Statistical judgments are influenced by the implied likelihood that samples represent the same population

Dana L. Chesney · Natalie A. Obrecht

Published online: 29 October 2011
© Psychonomic Society, Inc. 2011

Abstract When sample information is combined, it is generally considered normative to weight information based on larger samples more heavily than information based on smaller samples. However, if samples appear likely to have been drawn from different subpopulations, it is reasonable to combine estimates of these subpopulation means (typically, the sample means) without weighting these estimates by sample size. This study investigated whether laypeople are influenced by the likelihood of samples coming from the same population when determining how to combine information. In two experiments we show that (1) implied binomial variability affected participants' judgments of the likelihood that a sample was drawn from a given population, (2) participants' judgments were more affected by sample size when samples were implied to be drawn randomly from a general population, compared to when they were implied to be drawn from different subpopulations, and (3) people higher in numeracy gave more normative responses. We conclude that when determining how to weight and combine samples, laypeople use not only the provided data, but also information about likelihood and sampling processes that these data imply.

Keywords Judgment · Reasoning · Inductive reasoning · Mathematical cognition · Individual differences

D. L. Chesney (✉)
Department of Psychology, University of Notre Dame,
118 Haggard Hall,
Notre Dame, IN 46556, USA
e-mail: Dana.Chesney.3@nd.edu

N. A. Obrecht (✉)
Department of Psychology, William Paterson University,
300 Pompton Road,
Wayne, NJ 07470, USA
e-mail: obrecht@wpunj.edu

When an inference is made from a sample to a population, that sample's representativeness is of foremost concern. The more representative the sample, the better one can judge the true nature of the population. Humans are more trusting of data they believe is more representative (Kahneman & Tversky, 1972). Numerous studies highlight the importance of the law of large numbers in establishing representativeness: Humans are intuitively aware that larger samples typically provide more reliable information (e.g., Evans & Dussior, 1977; Evans & Pollard, 1985; Irwin, Smith & Mayfield, 1956; Jacobs & Narloch, 2001; Masnick & Morris, 2008; Nisbett, Krantz, Jepson & Kunda, 1983; Obrecht, Chapman & Gelman, 2007; Peterson & Beach, 1967; Sedlmeier & Gigerenzer, 1997; Sedlmeier & Gigerenzer, 2000). However, laypeople do not use sample size in what researchers consider to be a normative fashion. People's apparent confidence in samples' representativeness increases at a shallower rate than a normative use of sample size would predict (Obrecht, 2010). Also, mean difference has a stronger effect on people's judgments than does sample size (Obrecht et al., 2007). Furthermore, competing factors such as anecdotal descriptions and encounter frequency may lead to sample size being overlooked (Nisbett et al., 1983; Obrecht, Chapman & Gelman, 2009; Ubel, Jepson & Baron, 2001). Additionally, as illustrated by Kahneman and Tversky's (1972) classic "hospital problem," many people fail to recognize that larger sample sizes will reduce the chance that sample means will differ from a population mean (but see Evans & Dussior, 1977; Sedlmeier, 1998).

However, sample size is not the only factor that determines representativeness. The normative standard of weighting means by sample size assumes that samples are drawn randomly from a general population. When samples instead come from different subpopulations, ignoring sample size when combining estimates of subpopulation

means is statistically correct. For example, consider a case in which you wish to determine what proportion of the general population would like a movie: 90% of a sample of 500 men and 10% of a sample of 100 women liked the movie. Weighting means by sample size, one would infer that 77% of the general population would like the movie. This would incorrectly overweight men's opinions and underweight women's opinions, relative to their prevalence in the general population. Rather, the opinions of both subpopulations (as estimated from the samples) should be weighted by their frequency within the general population, or lacking this information, be given equal weight. Thus, one should infer that 50% of the general population would like the movie. This illustrates that sampling method is of concern when determining how sample information should be combined.

Laypeople are indeed sensitive to sampling method when determining samples' representativeness. For example, infants' inferences are affected by whether samples are shown to be randomly or nonrandomly selected (Xu & Denison, 2009), and they can detect that samples are unrepresentative of a known population, even when sampling method is hidden (Kushnir, Xu & Wellman, 2010). Furthermore, children are more likely to generalize attributes from samples to general populations when given diverse samples that are (presumably) more representative of a general population rather than homogeneous samples, which could represent a particular subpopulation (Rhodes, Brickman & Gelman, 2008; see also Osherson, Smith, Wilkie, Lopez & Shafir, 1990).

However, in many situations, sampling procedures and population parameters are both unknown. This sets up a tricky situation where samples are used to make inferences about a population but one needs to know about the population to determine whether the samples are representative. This becomes particularly problematic when determining how to combine information from different sources in order to make an inference about a population. One way of addressing this issue is to use the sample data themselves to judge the likelihood that samples came from the same population. Consider a case in which 32% of 100 people recommend a restaurant on one Web site, and 40% of 100 recommend it on another. There is a 30% chance of drawing samples of size 100 at least this different from the same population. If the recommendation percentages were, instead, 2% and 10%, there would be only a 3% chance of drawing such samples from the same population. The probability difference between these examples is due to differences in variance. Although the sample size ($N = 100$) and the difference between the percentages of recommendations ($40 - 32 = 10 - 2 = 8$) are identical, the sample variances are not. In a binomial distribution (only two outcomes; e.g., recommending vs. not recommending),

variability is a function of sample size (n) and the probability of obtaining a given outcome (π): $\sigma^2 = n\pi(1 - \pi)$. Holding sample size constant, variability is highest when $\pi = 50\%$ and decreases as π approaches 0% or 100%: distributions with higher variability (i.e., where π is closer to 50%) have a greater chance of producing disparate sample percentages than do distributions with lower variability (i.e., where π is further from 50%). Since it is more likely that the samples were randomly drawn from the same population in the 32% and 40% example than in the 2% and 10% example, it is reasonable to give more weight to sample size when combining sample information in the former case than in the latter. Thus, we argue that when it is unknown whether samples represent different subpopulations, it makes sense to use probabilistic inference to temper sample size use, giving less weight to sample size when it is less probable that samples were drawn from the same population.

Recent findings suggest that people may vary the extent to which they use sample size when combining sample information via just such a method. Obrecht (2010) found that participants' judgments were more influenced by sample size when sample sizes were smaller; recall that as sample size decreases, variability increases. Furthermore, participants gave greater consideration to sample size when samples had percentages closer to 50% (more variable) than when they were closer to 0% or 100% (less variable). Thus, participants gave more weight to sample size when combining information from samples that were more likely to be drawn from the same population. This likelihood would have had to have been intuited from sample size and percentage information, which, in a binomial distribution, determines variance. However, it is premature to posit that laypeople adjust the extent to which they weight sample size on the basis of the likelihood that samples have been drawn from the same population. First it must be demonstrated that laypeople's judgments of the likelihood that samples came from a given population are influenced by variance implied by sample size and percentage information.

