JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY 22, 339–354 (1986)

Prediction and the Partial Understanding of the Law of Large Numbers

ZIVA KUNDA

Princeton University

AND

RICHARD E. NISBETT

University of Michigan

Received May 13, 1985

We examined people's understanding of the implications of the law of large numbers for prediction of social behavior and abilities. We found that people possess a partial understanding of the law: They sometimes recognize that one can predict more confidently *from* larger samples, but do not recognize that one can predict more confidently *to* larger samples. This partial understanding was reflected not only in the predictions subjects made, but also in the explanations they constructed to account for their predictions. It appears that people's intuitions include the notion that increasing the size of the predictor sample of events increases predictability, whereas they do not include the notion that increasing the size of the predicted sample of events increases predictability. © 1986 Academic Press, Inc.

In the course of our daily lives we are constantly making predictions. We predict people's future behavior from their past behavior, we predict behavior in one domain from behavior in another domain, and we predict our own attitudes from those of other people. In fact, all of our expectations about the world, about other people, and about ourselves may be viewed as based on predictions.

It is therefore of great importance to determine how good people are at making predictions, and to what extent people appreciate and act upon the rules governing the predictability of events from other events. Failure to appreciate these rules could result in wrong beliefs, faulty decisions, and misguided behavior.

The research reported here was supported by NIMH Grant 1 R01 MH38466-01 and NSF Grant SES 85-07342. Requests for reprints should be addressed to Ziva Kunda, Psychology Department, Green Hall, Princeton University, Princeton, NJ 08544.

0022-1031/86 $3.00

One of the most basic rules governing predictions is the law of large numbers. The implications of this rule to prediction are quite straightforward: The confidence one may have in predictions increases with the size of the sample one is predicting from and with the size of the sample one is predicting to. Do people recognize this? Do people possess an intuitive understanding of the implications of the law of large numbers to prediction?

The evidence bearing on this question is mixed. In a recent paper we have argued that people sometimes do recognize that predictability increases with sample size (Kunda & Nisbett, 1986). We found that subjects sometimes understood that aggregation of the units of measurement increases the correlation magnitudes. This rule, termed the aggregation principle, was much more likely to be used when the events in question were relatively "codable," that is, when the events were capable of being unitized and interpreted clearly (See also Fong, Krantz, & Nisbett, 1986; Jepson, Krantz, & Nisbett, 1983; Nisbett, Krantz, Jepson, & Kunda, 1983). Relatively high codability exists for the domains of athletic and academic abilities. In these domains the units of measurement—games or tests—are clearly defined, and scores, once given, are straightforward and unambiguous. For such domains subjects made correlation estimates in accordance with the aggregation principle, realizing that, for example, grade in spelling for an entire term is better predicted by grade in spelling for another entire term than grade on a single spelling test is by grade on another single test.

Codability is considerably lower, however, in domains involving social behavior, for example, friendliness and honesty. In these domains there are no obvious units of measurement, and no well-defined and agreed-upon method of assigning scores to behaviors. How, for example, does one determine how many units of friendliness are contained in a conversation, or how much friendliness is expressed in a smile? In such domains people were less able to use statistical principles. Even for those latter domains, however, subjects were found to be able to use the aggregation principle to a degree when a "within" design was used that encouraged comparison of the degree of prediction to be obtained across levels of aggregation. With such designs subjects recognized, for example, that honesty over a large number of occasions is better predicted by honesty over another large number of occasions than is honesty on one occasion by honesty on another occasion.

The conclusion that people recognize that predictability of events increases with the size of the predictor class of events and the predicted class of events appears to be challenged, however, by some findings reported by Tversky and Kahneman (1980). These authors showed that people do not appear to appreciate the implications of the aggregation principle for cases in which an aggregate is predicted from a single event

and vice versa. Their subjects erroneously believed that an aggregate of events provides a better prediction of a single event than a single event does of an aggregate of events. Thus subjects believed that the score on a full length intelligence test was a better predictor of the score on an abbreviated test than the score on the abbreviated test was of the score on the full length test. Similarly, subjects believed that the probability of an athlete's winning the first event of the decathalon given that he won the decathalon was higher than the probability of the athlete's winning the decathalon given that he won the first event. These judgments are in error because the predictability of the aggregate from a single event is identical to the predictability of a single event from the aggregate. Since the magnitude of a correlation is independent of the direction of prediction, aggregation increases predictability equally whether it is performed on the predictor or on the predicted events.

In contrast to our results, these findings indicate that people do *not* understand the implications of the aggregation principle to prediction. Most disturbingly, Tversky and Kahneman's subjects' failure to appreciate these implications of aggregation occurred in the very domains where our subjects were best able to appreciate aggregation—the domains of academic and athletic abilities where event codability is high.

