Appalachian
STATE UNIVERSITY.
BOONE, NORTH CAROLINA

Sample Size Bias in Judgments of Perceptual Averages

Authors
Paul C. Price and Nicole M. Kimura, **Andrew R. Smith** and Lindsay D. Marshall

Abstract
Previous research has shown that people exhibit a sample size bias when judging the average of a set of stimuli on a single dimension. The more stimuli there are in the set, the greater people judge the average to be. This effect has been demonstrated reliably for judgments of the average likelihood that groups of people will experience negative, positive, and neutral events (Price, 2001; Price, Smith, & Lench, 2006) and also for estimates of the mean of sets of numbers (Smith & Price, 2010). The present research focuses on whether this effect is observed for judgments of average on a perceptual dimension. In 5 experiments we show that people's judgments of the average size of the squares in a set increase as the number of squares in the set increases. This effect occurs regardless of whether the squares in each set are presented simultaneously or sequentially; whether the squares in each set are different sizes or all the same size; and whether the response is a rating of size, an estimate of area, or a comparative judgment. These results are consistent with a priming account of the sample size bias, in which the sample size activates a representation of magnitude that directly biases the judgment of average.

Sample Size Bias in Judgments of Perceptual Averages

Paul C. Price and Nicole M. Kimura
California State University, Fresno

Andrew R. Smith and Lindsay D. Marshall
Appalachian State University

Previous research has shown that people exhibit a sample size bias when judging the average of a set of stimuli on a single dimension. The more stimuli there are in the set, the greater people judge the average to be. This effect has been demonstrated reliably for judgments of the average likelihood that groups of people will experience negative, positive, and neutral events (Price, 2001; Price, Smith, & Lench, 2006) and also for estimates of the mean of sets of numbers (Smith & Price, 2010). The present research focuses on whether this effect is observed for judgments of average on a perceptual dimension. In 5 experiments we show that people's judgments of the average size of the squares in a set increase as the number of squares in the set increases. This effect occurs regardless of whether the squares in each set are presented simultaneously or sequentially; whether the squares in each set are different sizes or all the same size; and whether the response is a rating of size, an estimate of area, or a comparative judgment. These results are consistent with a priming account of the sample size bias, in which the sample size activates a representation of magnitude that directly biases the judgment of average.

Keywords: judgments of average, perceptual judgment, size judgment, numerosity perception

People make judgments about averages in many different contexts and for many different purposes. For example, a teacher might judge the average mathematical ability of her students in deciding how best to teach them. Or a hospital patient might judge the average number of headaches he gets per month in response to a physician's question. Or a football coach might judge the average size or speed of an opposing defense in deciding what plays to call. Although a long line of psychological research on judgments of averages has shown that they tend to be accurate (Alvarez, 2011; Peterson & Beach, 1967), we have recently found that they also exhibit a curious bias. Specifically, they tend to increase as a function of the sample size. We have observed this sample size bias in judgments of average risk and likelihood for groups of people (Price, 2001; Price, Smith, & Lench, 2006) and also in estimates of the mean of sets of numbers (Smith & Price, 2010). In the present studies, we extend this basic result to judgments of averages on a perceptual dimension of a stimulus—the size of squares—and test several possible moderators of the effect. As in our previous research, we find not only that people exhibit the sample size bias but also that it is quite robust across a wide variety of conditions. We argue further that the robustness of the sample size bias across stimuli, stimulus presentation modes, dimensions of judgment, and response formats suggests that it is the result of a very basic and general cognitive process—most likely a form of priming. This, in turn, suggests possible connections among conceptually similar phenomena in the literatures on judgment and decision making and quantitative cognition and perception more generally.

The Sample Size Bias Phenomenon

The original impetus for studying the sample size bias was the social judgment phenomenon of unrealistic optimism. People generally judge themselves to be at lower risk than their peers for experiencing negative life events like developing cancer, being hurt in an accident, or getting divorced (e.g., Weinstein, 1980, 1987). In much of this research, however, the distinction between self and peers is confounded with sample size. Judgments about oneself are judgments about a small sample and judgments about one's peers are judgments about a large sample. Our goal was to eliminate this confound and study the effect of sample size on risk judgments directly. In one study, participants read a series of descriptions of the employees at fictional companies in terms of their risk factors for having a heart attack (Price, 2001). After reading descriptions of one, five, or nine employees at each company, participants judged the heart-attack risk of the typical employee at that company. As hypothesized, these risk judgments increased as a function of the number of employees. We then generalized this result in a number of ways in a series of follow-up studies (Price et al., 2006). For example, participants saw photo-graphs of groups of five, 10, and 15 peers and judged the likelihood that the average group member would experience various negative, neutral, and positive events. Again, as hypothesized, these likelihood judgments increased as a function of the number of people in the group. In the final study, the stimuli were groups of stick figures, the judgment was of their average height, and again a sample size bias was observed.

These results were intriguing given that earlier research on judgments of averages—primarily using numbers as stimuli—had

found such judgments to be quite accurate across a wide range of conditions (e.g., Anderson, 1964; Beach & Swenson, 1966; Levin, 1975; Spencer, 1961, 1963). Nothing like a sample size bias had ever been reported. (See Peterson & Beach, 1967, for a classic review of this work.) It seemed possible, therefore, that the sample size bias we had observed depended on our use of ambiguous concepts such as "risk," "likelihood," and the "average person." For this reason, we tested for the sample size bias by having people quickly estimate the means of samples of numbers—a relatively unambiguous task (Smith & Price, 2010). On each trial, participants saw samples of five, 10, 15, or 20 numbers with means of 20, 30, or 40. In one study, the numbers in each sample were presented simultaneously and in another they were presented sequentially. Although participants' estimates tracked the objective means fairly well— consistent with previous research and with the idea that participants correctly interpreted their task—there was also a clear sample size bias that accounted for approximately 10% of the variance in their estimates. This was true even among participants who consistently made the most accurate estimates.

Theoretical Considerations

One of the most notable features of the sample size bias has been its robustness across variations in the stimuli, the mode of stimulus presentation, the dimension of judgment, and the response scale. This is important because it casts doubt on many intuitively plausible theories that can explain it under some conditions but not others. For example, the sample size bias for risk judgments might be the result of a misunderstanding. Although participants are supposed to judge the average risk that the people in a group will experience a negative event, they might misunderstand their task as one of judging the risk that at least one person in the group will experience it. However, such misunderstandings seem much less likely for estimates of the average height of sets of stick figures or the mean of sets of numbers. As another example, the sample size bias might occur because people selectively attend to the most extreme individual stimuli (e.g., the riskiest looking people or the greatest numbers) or weight extreme stimuli more heavily in making their judgments. However, selective attention and weighting do not apply as neatly when the stimulus individuals are identical stick figures so that there are no extreme individuals (Price et al., 2006). As a final example, the anchoring-and-adjustment heuristic (Epley & Gilovich, 2004) might underlie the sample size bias. Specifically, people might use the sample size as a starting point for their judgment, and insufficiently adjust away from that anchor such that larger samples result in greater judgments. This explanation seems plausible when both the sample size and judgment of average are on the same order of magnitude but not when they are on different orders of magnitude—as when in one study sample sizes ranged from 1 to 15 but judgments were made on a 0-to-100 risk scale (Price et al., 2006).