In the following experiments, we investigate these points. In [Experiment 1](#) we demonstrate that participants' judgments of which population a sample was more likely to have come from were affected by probability differences resulting from implied binomial variability. In [Experiment 2](#) we show that, when integrating information from different sources, participants gave more weight to sample size when samples were implied to come from the same population, rather than from different subpopulations. These results suggest that ostensibly non-normative use of sample size may reflect statistically legitimate uncertainty as to whether samples are indeed representative of the same population.

Experiment 1

This experiment tested whether people use implied variance information when judging the likelihood of samples coming from populations. Similar to prior studies (Kahneman & Tversky, 1972; Nisbett et al., 1983), we used a dichotomous feature as the basis of comparison. Thus, variance was determined solely from sample sizes and the proportions of populations exhibiting that dichotomous feature: $\sigma^2 = n\pi(1 - \pi)$. Participants were given information about two different populations of trees and were asked to indicate which population a sample was more likely to have come from.

Method

Participants

Undergraduate students at the University of Notre Dame ($N = 266$) participated for course credit. Data from 45 participants were excluded for not completing the task within 5–120 min.

Design

This study was conducted online. Every trial involved comparing a sample to two different populations in order to determine from which population the sample was more likely to have been drawn. Participants were told what percentage of the sample had an outcome (*sample percentage*; e.g., 10% of trees in a grove of 100 trees have white flowers) and the chance of that outcome occurring (π ; e.g., 2% of Aoco trees have white flowers, 18% of Boco trees have white flowers) for two different populations. In each pair of π s, one was closer to 50% and one was further from 50%. We will refer to these as *inner* and *outer* π s, respectively. Sample percentages always fell between these two population π s. In order to determine how implied variability influences people's likelihood judgments while controlling other statistical factors, we manipulated population *centrality* (*central*, *extreme*), *spread* (*narrow*, *wide*), and *parity* (*low*, *high*), a $2 \times 2 \times 2$ within-subjects design. Additionally, we varied *sample location* (*close*, *middle*, *far*), and *order* (inner or outer π first), to create six unique trials in each of the eight resulting centrality, spread, and parity conditions, yielding a total of 48 pairs of questions. We explain these factors below.

Centrality Of primary interest, the *centrality* manipulation allowed us to test whether participants' judgments were influenced by differences in relative probability, when probability differences were detectable only via sensitivity to implied variance. Population pairs had π s that were either *central* (closer to the central value of 50%) or *extreme* (further from 50%). For example, π s of 2% and 18% are

extreme (relatively far from 50%), while π s of 32% and 48% are central (relatively close to 50%). In both cases, the π s are the same distance from each other ($2 - 8 = 32 - 48$). As was discussed above, in binomial distributions, variance increases as π s approach 50%. Thus, normatively, one should be more confident that a sample with a percentage of 10% came from a $\pi = 18\%$, rather than a $\pi = 2\%$, population than that a 40% sample came from a $\pi = 48\%$, rather than a $\pi = 32\%$, population, even though the sample percentages are equidistant from the population π s in both cases. Each participant was asked 24 pairs of questions that differed only in centrality. Absolute differences between π s and sample percentages, the presentation order of higher and lower π s, and whether the π of the more likely population was higher or lower than the sample percentage were otherwise identical in these pairs.

Spread In the *spread* conditions, we manipulated the absolute difference between sample percentages and population π s, which simultaneously manipulated variance. Population π s were more widely spaced in the *wide* (inner π - outer π = 29%) than the *narrow* spread conditions (inner π - outer π = 16%). Since populations with more central π s are more variable, inner populations are more variable in wide than in narrow spread conditions. For example, the variability of the inner population is greater in the extreme-wide-low condition (π s of 2% and 31%) than in the extreme-narrow-low condition (π s of 2% and 18%). Thus, one's confidence that a 16% sample came from a 31% population, rather than a 2% population, should be greater than one's confidence that a 10% sample came from an 18% population, rather than a 2% population. Indeed, the relative probability of samples being drawn from inner, rather than outer, populations was always greater in wide than in narrow conditions of the same centrality and parity (see Table 1).

We also planned to test whether the absolute difference between sample means and population π s would influence participants' judgments over and above effects of population variance. We matched the populations' likelihoods in the extreme-narrow conditions (where sample percentages were, on average, equidistant from the inner and outer π s) and the populations' likelihoods in the central-wide conditions (where samples were, on average, closer to the inner π s) as closely as possible, using whole number π s (see Table 1). For example, in the central-wide-high condition, the probability of drawing a 56% sample ($N = 100$) from the inner population ($\pi = 49\%$) is .030. This is similar to .032, the probability of drawing an 88% sample ($N = 100$) from the inner population ($\pi = 82\%$) in the extreme-narrow-high condition, despite the fact that 56% is further from 49% than 88% is from 82%.

Parity The *parity* manipulation balanced whether the inner population π was greater or less than the sample percentage

Table 1 Population π s and sample percentages used in constructing stimuli for Experiment 1 (columns 2, 3, and 4) and the absolute and relative probabilities of these sample percentages being drawn from populations with the given π s (columns 5, 6, and 7)

Population Centrality, Spread, and Parity	Inner Pop. π (s^{2*})	Outer Pop. π (s^{2*})	Sample % (Location)	P of Sample % Given Inner π	P of Sample % Given Outer π	Ratio of P s of Sample %s ($P_{\text{Inner}}/P_{\text{Outer}}$)
Extreme/narrow/low	18% (14.76)	2% (1.96)	8% (far)	2.4×10^{-3}	7.4×10^{-4}	3.3
			10% (middle)	1.1×10^{-2}	2.9×10^{-5}	3.8×10^2
			12% (close)	3.2×10^{-2}	7.3×10^{-7}	4.4×10^4
Extreme/wide/low	31% (21.39)	2% (1.96)	16% (far)	2.8×10^{-4}	1.6×10^{-10}	1.8×10^6
			20% (middle)	4.6×10^{-3}	1.1×10^{-14}	4.1×10^{11}
			24% (close)	2.8×10^{-2}	2.9×10^{-19}	9.7×10^{16}
Central/narrow/low	48% (24.96)	32% (21.76)	38% (far)	1.1×10^{-2}	3.7×10^{-2}	2.9×10^{-1}
			40% (middle)	2.2×10^{-2}	2.0×10^{-2}	1.1
			42% (close)	3.9×10^{-2}	9.0×10^{-3}	4.4
Central/wide/low	51% (24.99)	22% (17.16)	36% (far)	8.7×10^{-4}	5.2×10^{-4}	1.7
			40% (middle)	7.1×10^{-3}	2.3×10^{-5}	3.1×10^2
			44% (close)	3.0×10^{-2}	5.2×10^{-7}	5.2×10^4
Extreme/narrow/high	82% (14.76)	98% (1.96)	92% (far)	2.4×10^{-3}	7.4×10^{-4}	3.3
			90% (middle)	1.1×10^{-2}	2.9×10^{-5}	3.8×10^2
			88% (close)	3.2×10^{-2}	7.3×10^{-7}	4.4×10^4
Extreme/wide/high	69% (21.39)	98% (1.96)	84% (far)	2.8×10^{-4}	1.6×10^{-10}	1.8×10^6
			80% (middle)	4.6×10^{-3}	1.1×10^{-14}	4.1×10^{11}
			76% (close)	2.8×10^{-2}	2.9×10^{-19}	9.7×10^{16}
Central/narrow/high	52% (24.96)	68% (21.76)	62% (far)	1.1×10^{-2}	3.7×10^{-2}	2.9×10^{-1}
			60% (middle)	2.2×10^{-2}	2.0×10^{-2}	1.1
			58% (close)	3.9×10^{-2}	9.0×10^{-3}	4.4
Central/wide/high	49% (24.99)	78% (17.16)	64% (far)	8.7×10^{-4}	5.2×10^{-4}	1.7
			60% (middle)	7.1×10^{-3}	2.3×10^{-5}	3.1×10^2
			56% (close)	3.0×10^{-2}	5.2×10^{-7}	5.2×10^4

