Memory & Cognition
1984, 12 (4), 395-401

Beliefs underlying random sampling

ALEXANDER POLLATSEK, CLIFFORD E. KONOLD,
ARNOLD D. WELL, and SUSAN D. LIMA
University of Massachusetts, Amherst, Massachusetts

In Experiment 1, subjects estimated (1) the mean of a random sample of 10 scores consisting of 9 unknown scores and 1 known score that was divergent from the population mean and (2) the mean of the 9 unknown scores. The modal answer (about 40% of the responses) for both sample means was the population mean. The results extend the work of Tversky and Kahneman (1971) by demonstrating that subjects hold a passive, descriptive view of random sampling rather than an active-balancing model. This result was explored further in in-depth interviews (Experiment 2), wherein subjects solved the problem while explaining their reasoning. The interview data replicated Experiment 1 and further showed: (1) that subjects' solutions were fairly stable—when presented with alternative solutions, including the correct one, few subjects changed their answers; (2) little evidence of a balancing mechanism; and (3) that acceptance of both means as 400 is largely a result of the perceived unpredictability of "random samples."

There is at present a large body of evidence indicating that students believe that random samples resemble the population from which they are drawn. If the sample size is sufficiently large, then a random sample will, in fact, tend to be similar to the population from which it is drawn. The point at which the typical student apparently differs from the normative model of statistics is that he or she believes that small as well as large samples have a high probability of looking like the population. Tversky and Kahneman (1971) dubbed this misconception "the law of small numbers." They proposed that a heuristic or belief called "representativeness" underlies this misconception. "A person who follows this heuristic evaluates the probability of an uncertain event, or a sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated" (Kahneman & Tversky, 1972, p. 431).

One source of evidence for this misconception has come from investigation of what is popularly known as the "gambler's fallacy." A simple example of the gambler's fallacy is the belief that if a fair coin has come up heads a large number of times in a row, then there is an increased chance that it will come up tails on the next flip. The gambler's fallacy can be described as the belief that, in random sampling, the data that have already been sampled will influence the data that are yet to be sampled. This, of course, violates independence, which is a fundamental property of true random sampling. In real-life coin flipping, shaking the coin well between flips would guarantee some reasonable approximation of independence from one flip to another.

This research was supported by Research Grants SED-8016567 and SED-8113323 from the National Science Foundation. Requests for reprints should be sent to Alexander Pollatsek, Department of Psychology, University of Massachusetts, Amherst, MA 01003.

The prototypical problem used by Tversky and Kahneman (1971) to explore the gambler's fallacy is the following:

> The mean IQ of the population of eighth-graders in a city is known to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample?

If the sampling were random, then the best guess for the mean score of the next 49 children sampled is 100. Therefore, the best guess for the entire sample of 50 children is the weighted mean of 150 and 100, or 101. However, the typical answer to this problem is 100. This finding reflects the gambler's fallacy because the answer of "100" violates the assumption of independence. An answer of "100" logically implies that the mean of the next 49 children is influenced by the score of the first child sampled. It is not known whether subjects realize that this implication follows from their answer, or whether the implication is a critical component of the representativeness heuristic. Before discussing this question, we must briefly discuss other evidence for representativeness.

Bar-Hillel (1980) and Kahneman and Tversky (1972) have employed a second paradigm to demonstrate the heuristic of representativeness. Typically, the subject is shown two samples and asked to judge which is more likely. In their original work, Kahneman and Tversky (1972) dealt with events modeled by Bernoulli trials. They found, for example, that subjects thought that, for a sequence of six births, the exact order of G B G B B G (G = girl; B = boy) is more likely than the order B G B B B B, presumably because the sequence with five boys and one girl fails to reflect the proportion of boys and girls in the population. Subjects also estimated that the probability of a sequence like B B B G G G was less than that of G B B G B G, presumably

because the former appears less random. Bar-Hillel (1980) extended this research to determine which characteristics of samples to which subjects are attending when they judge samples to be more or less likely than others. She found that subjects think that a sample should have not only about the same mean as the population, but also about the same standard deviation.

