AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Decision

ISSN: 2325-9965

https://doi.org/10.1037/dec0000182

Relative Insensitivity to Sample Sizes in Judgments of Frequency Distributions: People are Similarly Confident in the Results From 30 Versus 3,000 Observations

Siran Zhan[1] and Krishna Savani[2, 3]

[1] School of Management and Governance, University of New South Wales, Sydney
[2] Department of Management and Marketing, The Hong Kong Polytechnic University
[3] Division of Leadership, Management, and Organization, Nanyang Business School,
Nanyang Technological University

OPEN DATA

OPEN MATERIALS

Six experiments test whether people are sensitive to variations in the sample size or relatively insensitive to sample sizes varying by one or two orders of magnitude. We posit that past studies found that people are increasingly confident in the sample mean as the sample size increases because variations in the sample size were likely highlighted by the within-participant designs used in many of these studies. Using between-participant designs, we find that people are only slightly sensitive to variations in the sample size by a factor of 50, 100, and 400 when they are making confidence judgments. Our finding suggests that the psychological *law of small numbers* applies not only to people's judgments about sample variances but also to their judgments of sample means. An intervention that provided people with the results of a statistical test about the extent to which the data are consistent with the null hypothesis versus the alternate hypothesis helped reduce people's insensitivity to variations in the sample size.

Keywords: sample size, law of small numbers, law of large numbers, confidence judgments, statistical inference

Supplemental materials: https://doi.org/10.1037/dec0000182.supp

Siran Zhan https://orcid.org/0000-0003-3279-2797
Krishna Savani https://orcid.org/0000-0002-6934-1917

This research was supported by a Special Research Grant by University of New South Wales to Siran Zhan. The authors thank Li Shi Tan, Ee Hwee Lau, Velvetina Lim, Dayana Bulchand, Andrea Low, and Sylvia Chin for their invaluable research assistance.

This study was approved by University of New South Wales ethics committee under protocol HC180012, titled "What counts as evidence? People believe data from a sample of 30 nearly as much as data from a sample of 3,000."

The ideas and data presented in this article were previously presented at the Academy of Management 2018 Annual Meeting.

Siran Zhan played lead role in data curation, formal analysis, investigation, project administration, and writing of original draft and equal role in conceptualization and methodology. Krishna Savani played lead role in funding acquisition, resources, and writing of review and editing and equal role in conceptualization and methodology.

The data are available at https://osf.io/9qxzh/?view_only=16bd77b425384ba6970fa9f95263f8eb

The experimental materials are available at https://osf.io/9qxzh/?view_only=16bd77b425384ba6970fa9f95263f8eb

Correspondence concerning this article should be addressed to Siran Zhan, School of Management and Governance, University of New South Wales, Sydney, College Road, Kensington, NSW 2052, Australia or Krishna Savani, Department of Management and Marketing, The Hong Kong Polytechnic University, Li Ka Shing Tower M857, Hung Hom, Kowloon, Hong Kong. Email: siran.zhan@unsw.edu.au or krishna.savani@polyu.edu.hk

Imagine a financial advisor informing an investor: "Moderna just reported that its experimental COVID-19 vaccine was safe and effective in eight healthy volunteers. This is promising—you should buy their shares before it is too late." Or imagine a senior journalist telling a national newspaper editor: "The COVID-19 case numbers decreased by 10% yesterday compared to the previous day, after climbing steeply for many days. We are finally flattening the curve. Let's broadcast the good news." Should the investor buy shares of the pharmaceutical company, and should the editor broadcast the "good news"? Although a person with statistical training may quickly realize that data from eight volunteers or a single day are insufficient to make statistically sound conclusions, decision-makers in the two preceding examples did indeed act based on these data: investors drove up the price of Moderna by 24% (Grady, 2020), and the national newspaper celebrated the "flattening of the curve" (ABC News, 2020). In the present research, we ask whether people have a statistical intuition that larger samples are more reliable. Moreover, if they do not, how can we debias them to help them make better decisions? We aim to address these questions by revisiting a longstanding debate about whether people's statistical intuition is consistent with the law of large numbers. On one side of the debate are researchers measuring people's judgments of *frequency distributions*, that is, how people's confidence in the sample mean changes with the sample size; on the other side are researchers measuring people's judgments of *sampling distributions*, that is, how people's estimate of the likelihood that the sample mean would deviate from the population mean changes with the sample size (Sedlmeier & Gigerenzer, 1997).

Two Streams of Research

A landmark article found that when making judgments about sampling variability (i.e., the extent to which the sample mean is likely to deviate from the population mean), people appear to subscribe to the *law of small numbers*, the idea that even small samples would be representative of the population (Tversky & Kahneman, 1971). For example, trained psychologists erroneously believed that samples of 20 and 40 observations would be similarly representative of the population and of each other (Tversky & Kahneman, 1971). For instance, participants estimated that a large hospital with 45 births daily and a small hospital with 15 births daily would have a similar number of days on which more than 60% of births were boys, whereas in reality, the larger hospital is much less likely to deviate from the mean of 50% than the smaller hospital (Kahneman & Tversky, 1972). People fail to realize that the means from smaller samples are likely to be more variable than means from larger samples drawn from the same population, indicating that people are largely insensitive to variations in the sample size when making judgments about variability in the sample mean (Kahneman & Tversky, 1972; Kutzner et al., 2016). Research on decision-making based on experiences has also found that people have higher confidence in data from single, isolated observations or anecdotes than data from repeated observations summarized in the form of statistical evidence (Obrecht et al., 2007), indicating that people do not take the sample size into consideration when interpreting data.

On the other hand, Jacob Bernoulli claimed that, "even the stupidest man [understands the law of large numbers]" (Sung, 1966), the idea that "a large random sample from a population will have a distribution that closely resembles that of the overall population" (Rabin, 2002, p. 775). More precisely, for any $\varepsilon > 0$, there exists a number n such that for all sample sizes greater than n, the probability that the difference between the sample mean and the population mean would exceed ε is zero (Feller, 1957). Consistent with Bernoulli's claim, many studies have found that people subscribe to the law of large numbers (Obrecht, 2019). For instance, when asked to infer about the average value of a deck of cards, people were more confident in their judgment after viewing a sample of 20 cards than after viewing 10 cards (Irwin et al., 1956). When judging the chances of an animal being aggressive based on a sample, participants weighed larger samples more and had higher confidence in their judgments from larger samples (Obrecht & Chesney, 2013). Over 80% of participants had more confidence in the results of a larger sample of 1,000 than in a smaller sample of 400, whereas almost none had more confidence in the results of the smaller sample (Bar-Hillel, 1979). Participants' estimates of an event's true likelihood of occurrence based on its occurrence in a sample increased as the sample size increased (Evans & Pollard, 1982).

College students (but not young children) were more confident in their judgment as the sample size increased from 1 to 6, suggesting that people may develop sensitivity to sample sizes as a part of cognitive development (Masnick & Morris, 2008). When asked to infer causation from correlation, college students rated the likelihood of a causal relationship to be higher when the sample size was larger (Koslowski et al., 1989). Furthermore, when asked whether they expect the mean height of 25 men or of 100 men to be closer to the national average, a majority of participants chose the higher sample size (Well et al., 1990). In a related literature on aggregation of opinions, Budescu and colleagues (Budescu et al., 2003; Budescu & Rantilla, 2000; Budescu & Yu, 2007) found that decision-makers were more confident in the modal opinion when the sample size increased (i.e., as more people's opinions were provided), and when there was more overlap or agreement among the opinion providers.

This inconsistency in the literature has been repeatedly observed by researchers. For instance, Well et al. (1990) referred to the ongoing debate as "the rather uncomfortable position of knowing that people sometimes use information about sample size appropriately and sometimes they do not" (p. 291). Sedlmeier and Gigerenzer (1997) concluded, "From one group of studies, it has been argued that people are good 'intuitive statisticians' who properly take sample size into account; from another group of studies the opposite claim has been made" (p. 33). More recently, Obrecht and Chesney (2013) noted that each camp has reported some empirical support for their assertion, yet "the literature is somewhat mixed in regard to people's ability to use sample size when making judgments" (p. 371).

Even though many researchers acknowledged this inconsistency in the literature, few have attempted to understand why this is the case. One exception is Sedlmeier and Gigerenzer (1997), who argued that a key distinction is whether people are presented with questions about the sampling distribution (i.e., judgments about how likely the sample mean is to deviate from the population mean) versus the frequency distribution (e.g., judgments about the extent to which a sample mean is representative of the population mean). Specifically, Sedlmeier and Gigerenzer (1997, Exhibit 2, p. 9) indicate that 19 out of 24 studies (i.e., 79% of the studies) that employed frequency distribution tasks found that people were sensitive to the sample size. Among these 19 studies, the average effect size is .43, a medium to large effect by Cohen's standard. In contrast, most participants were insensitive to the sample size in 25 out of 29 studies (i.e., 86% of the studies) that employed sampling distribution tasks (Sedlmeier & Gigerenzer, 1997, Exhibit 1, p. 7).