P refers to the probability of drawing a sample ($N = 100$) with a given % from a population with a given π : $P(\text{sample\%}|\pi)$

*Variance of the sampling distribution from a dichotomous population with the given π where $N = 100$

(see Table 1). Half of the population π pairs were centered below 50% (*low-parity*), while the other half were centered above 50% (*high-parity*). Values of population π s and sample percentages were reflected under (low) and over (high) 50%. For example, if one question referred to π s of 2% and 18%, with a sample percentage of 10% (low parity), another referred to π s of 98% and 82%, with a sample percentage of 90% (high parity). High-parity conditions might be thought of as negative parity versions of low-parity conditions: “98% of Doco mango trees have white flowers” is logically equivalent to “2% of Doco mango trees do not have white flowers.” Populations with π s equally distant from 50% (e.g., 25% and 75%) are also equally variable. Thus, normatively, this manipulation should not affect participants’ judgments.

Sample location and presentation order Three different sample locations (*close, middle, far*) were used in each of the

eight centrality $\times$ spread $\times$ parity conditions. Sample percentages were closer to the inner π than to the outer π (*close*), further from the inner π than from the outer π (*far*), or in the middle between those two locations (*middle*) (see Table 1). While the relative likelihood that samples were drawn from the inner population was greater when sample percentages were closer to the inner π , this was not a true manipulation, since, as was discussed above, these proximity differences were varied between spread conditions. We also varied *presentation order*. For half of the trials, the inner population was described first, while for the other half, the outer population was described first. This yielded six questions sets per centrality $\times$ spread $\times$ parity condition (see Table 1).

Other controls Participants were always told that there were 100 trees in the grove and that groves of either population occurred with equal frequency.

Procedure

Participants followed a link to a Web page that included the following text:

In this study you will be given information about different types of trees. For example, Ukon cherry trees tend to have yellow blossoms. In contrast, Kanzan cherry trees tend to have pink blossoms. Suppose you see a grove where someone planted either all Ukon or all Kanzan trees. If you did not know which kind of tree was planted, you could use the color of the blossoms in the grove to make an inference. For example, if the blossoms were mostly yellow, you might guess that Ukon, rather than Kanzan, trees were planted. In this study you will be asked to make inferences about which of two types of trees seems more likely to have been planted in a grove based on the percent of blossoms that are a certain color.

Next, participants were given 48 pairs of randomly presented questions, constructed from the sets of π s and sample percentages described in Table 1. Each pair of questions was preceded by an information block, like the one below:

Mango trees can have either white or yellow flowers. 2% of Aoco mango trees have white flowers. 18% of Boco mango trees have white flowers. There are equal numbers of Aoco and Boco groves. You see a grove of 100 mango trees. This grove consists of either all Aoco trees or all Boco trees. You see that 8% of the trees have white flowers.

Participants then indicated which kind of grove this was more likely to be (e.g., “Is this more likely to be an Aoco mango grove or Boco mango grove?”). They also rated on a scale of 1–7 how sure they were of their answer. After completing these 48 question sets, participants completed a 10-question multiple-choice numeracy evaluation adapted from Lipkus, Samsa and Rimer (2001) that required conversions between percentages, proportions, and frequencies. Participants were also asked their math SAT scores and what math and/or statistics classes they had taken.

Results and discussion

Numeracy scores

Participants scored highly on the numeracy assessment ($M = 8.91$, $SE = 0.088$, mode = 10). Since the distribution of scores was highly skewed, it was necessary to divide the participants into two approximately even groups on the basis of their numeracy performance: those with perfect numeracy scores ($N = 89$, 40%) and those who made errors ($N = 132$, 60%). Individuals with perfect numeracy scores had higher SAT

scores than those who made errors ($M = 727$, $SE = 4.9$, vs. $M = 689$, $SE = 6.1$; $t = 4.25$, $p < .005$; 31 participants did not report SAT scores). Individual differences in numeracy have been shown to influence how people make use of statistical information (Reyna, Nelson, Han & Dieckmann, 2009) and, thus, might be expected to influence participants' judgments in this study.

Coding

Since analyses of both dependent measures yielded highly similar effects, we report only the forced choice data. We wished to determine (1) whether the frequency with which participants judged that a sample came from a population was influenced by the relative probability that the sample was drawn from that population and (2) whether factors that did not influence probability (parity and numeracy) moderated these effects. As can be seen in Table 1, the samples described in the stimuli were typically more likely to come from the inner, rather than the outer, population: Normatively, participants should select inner, rather than outer, populations in most cases. Thus, we coded participants' forced choice responses according to whether or not they indicated that a sample was more likely to have come from the inner population than from the outer population.

We predicted that the rate at which participants chose the inner populations would reflect the relative likelihoods of the samples with the stated percentages coming from the inner versus the outer populations. The relative probabilities of samples coming from the inner populations were greater when centrality was extreme, rather than central, since the difference in variance between the inner and outer populations was greater in the extreme condition. Additionally, the relative probabilities of samples coming from the inner populations were greater when spread was wide, rather than narrow, due both to absolute differences between sample percentages and population π s and to differences in population variance. In contrast, parity did not affect the probability of samples coming from a population. Thus, effects of centrality can be attributed to the normative influence of variance, effects of spread can be attributed to the normative influence of both variance and proximity, but effects of parity are nonnormative.

Analysis

We ran a mixed model ANOVA with centrality (central, extreme), spread (narrow, wide), and parity (low, high) as within-subjects factors and numeracy (perfect, imperfect) as a between-subjects factor. Participants were given scores corresponding to the proportion of the six forced choice questions in each of these eight conditions (three sample percentages $\times$ two presentation orders) for which they responded that the sample was more likely to have been

drawn from the inner population. All statistics discussed in this experiment refer to this analysis unless otherwise noted.

Variance influenced likelihood judgments Individuals were more likely to choose the inner population in the extreme than in the central conditions, $F(1, 219) = 52.2$, $p < .0005$, $\eta_p^2 = .19$, indicating that, indeed, variance influenced participants' likelihood judgments (see Fig. 1). Furthermore, 137 of 221 participants chose the inner population more often in the extreme condition, while only 42 showed the opposite pattern (binomial test, $p < .0005$). This offers evidence that (1) people have some sensitivity to the likelihood of samples being drawn from a given population and (2) people can use variance when making such determinations. However, although variance shifts people's behavior in the normative direction, an ideal observer would have selected the inner population in all of the extreme trials and 80% of the central trials. This was not the case (see Fig. 1).

