SRRI #247

ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 47, 289–312 (1990)

Understanding the Effects of Sample Size on the Variability of the Mean

ARNOLD D. WELL, ALEXANDER POLLATSEK, AND SUSAN J. BOYCE

University of Massachusetts

In the first three experiments, we attempted to learn more about subjects' understanding of the importance of sample size by systematically changing aspects of the problems we gave to subjects. In a fourth study, understanding of the effects of sample size was tested as subjects went through a computer-assisted training procedure that dealt with random sampling and the sampling distribution of the mean. Subjects used sample size information more appropriately for problems that were stated in terms of the accuracy of the sample average or the center of the sampling distribution than for problems stated in terms of the tails of the sampling distribution. Apparently, people understand that the means of larger samples are more likely to resemble the population mean but not the implications of this fact for the variability of the mean. The fourth experiment showed that although instruction about the sampling distribution of the mean led to better understanding of the effects of sample size, subjects were still unable to make correct inferences about the variability of the mean. The appreciation that people have for some aspects of the law of large numbers does not seem to result from an in-depth understanding of the relation between sample size and variability. © 1990 Academic Press, Inc.

The "law of large numbers" states that as sample size increases, statistics of a sample become less variable and more closely estimate the corresponding parameters of the population from which the sample was selected. There have been a number of demonstrations that naive subjects ignore information about sample size when making judgments about samples and populations, leading to the suggestion that they do not have an intuitive understanding of the law of large numbers.

In a seminal series of experiments, Kahneman and Tversky (1972) demonstrated that high-school students ignored sample size when they were given information about some characteristic of a population and then were asked to make judgments about the corresponding characteristic in samples that varied in size. An example of one kind of problem, in which subjects were asked to judge which of three multiple-choice options was correct, is as follows:

The research described in this paper was supported by NSF Grant BNS85-09991. Requests for reprints should be sent either Arnold Well or Alexander Pollatsek at the Department of Psychology, University of Massachusetts, Amherst, MA 01003.

0749-5978/90 $3.00
Copyright © 1990 by Academic Press, Inc.
All rights of reproduction in any form reserved.

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
>
> For a period of one year, each hospital recorded the days on which (more/less) than 60% of the babies born were boys. Which hospital do do you think recorded more such days?
>
> The larger hospital
> The smaller hospital
> About the same (i.e., within 5% of each other)

The proportion of boys would be expected to be more variable for the smaller hospital. However, most subjects chose the answer "about the same," suggesting that they were unaware of how sample size influences variability. Similar effects were found for other cover stories using continuous distributions and other characteristics (mean, median) of samples.

In another kind of task, subjects were asked to generate the "subjective sampling distribution" for a sample characteristic. For example, told that the average height of males was 170–175 cm and that a regional induction center tested N men daily, subjects were asked to estimate the probability with which the average height of these N men fell into seven categories around the population average (up to 160 cm, 160–165 cm, . . . , more than 185 cm) for $N = 10$, 100, and 1000.

Sampling theory dictates that the variance of the sampling distribution of the mean should decrease with increasing N. The median distributions produced by subjects were, however, virtually the same for all three values of N. Kahneman and Tversky (1972) concluded that "the notion that sampling variance decreases in proportion to sample size is apparently not part of man's repertoire of intuitions." They explained this insensitivity to sample size in terms of the "representativeness" heuristic, according to which judgments about populations and samples are made on the basis of the similarity between their essential features. Because sample size is a characteristic of samples but not populations, it cannot be used as one of the essential features by which similarity is judged.

Kahneman and Tversky's "subjective sampling distribution" problems have subsequently been criticized as being very difficult for subjects to understand (Evans & Dusoir, 1977; Olson, 1976; Teigen, 1974) and replications have found a great deal of variation in the distributions produced by different subjects, which seems to suggest a lack of understanding. More importantly, there have been demonstrations that subjects do sometimes use sample size information appropriately, although, apparently, not always for the right reasons (Bar-Hillel, 1979, 1982; Evans & Dusoir, 1977; Evans & Pollard, 1985). Subjects seem to do particularly well on

what we might call *accuracy* questions in which subjects are asked which of two samples more accurately estimates some characteristic of the population. Bar-Hillel (1982), for example, found that over 80% of her subjects correctly chose the larger sample in the following problem:

> Two pollsters are conducting a survey to estimate the proportion of voters who intend to vote YES on a certain referendum. Firm A is surveying a sample of 400 individuals. Firm B is surveying a sample of 1000 individuals. Whose estimate would you be more confident in accepting?
>
> Firm A's____ Firm B's____ About the same____

Nisbett and his associates have used a different kind of testing procedure. They present subjects with a brief (several paragraphs) description of some scenario and ask them to make written comments about the reasoning of one of the characters described in the scenario; these comments are then scored for understanding of statistical principles. They claim that people who have not had any formal statistical training often use the law of large numbers when solving problems in everyday life and that peoples' understanding of statistical rules are enhanced even by brief and highly abstract training (Fong, Krantz, & Nisbett, 1986; Nisbett, Krantz, Jepson, & Kunda, 1983).

We are therefore left in the rather uncomfortable position of knowing that people sometimes use information about sample size appropriately and sometimes they do not, but not knowing a great deal about why. Do people actually understand the law of large numbers but sometimes fail to realize that it should be applied to a particular problem, thus making what Kahneman and Tversky (1982) would term an "error of application"? Or do people have an appreciation for some aspects of the law of large numbers but lack an in-depth understanding—so that they are capable of using sample size information appropriately in some kinds of problems but not in others? In fact, we know little about what form subjects' understanding of the effects of sample size takes and what rules or heuristics they use in choosing correct or incorrect answers to problems.