Given that Sedlmeier and Gigerenzer (1997) review of the literature was thorough up to year 1997 (or a few years prior to that, accounting for the time needed to write and publish the article), we reviewed more recent research on this topic (see Supplemental Materials for details). We found that 16 out of 19 studies (84%) employing frequency distribution tasks (i.e., the kinds of tasks that we used in our studies) found that people were sensitive to variations in the sample size, whereas two out of three studies (i.e., 67%) using a sampling distribution task found that people were *not* sensitive to the sample size. These findings are consistent with those from Sedlmeier and Gigerenzer (1997) review.

Our review of the literature also found that researchers appear to have reached the consensus that people's statistical intuitions about the relationship between sample means and population means are consistent with the law of large numbers, even if not perfectly commensurate with it, as long as the task involves judgments of frequency distributions. In particular, Sedlmeier and Gigerenzer (1997) argued that people's decisions are consistent with the *empirical law of large numbers*, the intuition that larger samples generally lead to more accurate estimates of the population mean. However, we argue that this conclusion is premature because a key methodological limitation of nearly all prior studies on frequency distribution tasks has been overlooked.

Methodological Concerns

Most past studies claiming that people are sensitive to sample sizes when judging the extent to which samples represent the population presented people with data from multiple samples, either simultaneously or sequentially (e.g., Griffin & Tversky, 1992; Masnick & Morris, 2008; Obrecht & Chesney, 2013; Obrecht et al., 2007). Indeed, Griffin and Tversky (1992, p. 418) stated that nearly all studies employed a within-subject design, in which "both the strength of the evidence (e.g., sample mean) and the mitigating variable (e.g., sample size) are varied within subjects." Within-participant designs are certainly useful for testing theory, particularly those pertaining to decision-making from experience (Hadar & Fox, 2009; Kudryavtsev & Pavlodsky, 2012) and aggregation of opinions (Budescu et al., 2003; Budescu & Rantilla, 2000; Budescu & Yu, 2007) where multiple experiences or sources of information are available to decision-makers. However, within-participant designs are less useful for studying decision-making from description, another important class of real-life decisions (Hadar & Fox, 2009; Jarvstad et al., 2013; Kudryavtsev & Pavlodsky, 2012). In these, a summary of the key statistical information is explicitly described (e.g., past mean return and variance of a mutual fund). In description-based decision-making, decision-makers are not asked, "Here are the findings from 30 samples, and here are the findings from 3,000 samples. What do you want to do next?" Instead, decision-makers hear only one version: "Here are the findings based on 30 (or 3,000) samples. What do you want to do next?"

A within-subject design is not conducive for studying people's sensitivity to sample size in description-based decision-making because it "may underestimate the dominance of strength because people tend to respond to whatever variable is manipulated within a study whether or not it is normative to do so" (Griffin & Tversky, 1992, p. 418). In other words, stimuli showing multiple sample sizes together to the same participant would "enhance the salience of sample size" (Bar-Hillel, 1979, p. 249). This is a significant concern because recent research has found that the likelihood or degree to which people make use of a statistical feature in their judgment and decision-making depends on how salient that feature is either in the way it was presented or by intuition. For example, Obrecht et al. (2007) found that when making decisions based on the sample mean, standard deviation, and sample size (all of which contribute to statistical power and thus should be jointly considered), the sample mean was significantly more salient to participants than the sample size and the standard deviation. Additionally, Morris and Masnick (2015) found that even in a within-participant design, participants' confidence in the sample was not sensitive to sample size when other more salient statistical features (i.e., sample mean and coefficient of variance) were presented. Thus, past research has found that people are sensitive to variations in the sample size when the sample size is highlighted in with-participant designs, and when other more salient pieces of information are not available.

These observations are consistent with two established theories in judgment and decision-making. First, the comparative ignorance hypothesis predicts that people are ambiguity averse when they evaluate options jointly but less so when they do so separately (Fox & Tversky, 1995). As larger samples are less ambiguous, this idea predicts that people would be more likely to trust larger samples in within-participant designs than in between-participant designs. Second, the distinction bias argues that people discern more differences between options that are presented jointly rather than separately (Hsee & Zhang, 2004), which would enhance people's susceptibility to sample sizes. By the same token, within-participant designs can suffer from experimenter demand effects—participants might infer that if the experimenter is presenting them with data from samples of different sizes, the experimenter likely wants them to take the sample size into account. We argue that evidence from experiments with within-participant designs can at best suggest that people conform to the law of relatively larger numbers—when presented with both large and small samples, people believe that larger samples more accurately represent the population. To address this limitation, we employ a between-participant design in which we systematically varied the sample size across participants to understand whether individuals are intuitively sensitive to variations in sample sizes and use this information to inform their decisions. We hypothesized that when the sample size is not made salient in a within-participant design, participants would be barely sensitive to sample sizes varying across multiple orders of magnitude. In other words, we hypothesize that the sample size does not feature prominently in people's statistical intuitions even when they are making judgments about the frequency distribution.

Overview of Studies

We conducted six experiments to test our hypothesis. Experiment 1 tested whether people are relatively insensitive to variations in sample sizes differing by a factor of 100 when making confidence judgments. Experiments 2A and 2B included the sample mean as an additional independent variable to test whether people's confidence judgments are sensitive to changes in sample means but not to changes in sample sizes. Experiment 2B further used an incentive compatible design and operationalized the dependent variable as a binary decision rather than a confidence judgment. Experiment 3 tested whether people are sensitive to variations in sample sizes in within-participant comparisons (which have been used in past research) but not in between-participant comparisons. By varying the sample size from 3 to 1,200 across participants, Experiment 4 tested whether there exists a threshold beyond which people's intuitions suggest that the sample size is no longer relevant. Finally, Experiment 5 assessed whether an intervention providing people the likelihood that the sample results conform to the null hypothesis versus the alternate hypothesis would reduce people's insensitivity to sample sizes.

We targeted a sample size of at least 100–200 per condition, which would give us 80% and 98% power to detect a medium effect size (i.e., Cohen's $d = .40$; Funder & Ozer, 2019). To provide an even more conservative test of our predictions, we recruited a sample size of 400 per condition in Experiments 1 and 3, which gives us 80% power to detect a small effect size (i.e., Cohen's $d = .20$). No participants were dropped from the analyses in any experiment. We report all conditions and decision confidence measures. In each experiment, data were analyzed only after the target sample size was met. The data (with variable legends) and experimental stimuli for all experiments are available at https://osf.io/9qxzh/?view_only=16bd77b425384ba6970fa9f95263f8eb. We also conducted Bayesian analyses for each experiment, where possible, using the JASP (Jeffreys's Amazing Statistics Program, Wagenmakers et al., 2018). Bayesian results are reported in the Supplemental Materials (section titled "Bayesian results" on p. 18).

Across all experiments, participants were not substantially more confident in findings from a small sample versus a much larger sample (e.g., 50–200 times larger). However, in nearly every experiment, participants had slightly more confidence in the large sample than the small sample. Importantly though, the difference in confidence across the two sample sizes is generally of very small magnitude (Cohen's $d = .01–.12$); yet, according to statistical principles, people should have very low confidence in the small sample condition and very high confidence in the large sample condition, which should lead to large gaps in confidence between the two conditions.

Experiment 1

Experiment 1 was designed to provide an initial test of our hypothesis that people are largely insensitive to variations in sample sizes when tested in a between-participant design. This study was preregistered (see the pre-registration file in our OSF project folder).

Method

Participants

The target sample size was 400 participants per condition, which would give us 80% power to detect a small effect size (Cohen's $d = .20$). A survey seeking 800 U.S. residents was posted on Amazon's mechanical turk (MTurk) using the CloudResearch platform; we sought participants using a computer (not a mobile phone or a tablet), had completed at least 100 assignments on MTurk and received an approval rating of at least 97%. We further used CloudResearch's "block low-quality participants" feature to ensure we sampled high-quality participants. In response, 805 participants (423 women, 376 men, 4 other, and 2 unreported; $M_{\text{age}} = 42.78$ years) completed the survey. Participants were randomly assigned to either the small sample size or the large sample size condition.

