconsistent (Fig. 7B). This pattern emerged when N_{max} was 50. A low prior is inconsistent with an overlap of 0/5 because the prior suggests a smaller population while the overlap suggests a larger population. This inconsistency resulted in descriptively, but not statistically, greater variability relative to a high prior, which is consistent with an overlap of 0/5 [$Var_{\text{Low Prior, 0/5 Overlap}}$ = 107.85 vs. $Var_{\text{High Prior, 0/5 Overlap}}$= 76.51; $\chi^2(1) = 3.42$, P = 0.06]. A low prior is consistent with an overlap of 4/5 because both the prior and overlap suggest a smaller population. This consistency resulted in lower variability relative to a high prior, which is inconsistent with an overlap of 4/5 [$Var_{\text{Low Prior, 4/5 Overlap}}$ = 87.35 vs. $Var_{\text{High Prior, 4/5 Overlap}}$ = 159.21; $\chi^2(1) = 14.67$, P = 0.0001]. Similar effects emerged when N_{max} was 100 (see Supplemental Materials for inferential statistics).

Although different priors affected estimate magnitude and variability, this manipulation hardly influenced how confident participants were in their population size inferences. Regardless of whether the prior and overlap were consistent – thereby reducing uncertainty – or whether the prior and overlap were in conflict – thereby exacerbating uncertainty – participants expressed the

similar levels of confidence (Fig. 7C). This pattern emerged when N_{max} was 50. Despite the difference in consistency between a low vs. high prior and an overlap of 0/5, confidence ratings hardly differed [$M_{\text{Low Prior, 0/5 Overlap}}$ = 2.15 vs. $M_{\text{High Prior, 0/5 Overlap}}$= 2.18; b = -0.03, t(1139) = -0.21, P = 0.84, r = 0.006]. And despite the difference in consistency between a low vs. high prior and an overlap of 4/5, confidence ratings once again hardly differed [$M_{\text{Low Prior, 4/5 Overlap}}$ = 2.41 vs. $M_{\text{High Prior, 4/5 Overlap}}$ = 2.14; b = 0.28, t(1139) = 2.01, P = 0.05, r = 0.06]. The same invariant level of confidence emerged when N_{max} was 100 (see Supplemental Materials for inferential statistics). Together, these data further support a dissociation between people's inferences of population size and people's confidence in these inferences.

Fig. 7. Experiment 4. **(A)** Estimate averages. **(B)** Estimate variance. **(C)** Average confidence in estimated population size. Error bars are 95% confidence intervals.



General Discussion

When inferring the size of an unobservable population based on observable samples, participants were largely accurate when the overlap between samples indicated a large population. But when this limited information was parametrically manipulated to indicate a small population, participants erred by overestimating the size of the population. Participants also failed to recognize their success and limits: confidence was highest when accuracy was at or near its worst. And as confirmed by the final experiment in which uncertainty was manipulated, the cognitive ability to make inferences about the size of an unobservable population is dissociated from the metacognitive ability to assess these inferences.

Although the task participants completed is superficially a math problem embedded in a hypothetical scenario, this task captures the essence of common everyday experiences on two levels. First, at a more specific level, resampling of objects occurs spontaneously. Whether it is bumping into a colleague at a café, driving past the same vehicle again, or noticing the same neighborhood dog, samples are continually drawn and the number of objects resampled provides information that is diagnostic of the underlying population size. The current results indicate how people may perform in more realistic settings – both in terms of cognitive and metacognitive ability.