Thus, the evidence is compelling that subjects believe that even small samples should look like the population and that a random sample should look random. Our interest is in determining whether the heuristic of representativeness is a fundamental belief, or axiom, in the layman's theory of random samples, or whether it is deducible from some more basic mechanistic belief. This distinction will become clearer if we digress for a moment and speculate about how an expert thinks about large samples.

Presumably, an expert's fundamental conception of random variables and random sampling is a process model. Perhaps the most widely used model is the "urn-drawing," or "box," model, in which random sampling is viewed as isomorphic to the process of drawing labeled balls or slips of paper from an urn or box, replacing them, shaking well, and then drawing again. From this model, the idealization of which can be summarized by algebraic expressions, certain conclusions follow. These include the "law of large numbers," which says (roughly) that if a random sample is large enough, the relative frequencies of outcomes in the sample have a very high probability of being close approximations of those in the population. It is likely that, in dealing with large samples, the expert appeals simply to the property of representativeness derivable from the law of large numbers, rather than conceptualizing random sampling in terms of a process. However, if challenged, or if some absurd consequence arose from an attempted application of this intuitive version of the law of large numbers, the expert could go back to the more basic process model of sampling to check whether the consequence did in fact follow from probability theory.

The evidence shows that novices are likely to believe that small, as well as large, samples are representative. (There are data indicating that experts overapply representativeness as well; Tversky & Kahneman, 1971.) This belief could plausibly follow from one of two basic heuristics. The first possibility is that representativeness itself is the basic heuristic. In other words, the basic heuristic in thinking about random samples is *descriptive*: Random samples look approximately like the population, and further, random sequences of events look "random." There is, however, a second possibility. Subjects could have an erroneous *process* model of random samples from which representativeness of even small samples followed as a conclusion, just as the heuristic of representativeness for large samples could follow from the correct urn-drawing heuristic of the expert. What might such a process model be? One that has been suggested in statistics books (e.g., Freedman, Pisani, & Purves, 1978, chap. 16; Hays, 1981, chap. 1) is "active balancing" or "compensation," specifically, that some active process guarantees that things will even out in the long run. Apparently, such a belief is exposed in the coin-flipping example of the gambler's fallacy when the subject predicts that, following a run of tails, the next coin is likely to come up heads. The idea that things will "even out" suggests a notion of active balancing.

However, the heuristic of active balancing might be deduced from the heuristic of representativeness. If, in the coin example, the subject believes that samples should look like the population of outcomes of flips, then samples that are close to half heads and half tails will be the most representative. If one has already observed 9 heads and is predicting the outcome of the 10th flip, then presumably a sample of 9 heads and 1 tail will be more representative of the population than a sample of 10 heads, so that the outcome of "tail" on the 10th trial should be more likely than "head."

On what basis can one decide whether the representativeness (i.e., descriptive) or active-balancing heuristic is the more basic? In the coin example mentioned above, both heuristics would predict that a head would be more likely to turn up following a run of tails. However, there are situations in which the active-balancing and representativeness heuristics lead to different predictions. Consider the Tversky and Kahneman (1971) IQ example mentioned earlier. Again, both heuristics would predict an answer of 100. However, if asked to predict the mean IQ of the *last 49* students in the sample, subjects who thought that all samples should look like the population would give an answer of 100, but those who employed an active-balancing heuristic would give an answer smaller than 100 (so that the entire sample of 50 scores could average 100).

The present study extended Tversky and Kahneman's (1971) study by employing an additional follow-up question about the mean of the sample excluding the known score. Additionally, we were concerned that subjects might think of 101 as being approximately 100, and thus answer "100" even though they knew the mean would be slightly higher than 100. Accordingly, in our problems the sample size was made smaller so that the difference between the correct answer and the population mean would be more salient. Another feature of our experiments was to have some subjects "think out loud" so that we could better understand the heuristics they were employing.

EXPERIMENT 1

Method

Materials. Two problems were employed. One was a variant of the Tversky and Kahneman (1971) IQ problem stated above.