Procedure

We used an incentive compatible design, and trained participants to provide confidence ratings following Charness and Dufwenberg (2006) scoring rule. Specifically, we informed participants that if their estimate is within 5% of the true value as per statistical principles, then they would receive a bonus of 50 cents. As the dependent measure required participants to state a probability, we included a training round consisting of four practice trials. Specifically, participants were asked to state their estimated probability that a fair six-sided dice roll would roll an odd number, a number greater than 2, a number less than 6, and a number greater than 0. As the dependent measure was assessed on a 50%–100% probability scale, we asked participants to respond to the practice trials on a 50%–100% slider scale and ensured that the actual probability of the outcome in the training trials was indeed between 50% and 100%. Participants received feedback on each practice—we informed them of the correct probability and told them whether they would receive a payment according to the scoring rule. The details of the scoring rule, the training, and the practice trials and feedback are reported in Supplemental Materials, Experiment 1 Stimuli.

Upon completing the practice trials, participants were presented with three decision scenarios in which we varied the sample size between participants. Across the three scenarios, the large sample size was 400, 50, and 100 times bigger than the small sample size, respectively, to present a truly conservative test of our hypothesis that people are largely insensitive to variations in the sample size. The three scenarios were presented in a fixed order as randomizing the three scenarios is likely to increase the error variance. We provide a sample scenario below; the full scenarios are available in Supplemental Materials, Experiment 1 Stimuli. Table 1 presents a summary of the key variables that varied across scenarios.

> Imagine you are the HR Director of a large company with over 200,000 employees. Your company has a standard policy of giving all employees 15 days of leave per year (equivalent to 3 full weeks). Although employees cannot carryover their unused leave days into subsequent years, you noticed that most employees fail to take their allocated leave days year after year. To ensure that all employees get a break from work, you recently instituted a new policy of forced vacations: At the end of each year, all employees have to take their remaining annual leave vacation days off. You want to assess whether a majority of your employees (i.e., over 50%) are happy with the new policy. You decided to survey 10 (4,000) randomly selected employees. You found out that 60% of employees said that they are happy with the new policy, whereas the remaining 40% said that they were not happy with it.

Table 1

Key Information Presented in Each Scenario

Scenarios	Condition	Sample size	Sample mean	BF$_{10}$
Scenario 1	Large	4,000	60%	Infinity
	Small	10	60%	.44
Scenario 2	Large	2,000	60%	Infinity
	Small	40	60%	.43
Scenario 3	Large	3,000	60%	Infinity
	Small	30	60%	.40

Note. BF = Bayes Factor.

After reading the scenario, participants were asked, “Based on the above survey results, on a scale from 50% (probable at chance level) to 100% (probable without a doubt), what do you think is the probability that a majority (i.e., over 50%) of all your employees are happy with the new forced vacation policy?”[1]

Results

We first performed a repeated measure analysis of variance (ANOVA). Mauchly’s test indicated that the assumption of sphericity is violated, $\chi^2(2) = 36.51$, $p < .001$, so we corrected the degrees of freedom using the Greenhouse–Geisser estimate of sphericity ($\varepsilon = .96$). We did not find a within-participant difference among the three scenarios, $F(1.92, 257.53) = 1.70$, $p = .19$. More importantly, participants’ estimated probability of whether a true majority of the population preferred the option that a majority of the sample preferred did not differ statistically differently between the large and the small sample size conditions, $F(1, 556.91) = 1.13$, $p = .29$. The means, standard deviations, and 95% confidence intervals (CIs) are reported in Table 2. The interaction between experimental scenario and sample size was also nonsignificant, $F(1.92, 73.48) = .49$, $p = .61$. We also conducted a Bayesian repeated measure ANOVA using the JASP program (Wagenmakers et al., 2018), which is reported in Supplemental Materials, Bayesian Results, Experiment 1. The results of the Bayesian analysis were similar to those of the conventional repeated measure ANOVA.

We further noted that 60% (the percentage of the sample mean presented in all three scenarios) was the modal response in the first and second scenarios (see histograms in Figure 1 of the Supplemental Materials). There seems to be a learning effect (Charness et al., 2012) such that

[1] Participants used a slider bar anchored between 50% and 100% to indicate their probability estimate. Due to a programming error, the slider bar for the small sample size condition for Scenario 2 was set to “snap to grid.” As a result, participants’ responses were recorded at nearest intervals of 5% (e.g., 50%, 55%, and 60%) for this condition.

Table 2
Experiment 1 Descriptive

Scenarios	Condition	*M*	*SD*	*N*	95% confidence interval	Cohen's *d*
Scenario 1	Large	72.84%	15.75%	402	71.30%–74.38%	0.07
	Small	71.74%	15.38%	403	70.23%–73.25%	
Scenario 2	Large	72.14%	14.91%	402	70.68%–73.60%	0.09
	Small	70.84%	14.46%	403	69.42%–72.26%	
Scenario 3	Large	72.17%	13.90%	402	70.81%–73.53%	0.04
	Small	71.69%	13.42%	403	70.38%–73.00%	

substantially more participants responded with a probability estimate of 60% in the first and second scenarios, but this tendency disappeared in the third scenario. This result pattern suggests that participants might have learned through the first two scenarios that all sample means are held constant at 60%, so 60% is unlikely the correct probability estimate. Thus, participants moved away from responding with 60% in the final scenario. Furthermore, the frequency distribution for the dependent variable in the third scenario (DV3) shows that after excluding participants who responded 60% to all three scenarios, 63% is the mode. Moreover, those who chose 60% in Scenarios 1 and 2 were significantly more likely to respond with 63% in Scenario 3 than those who did not: chi-square likelihood ratio = 246.63, $p < .001$, *OR* = 30.51. This finding suggests that in addition to a learning effect, there appears a strong anchoring effect (Tversky & Kahneman, 1974). Specifically, many participants likely adjusted their estimated probability from the salient sample mean of 60% to derive a final estimate of 63%. We conducted a series of supplementary analyses to ensure the main results are robust to this interpretation. The detailed results are reported in the Supplemental Materials (Experiment 1 supplementary analyses).

Discussion

Overall, Experiment 1 found that when we used a between-participant design free of experimenter demand characteristics, contrary to the claims of prior studies (see Sedlmeier & Gigerenzer, 1997), people did *not* reliably demonstrate a sensitivity to sample sizes varying by a factor of 50, 100, and 400. Strikingly, although past research has documented that people exhibit some sensitivity to sample sizes when sample sizes vary within a limited range (e.g., from 1 to 33; Griffin & Tversky, 1992), this study is the first to document minimal sensitivity to sample sizes varying by multiple orders of magnitude.

The unexpected pattern of 60% being the modal response in Scenarios 1 and 2 can be interpreted in three ways. First, it could suggest that some participants did not understand what they were asked to do initially and thus responded with the most salient number, 60%, which was sample mean information in all three scenarios. Second, it can mean that many participants erroneously thought that 60% sample mean found in the sample translates into 60% probability for a majority preference in the population. This could be a similar error commonly found in people's and even scientists' interpretation of the *p* value, that is, a *p* value of .05 translates into a 5% chance of the finding being wrong (Goodman, 1999). The proportion of participants choosing 60% decreased over scenarios possibly due to a learning effect: participants started to notice over the scenarios that the sample mean in every scenario is held consistent at 60%, and reasonably suspected that the experimenter cannot be expecting the same answer in all scenarios. A third explanation could be that these participants understood the task requirement but were influenced by an anchoring effect—seeing the percentage 60% in the experiment stimuli "pulled" their response to percentage values within a small range of 60%. All three explanations indicate that participants paid attention to the sample mean information (i.e., 60%) in each scenario while failing to demonstrate sample size considerations in their responses. This finding is consistent with past research demonstrating that participants' subjective confidence in the findings from different samples was influenced more by variation in the sample mean than by variation in sample size (Griffin & Tversky, 1992, see also Masnick & Morris, 2008, Obrecht et al., 2007).

Experiment 2A

In Experiment 1, people seem to be influenced more by sample mean information than by variation in sample size. To verify this finding, in this study, we crossed sample size with sample mean in the experimental design. We also conducted a posttest in which we explicitly asked participants what information they considered when making inferences about the population from the sample. Finally, to ensure that the results of Experiment 1 were not due to participants failing to understand the task requirements, we used a simpler dependent measure in this study, that is, confidence ratings.

We designed stimuli such that the sample size, not the sample mean, determined whether participants would be justified in generalizing the finding from the sample to the population. That is, with a sample size of 3,000, sample means of both 67% and 57% are different from 50% (i.e., chance), $\chi^2(1) = 346.80$ and $\chi^2(1) = 58.80$, respectively, p's < .0001, BF_{10}'s = infinity. However, with a sample size of 30, neither mean is significantly different from 50%, $\chi^2(1) = 3.33$ and $\chi^2(1) = .53$, respectively, p's = .07 and .47, respectively, BF_{10}'s = 2.42 and .29, respectively. Thus, if participants were making inferences as per principles (either null hypothesis significance testing or Bayesian statistics), we should observe a significant main effect of sample size but no main effect of the sample mean. The sample means of 67% and 57% were chosen so that they seem substantially different to a layperson yet would yield statistically nonsignificant differences when the sample size is small (i.e., 30).