Second, at a more general level, people make rich, sophisticated inferences that go beyond the data they receive. Furthermore, these inferences are often remarkably flexible and fast, sparking fruitful lines of research seeking to characterize this hallmark of human intelligence (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Although imperfect, the inferences made by participants were likewise rich, sophisticated, and quickly performed over a wide range of scenarios. The data collected here cannot formalize the process by which these

579 inferences were made, but these data are consistent with robust evidence that, at some level,
580 people's inferences resemble Bayesian prescriptions (Kersten, Mammassian, & Yuille, 2004).
581 However, resemblance between people's inferences and Bayesian prescriptions does not
582 necessarily mean that the underlying cognitive process is Bayesian. As shown in Experiment 3,
583 people's success in making population size inferences is, in part, due to the anchoring heuristic.
584 However, people's failure – as illustrated by overestimations – cannot be attributed to this
585 heuristic. Two points are noteworthy about these results. First, although heuristics are often and
586 justifiably discussed in the context of errors and biases, researchers would be remiss to ignore the
587 fact that these mental shortcuts serve people well under many circumstances. Inferring the size of
588 a population based on two samples with a small overlap appears to be one of these
589 circumstances. Second, these findings suggest that different mechanisms underlie successful vs.
590 unsuccessful inferences. As parsimonious as it would be for a single mechanism to account for
591 human performance, the data instead indicate greater complexity.
592 So why might people display a tendency to overestimate the size of an unobservable
593 population when the overlap between samples indicates a small population? These
594 overestimations cannot be simply attributed to non-uniform priors that favor large population
595 sizes. Uniform priors resulted in close fits between observed and theoretical distributions when
596 the overlap indicated a large population, and it is implausible that a change in single value that is
597 orthogonal to priors would result in such a shift.
598 One possible explanation for this overestimation bias is extremeness aversion, the
599 tendency to prefer intermediate options to options at the extremes (Simonson & Tversky, 1992).
600 In the present studies, small estimates of N_{min} were extreme options that participants may have

found unattractive. However, estimates of N_{max}, which is also an extreme option, were common among participants, which is inconsistent with this account.

An alternative is that small populations sizes are prone to overestimation insofar as the surprise associated with resampling the same objects interferes with subsequent mental computations. The idiom "it's a small world" can be thought of as an expression of surprise when the same objects are resampled. Note, however, that there is no analogous idiom for the absence of resampling. This asymmetry in surprise might account for differences in performance, a possibility that is consistent with copious evidence illustrating the influence of emotion on human judgment (Clark & Isen, 1982; Clore, Schwartz, & Conway, 1993).

The presenting findings – which show variability in human performance in inferring a hidden population size – raise the question of how performance might be bolstered. Contrary to one intuitive prediction, advanced training in or experience with statistics appears to be of limited benefit. Before any experiments were conducted, the procedure was piloted on psychology graduate students and postdocs. Their inferences resemble the inferences of less quantitatively sophisticated participants on Amazon Mechanical Turk (Fig. S8). Furthermore, individual differences in numeracy – defined as the ability to understand and manipulate numerical information (Paulos, 1988) – weakly predict estimate quality (Fig. S9).

These findings may inspire investigation into how fundamental and early emerging these effects are. Work from developmental psychology may be of note. Infants as young as eight months have been described as intuitive statisticians for making rational inferences about a population from which a sample is drawn (Xu & Garcia, 2009). Whether young children can make similarly rational inferences based on the overlap between samples – and avoid the overestimation bias documented here among adults – remains to be seen.

Although adult inferences erred towards overestimation when the overlap between samples indicated a small population, people expressed relatively high levels of confidence in these inferences. This disconnect between cognition and metacognition dovetails with previous work showing overconfidence (Moore & Healy, 2008). This metacognitive failure could prevent people from adjusting unlikely estimates and maintaining likely estimates, two key benefits of well-calibrated metacognition (Metcalfe, 1996). This miscalibration can be costly. When consequential decisions depend on accurate inferences of population size, too much confidence can lead to suboptimal outcomes. Knowledge of this miscalibration may be a first step in countering its effects.

Fig. S8. (A) Means of theoretical posterior distributions. **(B)** Means of participants' estimates from Experiment 1, marbles condition. **(C)** Means of estimates by 15 psychology PhD students and postdocs.