In a recent review, Pollard and Evans (1983) identified the demand characteristics of the task and the salience of the size information as important variables in determining whether information about sample size will be used. Although these variables are undoubtedly important, we believe more attention should be paid to the specific *types of judgments* subjects are required to make. Successful demonstrations of subjects' sensitivity to sample size have generally required subjects to make judgments about the accuracy with which a sample characteristic estimates the corresponding population characteristic or the extent to which a large or small sample constitutes evidence for some state of the population characteristic. In contrast, Kahneman and Tversky's problems that resulted in poor performance gave people information about a population

and asked for judgments about whether characteristics of large or of small samples were more likely to take on certain values. If problems that require the latter type of judgment result in poorer performance, it is not clear whether this is because they (a) require a more sophisticated understanding of the law of large numbers, (b) elicit different and less appropriate heuristics, or (c) simply are harder to understand than problems that require the first type of judgment. The distinction raises the possibility that we can learn more about subjects' understanding of the effects of sample size by looking at how performance changes as the nature of the required judgment is changed.

In the first three experiments reported here, we made systematic changes in the problems we gave to subjects, in some cases asking subjects to judge which of two samples was more likely to fall closer to the population mean and in others, asking them to judge which of two samples was likely to deviate more from the population mean. In the third experiment, we also attempted to elicit subjects' comments about why they answered the way they did. In a fourth study, understanding of the effects of sample size was tested as subjects went through a computer-assisted training procedure that dealt with random sampling and the sampling distribution of the mean.

EXPERIMENT 1

In Experiment 1, we compared performance on "accuracy" and "tail" versions of problems in the same group of subjects. In the accuracy versions, subjects were asked whether the average of a large or of a small sample was more likely to be close to the population average. In the tail versions, subjects were asked to estimate how likely it was that the sample average exceeded a value that was at some distance from the population average. Experiment 1 employed two different cover stories: the post office problem, which was a modification of the induction center problem used by Kahneman and Tversky (1972) (see Table 1) and the weighing problem. In the former, the population average was the national average height of 18-year-old males and the variability that occurred in samples was due to individual differences among males. In the latter, the population average was the true weight of an object to be measured and variability was due to measurement error that resulted from using an unbiased but not completely accurate scale. In both cover stories, subjects were given information about the relevant population means and were required to make judgments about the averages of samples of heights or weighings that differed in size.

Method

Subjects. One hundred fourteen subjects were recruited from sections

TABLE 1
PROBLEMS USED IN EXPERIMENT 1

Post Office Problem

Common first paragraph
When they turn 18, American males must register for the draft at the local post office. In addition to other information, the height of each male is recorded. The national average height of 18-year-old males is 5 feet 9 inches.

Accuracy version:
Yesterday, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of the men who registered there that day.
1. Would you expect one of these recorded average heights to be closer than the other to the national average?
2. If so, which one?

Tail version:
Each day 25 men register at post office A and 100 men register at post office B. At the end of every day, a clerk at each post office computes and records the average height of the men who registered there that day.
1. On what percentage of days would you estimate that the average height recorded for post office A is greater than 6 feet?
2. On what percentage of days would you estimate that the average height recorded for post office B is greater than 6 feet?

Weighing Problem

Common first paragraph:
In a particular geology course students must frequently determine the weight of rock samples. To help develop the needed skills, one instructor has her students practice weighing an object several times on the same scale. The object, a metal disk, is known to weight exactly 1000 grams. Although the scale is not completely accurate, it is equally likely to read high as it is to read low. However, it is never off from the actual weight by more than 25 grams.

Accuracy version:
After a lot of practice, a student from one of her classes (section 1) weighed the disk 20 times, then computed and recorded the average of the 20 weighings. After a lot of practice, a student from another of her classes (section 2) weighed the disk 5 times, then computed and recorded the average of the 5 weighings.
1. Would you expect one of these recorded average weights to be closer than the other to the actual weight of the disk?
2. If so, which one?

Tail version:
After a lot of practice, each student in one of her classes (section 1) weighed the disk 20 times and then computed and recorded the average of the 20 weighings. In the other class (section 2), after a lot of practice, each student weighed the disk 5 times and then computed and recorded the average of the 5 weighings.
1. What is your estimate of the percentage of students in section 1 that recorded an average weight greater than 1010 grams?
2. What is your estimate of the percentage of students in section 2 that recorded an average weight greater than 1010 grams?

of an introductory psychology course designed for majors. None had previously taken a college statistics course.

Procedure. Subjects participated in groups of 5–15. The weighing and post office problems were the first two items given during a session in which subjects completed questionnaires that addressed several different issues in statistical reasoning. Subjects were given up to 15 min to complete the two items. For all subjects, the first item was a version of the weighing problem and the second a version of the post office problem. Half the subjects received the accuracy version first and the tail version second and the other half received the items in the reverse order. Therefore, although cover story was confounded with order, problem version was counterbalanced across both order and cover story.

Results and Discussion

Subjects' answers for each accuracy version were classified as *correct, reversed,* or *equal* depending on whether the subjects indicated that they would expect (a) the average of the larger sample to be closer to the population average than the average of the smaller sample, (b) the average of the smaller sample to be closer, or (c) neither average to be closer to the population average than the other, respectively. In addition, the two numerical answers for each tail version were classified as *correct, reversed,* or *unequal* depending on whether the subject indicated that the percentage of extreme sample means was (a) greater for the smaller samples, (b) greater for larger samples, or (c) the same for smaller and larger samples. The classifications for subjects on both the accuracy and tail versions are displayed in Table 2. Because the classifications were similar for the weighing and post office problems, we have collapsed over cover story to simplify the presentation.