We measured participants' subjective confidence in the findings instead of asking them to make probability judgments because we wanted to minimize the difficulty of the task, a criticism Sedlmeier and Gigerenzer (1997) made about tasks used by Kahneman and Tversky (1972). As subjective confidence rating is the most widely used dependent measure in the literature we are trying to contribute to (e.g., Budescu et al., 2003; Budescu & Yu, 2007; see Sedlmeier & Gigerenzer, 1997, for a review), we used this dependent measure.

Method

Participants

A survey seeking 400 U.S. residents was posted on Amazon's mechanical turk (MTurk); 482 participants (272 women, 205 men, 4 others, and 1 unreported; $M_{\text{age}} = 34.64$ years; 77.80% currently employed) completed the survey. Participants were randomly assigned to one cell of a 2 (large vs. small sample mean) by 2 (large vs. small sample size) between-person design.

Procedure

Participants were presented with the scenario described above (Experiment 1, Scenario 3) above with the sample mean (57% vs. 67%) and sample size (30 vs. 3,000) across conditions (see the Supplemental Materials, Experiment 2A Stimuli, for the scenario and measurement items reported verbatim). Instead of asking participants to state a probability estimate (as in Experiment 1), we asked them, "How confident are you that your Corporate Social Responsibility (CSR) campaign has been effective? That is, how confident are you that a majority of people (over 50%) now have a positive opinion of your company?" Participants rated their confidence on a scale from 0 (*not at all confident*) to 10 (*extremely confident*). We chose this rating scale because "'confidence' is a highly subjective construct" (Budescu et al., 2003, p. 181).

While designing the scenario, we intentionally presented the sample size first and sample mean second to rule out the possibility that we made the sample mean more salient than the sample size; the primacy effect would predict that people would be more sensitive to the first presented information, the sample size (Anderson, 1965).

Results

The histogram of the dependent variable across all conditions (see Figure 2a in Supplemental Materials) indicates a mode at 6. This indicates most participants had a confidence level of 6 out of 10. We submitted participants' confidence rating to a 2 (sample size) × 2 (sample mean) ANOVA. We found a significant main effect of the sample mean, $F(1, 478) = 27.89$, $p < .001$, Cohen's $d = .48$, $\eta^2 = .06$. Participants had significantly higher confidence in the effectiveness of their corporate social responsibility campaign when the percentage of positive opinion was 67%, $M = 6.56$, 95% CI [6.36, 6.77], $SD = 1.61$, $N = 240$, rather than 57%, $M = 5.76$, 95% CI [5.55, 5.98], $SD = 1.71$, $N = 242$. However, there was no statistically significant difference between participants' confidence in the effectiveness of their corporate social responsibility campaign across the large, M = 6.26, 95% CI [6.04, 6.48], SD = 1.75, N = 245, and the small, M = 6.06, 95% CI [5.85, 6.27], SD = 1.66, N = 237, sample size conditions, $F(1, 478) = 2.04, p = .15$, Cohen's $d = .12$, $\eta^2 = .00$. The sample mean × sample size interaction was nonsignificant, $F(1, 478) = .89$, $p = .35$, $\eta^2 = .00$. Results from a 2 (sample size) × 2 (sample mean) Bayesian ANOVA were consistent and are reported in Supplemental Materials, Experiment 2A Stimuli.

Posttest

One might argue that participants' confidence in this case is driven by other considerations, such as organizational norms (e.g., that organizations should always keep innovating to stay competitive) rather than a neglect of the sample size. To investigate such alternative explanations, in a posttest study, we presented 215 participants recruited from Mturk with the above scenario and asked them to provide a rationale for their decision. Two participants did not respond. The 213 responses were first assessed by the first author, who came up with six categories that appeared to cover all the responses provided (see Table 3).

Two independent raters categorized participants' responses in these six categories; each response could reflect more than one category. Of the 475 categories indicated for the responses, disagreements occurred 67 times.[2] Disagreements were resolved in consultation with the first author. A final total of 230 categories were found in the 213 reasons given, because some reasons included multiple rationale categories. Among them, 56.52% were based on the sample mean, 2.17% were based on the same size, 2.17% were based on nonstatistical business heuristics, 18.26% were based on nonstatistical information mentioned in the scenario, 4.35% were based on information mentioned in the scenario but without a clear rationale, and 16.52% were irrelevant. Thus, consistent with our argument, participants appear to focus primarily on the sample mean and virtually not at all on the sample size.

Discussion

Experiment 2A found that when making decisions based on samples, individuals are sensitive to variations in the sample mean but not to variations in the sample size. Together, these findings indicated that people tend to judge samples as having similar properties to the population without considering the sample size as a relevant factor. Thus, people's judgments are not consistent with the empirical law of large numbers (Sedlmeier & Gigerenzer, 1997).

Experiment 2B

Experiment 2B aimed to provide a conceptual replication of Experiment 2A with a few extensions. First, we used a different scenario. Second, rather than merely asking about participants' confidence in the findings from the sample, we asked them to make a decision based on the sample findings. As confidence in one's judgment is an indicator of one's willingness to act upon them (Budescu & Rantilla, 2000), our previous findings from confidence judgments should translate into decisions. These alternative ways of operationalizing the independent and dependent variables serve to triangulate the results found in the earlier studies, which is necessary as we are predicting a null effect. Finally, to ensure that participants are motivated to process the information presented and make the decision that they think is truly best, we offered an additional bonus if participants' decision was in the best interest of the company described in the scenario as per statistical principles.

Method

Participants

A survey seeking 200 U.S. residents was posted on MTurk; 216 participants (128 women, 80 men, 8 unreported; $M_{\text{age}} = 39.85$ years) completed the survey. Participants were randomly assigned to one cell of a 2 (large vs. small sample mean) × 2 (large vs. small sample size) between-person design.

Procedure

Participants were presented with a scenario adapted from that in Study 1 in which the sample mean (57% vs. 65%) and sample size (30 vs. 3,000)

[2] The raters accidentally missed three responses thus these were not counted as disagreements.

Table 3
Decision Rationale Participants Used to Determine Their Confidence Ratings

Category	Decision rationale	Example
Reason 1	Sample mean	For example, >50% or majority of users prefer it.
Reason 2	Sample size	For example, 40 users is too few to be reliable.
Reason 3	Business heuristics but not based on statistical evidence	For example, should always keep updating to stay relevant/competitive.
Reason 4	Information mentioned in the scenario that is, not statistical in nature	For example, I would implement the new website because consumers have complained that it was not user-friendly and the visitors count has not risen.
Reason 5	Information mentioned in the scenario but without a clear rationale	For example, there is a 10% shift in those who prefer it. I am taking the risk that eventually the others who prefer the old site will adapt and accept the new site.
Reason 6	Irrelevant, nonsensical	For example, I like to take risks.

were varied across conditions (see Supplemental Materials, Experiment 2B Stimuli for the full scenario). Participants indicated whether they would implement a change ("yes" vs. "no") based on the information provided.

Results

The histogram of the dependent variable across all conditions (see Figure 2b in Supplemental Materials) indicates a mode at 1. This indicates most participants decided to implement the new website. We conducted a logistic regression analysis with participants' decision to change the company website as the dependent variable, and the sample mean, the sample size, and their interaction as predictors. The descriptive statistics are reported in Table 4. We found a significant main effect of the sample mean, $B = .58$, 95% CI [.25, .93], $SE = .17$, *Wald* criterion = 3.35, $p < .001$, $OR = 3.20$. Significantly more participants decided to implement the new website when sample mean was large (86.49%) than when it was small (66.67%). However, there was no main effect of sample size, $B = -.0021$, 95% CI [−.35, .34], $SE = .17$, *Wald* criterion = −.01, $p = .99$, $OR = 1.02$. The proportion of participants deciding to implement the new website was nearly identical across the large sample size condition (76.99%) and the small sample size condition (76.70%). Finally, the sample mean × sample size interaction was nonsignificant, $B = -.0021$, 95% CI [−.35, .34], $SE = .17$, *Wald* criterion = −.01, $p = .99$. Consistent results were obtained from Bayesian contingency tables (see Supplemental Materials, Bayesian Results, Experiment 2B).

Discussion

Experiment 2B conceptually replicated the key finding of Experiment 2A—people made the same decisions irrespective of whether they were informed about the results from a small or a large sample, and even when they were provided with a monetary incentive to make the decision that would be the best for their company.