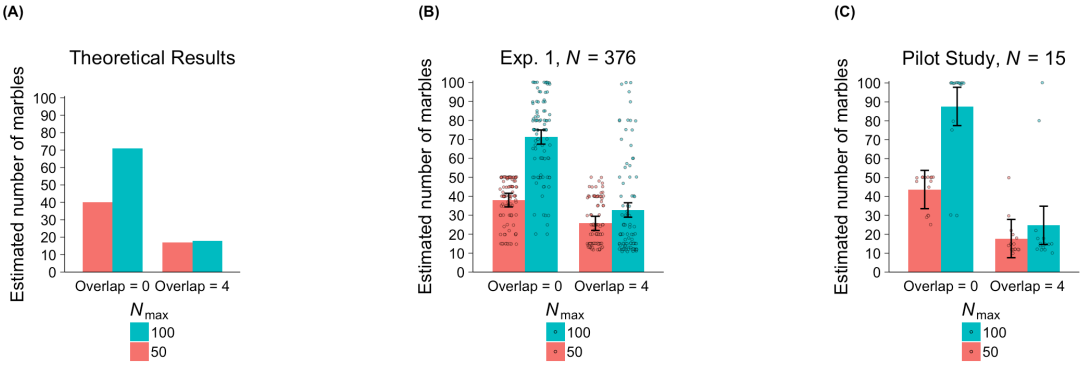
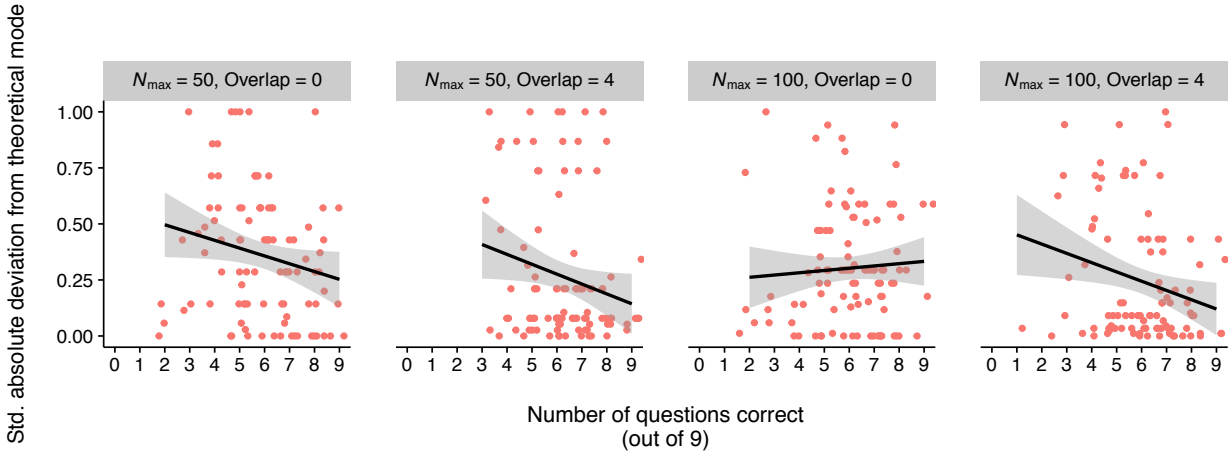


Fig. S9. Experiment 4, uniform prior condition. Relationship between estimate quality (y-axis; 0 = no difference between estimate and theoretical posterior mode; 1 = largest possible difference between estimate and theoretical posterior mode) and numeracy (x-axis; total number of questions correct). Small negative relationships emerged in three of the four conditions, indicating that greater numeracy is somewhat correlated with higher quality estimates [*b*s > |-0.05|, *t*s < |-2.35|, *P*s > 0.01, *r*s < 0.12].



Open Practices Statement

All data and code are available at [osf.io/7ady5]. All materials are included in Supplemental Materials section at the end of this manuscript.

References

Bai, H. (2018). Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. Retried from https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

Blake, P.R., & McAuliffe, K. (2011). I had so much it didn't seem fair: Eight-year-olds reject two forms of inequity. *Cognition*, *120*(2), 215 – 224.

Buehler, R., Griffin, D., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology, 67*(3), 366 – 381.

Cao, J., & Banaji, M.R. (2017). Social inferences from group size. *Journal of Experimetnal Social Psychology*, *70*, 204 – 211.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7*(4), 928 – 935.