As can be seen in Table 2, performance was markedly better for the accuracy versions than the tail versions. The difference (73% vs 43% correct classifications) is highly significant according to the McNemar

TABLE 2
SUBJECTS CLASSIFIED ACCORDING TO THEIR RESPONSES IN EXPERIMENT 1 (CELL ENTRIES ARE NUMBERS OF SUBJECTS)

Classification of answer in accuracy version	Classification of answer in tail version			
	Correct	Reversed	Equal	Total
Correct	40	18	25	83 (73%)
Reversed	2	2	4	8 (7%)
Equal	7	6	10	23 (20%)
Total	49 (43%)	26 (23%)	39 (34%)	114 (100%)

test for correlated proportions, $\chi^2(1) = 22.2$, $p < .001$. A second striking feature of the data is the difference in the number of "reversed" answers in the two versions: 7% in the accuracy version versus 23% in the tail version, again a highly significant difference, $\chi^2(1) = 7.26$, $p < .01$. Thus, we can conclude that the accuracy and tail versions elicit different patterns of responses.

In this experiment, subjects did reasonably well on all problems. Even for the tail versions, the modal classification of subjects' responses was "correct," rather than "equal" as in previous studies using similar problems (e.g., Kahneman & Tversky, 1972). One possible reason for this is that the tail versions used in Experiment 1 required that subjects provide numerical probability judgments. Although subjects do not tend to feel comfortable with the task of responding with numbers, having to do so seems to make the "equal" response a less convenient default answer than when the task is to choose among several multiple-choice options.

Despite the relatively good performance, more than half the subjects who chose the correct answer in the accuracy version chose one of the incorrect answers in the tail version. The fact that about 40% of these incorrect answers were classified as "reversed" rather than "equal" suggests that many errors resulted from the use of heuristics other than representativeness; a large number of subjects apparently thought that sample size information was important but used it incorrectly in the tail versions.

EXPERIMENT 2

In Experiment 1, performance was much better for the accuracy than for the tail versions. However, several differences between the versions, some more interesting than others, can be identified as possible causes of the effect: First, the accuracy versions focus subjects on the similarity or proximity of the sample and population means, whereas the tail versions focus subjects on how much they deviate from one another. It may be relatively easy to apply an unsophisticated and fuzzy heuristic such as "bigger samples are better" to questions about accuracy. If bigger samples are better, then means of bigger samples should be close to the population means. In contrast, the tail versions require subjects to draw the correct conclusion about the variability of the sample mean. At some level, they have to confront the concept of the "sampling distribution." Second, the accuracy version merely asked that the subject make simple choices, whereas the tail version required that the subject generate numerical probability estimates. Third, the wording of the tail version was somewhat more elaborate than that of the accuracy version, which also added to the complexity of the task.

Experiment 2 was designed to help discriminate among these alterna-

tives by determining whether the effect survived even if the second and third types of differences were eliminated. First of all, no numerical estimation tasks were employed and the wording of the accuracy and tail versions were made more parallel. All questions were couched in terms of a three-way choice in which subjects decided whether a particular condition was more likely to be met by the means of larger samples, or of smaller samples, or whether there was no reason to choose either the larger or smaller sample (i.e., size of the sample does not matter).

Also, an additional version of each problem was introduced. The *center* versions of the weighing and post office problems had wordings that were completely parallel to those of the tail versions except that they asked about the central part of the distribution of sample means rather than the tail. More specifically, the center versions asked whether the mean of the larger or of the smaller sample was more likely to be observed in an interval symmetrically placed around the population mean (between 5 feet 6 inches and 6 feet for the post office problem and between 990 and 1010 grams for the weighing problem).

If performance on the center and tail versions was similar and appreciably worse than that on the accuracy versions, this would suggest that whether or not the problem required some appeal to the concept of sampling distribution was a variable of major importance. If, on the other hand, performance on the center and accuracy versions was similar and better than that on the tail version, this would suggest that the major variable was whether the problem caused subjects to focus on the similarity between the sample mean and the population mean or on how much they deviated from one another.

Method

Subjects and procedure. The 151 subjects who served in Experiment 2a were recruited from sections of the general introductory psychology course and had not yet taken statistics. These subjects participated in small group sessions similar to those of Experiment 1. An additional 138 subjects served in Experiment 2b. These subjects were enrolled in two sections of an introductory statistics course and were given a questionnaire during the first day of classes. In both cases, the relevant problems were the first two items of a four-item questionnaire that dealt with statistical reasoning.

Stimuli and design. In Experiment 2a, six problems were employed: accuracy, center, and tail versions of both post office and weighing problems (see Table 3). Each subject received the accuracy version of one problem and either a center or tail version of the other. Thus, it was possible to make within-subject comparisons of the accuracy versions on the one hand and the tail and center versions on the other, but only

TABLE 3
POST OFFICE PROBLEMS USED IN EXPERIMENT 2

Common first paragraph:
When they turn 18, American males must register for the draft at a local post office. In addition to other information, the height of each male is obtained. The national average height of 18-year-old males is 5 feet, 9 inches.

Accuracy version:
Yesterday, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which would you expect to be true? (circle one)
1. The average height at post office A was closer to the national average than was the average height at post office B.
2. The average height at post office B was closer to the national average than was the average height at post office A.
3. There is no reason to think that the average height was closer to the national average at one post office than the other.

Tail version:
Every day for one year, 25 men registered at post office A and 100 men registered at post office B. At the end of each day, a clerk at each post office computed and recorded the average height of the men who had registered there that day.

Which would you expect to be true? (circle one)
1. The number of days on which the average height was 6 feet or more was greater for post office A than for post office B.
2. The number of days on which the average height was 6 feet or more was greater for post office B than for post office A.
3. There is no reason to expect that the number of days on which the average height was 6 feet or more was greater for one post office than for the other.

Center version:
The center version was identical to the tail version except that the words "6 feet or more" were replaced by the words "between 5 feet 6 inches and 6 feet" in the questions.

between-subject comparisons of the tail and center versions. Order of cover story and of version were both counterbalanced.

Experiment 2b was conducted in order to allow within-subject comparisons of the center and tail versions. Each subject was given a center version of one problem and the tail version of the other. As in Experiment 2a, the order of version and cover story was counterbalanced across subjects.