Experiment 3

Previous research had found that people were more confident in larger samples when samples sizes were varied in a within-participant design (Griffin & Tversky, 1992; Irwin et al., 1956; Masnick & Morris, 2008). However, we predict that when sample sizes are varied in a between-participant design, they would have minimal impact on people's judgment and decisions. Experiment 3 aimed to reconcile our findings with prior literature by testing whether both findings hold simultaneously. To do so, we varied sample sizes both within- and between-participants by varying whether participants saw a large sample size first or the small sample size first. This study was preregistered (see the preregistration file in our OSF project folder).

Method

Participants

A survey seeking 800 U.S. residents was posted on MTurk using the CloudResearch platform. We sought participants using a computer (not a mobile phone or a tablet), had completed at

Table 4
Experiment 2B Results: Dependent Variable by Experimental Condition

	Sample size		Sample mean	
	Small	Large	Small	Large
Decision	$N = 103$	$N = 113$	$N = 105$	$N = 111$
Implement	79 (76.70%)	87 (76.99%)	70 (66.67%)	96 (86.49%)
Not implement	24 (23.30%)	26 (23.01%)	35 (33.33%)	15 (13.51%)

Note. The frequency is followed by percentage values in parenthesis.

least 100 assignments on MTurk and received an approval rating of at least 97%, and belonged to the "CloudResearch approved participants" pool to ensure we sampled high-quality participants. Eight hundred and one participants (462 women, 335 men, and 4 other; M_{age} = 41.64 years) completed the survey.

Procedure

We followed the training procedure of Experiment 1. The details of the scoring rule, the training, and the practice trials and feedback are reported in Supplemental Materials, Experiment 1 Stimuli. Participants proceeded to the actual research scenarios upon completing the practice trials. The task was incentive compatible, similar to that in Experiment 1.

All participants were first presented with Scenario 2 from Experiment 1. Thereafter, participants read both a large and a small sample size version in a randomly assigned order (see Table 5) according to the experimental condition they were in. The sample mean was held constant at 60% across both versions. For each sample, participants were asked,

> Based on the above survey results, on a scale from 50% (probable at chance level) to 100% (probable without a doubt), what do you think is the probability that a majority (i.e., over 50%) of all your users find your new website more user-friendly than the old one?

Table 5
Experiment 3 Design

Between-participant conditions	Within-participant conditions	
	Condition 1	Condition 2
Between-participant condition 1	Sample size 30	Sample size 3,000
Between-participant condition 2	Sample size 3,000	Sample size 30

Participants were asked to indicate their probability estimate using a slider bar ranging from 50% to 100%, which could be adjusted in 5% increments (see Supplemental Materials, Experiment 3 Stimuli).

Results

We first conducted an independent sample *t* test to examine the effect of the sample size in the first scenario that participants received (i.e., a between-participant comparison). We found that participants' estimated probability of the population preference based on the sample preference did not differ between the large, M = 72.59, 95% CI [71.14, 74.04], SD = 14.75, N = 400, and the small, M = 71.62, 95% CI [70.19, 73.05], SD = 14.60, N = 401, sample size conditions, $t(799) = .93$, $p = .35$, Cohen's $d = .066$.

We then conducted a repeated measure ANOVA with sample size as a within-participant factor and sample size order as a between-participant factor (small sample first vs. large sample first). We did not find a statistically significant between-participant main effect of sample size order, $F(1, 881.06) = 2.35$, p = .13, Cohen's $d = .11$.[3] There was also no statistically significant within-participant main effect of the sample size, $F(1, 81.20) = 1.18$, $p = .28$, Cohen's $d = .06$. However, there was a significant interaction, $F(1, 2491.48) = 36.17$, $p < .001$, Cohen's $d = .41$. Specifically, when participants saw the small sample size first, they estimated a higher probability that the finding from the sample would generalize to the population in the

[3] We converted the effect size obtained in partial eta squared to Cohen's d using this online converter: https://www.psychometrica.de/effect_size.html.

subsequent large sample condition, M = 74.51, 95% CI [73.03, 75.99], SD = 15.12, N = 400, than in the prior small sample size condition, M = 71.56, 95% CI [70.13, 72.99], SD = 14.68, N = 400, $F(1, 796) = 13.84, p < .001$, Cohen's d = .26. In contrast, when participants saw the large sample size first, they estimated a similar probability in the subsequent small sample size condition, M = 70.53, 95% CI [69.03, 72.02], SD = 15.12, N = 398, as in the prior large sample size condition, M = 72.58, 95% CI [71.13, 74.02], SD = 14.68, N = 398, $F(1, 796) = .95, p = .33$, Cohen's d = .07. We obtained consistent results from a Bayesian independent sample t test and repeated measure ANOVA, which is reported in the Supplemental Materials, Bayesian Results, Experiment 3.

Similar to the pattern in Experiment 1, 60% (i.e., the sample mean percentage in the experiment stimulus) was the modal response (See Figure 3 in Supplemental Materials). The three plausible explanations provided in Experiment 1 hold here. As a robustness test, we excluded participants who responded with 60% to both scenarios in Experiment 3 (187 participants were excluded), but the results did not change. Specifically, participants' estimated probability of the population preference based on the sample preference did not differ between the large, M = 76.51, 95% CI [74.84, 78.18], SD = 14.85, and the small, M = 75.08, 95% CI [73.41, 76.75], SD = 14.99, sample size conditions, $t(612)$ = 1.19, p = .24, Cohen's d = .10.

Discussion

Experiment 3 conceptually replicated the key finding of the previous experiments that people are minimally sensitive to variations in sample sizes in a between-participant design—they perceived the results from a small sample to be similarly representative of the population as the results from a large sample in a between-person analysis. However, consistent with prior research (Griffin & Tversky, 1992; Irwin et al., 1956; Masnick & Morris, 2008), we found that people were sensitive to within-subject sample size variations such that they perceived the results from a large sample to be more representative of the population than those from a small sample if they saw the large sample *after* they saw the small sample. Although we had no reason to predict that the reverse should not hold, this effect was not statistically significant in our analysis.

A possible explanation is that it might be easier for participants to process the second set of information if it confirms, rather than challenges, the conclusion from the first set of information. For instance, when participants saw the small sample size first, they were leaning toward a favorable conclusion (i.e., that a majority of customers find the new website more user-friendly); subsequently seeing the large sample size confirmed their initial opinion and thus increased their probability estimate. When participants saw the large sample size first, they also leaned toward the same favorable conclusion. However, the results of small sample size would require them to potentially reverse their initial opinion to conclude that a majority of customers did not find the new website more user-friendly, which would require more cognitive processing. Thus, participants might dismiss the small sample size and insufficiently adjust their probability estimate based on the new information.

Experiment 4

The previous experiments documented that people were not very sensitive to variations in sample sizes by a factor of 100. However, our finding stands in contrast to past research that people are indeed sensitive to variations in sample sizes in the range of 1–6 (Masnick & Morris, 2008), 10–20 (Irwin et al., 1956), and 3–33 (Griffin & Tversky, 1992). This contrast raises the possibility that perhaps even in a between-participant design, people might be sensitive to variations in very small sample sizes but fail to take the sample size into account once it crossed a certain threshold. Indeed, recent research suggested that sensitivity to sample size may follow a curvilinear function (Obrecht, 2019). However, past research on this topic has focused on a relatively narrow range of sample sizes (typically up to the 30s), making it difficult to assess the threshold beyond which people are no longer sensitive to variations in the sample size. We address these limitations by using a between-participant design and varying sample sizes from 3 to 1,200. Anecdotal evidence from undergraduate, MBA, and EMBA classes indicates that many laypeople hold the notion that sample sizes of 30 are sufficient for making statistical conclusions, perhaps overgeneralizing the rule of thumb that once the sample size reaches 30, the sample mean is approximately normally distributed

(Kromer, 2015). Thus, in this study, we gradually varied the sample size across participants and tested whether people are sensitive to variations in sample sizes below a threshold (e.g., 30) but are no longer sensitive above that threshold (e.g., 30).

Method

Participants

Consistent with our previous experiments, we decided on a target sample size of 100 participants per condition. A survey seeking 1,200 U.S. residents was posted on MTurk; 1,231 participants (740 women, 473 men, 8 others, 10 participants did not report gender; $M_{\text{age}} = 36.77$ years) completed the survey. Participants were randomly assigned to one of the 12 sample size conditions.