We have also suggested that the robustness of the sample size bias implicates a very basic and general cognitive process—most likely a priming effect of sample size on judgments of averages that is independent of any conscious attempt to take the sample size into account (Smith & Price, 2010). There are two lines of evidence that give additional support to this interpretation. One is that there exist several examples of phenomena in which an irrelevant stimulus numerosity or frequency affects a quantitative

judgment. For example, Friedenberg and Limratana (2005) presented participants with displays consisting of several distinct clusters of equal numbers of dots. They found that judgments of the number of dots in a cluster were affected by the number of clusters and also that judgments of the number of clusters were affected by the number of dots in a cluster. Similarly, Pelham, Sumarta, and Myaskovsky (1994) found that the number of distinct elements in a stimulus affected a variety of quantitative judgments. For example, the number of wedges that a circle was divided into affected people's judgments of the total area of the circle. And Dormal and Pesenti (2007) have shown that the number of spots in each of two horizontal arrays affects people's ability to compare those two arrays in terms of their physical length. Specifically, if the longer array contains more spots, people make their comparisons faster and more accurately. But if the longer array contains fewer spots, people make their comparisons slower and less accurately. These researchers have also shown a similar effect of the number of spots in temporal sequences on people's ability to compare those two sequences in terms of their duration (Dormal, Seron, & Pesenti, 2006). In all of these examples, the number of stimuli in a set—whether the stimuli are distributed spatially or temporally— biased people's judgments of another quantity. Furthermore, these effects seem unlikely to be mediated by processes like miscommunication, selective attention, or anchoring and in- sufficient adjustment.

The second line of evidence comes from research on the cognitive neuroscience of quantitative cognition and perception. Specifically, there is considerable research showing that a variety of quantitative stimuli—including Arabic numerals, number words, sets of dots, and sequences of tones—activate a modality- independent representation of quantity or magnitude in the intra- parietal sulci (IPS; Cantlon, Platt, & Brannon, 2009; Dehaene, 2011; Dormal & Pesenti, 2009; Walsh, 2003; but see Matthews, Stewart, & Wearden, 2011, for an alternative interpretation). This same area is also involved in quantitative comparisons and simple computations (e.g., Chochon, Cohen, Van De Moortele, & De- haene, 1999; Dehaene, 2011). Dormal and Pesenti (2009) showed that both stimulus numerosity and stimulus length independently activate the IPS and suggested that this neural overlap might explain the effect of numerosity on judgments of length (among many conceptually similar effects). Thus, the key elements of a direct priming account of the sample size bias—that sample size activates a representation of quantity or magnitude, which in turn affects other quantitative judgments—are supported by research from other perspectives.

Judgments of Perceptual Averages

With this background, we decided to study the sample size bias for judgments of averages on a perceptual dimension: the size of squares. The primary reason is that it is not immediately clear that the sample size bias will generalize to such judgments. As with the early research on number averaging, research on perceptual aver-aging has shown it to be quite accurate across a wide range of conditions and nothing like a sample size bias has ever been reported or even suggested (see Alvarez, 2011, for a review). For example, Ariely (2001) conducted a study in which, on each trial, participants saw a sample of spots of varying sizes followed by a single test spot and then judged whether the test spot was larger or

smaller than the average size of the spots in the sample. With discrimination thresholds roughly in the range of 5 to 10%, he concluded that "the mean size of sets was known quite precisely" (Ariely, 2001, p. 160). Similar results have been reported by other researchers for judgments of average size (Chong & Treisman, 2005), and for other perceptual dimensions including brightness (Bauer, 2009), orientation (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001), motion (Watamaniuk & Duchon, 1992), and location (Alvarez & Oliva, 2008). Perceptual averaging also seems to occur at very short exposure times and does not require focal attention to any of the individual stimuli in the sample (e.g., Alvarez, 2011; Ariely, 2001; Parkes et al., 2001). These observations have suggested to some researchers the possibility of specific neural circuits that are responsible for the automatic computation of perceptual averages (e.g., Chong & Treisman, 2005). Thus, judgments of perceptual averages might not be open to the effects of misunderstanding, selective attention, anchoring and adjustment, or other processes that could explain the sample size bias for judgments of conceptual averages. On the other hand, it is not unreasonable to expect judgments of perceptual averages to be open to priming effects. After all, Dormal and Pesenti (2007) found a direct effect of numerosity on perceived length.

The present studies consist of five experiments focusing on people's judgments of the average sizes of sets of squares. Experiment 1 was the first strong test for a sample size bias for perceptual judgments of averages. The results of the study by Price et al. (2006), in which people judged the average heights of stick figures, was somewhat ambiguous because people might have interpreted the stick figures as *representations* of real people and based their judgments on their general knowledge about people's heights. In the first experiment, our approach was to present participants with sets of three, six, nine, and 12 squares and to ask them to rate the average size of the squares in each set. Then, in the next four experiments, we tested potential moderators of the sample size bias. In Experiment 2, we changed the response to an estimate of the area of the average square in terms of a standard unit of area. More important, we varied whether the size of the squares in each sample varied or was constant. Again, this is a way to test the idea that the sample size bias occurs because people focus on the most extreme individual stimuli when judging averages. In Experiment 3, we presented the squares in each set sequentially rather than simultaneously as a way of showing that it is the sample size rather than the spatial distribution of the squares that matters. In Experiments 4 and 5, we changed the response mode again to be more similar to previous research on perceptual averaging. Participants indicated whether the average square in a set or an individual comparison square was larger (Experiment 4) or smaller (Experiment 5). These studies were meant to test the possibility that the sample size bias is limited to quantitative judgments made on a numeric scale. Remarkably, the sample size bias was quite strong and consistent across every one of these conditions.

Experiment 1

The primary purpose of Experiment 1 was to test for the sample size bias in judgments of perceptual averages. The stimuli were squares presented on a computer screen, and the response was a rating of the average size of the squares.

Method

Participants. The participants were 35 undergraduate students (31 women and four men) at California State University, Fresno, who participated in this experiment as part of an introductory psychology course requirement.

Stimuli. The stimuli were 24 sets of gray squares presented on a white background. These sets varied in both sample size (3, 6, 9, and 12) and in average square size (small and large). In the small-square sets, there were equal numbers of squares that were 5, 11, and 17 mm on a side for a mean area of 145.00 mm^2. In the large-square sets, there were equal numbers of squares that were 13, 19, and 25 mm on a side for a mean area of 385.00 mm^2. For each combination of sample size and average square size, there were three sets of squares in different quasirandom spatial arrangements. Specifically, the squares were organized within a 12 X 8 cm rectangular area. For samples of size 3, three of the four corners of the rectangular area contained a square. For samples of size 6, 9, and 12, all four corners of the rectangular area contained a square and the remaining squares were distributed throughout the remaining space. This served as a partial control for the envelope area of the squares—the smallest polygon that contains all the squares.