Clark, M.S., & Isen, A.M. (1982). Toward understanding the relationship between feeling states and social behavior. In A.H. Hastorf & A.M. Isen (Eds.), *Cognitive Social Psychology* (pp. 73 –108). New York, NY: Elsevier/North-Holland.

Clayson, D.E. (2005). Performance overconfidence: Metacognitive effects of misplaced student expectations? *Journal of Marketing Education*, *27*(2), 122 – 129.

Clore, G.L., Schwarz, N., & Conway, M. (1993). Affective causes and consequences of social information processing. In R.S. Wyer & T.K. Srull (Eds.), *Handbook of Social Cognition*. Hillsdale, NJ: Erlbaum.

De Langhe, B., Fernbach, P.M., & Lichtenstein, D.R. (2016). Navigating by the stars: Investing the actual and perceived validity of online user ratings. *Journal of Consumer Research*, *42*, 817 – 833.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28 – 58.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science*, *17(4)*, 311 – 318.

Gregg, A.P., Seibt, B., & Banaji, M.R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1 – 20.

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

Kersten, D., Mammassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271 – 304.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105(2)*, 395 – 438.

Lee, M., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press: Cambridge, England.

Libertus, M.E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14(6)*, 1292 – 1300.

Metcalfe, J. (1996). *Metacognition: Knowing about knowing.* Cambridge, MA: MIT Press.

Moore, D., & Healy, P.J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502 – 517.

Obrecht, N.A., Chapman, G.B., & Gelman, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*, 1147 – 1152.

Paulos, J.A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York, NY: Hill and Wang.

Petersen, C.G.J. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station*, *6*, 5 – 84.

Pietraszewki, D., & Shaw, A. (2015). Not by strength alone: Children's conflict expectations follow the logic of the asymmetric war of attrition. *Human Nature*, *26*, 44 – 72.

Seber, G.A.F. (1982). A review of estimating animal abundance. *Biometrics*, *42*(2), 267 – 292.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, *29*(3), 281 – 295.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., & Goodman, N.D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279 – 1285.

Tulving, E. (1999). Study of memory: Processes and systems. In J.K. Foster & M. Jelicic (Eds.), *Debates in psychology*. *Memory: Systems, process, or function?* (pp. 11 – 30). New York, NY: Oxford University Press.

Tversky, A., & Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124 – 1131.

Ubel, P.A., Jepson, C., & Baron, J. (20011). The inclusion of patient testimonials in decision aids: Effects on treatment choices. *Medical Decision Making*, *21*, 60 – 68.

733 Xu , F., & Garcia, V. (2008). Intuitive statistics by 8-month old infants. *Proceedings of the*
734 *National Academy of Sciences of the United States of America*, *105*(13), 5012 – 5015.
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755

Supplemental Materials: Technical Details

The following problem, adapted from Wikipedia, is used to illustrate the technical details of inferring an unobservable population size from observable samples.

An ecologist wants to estimate the number of fish in a lake. She randomly samples 10 fish from the lake and applies paint to their scales. These 10 fish are released back into the lake. The ecologist then takes a second random sample of 15 fish. Of these 15 fish, 5 have paint on their scales. How many fish are in the lake?

In this example, the size of the first random sample, s_1, is 10; the size of the second random sample, s_2, is 15; and the overlap, o, between these two samples is 5. Let N be the total number of fish in the lake. Since both samples were random and 5 out of 15 fish in s_2 had paint on their scales, this proportion is equivalent to the proportion of all fish in the lake with paint on their scales. N denotes all fish in the lake, and s_1 denotes all fish in the lake with paint on their scales.

$$\frac{o}{s_2} = \frac{s_1}{N}$$

N is therefore equal to the product of s_1 and s_2 divided by o.

$$N = \frac{s_1 s_2}{o}$$

Plugging in the known values for s_1, s_2, and o therefore yields an estimate of the population size, N.