Nonnormative influence of parity The inner population was chosen more often in the low-parity than in the high-parity conditions, $F(1, 219) = 144.5$, $p < .0005$, $\eta_p^2 = .40$. Participants were also more strongly affected by centrality when parity was low, $F(1, 219) = 11.2$, $p < .0005$, $\eta_p^2 = .05$ (see Fig. 1). This may be due to differences in how well people are able to represent large and small numerical values. Representations of larger numbers are more variable, overlapping with representations of more neighboring numbers than do representations of smaller numbers (Gallistel & Gelman, 2005). Thus, performance may have been less normative for high- versus low-parity trials because the smaller relative differences between values used on high-parity trials made it more difficult for participants to discriminate between them and, subsequently, the probabilities they conveyed.

Proximity influenced likelihood judgments more strongly than variance Analyses of the different levels of spread

indicate that the influence of implied variance, while present, was not precisely normative. The inner population was chosen more often in the wide than in the narrow conditions, $F(1, 219) = 276.1$, $p < .0005$, $\eta_p^2 = .56$, an effect that can be attributed to a sensitivity to proximity as well as variance (see Fig. 1). Spread also interacted marginally with centrality, $F(1, 219) = 3.7$, $p < .06$, $\eta_p^2 = .02$; the effect of spread was stronger in extreme conditions. This is consistent with the relative probability difference described in Table 1. No interaction between spread and parity was seen.

As planned, a second analysis tested whether (1) variance or (2) absolute differences between sample percentages and π s had more influence on people's choices. We ran a 2 (narrow-extreme vs. wide-central) $\times$ 2 (parity) $\times$ 2 (numeracy) mix model ANOVA that compared narrow-extreme with wide-central conditions. As was previously mentioned, absolute and relative probabilities were closely matched in these conditions. Thus, normatively, no effect of the narrow-extreme versus wide-central condition should be expected. However, in narrow conditions, sample percentages were, on average, equidistant from the inner and outer population π s, while in wide conditions, sample percentages were, on average, closer to the inner population π s. Thus, if the participants were more strongly influenced by these proximity differences than by variance, they would tend to choose the inner population more frequently in the wide-central than in the narrow-extreme conditions. Indeed, this effect was observed, $F(1, 219) = 80.6$, $p < .0005$, $\eta_p^2 = .27$. The effect of parity also remained significant, $F(1, 19) = 119.6$, $p < .0005$, $\eta_p^2 = .35$.

Furthermore, the effect of narrow-extreme versus wide-central populations was stronger in the high-parity conditions, $F(1, 219) = 9.5$, $p < .005$, $\eta_p^2 = .04$. These findings align with previous research (Obrecht et al., 2007), indicating that differences in means have a stronger influence on people's decisions than do differences in

Fig. 1 Experiment 1: Mean proportions of inner population responses for questions in the eight different centrality-spread-parity conditions (see Table 1) for participants with both perfect and imperfect numeracy scores. Error bars represent standard errors

variance. However, one must consider that while sample percentages and π s were given explicitly in this experiment, variance information was not; instead, it had to be derived. Thus, it is possible that proximity had a greater effect than variance simply because it was more transparent. Of interest, while numeracy was not significant in this analysis, it did interact with parity, $F(1, 219) = 7.9, p < .01, \eta_p^2 = .03$. Less numerate individuals were more strongly influenced by parity (note: this is parallel to the effect of parity seen in the main analysis, discussed below). This analysis found no other significant interactions.

Individual differences Individual differences in numeracy have been shown to influence how people make use of statistical information (Reyna et al., 2009). Participants with perfect scores were more likely to choose the inner population, typically the normative response, $F(1, 219) = 4.4, p < .05, \eta_p^2 = .02$. Furthermore, such participants were more strongly influenced by numerical factors that affected likelihood [interaction between numeracy and spread, $F(1, 219) = 8.3, p < .005, \eta_p^2 = .04$; interaction between numeracy and centrality, $F(1, 219) = 3.4, p < .07, \eta_p^2 = .02$, marginally significant] and were less strongly influenced by factors that did not affect likelihood [interaction between numeracy and parity, $F(1, 219) = 7.6, p < .01, \eta_p^2 = .03$; see Fig. 1]. No other interactions were significant. Effects of numeracy were quite robust. When various other subdivisions of numeracy scores were tested, effects of numeracy and interactions between numeracy and spread and between numeracy and parity were still seen. In line with previous work (Obrecht et al., 2009), these findings suggest that more numerate individuals are more strongly influenced by numerical factors that affect probability than are less numerate individuals. Interestingly, another ANOVA found no effect of having taken a statistics course when this factor, rather than numeracy, was included as the between-subjects variable, $F(1, 215) = .01, p > .9$. It appears that statistical training did not affect performance on this task.

Conclusions and caveats

The results of this study support our hypothesis that laypeople have numerically based intuitions about the chances of samples coming from a given population. This was seen even when the relative probability of a sample having come from one population or another had to be derived solely from the variability of a binomial distribution as implied by given sample percentages and population π s. Furthermore, participants' numerical ability affected how statistical features influenced these judgments. Highly numerate individuals were more strongly influenced by numerical factors that affected likelihood (spread and

centrality) and were less strongly influenced by factors that did not affect likelihood (parity). However, we have yet to show that such intuitions can influence the way individuals make use of sample size when combining sample information. We explored this possibility in [Experiment 2](#).

Experiment 2

In [Experiment 2](#), we tested whether the implication that samples were or were not drawn from the same population affected how participants used sample size when combining sample information. When estimating π on the basis of samples drawn randomly from the general population, data should be weighted by sample size. When samples come from different subpopulations, one may, instead, combine estimates of subpopulations' π s (typically, the sample percentages) on the basis of the subpopulations' frequency within the general population or, lacking this information, give them equal weight. Therefore, it would make sense to vary the extent to which sample size is weighted on the basis of the likelihood that samples were drawn randomly from the same population. In this experiment, we manipulated whether samples were implied to have been drawn randomly from the same population or from different subpopulations in two ways: verbally and via statistical information.

Method

Participants

Undergraduate students at Rutgers University, New Brunswick ($N = 224$) participated for course credit. Participants completed the same numeracy assessment as that used in [Experiment 1](#) as part of a prescreening questionnaire that was administered to the undergraduate participant pool.

Design

This study was conducted online. Participants read six stories, each of which described the observations of six caretakers who collected samples about a given kind of animal. Within a story, participants were to combine information from the six observations to make an estimate of the percentage of animals that have a feature of interest. For instance, six caretakers gave data about how many leopards they had observed (sample size) and what percent of them had a feature (e.g., round markings). The six sample sizes provided within each story were always the same (1, 2, 3, 5, 80, 250). The percentage/sample-size combinations were possible given the sample size. Information from each caretaker was given on its own Web page.