Design and procedure. Participants were tested individually using desktop computers. All responses were size judgments made by using the mouse to click on one of the integers from 1 to 10 that were arrayed horizontally across the bottom of the screen. Anchor labels consisted of a small square (3 mm on a side) centered beneath the 1 at the left end of the scale and a large square (27 mm on a side) centered beneath the 10 at the right end of the scale. To ensure that participants were familiar with the rating scale, they were first presented with 13 individual squares ranging in size from 3 mm on a side to 27 mm on a side—in a random order—and judged the size of each one by clicking on a numeral on the rating scale. They made each of these judgments at their own pace while the stimulus square remained displayed on the screen. The main task was then explained to participants as one of using the same rating scale to judge the average size of the squares in each of several sets. Participants then made average-size judgments for two practice sets, had an opportunity to ask questions, and finally made average-size judgments for the 24 stimulus sets. Again, they made each of these judgments at their own pace while the stimulus set remained displayed on the screen, and no feedback was presented to them at any time about the accuracy of their judgments. The 24 stimulus sets were presented in a different random order for each participant, with the constraint that each block of eight trials contained one set with each combination of sample size and average square size.

Results and Discussion

For each participant, we computed the mean judgment for each combination of sample size and average square size. Figure 1 presents the means and standard errors of these mean judgments and shows a clear sample size bias, with the judgments increasing as a function of the sample size. To confirm this statistically, we submitted the mean judgments to a 4 (sample size) X 2 (average square size) repeated-measures
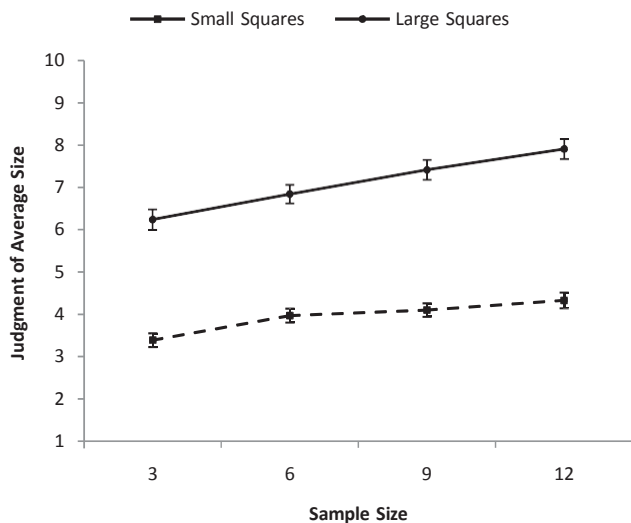
Figure 1. Means and standard errors of participants' judgments of average square size as a function of sample size and average square size in Experiment 1.

analysis of variance (ANOVA).[1] Not surprisingly, there was a main effect of average square size, which simply shows that participants distinguished the small-square sets from the large-square sets, $F(1, 34) = 302.44$, $p < .001$, partial $\eta^2 = .90$. Most important for present purposes, there was a linear effect of sample size, $F(1, 34) = 63.66$, $p < .001$, partial $\eta^2 = .65$. There was also an unexpected interaction between these two factors, with the linear effect of sample size being somewhat stronger for the large-square sets, $F(1, 34) = 15.03$, $p < .001$, partial $\eta^2 = .31$.

As a slightly different way of looking at these results, we regressed each participant's average size judgments onto the sample size to obtain both unstandardized and standardized regression slopes for each participant, where a positive slope indicates a sample size bias. The mean unstandardized regression slope was 0.14 ($SD = 0.11$), which is significantly greater than zero, $t(34) = 7.98$, $p < .001$, $d = 1.35$. This indicates that, on average, when the sample size increased by one square, the judged average size increased by 0.14 units on the 1-to-10 rating scale. The mean standardized slope was 0.25 ($SD = 0.15$). This indicates that, on average, when the sample size increased by one square, the judged average size increased by 0.25 standard deviations. Perhaps more remarkably, every one of the 35 participants had a positive regression slope.

Experiment 2

The results of Experiment 1 are consistent with previous research on the sample size bias and suggest that judgments of perceptual averages are biased by sample size just as judgments of conceptual and numerical averages are. In Experiment 2 we replicated this result while changing two important aspects of the design and procedure. The first is that we changed the response to be an estimate of area in terms of a standard unit that we provided (a purple circle that we defined as having an area of one unit). The second is that we manipulated whether the squares in each sample

varied in size or were all the same size as a way of testing the idea that the sample size bias requires selective attention to or selective weighting of more extreme individual stimuli. Recall that Price et al. (2006) observed a sample size bias in a study in which some of the sets consisted of identical stick figures and participants judged the average height of the stick figures. At first, this seems inconsistent with a selective attention explanation because there were no extreme individuals to selectively attend to. But, again, it is possible that participants interpreted the stick figures as representations of real people. A group of 10 stick figures might have prompted them to imagine a group of 10 real people—in which case they could still selectively attend to the taller imagined people or weight them more heavily in their judgments. The present study addresses this issue because it is clear to participants that they are judging the average size of the very squares they are looking at. Because all the squares are exactly the same size, there can be no selective attention to or selective weighting of larger squares.

Method

Participants. The participants were 25 undergraduate students (20 women and five men) at California State University, Fresno, who participated in this experiment as part of an introductory psychology course requirement.

Stimuli. The primary stimuli were 48 sets of black squares presented on a white background. The sets varied in terms of the sample size (3, 6, 9, and 12), the average square size (small and large), and the variability of the squares (variable or non-variable). In the small-square variable sets, there were equal numbers of squares that were 5, 11, and 17 mm on a side for a mean area of 145.00 mm^2. In the large-square variable sets, there were equal numbers of squares that were 13, 19, and 25 mm on a side for a mean area of 385.00 mm^2. In the small-square non-variable sets, the squares were all 11 mm on a side for a mean area of 121.00 mm^2. In the large-square non-variable sets, the squares were all 19 mm on a side for a mean area of 361.00 mm^2. For each combination of sample size, average square size, and variability, there were three sets in three different quasi-random arrangements in which each square was approximately 1 to 2 cm from its nearest neighbors. (We made no attempt to control the envelope area of the squares in this experiment.) In addition, a purple circle 6 mm in diameter (28.27 mm^2) appeared in the upper left corner of the screen throughout the experiment and was said to represent one unit of area.