Results and Discussion

As in Experiment 1, the responses to each version were coded as to whether the answer was correct (subjects chose the option associated

with the correct sample size), reversed (they chose the option associated with the wrong sample size), or equal (they chose option 3). Table 4 displays the number of correct, reversed, and equal responses for each of the versions. As responses were similar for the weighing and post office problems, we have again collapsed over cover story to simplify presentation.

The top panel of Table 4 indicates that subjects did about equally well in the center and accuracy versions (56 and 59% correct responses) and much better in those versions than in the tail versions (8% correct). The difference in percentage of correct responses was highly significant for the accuracy and tail versions using the test for correlated proportions, $\chi^2(1) = 13.47, p < .001$, and for the center and tail versions using a χ^2 test for independence, $\chi^2(1) = 21.28, p < .001$. Similar tests indicated a significantly greater percentage of reversed responses in the tail versions than in either the accuracy or center versions, $\chi^2(1) = 7.69, 11.14, p < .01$ and .001, respectively.

The poorer overall performance in Experiment 2a than in Experiment 1 was at least partly due to differences in the subject population from which the subjects were drawn—students enrolled in a one-semester introductory course as opposed to those enrolled in a more intensive two-semester sequence designed for majors. However, the fact that there was a large difference between the accuracy and the tail versions in both experiments strongly suggests that the phenomenon is real and robust with respect to changes in subject population and procedure.

Because the number of subjects who received the center and tail versions was relatively small in Experiment 2a, we presented both a tail and

TABLE 4
SUBJECTS' RESPONSES IN EXPERIMENT 2 (CELL ENTRIES ARE NUMBERS OF RESPONSES)

	Version of problem		
Classification	Accuracy	Center	Tail
Experiment 2a			
Correct	42 (56%)	23 (59%)	3 (8%)
Reversed	3 (4%)	0 (0%)	9 (25%)
Equal	30 (40%)	16 (41%)	24 (67%)
Total	75	39	36
Experiment 2b			
Correct		53 (38%)	30 (22%)
Reversed		14 (10%)	18 (13%)
Equal		71 (51%)	90 (65%)
Total		138	138

a center version to each member of another group of subjects in Experiment 2b. Although overall level of performance was poorer in Experiment 2b (probably because the subjects had not agreed in advance to participate in the study but were asked to complete these problems in a classroom setting), there was again a substantial difference in performance in favor of the center versions, 38% vs 22% correct responses, $\chi^2(1) = 11.25$, $p < .001$, by the test for correlated proportions.

We found the results of Experiment 2 to be somewhat counterintuitive. We had thought, as do the authors of many statistics books (e.g., Freedman, Pisani, & Purves, 1978), that a major difficulty students have in dealing with questions that address the variability of statistics such as the mean is that they do not have a concept of, and therefore cannot deal with, sampling distributions. However, we found that many subjects dealt adequately with the center versions of the problems, even though these were phrased in terms of the central part of the sampling distribution. Subjects thus appear to have access to a rudimentary concept of sampling distribution that is adequate to deal with some kinds of problems but not with others.

EXPERIMENT 3

The purpose of Experiment 3 was to gain additional understanding about why performance was so much better in the center than in the tail versions of problems. Although Experiment 2 suggested that the major variable was whether the question focused subjects on the similarity or difference between the sample and population means, it is possible that the asymmetry of the region of the sampling distribution asked about in the tail versions (the upper tail) constituted an additional source of difficulty.

To unconfound the similarity and symmetry variables, we introduced "two-tailed" versions of the problems in Experiment 3. In these versions, the subject was asked about the likelihood that the sample mean would fall more than a certain distance above *or* below the population mean. Because poor performance was obtained in Experiment 2b when subjects were given the problems in a classroom setting, Experiment 3 was conducted under more controlled laboratory conditions.

Also in Experiment 3, we attempted to have subjects indicate in a structured fashion why they chose the answers they did. In the previous experiments, we had left space below each problem for subjects to write comments about their answers and the instructions had contained the request that, if possible, subjects should indicate the justifications for their answers. Although the written comments were often brief and ambiguous, they tended to break down into six or seven general categories. We used these comments, along with those elicited during informal inter-

views with subjects, to generate possible rationales for answers that were presented to subjects in Experiment 3 (see Table 5).

Method

Subjects. The 120 undergraduates who participated were recruited from sections of the general introductory psychology course.

Procedure. Subjects served in groups of 5–15 and were given 30 min to complete four problems that dealt with statistical reasoning, of which the first and fourth dealt with the effects of sample size. In order to encourage subjects to take adequate time in deciding on their answers and to prevent them from changing answers to earlier problems on the basis of seeing later ones, subjects were given a fixed amount of time to complete each problem, after which the problem was collected and the next one was distributed. The center and one-tailed versions were exactly the same as in Experiment 2. The two-tailed versions were obtained by replacing the phrase "six feet or more" by "less than five feet six inches or more than six feet" in the post office problem and "1010 grams or more" by "less than 990 grams or more than 1010 grams" in the weighing problem.

Each student received a center version and either a one- or two-tailed version. The center version was presented first for half the subjects and

TABLE 5
POSSIBLE RATIONALES FOR ANSWERS PRESENTED IN EXPERIMENT 3 (BOLDFACE CAPTIONS WERE NOT SEEN BY SUBJECTS)

Alternatives for subjects who chose answer (1) or (2)

a. **Average influenced less by extreme values:** If a larger number of men register on a day, the average height recorded is less likely to be influenced by a few extremely tall or extremely short men.

b. **More opportunity for extremes:** If a larger number of men register on a day, there will be more opportunities to have extremely tall or short men and so the average height will be more variable.

c. **Larger samples more like the population:** Larger samples are more likely to be similar to the population they come from than smaller samples, so if more men register on a day, the average height recorded should be closer to the average of the population.

d. **Larger and smaller samples may come from different populations:** Larger post offices are likely to be in cities and cities are more likely to have diverse populations.