Procedure

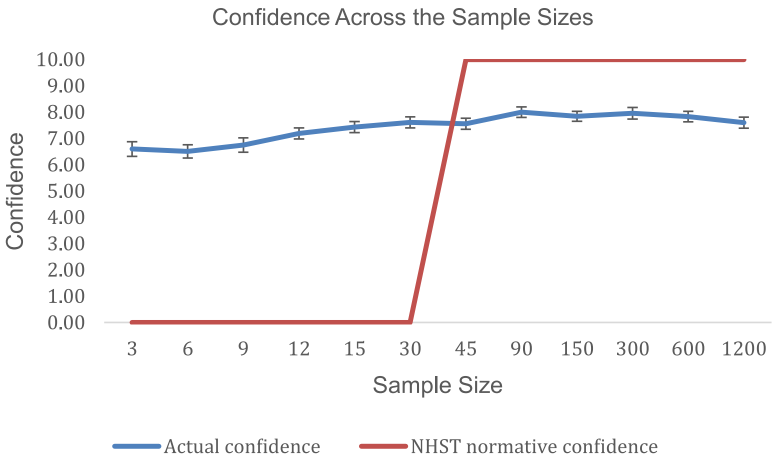
All participants read a scenario (see Supplemental Materials, Experiment 4 Stimuli) concerning assessment of the popularity of a new policy in a large organization via a survey of N employees, where $N \in \{3, 6, 9, 12, 15, 30, 45, 90, 150, 300, 600, 1{,}200\}$. All participants also read that the survey showed that 67% (X out of N employees sampled) supported the new policy, whereas the remaining 33% (Y of the N employees sampled) did not. After they read the scenario, participants were asked, "How confident are you that a majority of all your employees (over 50%) would want the new forced vacations policy to be instituted? Please rate on a scale from 0 (*not at all confident*) to 10 (*extremely confident*)."

Results

Figure 1 presents participants' mean confidence across the various sample sizes (log transformed). The standard error and 95% confidence interval in each condition can be found in Table S2 of the Supplemental Materials. We plotted the "normative" confidence based on the conventional rule of accepting/rejecting the null at a normative p value of .05 within the null hypothesis significant test (NHST) framework. We acknowledge that adopting a different assumption or rule of thumb will lead to a different "normative" confidence plot. As depicted in the figure below, for small samples ($N \leq 30$, people's confidence seems to be well above the normative confidence warranted by NHST statistical principles, but for large samples ($N > 30$), their confidence stays well below the normative confidence level. The histogram of the mean confidence across all conditions (see Figure 4 in Supplemental Materials) indicates a mode at 8. This indicates most participants had a confidence level of 8 out of 10.

Figure 1

Mean Confidence Across the Log-Transformed Sample Sizes



Note. Error bars represent the standard error of the mean. NHST = null hypothesis significant test. See the online article for the color version of this figure.

As evident in Figure 1 above, participants' confidence appeared to increase gradually until about a sample size of 30, and then largely plateaued. To test this idea statistically, we let a dummy variable c equal to 1 when $N < 30$ and 0 otherwise, and ran the following regression:

$$\begin{aligned} y' = b0 + b1 * (1 - c) + b2 * c * \ln(N) \\ + b3 * (1 - c) * \ln(N), \end{aligned} \tag{1}$$

where N is the sample size.

Here, b1 indicates the increase in confidence from $N < 30$ to $N >= 30$; b2 indicates the increase in confidence as the sample size increases below 30, and b3 indicates the increase in confidence as the sample size increases above 30. We expect statistically significant coefficients on b1 and b2, and a nonsignificant coefficient on b3. As shown in Table 6, the significant coefficient on b2 indicated a linear increase in confidence when the sample size increased from 3 to 30, b2 = 1.19, $p =$ .003. In contrast, the nonsignificant b3 indicates that people's confidence did not change when the sample size increased from 30 to 1,000, b3 = .05, $p = .74$.

We conducted a sensitivity analysis by shifting the threshold for the dummy coding above and below 30 (see Table 6 for the results). This analysis indicates that a sample size of 30 is the first sample size in which we obtain the expected pattern of results—a sharp jump in confidence between sample sizes below 30 versus equal to or above 30 (i.e., a significant b1), sensitivity to sample sizes up to 30 (i.e., a significant b2), and lack of to sample sizes above 30 (i.e., a nonsignificant b3). The first two rows of Table 6 indicate that participants were insensitive to variations in the sample size below 12 but were sensitive to variations above 12. The third row indicates that there is a jump in confidence when the sample size reaches 15, but no sensitivity to variations below 15 or above 15. The last two rows indicate that there is a jump in confidence when the sample size reaches 45 and 90, sensitivity to variation below 45 and 90, respectively, but no sensitivity to variations above.

Discussion

These findings helped resolve the seeming inconsistency between our findings that people are minimally sensitive to variation in the sample size by multiple orders of magnitude, and past research that people are sensitive to variations in sample sizes under 30. Experiment 4 found that although people are sensitive to changes in sample sizes from approximately 3–30, people's confidence does not increase as the sample size exceeds 30. Further, we found that even with a sample size of three, participants' mean confidence level was 6.6 out of 10, indicating that people have quite high confidence in data from incredibly small samples, consistent with prior research (Masnick & Morris, 2008). Additionally, for sample sizes above 45, as per NHST principles, people should have high confidence that the findings can be generalized from the sample to the population; however, their mean confidence ranged from 7.5 to 8, not substantially higher than the mean confidence of 6.6–7.6 for samples sizes for which the findings should not be generalized from the sample to the population as per NHST principles. A more fine-grained contrast between our finding and the findings for earlier studies is that we demonstrated a lack of sensitivity among very small sample sizes (i.e., 3, 6, and 9) and among very large sample sizes. Taken together, our results suggest that people might be sensitive only to sample size variations within a very narrow range.

Table 6
Experiment 4 Model Testing and Sensitivity Test Results

N	Coefficient b1 (*SE*)	Coefficient b2 (*SE*)	Coefficient b3 (*SE*)
<9	.17 (.71)	−.28 (1.05)	.36 (.10)***
<12	.77 (.55)	.24 (.65)	.23 (.11)*
<15	1.37 (.51)**	.84 (.49)	.14 (.13)
<30	1.82 (.52)***	1.19 (.40)**	.05 (.16)
<45	2.03 (.55)***	1.20 (.29)***	−.03 (.19)
<90	2.62 (.68)***	1.06 (.23)***	−.29 (.25)

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Experiment 5

Although extensive research has examined the deficits of people's statistical intuitions, few researchers have examined methods to debias people's inferences. If people's insensitivity to sample sizes is because of an inherent belief in the law of small numbers, the idea that beyond a small threshold, samples of all sizes are about equally representative of the population, then presenting statistics that explicitly specify the extent to which the data are consistent with the null hypothesis versus the alternate hypothesis could help reduce this bias. When behavioral scientists report data from samples, they typically report *p* values that indicate the extent to which the sample result would be obtained given a particular hypothesis about the population, and now increasingly, the Bayes Factor, which indicate the amount of evidence supporting the alternative hypothesis over the null. Analogously, we provided participants in this study with a scientifically accurate yet jargon-free interpretation of the *Bayes Factor*. We hypothesized that presenting sample means and sample sizes along with the *Bayes Factor* interpretation would reduce people's insensitivity to sample sizes.

Method

Participants

A survey seeking 400 U.S. residents was posted on MTurk; in response, 379 participants (206 women, 167 men, 6 others; $M_{\text{age}} = 38.21$ years; 76.50% currently employed) completed the experiment. Participants were randomly assigned to one cell of a 2 (sample size: small vs. large) by 2 (intervention: present vs. absent) between-participant design.

Procedure

We used Scenario 2 from Experiment 1 with a sample mean of 60%. We varied the sample sizes between conditions (30 vs. 3,000; see Supplemental Materials, Experiment 5 Stimuli). In the two intervention conditions, we added the following interpretations of the Bayes Factor:

> [Small sample size condition] Your company's statistician further explained that the survey results demonstrate that the evidence in favor of the conclusion that most people have no preference between the old and the new websites is 2.5 times as strong as the evidence in favor of the conclusion that most people prefer the new website.

> [Large sample size condition] Your company's statistician further explained that the survey results demonstrate that the evidence in favor of the conclusion that most people prefer the new website is over 100,000 times as strong as the evidence in favor of the conclusion that most people have no preference between the old and the new websites.[4]

For ease of interpretation, we presented participants with a graphic representation of this statement (see Supplemental Materials, Experiment 5 Stimuli). In the no-intervention condition, participants were not provided with the above information. After they read the scenario, participants indicated their confidence that a majority (i.e., >50%) of users prefer the new website to the old one on a scale from 0 (*not at all confident*) to 10 (*extremely confident*).