The verbally implied likelihood that samples were drawn from the same *population* was varied between subjects, while the *range* of the sample percentages and the *location* of the sample's weighted mean relative to the unweighted mean were varied within subjects. We explain these factors below. Additionally, counterbalancing variables were employed between subjects. Stories were presented in two different orders (*story order*), data sets and animal kinds were paired in two different ways (*data pairing*), and the order of sample presentation within datasets varied in two different ways (*data order*). This resulted in a 2 (population) $\times$ 3 (range) $\times$ 2 (weighted mean location) $\times$ 2 (story order) $\times$ 2 (data pairing) $\times$ 2 (data order) mixed design. Participants were randomly assigned to the different between subjects conditions.

Population We verbally manipulated whether samples were implied to have come from the same or different populations. We intended for the participants in the *same-population* condition to infer that samples had come from the same population. We intended for participants in the *different-population* condition to infer that the samples might represent different subpopulations. This difference was manipulated between subjects so as to not be apparent.

In the same-population condition, the data were provided from generic caretakers at the Mozambique Nature Preserve (e.g., “One of the caretakers tells you that of the 80 leopards he has seen, 97% had round markings. Another nature preserve worker says that . . .”). This condition replicated Experiment 3 in Obrecht (2010). The data descriptions in the different-population condition were identical to those used in the same population condition, except that caretakers were described as being from nature preserves in different countries (e.g., “An Egyptian caretaker tells you . . .”, “A Nepalese nature preserve worker says that . . .”). The population manipulation allowed us to test whether people are sensitive to issues of sampling, as implied by verbal information, when integrating statistical data.

Range The range of the sample percentages presented within a story was *low* (0%–33%), *central* (33%–66%), or *high* (66%–100%) for sample sizes greater than 2 (see Table 2). Recall that binomial populations are more variable when percentages are closer to 50%. Thus, samples were more likely to have come from the same population in the central-range condition than in the low- or high-range condition. This allowed us to test whether people are sensitive to sampling issues when integrating statistical information when the possibility of sampling from different subpopulations is implied by numerical information alone—specifically, by implied population variance.

Weighted mean location We manipulated whether the weighted mean of sample percentages was *higher* or *lower* than the unweighted mean (only samples' sizes and percentages were presented, not the means themselves). In half the data sets, high percentages were paired with high sample sizes, and thus the weighted mean of the six percentages was *higher* than the unweighted mean (the mean where sample size was ignored). In the rest of the data sets, high percentages were paired with lower sample sizes, and thus the weighted mean percentages were *lower* than the unweighted mean percentages (see Table 2 and equations below). This manipulation allowed us to further test whether numerically implied variability affects sample size use. Participants should give probability judgments that reflect greater use of sample size when a weighted mean is closer to 50% (and thus indicative of a more variable population) than the unweighted mean.

Procedure

At the start of the experiment, participants read, “Imagine that you work at the Mozambique Nature Preserve where many animals live. You are interested in learning more about the animals that you help.” Participants assigned to

Table 2 Data sets used for stories in the various range and weighted mean location conditions in Experiment 2, along with resulting weighted and unweighted means

Range	Weighted Mean Location	Weighted Mean of Stimuli	Unweighted Mean of Stimuli	Stimuli: Sample Percents Paired With Sample Sizes (<i>N</i>)					
				<i>N</i> = 1	<i>N</i> = 2	<i>N</i> = 3	<i>N</i> = 5	<i>N</i> = 80	<i>N</i> = 250
Low	Lower	5.4%	19.4%	0%	50%	33%	20%	10%	3%
Low	Higher	20.2%	5.8%	0%	0%	0%	0%	11%	24%
Central	Lower	40.2%	59.8%	100%	50%	66%	60%	44%	38%
Central	Higher	59.6%	40.8%	0%	0%	66%	60%	57%	61%
High	Lower	80.2%	95.2%	100%	100%	100%	100%	97%	74%
High	Higher	94.8%	80.4%	100%	50%	66%	80%	88%	98%

the same-population condition then read, “At a staff meeting you get a chance to talk to other nature preserve workers who have carefully recorded the chances of various outcomes.” Those assigned to the different-population condition instead read, “At a zoology conference you get a chance to talk to other nature preserve workers who have carefully recorded the chances of various outcomes.” Throughout the experiment, participants in the same-population condition were told of their co-workers observations. Those in the different-population condition were, instead, given sample data from individuals from various countries. Since the samples would thus have been drawn not only from different nature preserves, but also from different geographic locations, we expected that this manipulation would increase the participants’ suspicions that these samples might not be representative of the same population.

Next, participants read six stories in which an animal could have one of two possible outcomes for a given characteristic. Within each story, six different nature preserve caretakers stated how many animals they had seen with the outcome of interest. Participants were then asked to estimate the chances that an animal (e.g., leopard) born at the Mozambique nature preserve would have the outcome of interest (e.g., round markings). Participants gave a percent estimate from 0 to 100. Additionally, they were asked to rate this probability on a 9-point scale from *extremely unlikely* to *extremely likely*.

Results and discussion

Distribution of responses

The probability ratings were redundant with percent estimates; thus, we report only the latter measure. Responses tended to diverge from the weighted mean in the direction of the unweighted mean (see Table 3). Since this was the case even when the weighted mean was more

central than the unweighted mean, this response pattern cannot be attributed to people’s tendency to overestimate small probabilities and underestimate large probabilities (e.g., Gonzalez & Wu, 1999).

Under that account, responses would be predicted to be more central than the weighted mean, even in cases where the weighted mean was more central than the unweighted mean. Additionally, responses were more like the unweighted mean when the weighted mean was more extreme (see Table 3). This is consistent with our prediction that participants would give less weight to sample size when samples were less likely to have come from the same population. Analyses were run to confirm that this was the case.

Scaling

In the following analyses, we considered two possible methods by which participants might average sample information to produce population estimates: (1) weighting percentages according to their sample sizes, as would be correct if samples were drawn randomly from the same population (see Eq. 1), and (2) giving all percentages equal weight, as would be correct if samples were, instead, drawn from different subpopulations whose frequencies in the general population were unknown and one estimated the π s of these subpopulation as the sample means (see Eq. 2).

$$\text{Weighted mean} = \sum_i^6 \text{percent}_i \left(\frac{N_i}{\sum N_i} \right) \quad (1)$$

$$\text{Unweighted mean} = \frac{\sum \text{percents}}{\text{number_of_samples}} \quad (2)$$

Accordingly, participants’ percent estimates were scaled to indicate the extent to which they were consistent with the weighted mean versus the unweighted mean. Responses equal to the weighted mean were coded as 1, and responses

Table 3 Central tendency of responses in Experiment 2 and the number of responses within 1% of the weighted mean and unweighted means, in the various range and weighted mean location conditions