Alternatives for subjects who chose answer (3)

a. **No reason to choose:** The question is worded in such a way that the effect of having more or fewer men registering on a day is unclear. There does not seem to be any reason to choose answer (1) over answer (2) or answer (2) over answer (1).

b. **Number of men is not a factor:** The number of men who register each day at a post office is not a factor because the problem deals with averages.

c. **Not enough difference to matter:** The number of men who register each day may make some difference, but there is not enough difference between 25 men per day and 100 men per day to matter very much.

the tail version first for the other half. The weighing problem was used as the first cover story for all subjects. Thus, as in Experiment 1, order was confounded with the cover story, but version was counterbalanced over both cover story and order.

After subjects completed answering their last problem (a version of the post office problem), the problem sheets were collected and another packet was distributed. Subjects were instructed to turn to one page (which contained the rationales presented in the top panel of Table 5) if they had chosen either answer (1) or (2) and to another page (which contained the rationales presented in the bottom part of Table 5) if they had chosen answer (3). They were asked to place a check mark by one of the rationales if it seemed to be similar to the type of reasoning they had used in arriving at their answer. If none of the rationales were similar to their reasoning, they were instructed to check off an option marked "other" and to try to indicate in the space provided what their reasoning was.

Results and Discussion

Response classifications. As can be seen in Table 6, the percentage of correct responses for the center versions, 48%, was greater than for either the one- or two-tailed versions, 25 and 34%, respectively. Both differences are significant by the test for correlated proportions, $\chi^2(1) = 6.50$, $p < .025$ and $\chi^2(1) = 4.50$, $p < .05$, respectively. The percentage of correct responses was not significantly different for the one- and two-tailed versions, $\chi^2(1) = 1.27$, so that we do not have sufficient evidence to conclude that asymmetry contributes to the difficulty of the tail version. Again, there seemed to be evidence that the tail versions led some subjects to use information about sample size inappropriately. The percentage of reversed answers was significantly greater even in the two-tailed versions (17%) than in the center versions (9%), $\chi^2(1) = 4.00$, $p < .05$.

Therefore, the results of Experiment 3 provide additional evidence of

TABLE 6
SUBJECTS' RESPONSES TO PROBLEMS IN EXPERIMENT 3

	Version of problem		
Classification	Center	Two-tailed	One-tailed
Correct	57 (48%)	20 (34%)	15 (25%)
Reversed	11 (9%)	10 (17%)	12 (20%)
Equal	52 (43%)	29 (49%)	34 (56%)
Total	120	59	61

differences in response patterns to the center and tail versions but do not strongly suggest that the asymmetry of the region of the sampling distribution in question is a major consideration.

Subjects' reasons for their answers. Subjects' choices among the possible rationales are presented in Table 7. Twelve of the 15 subjects who chose the reversed answer for the tail version (i.e., the average of the larger sample is more likely to be far from the population mean) chose rationale (b), namely, that larger samples are more variable because they are more likely to contain extreme scores. The tail versions focus subjects on the deviations between the sample and population means and may cause them to think about extreme scores. If subjects correctly note that large samples are more likely to contain extreme scores but fail to distinguish between the number of extreme scores in a sample and the effect of these scores on the sample average, they may find it reasonable to choose the reversed answer. Only six subjects given the center version chose the reversed answer and only two of them chose rationale (b).

Of the 40 subjects who chose the correct answer for the center version of the post office problem, 18 chose rationale (c) (i.e., larger samples are more likely to be similar to the population) and 12 chose rationale (a), a variation of the "swamping" explanation often encountered in statistics textbooks (i.e., that an extreme score will have less of an effect on the average of a large sample). Nine of the remaining 10 subjects who gave the correct answer chose rationale (b), which if understood in the way we intended, is not compatible with the correct response unless combined with the idea of extreme scores being more likely to cancel out if there are more of them.

A concern about the kinds of multiple-choice problems used in Experiments 2 and 3 is that answer (3), the "equal" response, might not only

TABLE 7
CHOICE OF RATIONALES IN EXPERIMENT 3

	Rationale chosen			
Classification	Average less influenced by extremes	More opportunity for extremes	Larger samples more like population	None or other
Correct (combined)	12	9	18	1
Reversed (center version)	2	2	2	0
Reversed (tail version)	2	12	1	0
	No reason to choose	Number of men not a factor in average	Not enough difference to matter	None or other
Equal (combined)	7	28	4	20

reflect insensitivity to sample size but may also represent a convenient default answer that could be chosen when subjects are unable to decide between answers (1) and (2). However, from subjects' choices of rationales, it appears that few of those who made an equal response thought of their answer as a default; Only 7 of the 59 subjects who chose the answer selected rationale (a) (see the bottom panel of Table 5), a vaguely worded alternative that represented our conception of what might be considered a default answer. Almost half selected rationale (b), which stated that if one deals with averages, the number of scores in each sample does not matter. However, more than one-third of the subjects did not select any of the rationales that we provided. From written comments and from informal discussions with these subjects, it seems that some subjects who gave the equal answer had difficulty keeping straight that the questions asked about the distributions of *averages* of large and small samples rather than the distributions of *scores* in large and small samples. If the problem is interpreted as asking about the *proportion of scores in a sample* expected to fall in a certain range, then it is correct to conclude that sample size does not matter. This confusion is of some theoretical interest. We do not believe it should be thought of as simply resulting from carelessness on the part of the subjects; rather, it points out the difficulty of making and retaining certain distinctions without an understanding of the basic underlying concept (in this case, that of sampling distribution).

The picture that seems to be emerging is not a simple one. Errors may occur because subjects (a) cannot access or apply appropriate heuristics to the problem, (b) access inappropriate heuristics, or (c) misunderstand the problem. The center and accuracy versions of problems may be answered correctly more often because they map onto the heuristic that larger samples are more similar to the population. Performance on the tail versions may be poorer not only because they do not map onto the "bigger is better" heuristic very well but also because focusing subjects on deviations between samples and the population might elicit heuristics that deal with the occurrence of extreme scores. Also, some subjects make equal responses because they believe that the required judgment is about the proportion of scores in a sample rather than the proportion of sample averages.