These apparently conflicting results could be reconciled if people were shown to have only a partial understanding of the aggregation principle. People may realize that aggregation of the predictor class of events increases predictability, and at the same time fail to realize that aggregation of the predicted class of events also increases predictability. In other words, people may realize that it is easier to predict *from* larger samples, yet fail to realize that it is easier to predict *to* larger samples.

Informal conversations with our acquaintances suggest that the notion that predictability increases with the size of the predicted sample of events is extremely counterintuitive. We have found that laypeople as well as many of our statistically sophisticated colleagues tend to respond with incredulity when we propose this rule to them. "If I can't predict behavior in a single situation," people often protest, "how can I be expected to predict behavior over *20* situations?" The answer, of course, is that 20 situations allow for a much more stable estimate of true score than does a single situation. Therefore the larger the size of the predicted sample of events, the larger the correlation is expected to be.

A partial understanding of the aggregation rule could account for subjects' performance both in our studies (Kunda & Nisbett, 1986) and in Tversky and Kahneman's studies (1980). In our studies, subjects predicted aggregates from aggregates (i.e., total-to-total correlations) and single events from single events (i.e., item-to-item correlations). Thus their recognition that total-to-total correlations were higher than item-to-item correlations could have been produced by relying only on the left side of the equation,

that is, on the size of the predictor. The Tversky and Kahneman subjects, on the other hand, made judgments about total-to-item and item-to-total correlations. Here, too, the (mistaken) belief that total-to-item correlations were greater than item-to-total correlations could have been produced by relying only on the size of the predictor, which is greater in the first case, while ignoring the size of the predicted class of events, which is greater in the second case.

In order to determine whether people do indeed possess such a partial understanding of the aggregation principle, it is necessary to ask the same subjects to make all four types of predictions: single event to single event, single event to aggregate, aggregate to single event, and aggregate to aggregate. By varying the size of the predictor and predicted classes of events independently in this way, it will be possible to assess the role of each in producing judgments about predictability.

STUDY 1: ADHERENCE TO A RULE OF ASYMMETRIC AGGREGATION

In the first study we examined subjects' estimates of correlation as a function of the size of the predictor class of events and the size of the predicted class of events. We studied estimates both for the relatively codable domain of abilities and for the less codable domain of traits. We also compared peoples' estimates of correlation to the actual correlations in these domains, in order to determine accuracy. Subjects' estimates were examined in both a within design and a between design.

Method

Subjects were 131 University of Michigan undergraduates of both sexes who were enrolled in introductory psychology. They participated in groups of three to eight subjects of the same sex. As sex did not affect any of the dependent measures in this or the next study, it will not be discussed any further. Subjects were randomly assigned to conditions. They responded first in a between design, in which each subject provided estimates for both abilities and traits at only one level of aggregation. The numbers of subjects in the item-to-item, item-to-total, total-to-item, and total-to-total conditions were, respectively, 36, 37, 39, and 19. Next, they responded to a series of intervening, unrelated questionnaires, and finally, each subject provided estimates of all four levels of aggregation for one of the abilities or traits, thereby providing data for the within-subjects design. Thus, the between and the within designs were both 4 (content) × 4 (level of aggregation). Nine subjects did not respond to the within questionnaire, leaving a total of 122 subjects for the within design.

Using the same question format employed by Kunda and Nisbett (1986), subjects in the within version estimated correlations either for one of two abilities—spelling and basketball—or for one of two traits—honesty and friendliness. They were asked to estimate the probability that, for a given trait or ability, two individuals would maintain their relative rankings from one situation (or aggregate of situations) to another situation (or aggregate of situations). Each within design subject assessed consistency for one of the abilities or one of the traits at each of the four levels of aggregation: item-to-item, item-to-total, total-to-item, and total-to-total. [Estimates about item-to-item and total-to-total correlations constitute a replication of a study reported by Kunda and Nisbett (1986).] Subjects in the between version estimated

the correlations for both abilities and traits at only one of the four levels of aggregation. [Data for item–item and total–total correlations in the between design were also reported by Kunda and Nisbett (1986).]

The item for spelling was a single test and the total was the average of the 20 tests of a term. For basketball, the item was the number of points scored in a single game, and the total was the number scored over 20 games. For the two traits, the item was behavior in a single situation and the total was behavior over 20 situations, on average. The item-to-item questions for spelling and for friendliness read as follows:

> In Jefferson school students are required to take spelling tests each week in the 21 week term. Suppose you knew that Johnny got a higher grade than Danny on one such test. What do you suppose is the probability that Johnny would get a higher grade than Danny's again a few weeks later, on the last test taken that term?
>
> Suppose you observed Tom and Joe in a particular situation and found that Tom was more friendly than Joe. What do you suppose is the probability that in the next situation in which you observe them you will also find Tom to be more friendly than Joe?