IQ Problem

The average IQ of the population of eighth-graders in a city is known to be 100. You have selected a random sample of 10 children for a study in educational achievement. The first

child tested has an IQ of 150. What do you expect the average IQ to be for the whole sample?

What do you expect the average IQ to be for the next 9 children, not including the 150?

(The correct solution to the first question is 105; that to the second is 100.)

The second problem that was employed is similar, using an SAT instead of an IQ cover story.

SAT Problem

The average SAT for all the high school students in a large school district is known to be 400. You have randomly picked 10 students for a study in educational achievement. The first student you picked had an SAT of 250. What do you expect the average SAT to be for the entire sample of 10?

What do you expect the average SAT to be for the next 9 students, not including the 250?

(The correct solution to the first question is 385; that to the second is 400.)

Subjects. The subjects were undergraduates at the University of Massachusetts who were enrolled in psychology courses. The 31 subjects who were interviewed were selected from a pool of student volunteers and received bonus class credit for their participation. The 205 students who filled out questionnaires did so during a regular class session and were told that they would be helping us to understand how people think about statistics. No subject participated in both the questionnaire and interview phases. Both phases contained approximately equal numbers of males and females.

Procedure. The questionnaire was administered to four undergraduate psychology statistics classes and took about 10 min to complete. The SAT problem was the first of three problems on the questionnaire, and both parts of the SAT problem appeared together on a single page.

In the interview phase, the subject was given either the SAT or the IQ problem, as well as several other unrelated problems that will not be discussed in this paper. A subject was given a sheet of paper on which appeared the first paragraph of the problem and was asked to read it aloud, so that the experimenter knew that it had been read correctly. The subject then answered the first question, thinking aloud as much as possible. When he or she had given an answer, the interviewer orally presented the second part of the problem. The interviewer then asked follow-up questions designed to further elucidate what the subject was thinking. The session lasted about 1 h, and approximately 10 to 15 min were spent on one of the two problems discussed here.

Results and Discussion

The data are displayed in Table 1. For the questionnaire subjects, the numerical answers were tabulated. For the interview subjects, the numerical answers, before any interviewer intervention, were obtained from videotapes. Several features are apparent. First, the answer predicted by representativeness, namely, that the means of both samples are equal to the population mean, is the modal answer. It was given by 33% of the subjects who answered the questionnaires and 48% of the subjects who were interviewed. Second, there is considerable variation in the answers given by the subjects. Twenty-one percent of the subjects gave the correct solution, and only 13% of the subjects gave an answer consistent with a balancing heuristic.

In addition, 33% of the questionnaire subjects and 13% of the interview subjects gave answers inconsistent with the correct solution, representativeness, or balancing. The fact that most of these "deviant" answers occurred in the questionnaire situation suggests that many of them resulted from the subjects' not having read the question carefully enough and thus misunderstanding it on a trivial level. Many of these subjects reported a best guess of greater than 400 for the sample of 10, which seems uninterpretable except as a misreading of the question. However, one pattern (labeled "Trend" in Table 1) deserves some comment, because it appeared in the interviews and has a plausible underlying rationale. In this pattern, the subjects thought (correctly) that the mean of the sample of 10 would be lower than 400. In addition, the two means they gave were consistent, in that the mean of 10 could be the average of the first observation and the average of the next 9 observations. However, it departed from the correct statistical answer in that the mean of the next 9 students was also thought to be less than 400. Comments from the two subjects in the interviews who showed this pattern of responses indicated that the divergent first score led them to believe that the population mean was not actually 400 as stated in the problem.