EXPERIMENT 4

Experiments 1–3 indicated a strong tendency for subjects to use information about sample size more appropriately when asked about the accuracy of sample means or about the central region of the sampling distribution than when asked about the tail of the sampling distribution. However, it is difficult to determine the extent to which incorrect responses represent true insensitivity to the effects of sample size and to

what extent they reflect confusion about the required judgment. For example, some incorrect answers may reflect difficulty distinguishing between and understanding the distinction between the distribution of scores in a sample and the distribution of sample averages. In Experiment 3, we tried to address this issue by analyzing subjects' written comments and choices of rationales. In Experiment 4, we adopted a different strategy. We provided subjects with instruction about the concept of sampling distribution and then elicited different kinds of judgments about large and small samples in an attempt to determine to what extent appropriate use of sample size information follows from understanding aspects of the concept of sampling distribution.

More specifically, we (a) developed a computer-assisted training procedure that we thought should clarify the concept of sampling distribution and clearly distinguish questions about sample means from those about the scores within a sample, (b) selected a group of subjects who had given incorrect answers to a version of the post office problem during a pretest and had them go through the training procedure, and (c) tested their understanding about the effects of sample size during the pretest and at several points during the training procedure.

Our primary interest in the training procedure was not simply to demonstrate that performance with sample size problems improved after subjects were taught something about the law of large numbers. Rather, we viewed training as an opportunity to test subjects' intuitions when we were sure that they understood the questions we were asking.

Method

Subjects. Twenty-seven subjects recruited from the department of psychology subject pool took the pretest. Six subjects answered correctly and did not participate further. The remaining 21 subjects participated in the training session.

Procedure. Each subject was run individually and both the pretest and training sessions took the form of structured interviews that together took about 45 min. The subject and the interviewer sat across a table from one another and were both videotaped during the entire session.

Pretest. At the beginning of the pretest, the subject was handed a sheet of paper containing a tail version of the post office problem that was identical to those used in Experiments 2 and 3 except that the sample sizes mentioned were 10 and 100 (instead of 25 and 100). Subjects were asked to mark the answer they thought was correct and then to indicate to the interviewer why they chose their answer. When the subject had finished, the interviewer removed the sheet of paper containing the problem and the subject's answer and asked the subject to restate in his or her

own words what the problem had asked. This constituted the test of understanding before the training session.

Training session. The goal of the training session was to provide subjects with an understanding of random sampling and of the sampling distribution of the mean while at the same time testing subjects' intuitions about the effects of sample size on the variability of sample means.

The training session began with a brief explanation to the effect that subjects often had difficulties understanding problems of the type that had been encountered in the pretest and that the training session was designed to help them understand exactly what such questions were asking.

The subject was then given some instruction about frequency distributions and random samples. The subject was seated in front of a microcomputer. The lower part of the computer screen contained a symmetric frequency distribution that was supposed to represent the heights of a population of 400 adult men in a town. The distribution consisted of 400 Xs and was centered at 69 inches. The upper part of the screen contained a similar scale but initially no Xs (it was to be later used as the scale for the sampling distribution of the mean). First, to make sure that the subject understood the frequency distribution and to provide a baseline for future estimates, the subject was asked to estimate the proportion of men in the town with heights greater than 69, 70, and 71 inches. The interviewer then produced a box containing 400 slips of paper with heights written on them corresponding to the Xs on the screen. The interviewer explained the model of random selection of heights in terms of repeated selection of slips of paper from the box and indicated that the computer was programmed to select heights randomly from the population in an analogous fashion. The subject was then instructed to press the space bar of the computer keyboard to select one of the heights in the population randomly. When the space bar was pressed, one of the Xs in the histogram was covered with a rectangle to indicate it was selected, and the height corresponding to the X was displayed on the top of the screen. This procedure was repeated until a sample of 10 heights had been selected (and 10 Xs in the population distribution had been covered). The subject was then asked to compute the mean of the 10 scores to emphasize that we were concerned with sample means. After the subject calculated the mean of the first sample, the space bar was pressed once more and the number representing the mean of the 10 selected heights was displayed next to the 10 numbers representing the individual heights. In addition, an X was displayed on the scale in the upper part of the screen to represent the mean of the sample that had just been selected. The interviewer indicated that the procedure of randomly selecting a sample of 10 observations and computing its mean would be repeated 100 times and that

each time the computer computed a sample mean it would plot an X in the upper distribution to maintain a continuing record of sample means.

Subjects' thoughts about how much variability should occur in the distribution of sample means were then tested in several ways. First, subjects were asked how far a sample mean would have to be from the mean of the population before they would be "surprised." Second, subjects were asked to estimate the proportion of means of samples of size 10 that would be expected to exceed 71 inches (this estimate could be compared with their previous estimate of what proportion of men in the population had heights exceeding 71 inches). Finally, they were asked to draw a diagram that indicated what they thought the distribution of sample means would look like.

Subjects then sampled 100 samples of size 10. For the first 5 samples, subjects had to press the space bar 11 times, 10 times to select heights from the population and an eleventh time to display the numerical value of the mean and place an X in the corresponding position in the sampling distribution in the upper part of the screen. After the first 5 samples, the interviewer checked to make sure that subjects understood that the Xs in the upper distribution represented means of samples while the Xs in the lower distribution represented individual heights. (This appeared to pose little problem.)

For the remaining 95 samples, a single key press selected an entire sample of 10 heights and displayed the mean of the sample in the upper (sampling) distribution. After the 100 samples had been selected, the interviewer again made sure that subjects understood the difference between the upper and lower distributions and that all subjects noted that the sampling distribution of means was less variable than the population distribution that was displayed below it.