Weighted Mean of Stimuli	Unweighted Mean of Stimuli	Central Tendency			# Responses Within 1% of Weighted Mean ($N = 224$)	# Responses Within 1% of Unweighted Mean ($N = 224$)
		Mean Response	Median Response	Modal Response		
5.4%	19.4%	23.6%	20%	10%	30	22
20.2%	5.8%	17.9%	15%	20%	41	23
40.2%	59.8%	49.7%	50%	40%	47	27
59.6%	40.8%	52.7%	57%	60%	64	14
80.2%	95.2%	85.7%	90%	90%	30	38
94.8%	80.4%	81.2%	85%	90%	24	39

equal to the unweighted mean were coded as 0. All other responses were scaled such that the scaled values' relative proximity to 1 and 0 mirrored the responses' relative proximity to the weighted and unweighted means. For example, if the weighted and unweighted means for a data set were 20% and 5% (respectively), a response of 8% would be scaled as 0.2 ($20\% - 5\% = 15\%$; $8\% - 5\% = 3\%$; $3\%/15\%=0.2$); indicating that the participant gave 0.2 as much weight to sample size as he should have, assuming that the data were randomly selected from the same population. Similarly, a response of 17% would be scaled as 0.8, and 23% would be scaled as 1.2 (indicating overweighting sample size). Conversely, if the unweighted mean was 20% and the weighted mean was 5%, then 8% would be scaled as 0.8, 17% would be scaled as 0.2, and 23% would be scaled as -0.2. Thus, the scaled responses become more like the weighted mean and less like the unweighted mean as values increase. This allowed us to directly compare across conditions the extent to which participants' responses differed from those predicted by the two averaging models.

Numeracy scores

Participants on average answered 7.66 ($SE = 0.144$) of the 10 numeracy questions correctly; the modal score was 9. Since the distribution of scores was less skewed than in Experiment 1, we divided participants into five groups on the basis of the quintile numeracy scores. We henceforth refer to these 5 numeracy levels as *lowest* (1–5 correct, $N = 32$), *low-middle* (6–7 correct, $N = 47$), *middle* (8 correct, $N = 42$), *high-middle* (9 correct, $N = 68$), and *highest* (10 correct, $N = 35$).

Analysis

A 2 (population) $\times$ 3 (range) $\times$ 2 (weighted mean location) $\times$ 2 (data-pairing) $\times$ 2 (data order) $\times$ 5 (numeracy level) mixed model ANOVA was run on the scaled percent estimate responses. There was an uneven distribution of participants with a given numeracy level among the various counterbalancing conditions. Thus, to ensure that there were no missing cells, the story order counterbalancing factor was not included in the model. We determined that this would be the factor least likely to affect responses, since it did not alter the presentation of the individual stories. Indeed, an analysis including this factor (and hence, with missing cells) yielded results similar to those described below. Note that all means reported are marginal means.

Population condition influenced responses As was predicted, participants in the same-population condition gave responses that were more consistent with the weighted

mean model than did participants in the different-population condition, $F(1, 181) = 4.1, p < .05, \eta_p^2 = .022$. This suggests that people indeed give greater consideration to sample size when samples are suspected to come from the same population, rather than from different subpopulations. In the same-population condition, participants' mean scaled response was .508 ($SE = .047$); this is not close to the normative value of 1 but is closer than .369 ($SE = .050$), the mean scaled response in the different-population condition (see Fig. 2).

To ensure that collapsing across raw numeracy scores and story order conditions did not yield erroneous findings, we confirmed this result via a nonparametric test. Mean scaled responses of participants in the two population conditions were compared for each combination of range, weighted mean location, story order, data pairing, data order, and raw numeracy scores. For 167 of these conditions, participants in the same-population condition favored the weighted mean model more than did participants in the different-population condition; the reverse pattern was found in only 98 cells (binomial test, $p < .03$).

Numerical variance influenced participants' responses

More varied populations yield more varied samples. Thus, when weighted means indicated more varied populations, the samples were more likely to come from those indicated populations. We predicted that participants would give more weight to sample size in such conditions. Indeed, responses were most consistent with the weighted mean model in the central-range condition, where weighted means were closest to 50% and, thus, denoted more variable populations [scaled response $M_{\text{central}} = .564, SE = .038$; $M_{\text{low}} = .324, SE = .057$; $M_{\text{high}} = .427, SE = .047$; $F(2, 362) = 8.6, p < .0005, \eta_p^2 = .045$; see Figs. 2 and 3]. A binomial test confirmed this effect: 98 of the 221 participants who responded to all stories had their highest

Fig. 2 Experiment 2: Marginal means of scaled percent estimates as a function of population and range. Error bars represent standard errors

Fig. 3 Experiment 2: Marginal means of scaled percent estimates as a function of range and weighted mean location. Error bars represent standard errors

mean scaled responses in the central- (rather than the low- or high-) range conditions [chance = 74 (33.3%), $p < .001$]. Furthermore, participants in the different-population condition were more affected by the range of the percentages than were those in the same-population condition, $F(2, 362) = 3.2$, $p < .05$, $\eta_p^2 = .017$ (see Fig. 2). This suggests that people may be particularly likely to use numerical information in determining how to weight sample size when nonnumerical information implies that samples may come from different subpopulations.

Additionally, responses were least consistent with the weighted mean when it was furthest from 50% and, consequently, indicated less variable populations. When the weighted means were 5.4% (low-range-lower weighted mean condition) and 94.8% (high-range-higher weighted mean condition), participants gave estimates that most closely resembled the unweighted mean [interaction between range and weighted mean location: $F(2, 362) = 79.8$, $p < .0005$, $\eta_p^2 = .306$; see Fig. 3].

Effect of weighted mean location Responses were more consistent with the weighted mean model when the weighted mean was higher than the unweighted mean, $F(1, 181) = 16.7$, $p < .0005$, $\eta_p^2 = .085$ (see Fig. 3). A binomial test confirmed this: 131 of the 221 participants that gave percent estimates for all stories had greater mean scaled responses in the higher conditions than in the lower conditions [chance = 111 (50%), $p < .01$]. We speculate that this may be because, since representations of larger numbers are more variable (Gallistel & Gelman, 2005), participants may ascribe more variance to populations whose means are described by larger numbers.

Individual differences Participants' responses were significantly influenced by numeracy, $F(4, 181) = 7.0$, $p < .0005$, $\eta_p^2 = .134$). Participants at the highest and high-middle numeracy levels had mean scaled percent responses that were most consistent with the weighted mean model (.686, $SE = .079$, and .551, $SE = .073$, respectively), while participants at the middle and low-middle numeracy levels had responses that were most consistent with the unweighted mean (.415, $SE = .074$, and .176, $SE = .068$, respectively). This suggests that, in accordance with previous results (Obrecht et al., 2009), more numerate participants used sample size more extensively. However, scaled responses of the participants at the lowest numeracy level (.364, $SE = .087$) were more consistent with the weighted mean model than were those in the low-middle group. We speculate that perhaps these lowest numeracy scores are indicative of participants' not engaging in the numeracy assessment, rather than indicative of low numeracy. Indeed, the answering pattern of 31% of the participants in that group was highly suspect, either giving the same response to most numeracy questions or alternating between two responses.