In summary, the present results replicate those of Kahneman and Tversky (1972) in that the modal esti-

Table 1
Frequency of Solution Types, Experiment 1

Solution Type*			Questionnaires:	Interviews:		
Mean of 10 Scores	Mean of 9 Scores	Label	SAT Problem	IQ Problem	SAT Problem	Combined
Less than 400	400	Correct Solution	44 (21%)	3 (30%)	3 (14%)	6 (19%)
400	400	Representative	68 (33%)	6 (60%)	9 (43%)	15 (48%)
400	400+	Balancing	25 (12%)	1 (10%)	5 (24%)	6 (19%)
400−−**	400−	Trend	18 (9%)	0 (0%)	2 (10%)	2 (6%)
		Unclassified	50 (24%)	0 (0%)	2 (10%)	2 (6%)
Totals			205	10	21	31

**Numerical values are answers for the SAT problem. Classification of responses for the IQ problem is analogous.* ***For the trend solution, mean of 10 scores < mean of 9 scores < 400.*

mate of the mean of the sample of 10 was the population mean. More importantly, 71% of the 95 questionnaire subjects and 71% of the 21 interview subjects who gave the population mean for the mean of the sample of 10 also gave the population mean as their best guess of the mean of the nine unknown scores. The percentage for each group was significantly greater than 50% [$\chi^2(1) = 26.5$, $p < .001$, and $\chi^2(1) = 3.86$, $p < .05$, respectively]. This answer was inconsistent with a balancing heuristic and indicated that these subjects thought that both the sample of 10 students and the sample of 9 students were representative. Moreover, representativeness could even be the fundamental heuristic for subjects classified as "balancers." Using the argument in the introduction, one could claim that these subjects took the sample of 10 as fundamental, believing that it should be representative, and then demanded enough consistency of their predictions to make the mean of the sample of 9 consistent with their answer for the mean of the sample of 10. On the other hand, it is possible that subjects who give answers that are consistent with a balancing solution think fundamentally differently about the problem from the way in which subjects who give representativeness answers do.

We had hoped that in-depth analyses of the interview videotapes would provide further insights into subjects' heuristics. Unfortunately, audio problems with the recording equipment made evaluating some protocols extremely difficult. Accordingly, a second set of interviews was conducted with new equipment. In these interviews, a relatively standardized set of probe questions was developed on the basis of an analysis of the most informative probes used in the first set of interviews. The focus of the more standardized interviews was to confront subjects with solutions different from their own. We believed that information that would be difficult to obtain from a more objective format could be obtained from this confrontation. First, the strength of subjects' confidence in their answers could be assessed. If they maintained their solutions after being shown reasonable alternatives, then one could conclude that their original answers were not frivolous. Second, since subjects were given only the alternative numerical solutions and were asked what they thought the rationale was for those solutions, their understanding of the problem could be assessed more fully.

EXPERIMENT 2

Method

Subjects. The subjects were 26 students recruited from undergraduate psychology classes who participated in the experiments for extra credit. The interview of 1 subject, whose data are not reported, was stopped in the middle because she appeared to be very anxious in the interview situation.

Materials. The SAT problem was used for all subjects. For Subjects 1-11, the problem was identical to the one cited in Experiment 1. For subjects 12-25, the only difference in the problem was that the first person sampled was said to have an SAT score of 550 instead of 250 (correct answer = 415 for the mean of the sample of 10).

Procedure. The general interview procedure was the same as that in Experiment 1. The subject read the first question, which asked for the best guess for the mean of the sample of 10, and answered it, being encouraged to think aloud as much as possible. After the subject's answer, the interviewer asked for the best guess for the mean of the sample of 9. Up to the point of the subject's answering this second question, the interviewer did not intervene except to clarify parts of the problem on request, to correct the subject if he or she misread the question, or to encourage the subject to think aloud. The subject's answer (assuming the first score was 250) was classified by the interviewer as: (1) demonstrating the correct rationale (if the answers to the questions were less than 400 and 400); (2) demonstrating representativeness (if both answers were 400); or (3) demonstrating balancing (if the answers were 400 and greater than 400).