Design and procedure. Participants were tested individually using desktop computers. They began by reading a detailed set of instructions that described their task in a general way and explained how to make area judgments in terms of the standard unit of area. Specifically, it was explained that the purple circle covered one unit of area on the screen, and an example was presented to show that a square that covered the same amount of area as three purples circles would have an area of three units. Another example was presented to show how the areas of four different sized squares could be combined mathematically to find the average

---

[1] In reporting our ANOVA results, we focus on the linear effect of sample size. Results pertaining to the quadratic and cubic effects for all experiments are presented in the Appendix. The only one of these effects that was statistically significant was the quadratic effect in Experiment 1.

(arithmetic mean) area of the squares. The instructions then explained that the participants' goal was not to compute the average area precisely, but to make an intuitive estimate of the average area in no more than about 10 s (although no time limit was actually enforced). The instructions also explained that participants' judgments would be limited to the integers from 1 to 20 because all of the averages were within this range.

After making three practice judgments and having an opportunity to ask questions, participants saw the 48 sets of squares in a random order and estimated the average area of each set by typing an integer from 1 to 20. They made these judgments at their own pace while the stimulus sets remained displayed on the screen, and they received no feedback about their accuracy.

Results and Discussion

Figure 2 presents the means and standard errors of participants' estimates of average square area as a function of sample size, square size, and square variability. The figure shows a clear sample size bias under all conditions. To confirm this statistically, we computed the mean estimate for each of the 16 combinations of sample size, average square size, and square variability for each participant. Then we submitted these mean estimates to a 4 (sample size) $\times$ 2 (square size) $\times$ 2 (square variability) repeated-measures ANOVA. As in Experiment 1, there was a main effect of square size, which shows that participants reliably distinguished the small-square sets from the large-square sets, $F(1, 24) = 53.47$, $p < .001$, partial $\eta^2 = .69$. There was no main effect of square variability, $F(1, 24) = 0.90$, $p = .77$, partial $\eta^2 = .004$, which shows that participants did not distinguish the variable sets from the non-variable sets (even though there was actually a small difference in their mean areas). Most important for present purposes, there was a linear effect of sample size, $F(1, 24) = 11.44$, $p = .002$, partial $\eta^2 = .32$. Unlike in Experiment 1, in this experiment there was no interaction between sample size and average square size, $F(1, 24) = 0.26$, $p = .62$, partial $\eta^2 = .01$. There was also no interaction between sample size and square variability, $F(1, 24) = 0.02$, $p = .90$ partial $\eta^2 = .001$. This is particularly important theoretically because it confirms that the sample size bias does not depend on there being variability among the individual items (see also Price et al., 2006). This, in turn, provides evidence against the idea that the sample size bias depends on selective attention to or weighting of the most extreme stimuli in the set. Finally, there was no interaction between aver- age square size and variability, $F(1, 24) = 0.33$, $p = .57$, partial $\eta^2 = .014$, nor was there a three-way interaction among the linear effect of sample size, average square size, and square variability, $F(1, 24) = 0.59$, $p = .45$, partial $\eta^2 = .024$.

The nature of the response scale in this experiment makes it possible to examine the accuracy of participants' estimates of average area. The dotted lines in Figure 2 show the objectively correct areas for the small-square sets and large-square sets in terms of the standard unit of area that participants used. For the small-square sets, the mean response is a slight underestimate of the objective value for samples of three squares, a slight overestimate for samples of six squares, and then an increasingly greater overestimate as the sample size increases to nine and 12 squares. For the large-square sets, the mean response is a substantial underestimate for samples of three squares and then a smaller underestimate
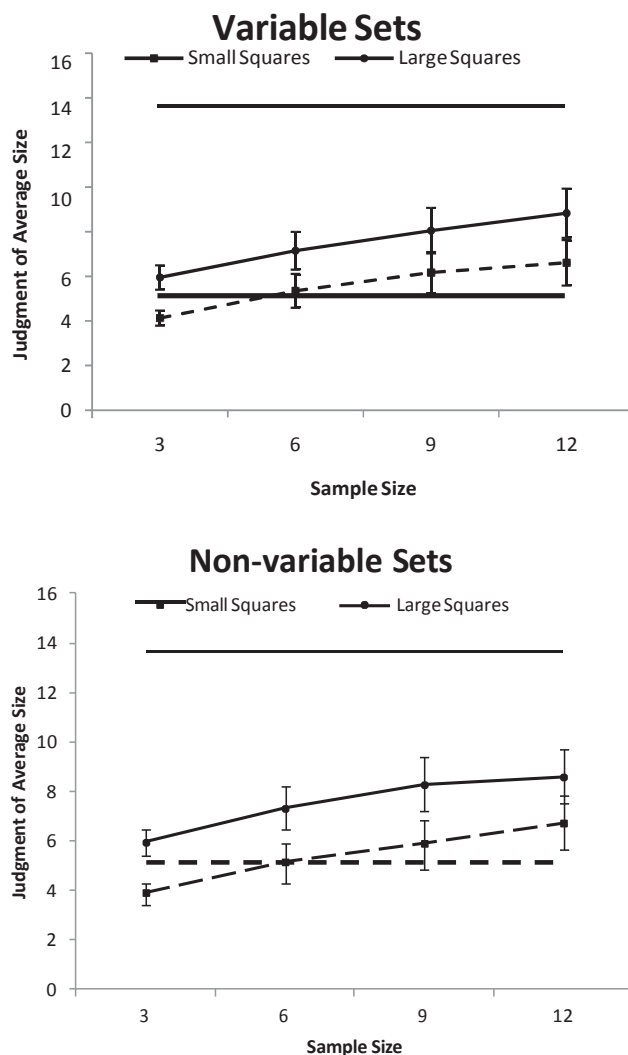

Figure 2. Means and standard errors of participants' estimates of average square area separately by sample size, average square size, and square variability in Experiment 2. The judgments were made in terms of a standard unit of area represented by a circle with an area of 28.27 mm². The dotted horizontal lines represent what would be objectively correct judgments of area in terms of the standard unit.

as sample size increases to six, nine, and 12 squares. These results illustrate that there is no fixed relationship between sample size and accuracy.

Finally, we also regressed each participant's average size estimates onto the sample size to obtain both an unstandardized regression slope and a standardized regression slope for each participant—where a positive slope indicates sample size bias. The mean unstandardized regression slope was 0.30 ($SD = 0.44$), which was significantly greater than zero, $t(24) = 3.37$, $p = .002$, $d = 0.66$. This indicates that, on average, when the sample size increased by one square, the estimated average area of the squares increased by 0.29 units of area (8.20 mm²). The mean standardized regression slope was 0.29 ($SD = 0.27$). This indicates that, on average, when the sample size increased by one standard devia-

tion, the estimated average area of the squares increased by 0.29 standard deviations. Note that the standardized regression coefficients can be directly compared across Experiments 1 and 2 and indicate a similarly strong sample size bias. It is also worth noting that 24 of the 25 participants (96%) produced positive regression slopes.