Numeracy level also interacted with weighted mean location, $F(4, 181) = 2.6$, $p < .05$, $\eta_p^2 = .055$. Similar to the interaction between parity and numeracy in Experiment 1, the effect of weighted mean location was least pronounced for the participants with high-middle and high numeracy scores. Additionally, there were interactions between numeracy, population, and range, $F(8, 362) = 2.0$, $p < .05$, $\eta_p^2 = .042$, between numeracy, population, and weighted mean location, $F(4, 181) = 2.1$, $p < .05$, $\eta_p^2 = .064$, between numeracy, range, and weighted mean location, $F(8, 362) = 2.3$, $p < .05$, $\eta_p^2 = .048$, and between numeracy, population, range, and weighted mean location, $F(8, 362) = 2.3$, $p < .05$, $\eta_p^2 = .048$. Participants at lower numeracy levels showed greater differences in their scaled responses for different levels of population variability implied by the means than did participants at higher numeracy levels, particularly those in the different-, rather than the same-, population condition. Perhaps more numerate participants were more likely to weight means by sample size (as they have been taught to do), regardless of implied probability that samples are not representative (see Table 4). Additionally, there were interactions with the counterbalancing factors that we believe are attributable to individual differences between groups' participants (see the Appendix). No other main effects or interactions were significant.

Conclusions

The results of this experiment support our hypothesis that laypeople adjust the extent to which they weight sample

Table 4 Experiment 2: Marginal means of the scaled percent responses in the various population, range, weighted mean location, and numeracy conditions

Population	Range	Weighted Mean Location	Marginal Mean (<i>SE</i>) of Scaled Responses for the Various Numeracy Condition				
			Lowest	Low-Middle	Middle	High-Middle	Highest
Same	Low	Lower	−0.720 (0.318)	−0.588 (0.27)	0.038 (0.319)	−0.308 (0.356)	0.194 (0.322)
			0.902 (0.242)	0.844 (0.205)	0.914 (0.243)	1.659 (0.271)	1.278 (0.245)
		Higher	0.321 (0.154)	0.399 (0.130)	0.611 (0.154)	0.570 (0.172)	0.660 (0.156)
			0.725 (0.179)	0.580 (0.152)	0.565 (0.180)	0.303 (0.201)	0.763 (0.182)
	Central	Lower	0.642 (0.190)	0.496 (0.161)	0.858 (0.191)	1.915 (0.213)	1.015 (0.193)
			0.063 (0.242)	−0.406 (0.205)	0.458 (0.243)	−0.012 (0.271)	0.488 (0.245)
		Higher	1.009 (0.321)	0.744 (0.235)	0.645 (0.240)	0.632 (0.201)	1.349 (0.267)
			0.470 (0.204)	0.070 (0.150)	0.460 (0.152)	0.745 (0.128)	0.617 (0.170)
	High	Lower	0.799 (0.238)	0.680 (0.175)	0.594 (0.178)	0.646 (0.149)	0.697 (0.198)
			1.650 (0.253)	0.431 (0.185)	0.265 (0.189)	0.579 (0.158)	0.564 (0.210)
		Higher	−1.000 (0.322)	−0.082 (0.236)	0.139 (0.240)	−0.027 (0.202)	0.511 (0.268)
Different	Low	Lower	−0.49 (0.423)	−1.057 (0.310)	−0.565 (0.315)	−0.088 (0.265)	0.096 (0.352)
			1.009 (0.321)	0.744 (0.235)	0.645 (0.240)	0.632 (0.201)	1.349 (0.267)
		Higher	0.470 (0.204)	0.070 (0.150)	0.460 (0.152)	0.745 (0.128)	0.617 (0.170)
			0.799 (0.238)	0.680 (0.175)	0.594 (0.178)	0.646 (0.149)	0.697 (0.198)
	Central	Lower	1.650 (0.253)	0.431 (0.185)	0.265 (0.189)	0.579 (0.158)	0.564 (0.210)
			−1.000 (0.322)	−0.082 (0.236)	0.139 (0.240)	−0.027 (0.202)	0.511 (0.268)
		Higher					
	High	Lower					
		Higher					

size when combining sample data on the basis of the implication that those data have been drawn from the same population. Participants who were told that all samples came from the same location (and presumably the same population) gave more weight to sample size when combining sample information than did participants who were told that samples came from various locations (and presumably, different subpopulations). Furthermore, participants gave less weight to sample size when numerical information alone implied that the samples were unlikely to come from the same population.

General discussion

Summary of findings

In Experiment 1, we demonstrated that laypeople have numerically based intuitions about the probability of samples' coming from populations, even when such intuitions must be derived solely from implied binomial

variability. In Experiment 2, we demonstrated that laypeople's use of sample size is moderated by the implied likelihood that samples were randomly drawn from the same population. Participants who were told that all samples were drawn from the same location gave more weight to sample size when combining sample information than did participants who were told that samples had been drawn from different locations. This experiment also replicated the results of Obrecht (2010), which suggested that participants gave less weight to sample size when numerical information implied that samples were unlikely to come from the same population. Furthermore, sensitivity to these statistical factors reflects individual differences in numerical ability. Taken together, these findings suggest that people's sensitivity to the representativeness of samples may be much richer than previously thought. People do not only use sample size to determine how representative a sample may be. They are also influenced by probabilistic information that is indicative of the sampling method, which subsequently influences the extent to which sample size is utilized when statistical judgments are made.

Implications for the interpretation of prior research

Weighting means by sample size is statistically correct when samples have been drawn randomly from the population of interest, but not when samples come from different subpopulations. Often, sampling methods are unknown. Thus, it is reasonable to use both numeric and nonnumeric cues to judge the likelihood that samples come from different subpopulations, in order to determine the relevance of sample size. Consequently, a person who varies the weight they give to sample size on the basis of this certainty would give less weight to sample size than would be considered normative. This may explain prior findings that sample size is underweighted relative to normative standards (e.g., Obrecht, 2010), as well as findings that sample size is given little or no weight when anecdotes or individual cases are presented along with summary data (Obrecht et al., 2009; Ubel et al., 2001). If laypeople think that presenting samples separately implies that samples come from separate subpopulations, “underweighting” sample size is justifiable. Furthermore, given the findings of Experiment 1 and 2, one might predict that laypeople would be particularly likely to conclude that samples come from different subpopulations when numeric information indicates that the data are less likely to be drawn randomly from the same population.

To illustrate, consider Obrecht et al. (2009). In this study, participants were told that a report found that 30 out of 1,000 radios tested (3%) broke within a year and also were given statements from radio owners of whom 3 out of 4 or 12 out of 16 claimed that the product broke. When judging whether radios would break, participants did not weight percentages by sample size, as the authors considered normatively correct. However, the normative method of weighting means by sample size assumes that data represent a random sampling of the population. The results of Experiment 1 and 2 suggest that laypeople can intuit that the randomness of this sampling is in question just from the numbers: With a population π of 3%, there would be less than a 0.1% chance of even 3 out of 4 sampled radios breaking, and less than a 0.0000000001% chance that 12 out of 16 would break. Thus, it would be legitimate for people to conclude that the radios tested in the report were different from those that the customers were buying.