The interviewer (Konold) then told the subject that the problem had been given to many other students and that he was going to present some answers that other students had given. The subject was presented with one of the two patterns of answers that he or she had not given and was asked to comment on it. The subject was then provided with the remaining pattern and asked to comment on that. For example, if a subject gave 400 as the answer to both questions, he or she would be classified as "representative." The interviewer would then say that some people had answered that the best guess for the mean of 10 was less than 400 while the best guess for the mean of 9 was 400 (i.e., the correct solution). The subject was asked if he or she could figure out how someone would have arrived at such an answer, and then was asked what he or she thought of the answer. In the next segment, the interviewer said that some subjects' best guess for the mean of the sample of 10 was 400, while for the sample of 9 it was greater than 400 (the balancing solution). The same series of questions ensued. At the end, the interviewer asked the subject explicitly what the best answer to the question was. (The suggestion that subjects might want to reconsider their original answer is, of course, implicit in presenting alternative answers.) The order of presentation of the two patterns of alternative answers was approximately counterbalanced over subjects. Analogously, subjects who gave the correct solution were presented with the representative and balancing solutions, and the balancers were given the correct and representative solutions. (One subject who demonstrated the "trend" strategy and one whose original answer was confusing were given all three alternative patterns.) The correct answer was never identified as such.

The SAT problem was part of a 1-h-long interview that included several other statistics problems. For Subjects 1-11, the SAT problem was the first problem in the interview, and for Subjects 12-25, it was the third or fourth. The interview on this problem lasted about 10 to 15 min.

Results and Discussion

As described above, the interview consisted of two parts. In the first, the interviewer assumed a passive role, allowing the subjects to independently arrive at answers. In a few cases, the subjects gave more than one answer and seemed undecided about which was correct. Accordingly, two answers are considered in the subsequent discussion. The first is the answer that the subject settled on before the experimenter presented the alternative solutions; the second is the answer that the subject settled on at the end of the interview.

Many subjects hedged their numerical answers with the qualifier "about" or with numerical ranges (see later discussion). Because we were concerned that the subjects might view a best guess of 415 as "about" 400, the interviewer specifically asked these subjects whether the mean would be any more likely to be above

or below 400. An answer was coded as "400" only if the subject thought that there was no tendency in either direction.

Final solution before intervention. The results closely replicated those of Experiment 1 (see Table 2). The final answers subjects gave before the second phase of the interview are referred to as "Answer No. 1." The representative solution was again the modal response (56%), whereas 20% chose the correct solution, 12% chose the balancing solution, 4% chose a "trend" answer, and 8% of the responses fell into an unclassified category (see Table 2). The two unclassified subjects will not be discussed further. One did not appear to understand the question, and the other had several fairly incoherent approaches to the problem, making it impossible to determine what he really believed.

Reactions to alternative solutions. The most striking aspect of the data is that the pattern of results at the end of the interview ("Answer No. 2") was not very different from that before interviewer intervention (see Table 2). There appeared to be a slight movement away from representativeness and toward balancing. However, of the 23 subjects of interest, only 4 changed their answers as a result of considering the alternative solutions. We can conclude that the representative answers were not merely hasty answers to the problem, since when confronted with the correct and balancing answers, 12 of the 14 subjects maintained their representative answers. (The other 2 changed to balancing solutions, 1 subject changed from a correct solution to a balancing solution, and the trend subject changed to a balancing solution.)

We also examined the subjects' reactions to the alternative solutions to determine how well they understood them. As mentioned earlier, after the subjects had been presented with an alternative solution, they were asked how somebody might have arrived at that solution. On the basis of the subjects' comments, understanding of the rationale for the alternative solution was rated independently by two of the authors on a scale from 1 (no understanding) to 10 (excellent understanding). The correlation (r) between the two sets of ratings was .75, and there were only seven cases in which the ratings differed by more than three. As can be seen in Table 3, a majority of subjects showed reasonable comprehension of alternative solutions. Of particular interest is the fact that a majority of subjects who had given representative answers understood the balancing and correct solutions.

Verbal expression of heuristics. Having classified the subjects according to the numerical answer they gave, we wished to explore the extent to which the subjects who gave representative and balancing answers made verbal comments consistent with these heuristics. Unfortunately, few subjects made comments that indicated that they had consciously adopted either heuristic. All of the subjects were asked to explain their numerical answers. Twenty-two of the 23 subjects of interest gave at least one answer of 400. Eleven gave no clear rationale for their answer of 400. Of the remaining 11, 2 gave answers that strongly implicated representativeness, for example, "this random sample is giving you something about the whole community, so it would still be that [points to 400]," and 7 gave justifications that suggested a representativeness heuristic, for example, "if you made sure you were picking totally randomly, it's supposed to come up around the mean." The other two subjects gave an "equal ignorance" argument, consistent with either representativeness or balancing, that is, that there was no reason to expect the sample mean to be either higher or lower than the population mean.