Experiment 3

Experiment 3 was essentially a replication of the variable-squares condition of Experiment 2 with one very important difference. The squares in each set were presented sequentially rather than simultaneously and the estimates of average area were made immediately after all the squares in the set had been presented. Previous research on the sample size bias has demonstrated it under sequential presentation conditions (Price, 2001; Smith & Price, 2010), which strongly suggests that we should observe it here as well. Most important, sequential presentation controls for the spatial distribution of the squares. In Experiments 1 and 2, the more squares there were in a set, the more total area those squares covered and the larger the envelope area of those squares was. Thus, it is possible that either the total area or the envelope area—rather than the sample size—was what is driving the sample size bias. But a sample size bias with sequential presentation of the squares rules out both of these possibilities.

Method

Participants. The participants were 21 undergraduate students (15 women and six men) at California State University, Fresno, who participated as part of an introductory psychology course requirement.

Stimuli, design, and procedure. The design and procedure were essentially the same as for Experiment 2. However, only the variable sets of squares were used and the squares in each set appeared one at a time in a random order in the center of the screen. Each square appeared for 1,000 ms with a 500-ms interval between squares. Immediately after the last square in the sample was presented, participants were prompted to enter their average area estimate, which again they did at their own pace.

Results and Discussion

Figure 3 presents the means and standard errors of participants' estimates of average square area as a function of sample size and average square size. Again, there was a main effect of average square size, which simply shows that participants distinguished the small-square sets from the large-square sets, $F(1, 20) = 17.66$, $p < .001$, partial $11^2 = .47$. Most important for present purposes, there was also a linear effect of sample size, $F(1, 20) = 21.02$, $p < .001$, partial $11^2 = .51$. There was no interaction between sample size and average square size, $F(1, 20) = 2.55$, $p = .13$, partial $11^2 = .11$. In terms of accuracy of participants' estimates, the pattern was similar to that from Experiment 2.

Again, we also regressed each participant's average-area estimates onto the sample size to obtain both an unstandardized regression slope and a standardized regression slope for each participant. The mean unstandardized regression slope was 0.38 ($SD = 0.38$), which was significantly greater than zero, $t(20) =$
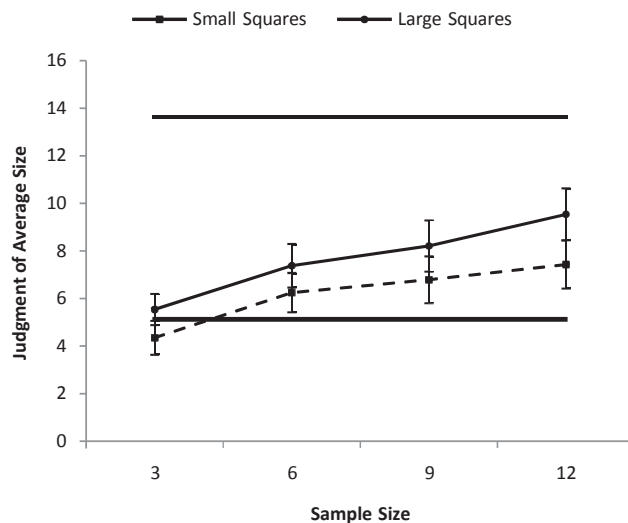

Figure 3. Means and standard errors of participants' estimates of average square area separately by sample size and average square size in Experiment 3. The squares in each set were presented sequentially and judgments were made in terms of a standard unit of area represented by a circle with an area of 28.27 mm². The dotted horizontal lines represent what would be objectively correct judgments of area in terms of the standard unit.

4.58, $p < .001$, $d = 1.00$. This indicates that, on average, when the sample size increased by one square the estimated average area of the squares increased by 0.38 units (10.75 mm²). The mean standardized regression slope was 0.37 ($SD = 0.28$). This indicates that, on average, when the sample size increased by one standard deviation, the estimated average area of the squares increased by 0.37 standard deviations—a slightly stronger effect than in Experiments 1 and 2. Furthermore, 19 of the 21 participants (90%) had positive slopes.

Again, the sequential presentation of the squares in Experiment 3 rules out the possibility that either the total area of the squares or the envelope area of the squares—as opposed to the number of squares—is what is driving the sample size bias. Of course, the design used in Experiment 3 introduces a new confounding variable—the total amount of time it takes to present the squares in a set. So it is possible that total presentation time is driving the sample size bias here. But this alter- native explanation has two difficulties. One is that, in their experiments on number averaging, Smith and Price (2010) observed the sample size bias even in a sequential-presentation condition in which total presentation time was controlled by varying the time between stimulus numbers within a set. The second is that it is more parsimonious to assume that the sample size bias is driven by sample size for both simultaneous and sequential presentation rather than being driven by total area or envelope area for simultaneous presentation and by total duration for sequential presentation.

Experiments 4 and 5

In Experiments 4 and 5 we addressed the question of whether there is a sample size bias in comparative judgments of size. This is important because all of our previous research on the sample

size bias has focused on ratings and estimates, while research on perceptual averaging has tended to focus on comparative judgments (e.g., Ariely, 2001; Chong & Treisman, 2005). This leaves open the possibility that the sample size bias is introduced only at the point of generating a quantitative response on a numeric scale. The internal representation of average might remain unaffected. In Experiment 4, we tested for this possibility by asking participants on each trial to compare a set of squares with an individual comparison square and choose which was larger: the average of the set of squares or the comparison square. Note, however, that a tendency to choose the average of the set could indicate that that sample size is affecting the representation of the average, but it could also indicate a simpler bias toward choosing physically larger stimuli over physically smaller stimuli (Silvera, Josephs, & Giesler, 2002). In Experiment 5, therefore, we asked participants to choose which was smaller: the average of the set of squares or the comparison square. Here the sample size bias should be reflected in a tendency to choose the comparison square, which would rule out the possibility that participants are simply choosing the physically larger stimulus.

Method

Participants. The participants were 110 undergraduate students at Appalachian State University, who participated in this experiment as part of an introductory psychology course requirement. There were 58 in Experiment 4 and 52 in Experiment 5.

Stimuli. The primary stimuli were sets of black squares on a white background, which varied in terms of their sample size (3, 6, 9, 12, and 15). Each set contained an equal number of squares that were 9.09, 11.69, and 14.29 mm on a side for a mean area of 141.16 mm^2. The squares in each set were presented simultaneously in the upper two-thirds of the screen and their positions were determined quasirandomly on each trial, with the constraint that no squares could overlap.

Design and procedure. Participants were tested individually using desktop computers. On the first trial of the experiment, they were presented with a sample of nine squares. At the same time, six individual squares ranging from 7.79 mm to 16.89 mm on a side were arrayed from left to right across the bottom of the screen.