Results

The histogram of the mean confidence across all conditions (see Figure 5 in Supplemental Materials) indicates a mode at 8. This indicates most participants had a confidence level of 8 out of 10. We first submitted participants' confidence rating to a 2 (sample size) × 2 (intervention) ANOVA. We found a significant main effect of the sample size, $F(1, 378) = 55.19, p < .001$. The main effect of the intervention condition was also statistically significant $F(1, 378) = 18.79$, $p < .001$, $\eta^2 = .13$. Furthermore, there was a statistically significant interaction between sample size and intervention condition, $F(1, 378) = 44.25$, $p < .001, \eta^2 = .11$. A 2 (sample size) × 2 (sample mean) Bayesian ANOVA replicated these findings (see Supplemental Materials, Bayesian Results, Experiment 5).

Follow-up contrasts showed that sample size made no difference to participants' confidence ratings in the no-intervention condition, replicating the finding from the earlier experiments: Participants' confidence in the effectiveness of their new website did not differ statistically significantly between the large sample size—no-intervention condition, $M = 7.34$, 95% CI [6.90, 7.77], $SD = 2.15$, $N = 95$, and the small

[4] The Bayes Factor for each intervention message is computed using the proportionBF function in *R* with rscale = 1 (consistent with the specification used in Experiment 1).

sample size—no-intervention condition, M = 7.17, 95% CI [6.74, 7.60], SD = 2.16, N = 97, F(1, 375) = .31, p = .58, Cohen's d = .08. In contrast, participants had higher confidence in findings in the large sample size—intervention condition, M = 7.85, 95% CI [7.41, 8.29], SD = 2.15, N = 93, than in the small sample size—intervention condition, M = 4.73, 95% CI [4.30, 5.17], SD = 2.15, N = 94, F(1, 375) = 97.85, $p <$ .001, Cohen's d = 1.38. The means are plotted in Figure 2 below, which indicates that the main effect of sample size is entirely driven by the intervention condition.

Discussion

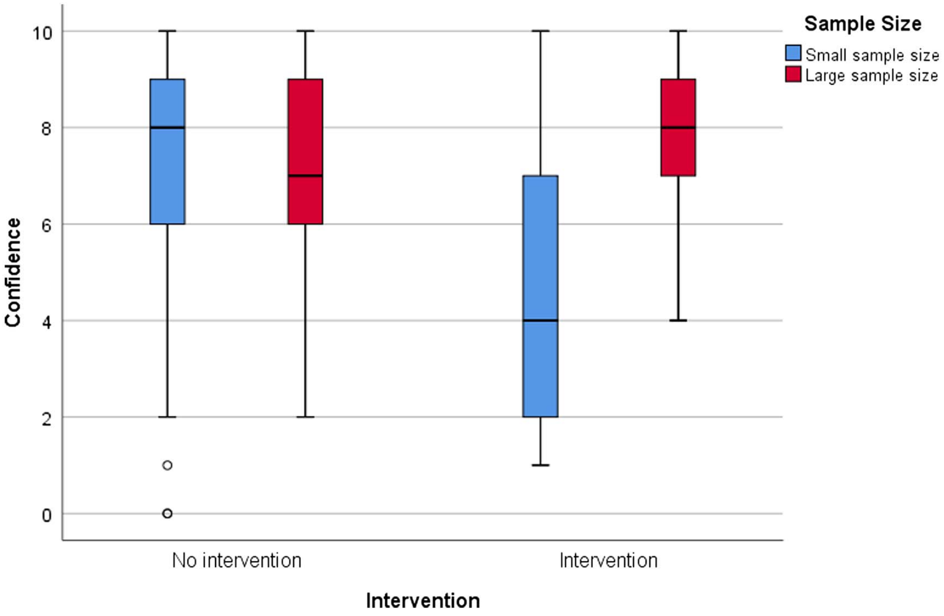
Experiment 5 identified a simple intervention that makes people more sensitive to sample sizes—specifying the Bayes factor, along with an easy-to-understand interpretation of the Bayes factor—when presenting findings from samples. This finding has an important practical implication: Whenever providing findings from samples, the information provider should provide p values or Bayes factor and their interpretation to help decision-makers evaluate the information and make decisions. Note that the intervention did not completely de-bias people—in the large sample condition, participants' mean confidence was still far from the upper limit of the response scale, and in the small sample condition, it was far from the lower limit of the response scale; nevertheless, it substantially reduced people's lack of sensitivity to sample sizes.

General Discussion

Six studies documented a general lack of sensitivity to sample sizes varying by multiple orders of magnitude in judgments and decisions based on a single sample. In contrast to prior research that claimed people's confidence in samples increased with the increase of sample sizes in within-participant designs (Griffin & Tversky, 1992; Irwin et al., 1956; Masnick & Morris, 2008), Experiment 1 found that people's estimated probability that the sample mean corresponds to the population mean is minimally sensitive to large differences in sample sizes in a between-participant design. Experiment 2A

Figure 2

Interaction Between Sample Size and Intervention Condition (Whether Intervention Is Present or Absent) on Participants' Confidence Rating



Note. See the online article for the color version of this figure.

found that when making decisions based on samples, people are strongly influenced by information about the sample mean—they were more confident in the findings of the sample when the mean was 67% rather than 57%. However, people's confidence in the findings were nearly identical irrespective of whether the findings came from a small sample of 30 or a much larger sample of 3,000. Experiment 2B replicated the findings of Experiment 2A while measuring people's decisions rather than their confidence judgments. People were more likely to decide to change their company website when told that 67% rather than 55% of their sample preferred the new website, irrespective of whether the sample size was of 30 or 3,000.

Experiment 3 partially replicated prior research by demonstrating that people are sensitive to within-person variations in sample sizes. We found that participants who were first exposed to a small sample were subsequently more confident to a large sample, but not vice-versa. One possible explanation is that participants update their beliefs if the second set of information confirmed, rather than challenged, the conclusion that they drew from the first set of information. Experiment 3 also found that participants were minimally sensitive to variations in the sample size in a between-subject comparison, conceptually replicating the results from our previous studies. Experiment 4 demonstrated that people's confidence in the findings of a sample increased as the sample size increased from 3 to 30, but increases in sample size from 30 to 1,200 did not lead to any further increase in confidence. This finding indicated that laypeople hold the notion that sample sizes of 30 are sufficient for making statistical conclusions. Experiment 5 identified a simple intervention that reduced people's insensitivity to sample sizes—providing people with a layperson interpretation of the Bayes Factor.

With each experiment taken by itself and with little assumption about prior odds, the Bayes Factor (see Supplemental Materials, Bayesian Results) generally indicated greater evidence for the null hypothesis, consistent with our hypothesis that people are more likely insensitive rather than sensitive to variations in the sample size. Bayesian analyses using an informed prior for the series of experiments suggest that the 95% credibility interval for the effect size of the series of experiments is narrowly distributed around zero (see Supplemental Materials). However, in nearly every experiment (except Experiment 1 Scenario 1 and Experiment 3), participants had slightly more confidence in the large sample than the small sample. Therefore, people are *nearly* insensitive to large variations in the sample size, not *completely* insensitive. Nevertheless, across all experiments using Likert scale dependent measures, the magnitude of difference in participants' confidence between the small and large samples is less than 0.5 units on a 10-point scale, when rationally speaking, the actual difference should have been substantially greater (i.e., participants should have very little confidence in the small sample and near perfect confidence in the large sample). Similarly, the overall effect size of people's sensitivity to sample size across the experiments is tiny at Cohen's $d = .07$ despite our experiments being well powered and despite our sample sizes being drastically different across conditions, indicating that people's very slight sensitivity to sample size is incommensurate with statistical principles.

Constraints on Generality

Our participants came from diverse cultural backgrounds and diverse walks of life (e.g., working professionals, students, retirees). They were not people who do data analysis for a living, so it is possible that expert data analysts will be less susceptible to this bias. As statistical training becomes a more central part of education, the bias identified in the current research might weaken over time.

Theoretical Implications and Future Directions

This research addresses a debate on the extent to which people take sample sizes into account when making judgments and decisions. Initial research indicated that people are not sensitive to sample sizes when deciding whether large or small samples are more likely to deviate from the population mean, which led researchers to conclude that people believe in the *law of small numbers* (Tversky & Kahneman, 1971), the idea that samples of any given size should be representative of the population. However, subsequent research has concluded that people believe in the *empirical law of large numbers* (Sedlmeier & Gigerenzer, 1997), that larger samples are more representative of the population, based on the finding that when large and small samples were juxtaposed, people placed higher confidence in larger samples (e.g., DuCharme & Peterson, 1969; Evans & Pollard, 1982; Griffin & Tversky, 1992; Irwin et al., 1956; Koslowski et al., 1989; Kunda & Nisbett, 1986; Levin, 1975; Masnick & Morris, 2008). The present research documents that this assumption is premature, as the studies on which this conclusion is based likely suffer from experimenter demand effects because researchers have nearly exclusively used within-participant designs. Using a between-participant design free from experimenter demand effects, we found evidence suggesting that people operate on the law of small numbers such that they tend not to differentiate findings from samples varying by a factor of 100.