To try to find evidence for balancing heuristics, the entire set of interviews was searched for any statement suggestive of balancing. Only two subjects (one of whom had a representative solution) gave what could be construed as balancing rationales, saying either that there were usually as many scores above the mean as below or that there should be a higher score that would "compensate" for the lower one. Thirteen additional subjects did mention that there should be scores in the sample of nine in the opposite direction from the known score, but this statement was mentioned in passing or paired with a statement that some scores would also be in the same direction.

Also of interest was the possibility that subjects may not have considered the implications of sampling from a large population and consequently may have been concerned about sampling without replacement. Only four subjects made comments indicating that they had considered implications of the fact that sampling was done without replacement, and in only one case did it seem to be part of an eventual balancing solution. One subject brought up the issue and then said it would not matter as the population was large. Two others mentioned

Table 2
Frequency of Solution Types, Experiment 2

Position in Interview	Solution Type*				
	Correct	Representative	Balancing	Trend	Unclassified
Final answer before alternative solutions were presented (Answer No. 1)	5 (20%)	14 (56%)	3 (12%)	1 (4%)	2 (8%)
Answer at end of interview (Answer No. 2)	4 (16%)	12 (48%)	7 (28%)	0 (0%)	2 (8%)

**See Table 1 and text for an explanation of these labels.*

Table 3
Mean Understanding Scores for Subgroups of Subjects

Answer No. 1	Number of Subjects	Understanding of Representative Mean	Representative SD	Balancing Mean	Balancing SD	Correct Mean	Correct SD	Overall Mean	Overall SD
Representative	14			6.75	3.28	6.28	2.70	6.52	2.81
Balancing	3	8.83	0.85			4.50	2.86	6.67	1.43
Correct	5	7.60	1.69	9.80	0.24			8.70	0.93

sampling without replacement only when they were presented with the balancing answer and were asked to hypothesize why other students may have given such an answer.

Three subjects gave a "trend" answer initially, although two spontaneously changed their answers. They seemed to arrive at their estimates for the mean of the nine scores through a quasi-Bayesian rationale in which the divergent first score influenced their estimates of the mean of the population. A related phenomenon was the curious protestations of four subjects that the discrepant score would not change the population mean: for example, "Well, if they've determined that mean from a large school district, then I would certainly put a fair amount of faith in it, and I wouldn't vary it on just one drawing. I wouldn't vary it on a sample of 10 either." These statements all suggested that the population mean was not fixed but that the sample evidence was insufficient to alter their estimate of it. It is possible that these seven subjects thought that there was a larger, unstated, population of which the school district was only a (possibly nonrandom) sample.

Consistency. The verbalizations of the subjects who gave balancing answers thus showed little evidence that they had more of a process view of random sampling than those who gave representative answers. The two groups appeared to differ chiefly in their beliefs about whether the means of the samples should be consistent (i.e., that the mean of the sample of 10 be equal to the weighted average of the first score and the mean of the last nine scores). When the subjects who gave correct answers and those who gave balancing answers were shown the representative answers, most immediately rejected them with comments like "mathematically, it wouldn't work out," or "if they knew anything about math, it [the 550 score] would increase the score [the average of 10]." All three subjects who gave a balancing answer gave a clear rationale for rejecting the representative answer on these grounds.

The representative answer may seem reasonable to many subjects because the question asks for the best guess of the means of two *hypothetical* random samples. Subjects may believe that a lack of consistency is possible for hypothetical random samples, since a best guess for the mean is not necessarily the mean of any particular set of scores. At the end of the interview, those of Subjects 1-11 who gave a representative answer were asked whether both means could be 400 if one was dealing with observed scores. Only one said yes, and it was not clear that she understood that the interviewer was asking about actual scores. The others seemed to believe that both means could not be 400 with actual observations, but could if you were making predictions: "Because I don't know the actual mean of the sample. This is probability, not fact"; "It seems like a contradiction, but I still think that the best guess is 400 because it's random." Although Subjects 12-25 were not explicitly asked this question due to an error in procedure, many of them dealt with its implications at some point, usually in responding to the balancing solution: for example, "They think that the other nine will come out to make it a perfect 400, but when you're picking samples, you're not going to come out with an exact figure."