This highlights a basic methodological issue inherent in such experimental designs. In order to conclude that participants are not weighting data by sample size, there must be a distinction in the means produced when data are weighted by sample size and when they are not. The more different these means are, the better a researcher would be able to distinguish the extent to which participants utilize sample size. However, as the difference between these means increases, the likelihood that

these samples would be randomly drawn from the same population decreases. Consequently, scenarios where one can best detect that sample size is ignored are also scenarios where it is most likely that sample size should be ignored.

Future directions

Although we have shown that people have intuitions about the likelihood that samples were drawn randomly from a population, we have not addressed how various factors may influence this perceived likelihood. It appears that variance and numeracy are influential, but this is not an exhaustive list. Additionally, further research is needed to determine how the perceived likelihoods that samples came from populations moderate how sample sizes influence judgments. For example, we do not know whether people have some likelihood threshold below which sample size is not considered or whether, rather, sample size is continuously given less weight as perceived likelihoods decrease. Furthermore, merely considering the extent to which participants may weight information by sample size fails to take into account the range of possible methods by which people can and should combine data. For example, when sample data come from a subpopulation whose base rate is known, this base rate information should impact how samples are integrated. Also, it may not always be valid to assume (as our unweighted mean model does) that people’s estimates of subpopulations’ means are equal to sample means. People may base such estimates on prior knowledge, particularly when sample sizes are small. More studies are needed to address these issues.

Author Note This research was partially supported by a Summer Research Stipend awarded to the second author by the Research Center for the Humanities and Social Sciences at William Paterson University.

We thank A. Chapman, R. Gelman, P. Mathews, N. McNeil, and L. Peterson for their help and support.

Appendix

The data order and data pairing counterbalancing conditions neither yielded main effects nor interacted with each other, but did interact with other factors in the model (see Table 5). However, analyses using different inclusion criteria reverse the direction of these effects, although they remain significant. In contrast, effects of range, location, numeracy, and population remain stable. Thus, we speculate that these interactions reflect individual differences between groups’ participants. While it is necessary to account for such factors in analyses, they are not presently of interest.

Table 5 Interactions of various experimental factors in Experiment 2 with the data paring and data order counterbalancing conditions

Interacting Factors	× Data paring $F(\eta_p^2)$	× Data order $F(\eta_p^2)$	× Data paring × data order $F(\eta_p^2)$
Range	$F(2, 362) = 2.7 (.02)$	$F(2, 362) = 0.4 (.00)$	$F(2, 363) = 6.0^{**} (.03)$
Population × range	$F(2, 362) = 2.6 (.01)$	$F(2, 362) = 0.6 (.00)$	$F(2, 362) = 4.7^{**} (.03)$
Population × location	$F(1, 181) = 0 (.00)$	$F(1, 181) = 0.7 (.00)$	$F(1, 181) = 5.1^* (.03)$
Range × location	$F(2, 362) = 3.4^* (.02)$	$F(2, 362) = 1.9 (.01)$	$F(2, 362) = 0.9 (.00)$
Range × numeracy	$F(8, 362) = 2.0 (.04)$	$F(8, 362) = 1.9 (.04)$	$F(8, 362) = 2.6^{**} (.06)$
Location × numeracy	$F(4, 181) = 2.7^* (.06)$	$F(4, 181) = 2.0 (.04)$	$F(4, 181) = 3.5^{**} (.07)$
Population × range × location	$F(2, 362) = 0.9 (.00)$	$F(2, 362) = 6.6^{**} (.03)$	$F(2, 362) = 1.9 (.01)$
Population × range × numeracy	$F(8, 362) = 1.7 (.04)$	$F(8, 362) = 2.0^* (.04)$	$F(8, 362) = 1.2 (.03)$
Population × location × numeracy	$F(4, 181) = 3.8^{**} (.08)$	$F(4, 181) = 1.6 (.03)$	$F(4, 181) = 2.2 (.05)$
Range × location × numeracy	$F(8, 362) = 2.2^* (.05)$	$F(8, 362) = 2.7^{**} (.06)$	$F(8, 362) = 1.2 (.03)$
Population × range × location × numeracy	$F(8, 362) = 2.0^* (.04)$	$F(8, 362) = 2.0^* (.04)$	$F(8, 362) = 2.8^{**} (.06)$

Main effects and interactions not described in the table or in the text are not significant.

* $p < .05$

** $p < .01$

References

- Evans, J. St. B. T., & Dusiior, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgments. *Acta Psychologica*, 41, 129–137. doi:10.1016/0001-6918(77)90030-0
- Evans, J. St. B. T., & Pollard, P. (1985). Intuitive statistical inferences about normally distributed data. *Acta Psychologica*, 60, 57–71. doi:10.1016/0001-6918(85)90013-7
- Gallistel, C. R., & Gelman, R. (2005). Mathematical cognition. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 559–588). Cambridge: Cambridge University Press.
- Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38, 129–166.
- Irwin, W. F., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an “expanded judgment” situation. *Journal of Experimental Psychology*, 51, 261–268. doi:10.1037/h0041911
- Jacobs, J. E., & Narloch, R. H. (2001). Children’s use of sample size and variability to make social inferences. *Journal of Applied Developmental Psychology*, 22, 311–331. doi:10.1016/S0193-3973(01)00086-7
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454. doi:10.1016/0010-0285(72)90016-3
- Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science*, 21, 1134–1140. doi:10.1177/0956797610376652
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21, 37–44. doi:10.1177/0272989X0102100105
- Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development*, 79, 1032–1048. doi:10.1111/j.1467-8624.2008.01174.x
- Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90, 339–363. doi:10.1037/0033-295X.90.4.339
- Obrecht, N. A. (2010). *Sample size weighting in probabilistic inference* (Unpublished doctoral dissertation). New Brunswick, NJ: Rutgers University.
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t*-tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, 14, 1147–1152. doi:10.3758/BF03193104
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition*, 37, 632–643. doi:10.3758/MC.37.5.632
- Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category based induction. *Psychological Review*, 97, 185–200. doi:10.1037/0033-295X.97.2.185
- Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, 68, 29–46. doi:10.1037/h0024722
- Reyna, V. F., Nelson, W., Han, P., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, 135, 943–973. doi:10.1037/a0017327
- Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typically in inductive reasoning: Evidence for developmental change. *Cognition*, 108, 543–556. doi:10.1016/j.cognition.2008.03.002
- Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making*, 11, 281–301. doi:10.1002/(SICI)1099-0771(1998120)11:4<281::AID-BDM302>3.0.CO;2-U.
- Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, 10, 33–51. doi:10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6.
- Sedlmeier, P., & Gigerenzer, G. (2000). Was Bernoulli wrong? On intuitions about sample size. *Journal of Behavioral Decision Making*, 13, 133–139. doi:10.1002/(SICI)1099-0771(200001/03)13:1<133::AID-BDM352>3.0.CO;2-H.
- Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decisions aids: Effects on treatment choices. *Medical Decision Making*, 21, 60–68.
- Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 112, 97–104. doi:10.1016/j.cognition.2009.04.006