At this point, the interviewer conducted another key test of subjects' understanding of the effect of sample size on the variability of sample means. She indicated that the same type of sampling procedure would now be conducted for samples of size 100. She asked subjects to indicate how variable they thought the sampling distribution for means of size 100 would be, both by estimating what proportion of samples of size 100 would have means greater than 71 inches and by drawing on a sheet of paper what they thought the sampling distribution would look like. Subjects were also explicitly asked whether the variability in the distribution of means of samples of size 100 would be about the same, or different from the distribution for means of samples of size 10.

Subjects then selected 100 samples of size 100 and noted that the variability in the resultant sampling distribution was dramatically smaller than was the case for samples of size 10. Subjects were then asked about why

they thought the sampling distributions were so different and were then given several transfer problems.

Results and Discussion

In Experiment 4 we were primarily interested in determining for the subjects who had answered incorrectly in the pretest (a) whether they clearly understood that the problem they received during the pretest asked about the distribution of sample means rather than about sample scores; (b) whether, after subjects had learned during the training session what the distribution of sample means was, but before they had learned exactly what it would look like, subjects understood that the variability of sample means should be less than that of population scores; and (c) whether, after having constructed a sampling distribution for the mean of samples of size 10, subjects understood how the variability in the sampling distribution would change when the size of the sample was changed from 10 to 100. Classifications of subjects' responses were based on coding the videotaped interviews. Each tape was coded independently by the interviewer (S.J.B.) and at least one of the other authors.

Original understanding of post office problem in the pretest. Subjects' initial explanations of why they chose their answers were, for the most part, vague and of little use in understanding their reasoning. More revealing were subjects' attempts to restate the problem in their own words after selecting their answer. Only 3 of the 21 subjects who answered the problem incorrectly restated the problem accurately. Eleven of these subjects restated the problem as asking about the *percentage of men* who had heights greater than 6 feet, rather than about the percentage of sample means greater than 6 feet. Although we cannot be sure that subjects were similarly confused when they had the problem sheet in front of them and were deciding on their answer, this result underscores the difficulty of distinguishing between distributions of scores and of averages.

Understanding that the variability in the distribution of sample means should be smaller than the variability in the population distribution. As indicated above, we attempted to assess subjects' understanding about the variability of sample means in several ways. The most useful indicator of subjects' intuitions about the variability of the sample mean after they had selected their first sample of 10 scores was the comparison between subjects' estimates of the *proportion of scores in the population* greater than 71 inches and their estimates of the *proportion of sample means* that would be greater than 71 inches. Eleven of the 21 subjects produced estimates that were smaller for sample means than for population scores (with the remaining 10 subjects indicating they should be the same), suggesting that some subjects had an appreciation of the fact that sample

means are less variable than individual observations. However, for some subjects this correct intuition depended on thinking in terms of sampling without replacement. A number of subjects stated that the distribution of means should be less variable because there are not enough of the most extreme scores to fill an entire sample, so that it would be unlikely or impossible to get a mean as extreme as the most extreme scores in the population.[1]

Understanding that the variability in the sampling distribution of the mean should be smaller for samples of size 100 than for samples of size 10. After subjects had generated 100 samples of size 10 and had observed the distribution of sample means, they were told that the next part of the procedure would involve selecting a large number of samples of size 100 and plotting the distribution of their means. Subjects were asked to estimate the proportion of the means of samples of size 100 that would be larger than 71 inches and were explicitly asked to indicate whether this proportion would differ from what was obtained for samples of size 10. The surprising result was that 16 of the 21 subjects indicated both by their estimates and by their statements about the variability of the two sampling distributions that they believed the variability would be about the same for the means of samples of size 10 and those of samples of size 100.

Thus it seems that even after these subjects had become familiar with the concept of the sampling distribution, they did not have a good understanding about the effect of sample size on the variability of the sample mean. After the sampling distribution for the larger samples had been generated, 11 of the 21 subjects indicated that they had been surprised by how much less variable the second sampling distribution was. Although they had not anticipated the result, none of these subjects indicated that they had misunderstood the previous request to compare the two sampling distributions.

Although few subjects anticipated that the variability of the second sampling distribution would be much smaller than the first, when the computer generated the second distribution, they accepted the result. When asked why the variability of the second distribution was smaller, 18 of the 21 subjects provided explanations. Nine subjects gave appropriate

[1] The other two procedures we employed did not provide very useful data. Our question about how far a sample mean would have to be from the population mean before the subject would be "surprised," which we hoped might give us some idea of the subject's intuitive confidence intervals, was not well understood. Also, subjects' attempts to draw the distribution of the sample means was complicated by the fact that the sampling distribution was to consist of only 100 means in contrast to the population distribution which consisted of 400 scores. Subjects realized that the former distribution would be "flatter" and found it difficult to represent the relative variability in the two distributions.

explanations such as swamping (an extreme score will affect the mean of a larger sample less than the mean of a smaller sample) or balancing (larger samples provide more opportunity for large and small scores to balance out). Six subjects indicated that the variability was less because "bigger samples are better" and three additional subjects indicated that the larger sample would be more like the population because it represented a greater proportion of the population.

At the end of session when asked whether the mean of a larger or smaller sample would be a more accurate indicator of the population mean, all 21 subjects replied that the mean of the larger sample would be more accurate. Eighteen subjects now gave the correct answer to the version of the post office problem that they had seen in the pretest and 16 provided the correct answer to the tail version of the weighing problem.

The data of Experiment 4 suggest that a large proportion of naive subjects who give incorrect answers to questions about the variability of sample means have difficulty making and retaining the distinction between distributions of scores and distributions of averages. However, our results indicate that even if such subjects are trained to the point that they understand aspects of the concept of sampling distribution and can clearly distinguish between questions about sample scores and sample averages, many of them will not be able to make appropriate conclusions about the how sample size affects the variability of the sample average. Most of the subjects in Experiment 4 failed to realize that averages of larger samples are less variable than those of smaller samples, even when they seemed to have a clear understanding of the distinction between sample and sampling distributions and had observed that the variability in the distribution of the averages of samples of size 10 was less than the variability in the population.