Many subjects showed discomfort in predicting a single value for the mean of a sample. Some subjects explicitly tied in variability or "randomness" with justification of the representativeness answers; others alluded to the "random" (i.e., indeterminate) nature of the sample and/or remarked that individual scores or even sample means "could be anything." Thirteen of the 23 classified subjects preferred either to preface their estimates of the sample means with hedges such as "about" or "around" or to give interval estimates. However, only 7 of these gave a representative solution.

To summarize, most of the representativeness subjects who were explicitly asked about consistency made it clear that they realized that both means could not be 400 if they were means of actual scores. Other representativeness subjects also commented that, because of variability or randomness in the sampling process, it did not have to work out neatly as in the balancing solution. Many of the subjects also showed discomfort with giving point estimates, indicating that the variability of the sampling process was very much on their minds and suggesting that a best guess for the mean of a hypothetical sample should not be treated the same as an actual sample mean. This discomfort may reflect Kahneman and Tversky's (1972) second meaning of representativeness (i.e., that a random sample should reflect the sampling process): A sample mean must be "random" and hence have considerable variability and uncertainty associated with it. The point is not, of course, that it is a misconception to be aware of the

variability of sample means. What may distinguish experts from novices is that, for the expert, a best guess and the variability of that guess are two separate concepts, whereas the novice has difficulty making this differentiation.

SUMMARY AND CONCLUSION

In the introduction, we raised the general question of whether the tendency of subjects to ignore the known score in giving the best guess for a sample mean was due to a descriptive heuristic such as representativeness or to a mechanistic one such as active balancing. In both studies, the preponderance of subjects who thought that the mean of the sample of 10 was the population mean believed that the mean of the sample of 9 was also the population mean—an answer incompatible with active balancing.

The interviews indicate that, for most subjects, the belief that the population mean was the best guess for both sample means was deeply held: They continued to believe that answer even after having been presented with alternative solutions, and in spite of the fact that they showed reasonably good comprehension of the rationales underlying those solutions. Moreover, detailed analysis of the subjects' explanations of their answers revealed little evidence for balancing imagery. The interviews further suggested that the subjects considered the representative answer reasonable, since they regarded best guesses for the means of random samples differently from the way in which they regarded means of known scores. Moreover, many of the subjects seemed uneasy about making a best guess for the mean of a random sample.

These results have some pedagogical implications. Many textbooks in statistics that discuss the law of large numbers attempt to dispel students' belief in the gambler's fallacy. However, they assume that the basic misconception students have is that of active balancing, and they oppose this mechanism with a correct one called "swamping," wherein the large amount of subsequent data swamps out the impact of the discrepant score on the mean (e.g., Hays, 1981). Our own attempts to teach the swamping conceptualization have usually proved unsuccessful. Our research suggests that such an approach is unfruitful because subjects do not have an incorrect process mechanism; indeed, they have virtually no mechanistic way of thinking about random samples. To refute active balancing is to refute a belief that students actually do not have, and this may confuse them. Since students' actual heuristic, representativeness, is so different in form from the appropriate mechanistic belief, it may not be easy to set up an appropriate confrontation between the two systems to effect any lasting change in students' beliefs about random samples.

REFERENCES

BAR-HILLEL, M. (1980). What features make samples seem representative? *Journal of Experimental Psychology: Human Perception and Performance, 6,* 578-589.

HAYS, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

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.

TVERSKY, A., & KAHNEMAN, D. (1971). The belief in the law of small numbers. *Psychological Bulletin, 76,* 105-110.

(Manuscript received October 3, 1983;
revision accepted for publication March 29, 1984.)