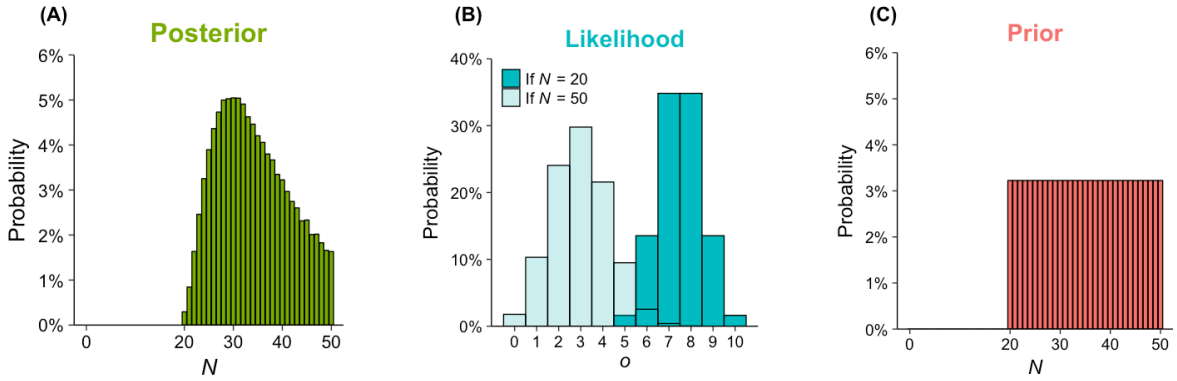
$$N = \frac{10 \times 15}{5} = 30$$

A Bayesian implementation, described on the next page, addresses three shortcomings of the math above. The first shortcoming is that the math is deterministic and thus ignores uncertainty in sampling. N must equal the product of s_1 and s_2 divided by o; however, any number greater than or equal to $s_1 + s_2 + o$ is possible because this is the number of unique fish observed. Second, the math cannot compute if the overlap, o, is zero. In this case, the denominator would be zero, making N nonreal. Third, prior knowledge cannot be integrated. To illustrate, there are surely more fish in a lake that spans hundreds of kilometers than there are in a smaller lake that covers only a few kilometers. However, the same values for s_1, s_2, and o can be obtained from both lakes, which would lead to the mistaken conclusion that there are an equal number of fish in both lakes.

Supplemental Materials: Technical Details, continued

The Bayesian implementation computes a posterior distribution over N by integrating a hypergeometric likelihood function with a categorical prior (see figure below). The prior is bounded between $s_1 + s_2 - o$, the minimum number of objects in the population, and N_{max}, the maximum number of possible objects in the population. In the example above, the minimum is 20 ($s_1 + s_2 - o = 10 + 15 - 5$), and for illustrative purposes, $N_{\max}$ is 50. A uniform prior over this range is updated to a posterior whose central tendency measures are close to 30 because this population size is most likely to produce an observed overlap, o, of 5. In the middle figure below, two likelihood distributions are plotted, one for the smallest and another for largest possible value of N. These likelihood functions illustrate the intuition that large overlaps are more likely when the population is small, whereas small overlaps are more likely when the population is large.

(A) Posterior distribution that results from integrating a hypergeometric likelihood **(B)** and uniform prior **(C)**.



Supplemental Materials: Stimuli for Experiment 1

Table 1 in the main text presents the main task participants completed, which was to estimate the total number of marbles in an urn (or spoons or bottle caps in a box). After providing their estimates, participants indicated their confidence in those estimates on a 1 to 5 Likert-type scale (*1 = Not at all confident … 5 = Extremely confident*).

> *How confident you do you feel about the answer you just gave about the number of marbles [spoons; bottle caps] inside the urn [box]?*

Lastly, particiapnts indicated their agreement with the statements below (*1 = Strongly disagree … 5 = Strongly agree*).

> *I feel confident in my ability to solve statistical problems.*
> *I have been well trained to deal with statistical problems.*
> *When it comes to solving statistical problems, I feel doubtful about my answers.* (R)
> *I feel that I'm unable to given accurate solutions to statistical problems.* (R)
>
> (R) = Reverse-coded

Supplemental Materials: Stimuli for Experiment 2

Stimuli for Experiment 2 were identical to those used in Experiment 1 except the value of the overlap between samples was manipulated to range from 0/5 to 5/5. See relevant sentence below:

…He shows you these 5 marbles, and you see that [none vs. 1 vs. 2 vs. 3 vs. 4 vs. all] of these 5 marbles have large red dots on them.