Participants were instructed to select the individual square from this array that was closest to the average size of the sample of nine squares. Unknown to the participants, this established a comparison square to be used in the rest of the experiment.[2]

Next came 19 trials on which a set of squares and a comparison square were presented simultaneously, and participants were asked to make their comparative judgments. In Experiment 4 they were asked to indicate which was larger: the average of the set of squares or the comparison square. In Experiment 5 they were asked to indicate which was smaller: the average of the set of squares or the comparison square. The set of squares appeared within the top two thirds of the screen and the individual comparison square was centered within the bottom third. A horizontal band separated the set of squares and the comparison square and contained both the judgment prompt ("Which is larger/smaller, the size of the average square in the group above or the square below?") and two buttons (one that said "Average of Group" and one that said "Individual Square"). These stimuli remained on the

screen until participants responded by clicking on one of the two buttons.

Fifteen of these 19 trials were critical trials on which each of the five sample sizes (3, 6, 9, 12, and 15) appeared three times each and the comparison square was the one that participants had selected on the first trial. The other four trials were filler trials. For two of the filler trials, the sample sizes were 3 and 15 and the comparison square was 7.79 mm on a side (slightly smaller than the smallest square in the sample). For the other two filler trials, the sample sizes were 6 and 12 and the comparison square was 16.89 mm on a side (slightly larger than the largest square in the sample). The presentation order of these 19 trials (15 critical and four filler) was determined randomly for each participant.

Results and Discussion

First, it is worth noting that participants made the correct choice on 97% of the filler trials in Experiment 4 and 98% of the filler trials in Experiment 5, indicating that they understood their task and could almost always make the correct choice when it was fairly obvious.

For each participant, we computed the percentage of critical trials on which he or she indicated that the average of the set of squares was larger than the comparison square (Experiment 4) or, analogously, that the comparison square was smaller than the average of the set of squares (Experiment 5). Figure 4 presents the means and standard errors of these percentages and again we see a clear sample size bias. We then analyzed these results separately for the two experiments. In Experiment 4, a repeated-measures ANOVA on the percentages showed a linear effect of sample size, $F(1, 57) = 38.06$, $p < .001$, partial $\eta^2 = .40$. As the sample size increased, participants were more likely to choose the average of the set as being larger than the individual comparison square. The mean of the unstandardized regression slopes was 0.03 ($SD = 0.04$), which was significantly greater than zero, $t(57) = 6.18$, $p < .001$, $d = 0.81$. The mean of the standardized regression slopes was 0.41 ($SD = 0.51$). Of the 58 participants, 43 (74%) had positive slopes, while 11 (19%) had negative slopes.

In Experiment 5, there was also a linear effect of sample size,

$F(1, 51) = 4.16$, $p < .05$, partial $\eta^2 = .08$. In other words, as the sample size increased, participants were again more likely to perceive the average of the set to be larger as indicated by their being more likely to choose the comparison square as smaller. The mean of the unstandardized regression slopes was 0.01 ($SD = 0.03$), which was significantly greater than zero, $t(51) = 2.03$, $p < .05$, $d = 0.28$. The mean of the standardized regression slopes was 0.18 ($SD = 0.53$). Of the 52 participants, 29 (56%) had positive slopes while 15 (29%) had negative slopes.

Given that these two experiments were conducted in the same lab within a few months of each other, it also made sense to compare them directly to see if there were any differences. To do so, we conducted a 5 (sample size) X 2 (question: larger vs.

---

[2] In a pilot study, we provided the comparison square instead of letting participants choose it. The problem with this approach was that some participants *always* perceived the comparison square to be smaller than the average of the set while others *always* perceived it to be larger. Thus, it was not possible to observe a sample size bias for these participants using this approach.
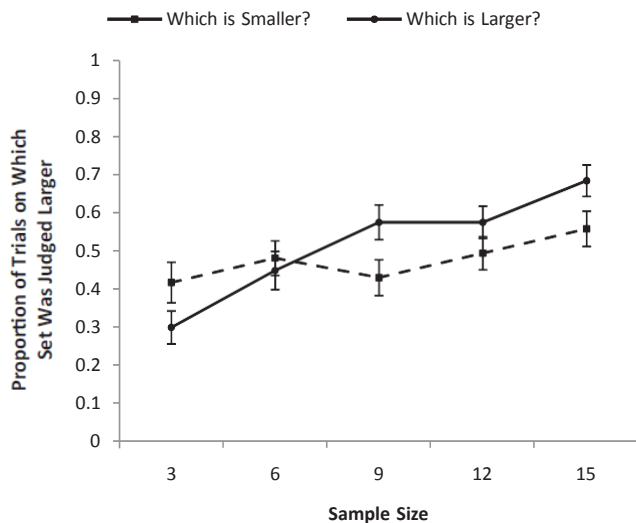
Figure 4. Means and standard errors of the proportion of trials on which participants indicated that the average of a set of squares was larger than an individual comparison square (Experiment 4; solid line) or that an individual comparison square was smaller than the average of ament 5; dashed line), separately by sample size.

smaller) ANOVA. Of course, this revealed a significant linear effect of sample size, $F(1, 108) = 33.55$, $p < .001$, partial $\eta^2 = .24$. There was no main effect of question, $F(1, 108) = 0.87$, $p = .35$, partial $\eta^2 = .008$, but there was an interaction, with the linear effect of sample size being stronger in the larger-question condition, $F(1, 108) = 8.56$, $p = .004$, partial $\eta^2 = .07$. One possible explanation of this difference is that it reflects another effect that operates independently—and in the opposite direction— of the sample size bias. One possibility is the aforementioned tendency for people to prefer larger stimuli over smaller ones (Silvera et al., 2002). Thus, when choosing whether the average of a set of squares or an individual comparison square is larger, both the sample size bias and the preference for larger stimuli would make them more likely to choose the set as the sample size increases. But when choosing whether the average of a set of squares or an individual comparison square is smaller, the sample size bias would lead them to choose the individual square as the sample size increases, but the preference for larger stimuli would lead them to choose the set as the sample size increases. This would result in a weakened overall sample size bias.

Nevertheless, the results of Experiments 4 and 5 show that the sample size bias does extend to comparative judgments of size. These results are also consistent with the idea that sample size affects the internal representation of the average. It is not simply a response bias that occurs when people make a quantitative judgment on a numeric scale, nor is it simply a response bias that involves choosing physically larger stimuli over physically smaller ones.

General Discussion

Summary

In this series of five experiments, we demonstrated a sample size bias on people's judgments of the average size of squares; the

more squares there were in a set, the greater people judged their average size to be. Furthermore, this effect occurred across a wide range of conditions. It occurred when the squares in each set were presented simultaneously and also when they were presented sequentially. It occurred when the squares in the sets varied in size and also when they were all the same size. It occurred when the judgments were ratings or estimates of the average size of a set of squares and also when they involved comparisons of the average size of a set of squares to the size of an individual comparison square. And for the comparisons, it occurred when participants had to indicate whether the average of the set or the comparison square was larger and also when they had to indicate whether the average of the set or the comparison square was smaller.