The findings of the current research are consistent with known cognitive and decision patterns. First, consistent with the Obrecht et al. (2007) and Morris and Masnick (2015), we found that people pay differential attention to different statistical features—they are highly sensitive to variations in the sample mean but minimally sensitive to variations in the sample size when both were presented within a single scenario. In addition, our findings are consistent with the representativeness heuristic (Kahneman & Tversky, 1972), gambler's fallacy (Lindman & Edwards, 1961), the tendency to make decisions based on limited experience (Hertwig & Pleskac, 2010), and undersampling of failure examples in management decision-making (Denrell, 2003), all of which suggest that people generally lack the statistical intuition that large samples are more reliable than small ones.

In addition to replicating prior findings that people tend to be overconfident in small samples, the present study also suggests that people are underconfident in large samples. Simultaneous overconfidence in small samples and underconfidence in large numbers may have contributed to the insensitivity to sample size we observed. The present study also uncovered some nuances in people's pattern of sensitivity to sample size variations. Particularly, Experiment 4 found that people's confidence did not differ across sample sizes 3, 6, and 9, or above 30. We observed a sharp increase in confidence for sample sizes below 30 versus equal to or above 30. In other words, taking a between-participant design that is free of experimenter demand effects, we found that people are neither completely insensitive nor highly sensitive to sample size variations. Instead, people might be sensitive to sample size variations within a narrow range of approximately 10–30. Taken together, this finding lends support to the anecdotal account that laypeople hold the notion that sample sizes of about 30 are sufficient for making statistical conclusions because they overgeneralize the rule of thumb that once the sample size reaches 30, the sample mean is approximately normally distributed (Kromer, 2015).

Although extensive research has focused on uncovering people's cognitive biases in the domain of statistical judgments and decision-making, little research has examined methods to correct people's biases in this domain (cf. Fong et al., 1986). The present research further contributes to the decision-making literature by proposing a simple yet effective intervention that helps alleviate people's insensitivity to sample size. We found that a nontechnical message that conveys the Bayes Factor and its interpretation can help people make better decisions. Although past research has found that extensive statistical training can also improve people's reasoning about everyday problems that require them to apply the law of large numbers (Fong et al., 1986), the training's effectiveness maybe limited due to factors such as memory decay, cognitive load, and fatigue. In contrast, our proposed invention may be a simpler and more effective way to nudge people toward better decision-making. Our research makes the strong suggestion that all reports of findings from samples be accompanied with a jargon-free interpretation of the Bayes Factor (or *p* value) that interprets the strength of the evidence.

Our findings that people are systematically underconfident about findings from large samples even though the findings are highly reliable statistically suggest that people's statistical intuition might be fundamentally against the idea that sufficiently large samples give a fairly accurate view of the population. Future research can test whether people believe that it is possible to make solid conclusions from large enough samples, or whether they believe that no matter how large a sample, it can never give an accurate picture of the population—to get an accurate picture, one needs to study the entire population (Bar-Hillel, 1979).

As researchers, we clearly realize that the same finding is much more believable from a sample of

3,000 than from a sample of 30. But shockingly, laypeople, do not appear to share this intuition. With the proliferation of statistics in the news media and in organizations that call for evidence-based decision-making, the current findings indicate that people might not have the correct intuition as to "what counts as evidence," making it difficult for them to correctly use statistics and research evidence to guide their inferences and decisions. To circumvent this common deficit and improve decision quality, the present study suggests that all statistics need to be accompanied with statistical inferences (whether in the null hypothesis significance testing or the Bayesian frameworks), along with layperson interpretations of these statistical inferences.

References

ABC News. (2020, March 26). *Number of new coronavirus cases in NSW drops, overall infections now at 1,219*. https://www.abc.net.au/news/2020-03-26/coronavirus-cases-in-nsw-increase-but-new-infections-down/12090784

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, *2*(1), 1–9. https://doi.org/10.1037/h0021966

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24*(2), 245–257. https://doi.org/10.1016/0030-5073(79)90028-X

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. https://doi.org/10.1016/S0001-6918(00)00037-8

Budescu, D., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Budescu, D., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177. https://doi.org/10.1002/bdm.547

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601. https://doi.org/10.1111/j.1468-0262.2006.00719.x

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. *Organization Science*, *14*(3), 227–243. https://doi.org/10.1287/orsc.14.2.227.15164

DuCharme, W. M., & Peterson, C. R. (1969). Proportion estimation as a function of proportion and sample size. *Journal of Experimental Psychology*, *81*(3), 536–541. https://doi.org/10.1037/h0027914

Evans, J. S. B. T., & Pollard, P. (1982). Statistical judgement: A further test of the representatives construct. *Acta Psychologica*, *51*(2), 91–103. https://doi.org/10.1016/0001-6918(82)90054-3

Feller, W. (1957). *An introduction to probability theory and its applications* (2nd ed.). Wiley.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, *18*(3), 253–292. https://doi.org/10.1016/0010-0285(86)90001-0

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603. https://doi.org/10.2307/2946693

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*, *130*(12), 995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Grady, D. (2020, May 18). Moderna coronavirus vaccine trial shows promising early results. *New York Times*. https://www.nytimes.com/2020/05/18/health/coronavirus-vaccine-moderna.html

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Hadar, L., & Fox, C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, *4*(4), 317–325.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237. https://doi.org/10.1016/j.cognition.2009.12.009

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*(5), 680–695. https://doi.org/10.1037/0022-3514.86.5.680

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an expanded judgment situation. *Journal of Experimental Psychology*, *51*(4), 261–268. https://doi.org/10.1037/h0041911

Jarvstad, A., Hahn, U., Rushton, S. K., & Warren, P. A. (2013). Perceptuo-motor, cognitive, and description-based decision-making seem equally good. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(40), 16271–16276. https://doi.org/10.1073/pnas.1300239110

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development*, *60*(6), 1316–1327. https://doi.org/10.2307/1130923

Kromer, T. (2015). *How to verify your assumptions at small sample sizes*. https://kromatic.com/blog/how-to-verify-your-assumptions-at-small-sample-sizes/

Kudryavtsev, A., & Pavlodsky, J. (2012). Description-based and experience-based decisions: Individual analysis. *Judgment and Decision Making*, *7*(3), 316–331.

Kunda, Z., & Nisbett, R. E. (1986). Prediction and the partial understanding of the law of large numbers. *Journal of Experimental Social Psychology*, *22*(4), 339–354. https://doi.org/10.1016/0022-1031(86)90019-3

Kutzner, F. L., Read, D., Stewart, N., Brown, G., Kutzner, F. L., Stewart, N., & Brown, G. (2016). Choosing the devil you don't know: Evidence for limited sensitivity to sample size—based uncertainty when it offers an advantage. *Management Science*, *63*(5), 1519–1528. https://doi.org/10.1287/mnsc.2015.2394

Levin, I. P. (1975). Information integration in numerical judgments and decision processes. *Journal of Experimental Psychology: General*, *104*(1), 39–53. https://doi.org/10.1037/0096-3445.104.1.39

Lindman, H., & Edwards, W. (1961). Supplementary report:unlearning the gambler's fallacy. *Journal of Experimental Psychology*, *62*(6), 630. https://doi.org/10.1037/h0046635

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development*, *79*(4), 1032–1048. https://doi.org/10.1111/j.1467-8624.2008.01174.x

Morris, B. J., & Masnick, A. M. (2015). Comparing data sets: Implicit summaries of the statistical properties of number sets. *Cognitive Science*, *39*(1), 156–170. https://doi.org/10.1111/cogs.12141

Obrecht, N. A. (2019). Sample size weighting follows a curvilinear function. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 614–626. https://doi.org/10.1037/xlm0000615

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t* tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, *14*(6), 1147–1152. https://doi.org/10.3758/BF03193104

Obrecht, N. A., & Chesney, D. L. (2013). Sample representativeness affects whether judgments are influenced by base rate or sample size. *Acta Psychologica*, *142*(3), 370–382. https://doi.org/10.1016/j.actpsy.2013.01.012

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, *117*(3), 775–816. https://doi.org/10.1162/003355302760193896

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, *10*(1), 33–51. https://doi.org/10.1002/(SICI)1099-0771(199703)10:1<33::AID-BDM244>3.0.CO;2-6

Sung, B. (1966). *Translations from James Bernoulli*. Harvard University.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes*, *47*(2), 289–312. https://doi.org/10.1016/0749-5978(90)90040-G

Received May 12, 2021
Revision received March 10, 2022
Accepted March 15, 2022 ■