Before starting Experiment 2, all participants were randomly assigned one of the following two manipulation checks. Participants had to answer "three" or "twelve" in order to continue in the experiment.

Before we start, please answer the following question. When the following numbers are arranged by their numerical value, which is the middle number: three, 6, or one? Please write your response in lowercase letters.

Before we start, please answer the following question. When the following numbers are arranged by their numerical value, which is the middle number: fourteen, eleven, or 12? Please write your response in lowercase letters.

After estimating the number of marbles in the urn and providing a confidence rating, participants answered the following two comprehension check questions. They had to answer "Marbles" and "Red" in order for their data to be included in the analysis.

The questions you just answered were about a specific type of object. What was the object?

- *Marbles*
- *Spoons*
- *Bottle Caps*
- *Pencils*

The person in the scenarios you just red used a permanent marker to draw colored dots. What color were those dots?

- *Red*
- *Blue*
- *Green*
- *Yellow*

These same comprehension check questions were used in Experiment 3, except for the question immediately above regarding the color of the dots. That question was replaced with the following question in Experiment 3. Participants had to select "Urn" in order for their data to be included in the analysis.

The question you just answered was about how many objects were in a specific type of container. What type of container was it?

- *Urn*
- *Box*

- *Jar*
- *Bag*

Supplemental Materials: Stimuli for Experiment 3

Stimuli for Experiment 3 were identical those used in Experiment 1 except in the N_{min} present conditions, an additional sentence was included to convey the minimum number of possible marbles. This additional sentence, highlighted in grey and surrounded by the preceding and following text, is below.

…He shows you these 5 marbles, and you see that [none vs. 4] of these 5 marbles have large red dots on them.

This means there are at least N_{min} marbles inside the urn.

At this point the person asks you to guess how many marbles are inside the urn.

How many marbles do you think are inside the urn? Please type in a number below.

Of course, the exact value for N_{min} depended on condition. When the overlap was 0/5, N_{min} was 15. When the overlap was 4/5, N_{min} was 11.

Supplemental Materials: Stimuli for Experiment 4

The stimuli used in Experiment 4 were identical to those in Experiment 1 except for sentences highlighted in grey below. These sentences comprise the prior belief manipulation.

Low prior condition:
Intrigued, you put on the headphones and then watch as the person thoroughly mixes up all the marbles in the urn. The noise-cancelling headphones aren't totally effective. You can still hear some sounds, and it seems like there's a small number of marbles moving around inside the urn.

High prior condition:
Intrigued, you put on the headphones and then watch as the person thoroughly mixes up all the marbles in the urn. The noise-cancelling headphones aren't totally effective. You can still hear some sounds, and it seems like there's a large number of marbles moving around inside the urn.

Uniform prior condition:
Intrigued, you put on the headphones and then watch as the person thoroughly mixes up all the marbles in the urn. The noise-cancelling headphones are totally effective. You can't hear anything, so you have no idea how many marbles are moving around inside the urn.

Supplemental Materials: Stimuli for Experiment 4, continued

The nine-item measure of probabilistic reasoning (delMas et al., 2007) is below. Correct answers are bolded.

1. Bob and Bill each bought one ticket for a lottery each week for the past 100 weeks. Bob has not won a single prize yet. Bob just won a $20 prize last week. Who is more likely to win a prize this coming week if they each buy only one ticket?
    a. Bill
    b. Bob
    c. They have an equal chance of winning

2. Two containers, labeled A and B, are filled with red and blue marbles according to the quantities listed in the table below. Each container is shaken vigorously. After choosing one of the containers, you will reach in and, without looking, draw out a marble. If the marble is blue, you win $50. Which container gives you the best chance of drawing a blue marble?