The robustness of the sample size bias is even more impressive when one takes into account that it has already been demonstrated when people judge the average risk of groups of people—whether the people are represented by written descriptions presented sequentially, photographs of real people presented simultaneously, or even identical stick figures (Price, 2001; Price et al., 2006). And it has been demonstrated when people quickly estimate the mean of sets of numbers presented simultaneously or sequentially (Smith & Price, 2010). Across all of these studies, the sample size bias has been strong and consistent, with the vast majority of participants exhibiting an effect in the expected direction.

Theoretical Considerations

As we argued in the introduction, the robustness of the sample size bias casts doubt on many plausible explanations that can account for it under some conditions but not others—and the present results continue this trend. For example, the nature of the task—judging average size— casts doubt on the idea that participants fundamentally misunderstand what they are supposed to do (e.g., summing rather averaging). The sample size bias for sequentially presented stimuli (Experiment 3) casts doubt on the idea that the effect is driven by the spatial distribution of the items in a set (as opposed to the number of items). The sample size bias for non-variable sets (Experiment 2) casts doubt on the idea that participants' judgments of average are based primarily on the most extreme stimuli in a set. The sample size bias for comparative judgments (Experiment 4) casts doubt on the idea that it only affects quantitative judgments on a numeric scale, and the sample size bias for comparisons in terms of smallness (Experiment 5) casts doubt on the idea that people are simply more likely to choose larger samples.

Yet another explanation that the present results cast doubt on is that people's judgments increase as a function of their response time. Although response times in the present experiments are difficult to interpret because participants were free to respond at their own pace, they still provide some insight into this issue. First, the mean response time generally did increase as a function of sample size in the present experiments. However, unlike the sample size bias itself, this result is not a strong one at the level of individual participants. For example, in Experiment 1 there were 18 participants who exhibited positive correlations between sample size and response time, but there were 17 participants who exhibited negative correlations. Recall also that every participant in Experiment 1 exhibited a positive correlation between sample size and judged average size. In other words, the sample size bias

was observed among participants who took more time to respond to larger samples and also among people who took less time to respond to larger samples. Although there is certainly more that can be learned about judgments of averages from response times in experiments designed specifically for that purpose, the present results suggest that the sample size bias is not closely related to them.

We believe that all of the present results are consistent with the theory we described in the introduction—that the sample size bias is a priming effect. Specifically, it seems likely that the sample size activates an internal representation of relative quantity or magnitude that directly affects the internal representation of the average and therefore affects the judgment of average. In addition to accounting for the robustness of the results, this theory is consistent with research showing that stimulus numerosity and frequency do seem to activate internal representations of magnitude (e.g., Dehaene, 2011; Dormal & Pesenti, 2009; Feigenson, Dehaene, & Spelke, 2004) and that irrelevant numerosities and frequencies do affect subsequent quantitative judgments (Dormal & Pesenti, 2007; Friedenberg & Limratana, 2005; Pelham et al., 1994).

Although our research has focused exclusively on judgments of average—this priming theory implies that the judgment does not have to be about an average. So, for example, the number of squares in a set should also affect judgments about the size of any individual square in the set or the total area covered by the squares in the set (cf. Pelham et al., 1994). In fact, from the priming perspective, the number of items in a set should affect judgments about entirely different stimuli. For example, if participants were exposed to different numbers of squares while estimating quantities like the length of the Mississippi River or the high temperature in Honolulu, the number of squares should affect these judgments too. Of course, there would have to be boundary conditions on these effects. Among them are that the sample size might require some minimal level of cognitive processing and the judgment might have to be made simultaneously with the presentation of the set or immediately afterward (cf. Wilson, Houston, Etling, & Brekke, 1996). Another potential boundary condition is that there might have to be a certain amount of uncertainty associated with the judgment. Clearly there is some uncertainty in judging average risk, number, and size. But what if people were to judge the average length of a set of yard sticks or the average weight of a set of 16-pound bowling balls? Here it seems likely that knowledge about the items being judged—along with their conceptual under- standing of averages—would lead them to the same (correct) answer regardless of the sample size.

Yet another factor that might moderate the sample size bias is whether sample size is varied within subject—as in all of our research to date—or between subjects. On the one hand, a within-subject design calls attention to the changing sample sizes, which may be important for producing the effect. In fact, research on the response of the intraparietal sulcus to the presentation of sets of stimuli shows habituation when the same number of items is presented repeatedly and renewed activation when there is a *change* in this number (Pinel, Piazza, Le Bihan, & Dehaene, 2004). For this reason, in a between-subjects design—or even a within-subject design with the stimuli blocked by sample size—it is possible that the sample size bias would be reduced or even eliminated. On the other hand, conceptually similar effects studied by researchers in judgment and decision making do not seem to require within-subject designs. For example, Wilson et al. (1996)—in a between-subjects design—found that an arbitrary ID number affected participants' subsequent estimates of the number of physicians listed in the telephone directory. Similarly, Oppenheimer, LeBoeuf, and Brewer (2008)—also in a between-subjects design—found that copying a set of short or long lines affected participants' subsequent estimates of the length of the Mississippi River and the average high temperature in Honolulu. Thus, it remains important to study the role of within-subject versus between-subjects designs in producing the sample size bias—not only because of its theoretical implications but because of its implications for understanding when the sample size bias is likely to occur outside the laboratory.

Although we believe that the priming account of the sample size bias is highly plausible and leads to many interesting and eminently testable predictions, we should emphasize that there are still other kinds of accounts that should be explored. One kind is that sample size affects some other variable that, in turn, affects quantitative judgments. We have already seen that it seems unlikely to be the spatial distribution of the stimuli or the time it takes to respond to them. But there are still other possibilities. For example, making judgments about larger samples might place a greater load on working memory, and quantitative judgments might increase as a function of working memory load. This idea could be tested by having participants make quantitative judgments while manipulating the working memory demands of a secondary task. Another kind of account is a psychophysical one. Perhaps to make their judgments of average, people form representations of the dimension under consideration for the individual items, mentally sum these representations, and then divide this sum by a representation of the sample size. The psychophysical function for sample size is almost certainly negatively accelerated (e.g., Feigenson et al., 2004; Hintzman, 1988), which could produce a sample size bias because participants would be dividing by a subjective sample size that increases too slowly relative to the objective sample size.

Conclusions

The sample size bias in judgments of averages appears to be an extremely robust phenomenon with important theoretical and practical implications. Theoretically, it is important now to identify the precise mechanism underlying it and to explore how it relates to conceptually similar effects from the study of quantitative cognition and perception and judgment and decision making. Practically, it is also important to study whether the sample size bias affects judgments of averages in contexts in which they frequently occur. These include supervisors' evaluations of groups of people (e.g., students, athletes), patients' judgments of average symptom frequency or severity, physicians' judgments of average treatment outcome, consumers' estimates of average price, and people's judgments of average completion time for a repeated task. Given the robustness of the sample size bias, it seems likely that it will turn up in many, if not all, of these domains.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15,* 122–131. doi:10.1016/j.tics.2011.01.003

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19,* 392–398. doi:10.1111/j.1467-9280.2008.02098.x

Anderson, N. H. (1964). Test of a model for number-averaging behavior. *Psychonomic Science, 1,* 191–192. doi:10.3758/BF03342858

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12,* 157–162. doi:10.1111/1467-9280.00327

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record, 59,* 171–186.