GENERAL DISCUSSION

Experiments 1–3 strongly indicate that people use information about sample size more appropriately when asked about the accuracy of sample means or about the central region of the sampling distribution than when asked about the tails of the sampling distributions. Collectively, the three experiments tested more than 500 subjects and, despite differences in problem wordings, procedures, and subject populations, strong effects were found in each experiment. The difference in the percentage of correct responses between the accuracy or center versions of problems, on the one hand, and tail versions on the other, was approximately 30, 50, 15, and 20% in Experiments 1, 2a, 2b, and 3, respectively. The fact that performance was about the same for accuracy and center versions suggests that the important variable was whether or not the problem focused

on the similarity between the population and sample mean, rather than whether or not it was stated in terms of the sampling distribution.

People appear to understand that the averages of larger samples are more likely to resemble the population average but do not understand the implications of this fact for the variability of the sample average. Moreover, tail problems focus subjects on the degree of discrepancy between the population and sample averages and, for some subjects, this apparently elicits an inappropriate heuristic about extreme scores, namely, that extreme scores are more likely to occur in large samples (which is true) and that, therefore, the averages of large samples will be more variable (which is not true). The poorer performance for the tail versions does not merely reflect complexity of wording, as the wording of the center versions is at least as complex as that of the tail versions. Nor does the asymmetry of the region of the sampling distribution asked about in the tail versions used in Experiments 1 and 2 seem to be an important factor.

Perhaps the most interesting findings were obtained in Experiment 4. We found in Experiment 3 and in the pretest portion of Experiment 4 that naive subjects have difficulty understanding and retaining the distinction between distributions of averages for large and small samples and distributions of scores in large and small samples. Although this confusion is itself of interest because it points out the difficulty of making the distinction without an understanding of the concept of sampling distribution, it raises the possibility that some of the equal responses may have arisen from this confusion rather than a "true" insensitivity to the effects of sample size. The major finding in Experiment 4 was that even when subjects had been instructed about sampling distributions, had generated a sampling distribution, and apparently clearly understood the distinction between distributions of sample averages and distributions of scores in a sample, many of them still did not understand how sample size influenced the variability of the sample mean. In particular, even after observing how the sampling distribution of the mean for samples of size 10 was generated from the population and noting that it had less variability than the population distribution, most subjects did not realize that the sampling distributions for samples of size 100 would have even less variability.

Taken together with the findings of others (e.g., Bar-Hillel, 1979, 1982; Evans & Dusoir, 1977; Evans & Pollard, 1985; Nisbett *et al.*, 1983), our results suggest that the appreciation that naive subjects have been shown to have for the law of large numbers often does not result from in-depth understanding of the relation between sample size and variability. Rather, correct answers may result from the application of low-level heuristics such as "bigger is better." Performance will be poorer on some problems because they do not map well onto the correct heuristics or because they

elicit inappropriate heuristics such as representativeness or the "extreme score" heuristic mentioned above.

In a recent discussion of research on statistical intuitions, Kahneman and Tversky (1982) made the distinction between *errors of application,* which occur when people know and accept the appropriate rule but fail to apply it to a given problem, and *errors of comprehension,* which occur when people do not "recognize the validity of the rule that they have violated." Although the distinction is valuable, we hope that thinking in terms of this dichotomy does not diminish our appreciation of the fact that people can have a partial understanding of a rule that allows them to answer some problems correctly but not others. A level of understanding adequate to answer some types of problems correctly might result in errors of comprehension for other types.

A final point concerns the comparison of research about statistical intuitions with the rapidly growing body of research dealing with intuitive or naive physics (Clement, 1982; McCloskey, 1983; McCloskey, Camarazza, & Green, 1980). In the latter domain, people often seem to have strongly held beliefs that are consistent with Aristotelian but not with Newtonian mechanics. These intuitive beliefs guide thinking about problems in physics and apparently represent a major impediment to instruction (e.g., Clement, 1982; Lin, 1979).

The situation seems to be different for statistical intuitions. The kinds of heuristics that naive subjects apply to statistical problems are often not so much wrong as they are crude, inadequate versions of appropriate statistical ideas and in this domain it may be possible, as Bar-Hillel (1984) puts it, "to educate intuition to make it more valid and useful." The educational implication of this distinction is that many statistical intuitions may not represent impediments that must be excised before effective learning can take place. Rather, they may represent opportunities for instruction: crude ideas that potentially may be developed and refined.

REFERENCES

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Decision Processes,* **11**, 277–282.

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge Univ. Press.

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica,* **55**, 91–107.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics,* **50**, 66–71.

Evans, J. St. B. T., & Dusoir, A. E. (1977). Proportionality and sample size as factors in intuitive statistical judgment. *Acta Psychologica,* **41**, 129–137.

Evans, J. St. B. T., & Pollard, P. (1985). Intuitive statistical inferences about normally distributed data. *Acta Psychologica,* **60,** 57–71.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology,* **18,** 253–292.

Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics.* New York: Norton.

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

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition,* **11,** 123–141.

Lin, H. (1979). Approaches to clinical research in cognitive processes instruction. In J. Lochhead & J. Clement (Eds.), *Cognitive process instruction: Research in teaching thinking skills.* Philadelphia: Franklin Institute Press.

McCloskey, M. (1983). Intuitive physics. *Scientific American,* **248,** 122–130.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science,* **210,** 1139–1141.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review,* **90,** 339–363.

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. *Journal of Experimental Psychology: Human Perception and Performance,* **2,** 599–608.

Pollard, P., & Evans, J. St. B. T. (1983). The role of 'representativeness' in statistical inference: A critical appraisal. In J. St. B. T. Evans (Ed.), *Thinking and reasoning.* London: Routledge & Kegan Paul.

Teigen, K. H. (1974). Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology,* **15,** 56–62.

RECEIVED: January 31, 1989