Container	Red	Blue
A	6	4
B	60	40

    a. Container A (with 6 red and 4 blue)
    b. Container B (with 60 red and 40 blue)
    c. Equal chances from each container

3. When two fair six-sided dice are simultaneously thrown, these are two possible results that could occur: Result 1: a 5 and a 6 are obtained in any order. Result 2: a 5 is obtained on each die. Which of the following statements is correct?
    a. The probability of obtaining each of these results is equal
    b. There is a higher probability of obtaining Result 1 (a 5 and a 6 in any order)
    c. There is a higher probability of obtaining Result 2 (a 5 on each die)
    d. It is impossible to give an answer

4. Suppose you read on the back of a lottery ticket that the chances of winning a prize are 1 out of 10. Select the best interpretation.
    a. You will win at least once out of the next 10 times you buy a ticket
    b. You will win exactly once out of the next 10 times you buy a ticket
    c. You might win once out of the next 10 times but it is not for sure

Supplemental Materials: Stimuli for Experiment 4, continued

5. You are about to roll 2 fair six-sided dice, hoping to get a double. (A double = both dice show the same value on top). Which double will occur the least often?
    a. 6 – 6
    b. 1 – 1
    c. 1 – 1 and 6 – 6 are both least likely to occur
    d. All doubles are equally likely

6. Colin is flipping a fair coin. Heads has just come up 5 times in a row! The chance of getting heads on the next throw is
    a. Less than the chance of getting tails since we are due for a tails
    b. Equal to the chance of getting tails since the flips are independent and the coin is fair
    c. Greater than the chance of getting tails since heads seem to be coming up

7. A game company created a little plastic dog that can be tossed in the air. It can land either with all four feet on the ground, lying on its back, lying on its right side, or lying on its left side. However, the company does not know the probability of each of these outcomes. They want to estimate the probabilities. Which of the following methods is most appropriate?
    a. Since there are four possible outcomes, assign a probability of 1/4 to each outcome
    b. Toss the plastic dog many times and see what percent of the time each outcome occurs
    c. Simulate the data using a model that has four equally likely outcomes
    d. None of the above

8. Two containers, labeled A and B, are filled with red and blue marbles according to the quantities listed in the table below. Each container is shaken several times. Which of the following outcomes has the smallest probability?

Container	A	B
Red	80	40
Blue	20	60

    a. Obtaining a blue marble from container A
    b. Obtaining a blue marble from container A and a blue marble from container B
    c. All of the above equally likely

Supplemental Materials: Stimuli for Experiment 4, continued

9. The local meteorologist claims that there is a 70% probability of rain tomorrow. Provide the best interpretation of this statement.
    a. Approximately 70% of the city will receive rain within the next 24 hours
    b. Historical records show that it has rained on 70% of previous occasions with the same weather conditions
    c. If we were to repeatedly monitor the weather tomorrow, 70% of the time it will be raining
    d. Over the next ten days, it should rain on seven of them

Supplemental Materials: Inferential Statistics for Experiment 4

As was the case when N_{max} was 50, estimate variability when N_{max} was 100 was greater when the prior and data were consistent and lower when the prior and data were inconsistent [$Var_{\text{Low Prior, 0/5 Overlap}} = 592.03$ vs. $Var_{\text{High Prior, 0/5 Overlap}} = 552.49$; $\chi^2(1) = 0.08$, $P = 0.77$; $Var_{\text{Low Prior, 4/5 Overlap}} = 447.14$ vs. $Var_{\text{High Prior, 4/5 Overlap}} = 673.87$; $\chi^2(1) = 16.19$, $P < 0.0001$].

And as was the case when N_{max} was 50, confidence ratings when N_{max} was 100 were relatively constant. This occurred both when the overlap between samples was 0/5 [$M_{\text{Low Prior}} = 1.98$ vs. $M_{\text{High Prior}} = 2.18$; $b = -0.21$, $t(1139) = -1.51$, $P = 0.13$, $r = 0.04$] and when the overlap between samples was 4/5 [$M_{Low\ Prior} = 2.26$ vs. $M_{High\ Prior} = 2.11$; $b = 0.16$, $t(1139) = 1.18$, $P = 0.24$, $r = 0.03$].