Beach, L. R., & Swenson, R. G. (1966). Intuitive estimation of means. *Psychonomic Science, 5,* 161–162. doi:10.3758/BF03328331

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences, 13,* 83–91. doi:10.1016/j.tics.2008.11.007

Chochon, F., Cohen, L., Van De Moortele, P., & Dehaene, S. (1999). Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience, 11,* 617–630. doi:10.1162/089892999563689

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45,* 891–900. doi:10.1016/j.visres.2004.10.004

Dehaene, S. (2011). *The number sense: How the mind creates mathematics.* New York, NY: Oxford University Press.

Dormal, V., & Pesenti, M. (2007). Numerosity-length interference: A Stroop experiment. *Experimental Psychology, 54,* 289–297. doi:10.1027/1618-3169.54.4.289

Dormal, V., & Pesenti, M. (2009). Common and specific contributions of the intraparietal sulci to numerosity and length processing. *Human Brain Mapping, 30,* 2466–2476. doi:10.1002/hbm.20677

Dormal, V., Seron, X., & Pesenti, M. (2006). Numerosity-duration interference: A Stroop experiment. *Acta Psychologica, 121,* 109–124. doi:10.1016/j.actpsy.2005.06.003

Epley, N., & Gilovich, T. (2004). Are adjustments insufficient? *Personality and Social Psychology Bulletin, 30,* 447–460. doi:10.1177/0146167203261889

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8,* 307–314. doi:10.1016/j.tics.2004.05.002

Friedenberg, J., & Limratana, W. (2005). Hierarchical number estimation. *Psychological Research, 69,* 211–220. doi:10.1007/s00426-003-0169-y

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Levin, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General, 104,* 39–53. doi:10.1037/0096-3445.104.1.39

Matthews, W. J., Stewart, N., & Wearden, J. H. (2011). Stimulus intensity and the perception of duration. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 303–313. doi:10.1037/a0019961

Oppenheimer, D. M., LeBoeuf, R. A., & Brewer, N. T. (2008). Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition, 106,* 13–26. doi:10.1016/j.cognition.2006.12.008

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4,* 739–744. doi:10.1038/89532

Pelham, B. W., Sumarta, T. T., & Myaskovsky, L. (1994). The easy path from many to much: The numerosity heuristic. *Cognitive Psychology, 26,* 103–133. doi:10.1006/cogp.1994.1004

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46. doi:10.1037/h0024722

Pinel, P., Piazza, M., Le Bihan, D., & Dehaene, S. (2004). Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron, 41,* 983–993. doi:10.1016/S0896-6273(04)00107-2

Price, P. C. (2001). A group size effect on personal risk judgments: Implications for unrealistic optimism. *Memory & Cognition, 29,* 578–586. doi:10.3758/BF03200459

Price, P. C., Smith, A. R., & Lench, H. C. (2006). The effect of target group size on risk judgments and comparative optimism: The more, the riskier. *Journal of Personality and Social Psychology, 90,* 382–398. doi:10.1037/0022-3514.90.3.382

Silvera, D. H., Josephs, R. A., & Giesler, R. B. (2002). Bigger is better: The influence of physical size on aesthetic preference judgments. *Journal of Behavioral Decision Making, 15,* 189–202. doi:10.1002/bdm.410

Smith, A. R., & Price, P. C. (2010). Sample size bias in the estimation of means. *Psychonomic Bulletin & Review, 17,* 499–503. doi:10.3758/PBR.17.4.499

Spencer, J. (1961). Estimating averages. *Ergonomics, 4,* 317–328. doi:10.1080/00140136108930533

Spencer, J. (1963). A further study of estimating averages. *Ergonomics, 6,* 255–265. doi:10.1080/00140136308930705

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7,* 483–488. doi:10.1016/j.tics.2003.09.002

Watamaniuk, S. N., & Duchon, A. (1992). The human visual system averages speed information. *Vision Research, 32,* 931–941. doi:10.1016/0042-6989(92)90036-I

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39,* 806–820. doi:10.1037/0022-3514.39.5.806

Weinstein, N. D. (1987). Unrealistic optimism about susceptibility to health problems: Conclusions from a community-wide sample. *Journal of Behavioral Medicine, 10,* 481–500. doi:10.1007/BF00846146

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125,* 387–402. doi:10.1037/0096-3445.125.4.387

Appendix

Additional Statistical Results: ANOVA Results for Involving Quadratic and Cubic Effects of Sample Size for All Experiments

Effect	df	F	p	Partial η^2
Experiment 1				
Quadratic effect	1, 34	4.60	.04	.119
Cubic effect	1, 34	0.83	.37	.024
Quadratic X Size[a]	1, 34	0.74	.40	.021
Cubic X Size	1, 34	1.38	.25	.039
Experiment 2				
Quadratic effect	1, 24	2.75	.11	.103
Cubic effect	1, 24	0.09	.77	.004
Quadratic X Size	1, 24	0.12	.73	.005
Cubic X Size	1, 24	0.22	.64	.009
Quadratic X Var[b]	1, 24	0.26	.61	.011
Cubic X Var	1, 24	0.02	.90	.001
Quadratic X Area X Var	1, 24	2.51	.13	.095
Cubic X Area X Var	1, 24	0.29	.59	.012
Experiment 3				
Quadratic effect	1, 20	3.12	.09	.135
Cubic effect	1, 20	1.50	.24	.070
Quadratic X Size	1, 20	0.74	.40	.036
Cubic X Size	1, 20	0.00	.97	.000
Experiment 4				
Quadratic effect	1, 57	2.69	.11	.05
Cubic effect	1, 57	1.75	.19	.03
Experiment 5				
Quadratric effect	1, 51	0.70	.41	.01
Cubic effect	1, 51	1.12	.30	.02
Experiment 4 and 5 combined				
Quadratic effect	1, 108	0.24	.63	.00
Cubic effect	1, 108	2.81	.10	.03
Quadratic X Experiment[c]	1, 108	2.99	.09	.03
Cubic X Experiment	1, 108	0.01	.91	.00

Note. ANOVA = analysis of variance.
[a] "Size" refers to the size of the squares in a set: smaller or larger. [b] "Var" refers to the variability of the squares in a set: variable or nonvariable. [c] "Experiment" refers to Experiment 4 in which participants were asked whether the average of the set or the comparison square was larger or Experiment 5 in which they indicated whether the average of the set or the comparison square was smaller.