The remaining questions used the combination of levels of aggregation appropriate for each condition. For example, the total-to-total question about friendliness substituted "20 different situations" for "a particular situation" and asked the subjects to suppose that Tom had been found to be more friendly "on the average."

This question format has been shown to provide a highly sensitive measure of subjects' beliefs about predictability, and was validated by the remarkable accuracy of subjects' estimates in some domains (Kunda & Nisbett, 1986). Such probability estimates have the additional virtue of providing a direct measure of Kendall's τ which, in turn, yields by derivation an estimate of Spearman's r:

$$E(r) = \sin \pi\tau/2 \quad \text{(Kendall, 1962, p. 124).}$$

We report results in terms of these derived correlations, although all statistical tests are based on the raw percentage estimates obtained from subjects. We do this to allow readers to compare the present results to related results presented in this form by Kunda and Nisbett (1986) and also because this is a convenient way of communicating with psychologists, who often think about prediction in terms of correlation coefficients.

Actual item-to-item correlations were obtained empirically by Kunda and Nisbett (1986). Correlations for honesty and friendliness are based on research investigating the cross-situational consistency of these traits (Bem & Allen, 1974; Chaplin & Goldberg, 1985; Hartshorne & May, 1928; Mischel & Peake, 1982). Correlations for basketball and spelling were obtained, respectively, from the scores of University of Michigan basketball players and from spelling test scores in two fifth-grade classes. Actual correlations at the remaining levels of aggregation for each trait and ability were obtained by applying the Spearman-Brown formula to the item-to-item correlations.

Results

We will present first the estimates obtained in the within-subjects design. In such designs subjects are most likely to appreciate statistical principles (Fischhoff, Slovic, & Lichtenstein, 1979; Kunda & Nisbett, 1986). Subjects' estimates of the consistency of the two abilities (spelling and basketball) did not differ from each other at any level of aggregation nor did their estimates of the consistency of the two traits (honesty and

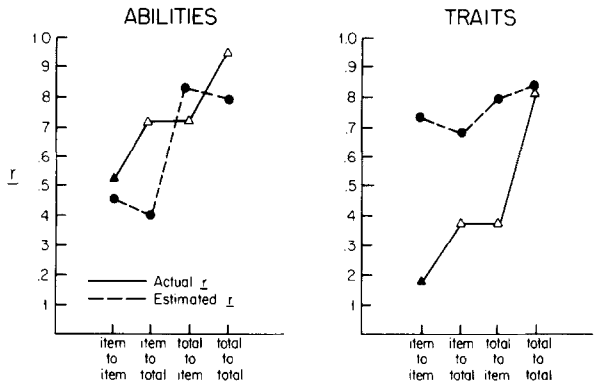

FIG. 1. Actual correlations and correlations estimated in a within-subjects design, at all four levels of aggregation, for traits and for abilities. Open triangles indicate actual *r*s predicted by Spearman-Brown from the actual item-to-item *r*s.

friendliness), so both ability and trait estimates were pooled together at each level of aggregation and so were the actual correlations (which were also close).

It may be seen in Fig. 1 that for both abilities and traits, subjects' application of the aggregation principle was based only on the size of the *predictor* sample of events. They entirely ignored the size of the predicted sample. In each case a 2 × 2 (size of predictor × size of predicted) ANOVA revealed a significant effect only for the size of the predictor, $F(1, 61) = 63.05$, $p < .001$ for abilities and $F(1, 59) = 4.20$, $p < .05$ for traits. There was no significant effect for the size of the predicted class of events. The $F(1, 59)$ for traits was approximately 1, while the effect for abilities, $F(1, 61) = 2.36$, $.10 < p < .15$, was in the counternormative direction.

Thus subjects did not differentiate in their estimates between item-to-item and item-to-total correlations, nor did they differentiate between total-to-item and total-to-total correlations. They expected the item-to-total correlation to be as low as the item-to-item one, and expected the total-to-item correlation to be as high as the total-to-total one. As a consequence, for abilities, where they were relatively accurate about item-to-item and total-to-total correlations, they underestimated the correlation when it was presented as item-to-total, $t(61) = 5.94$, $p < .001$, and overestimated it when it was presented as total-to-item $t(61) = 3.12$, $p < .01$.[1] For traits, where they grossly overestimated item-to-item correlations, they also greatly overestimated both item-to-total and total-to-item correlations, both *p*s $< .001$.

[1] For this test the actual correlation was converted into a percentage estimate and treated as μ.

For both abilities and traits, subjects' estimates of the correlation between an item and the total were asymmetrical: They estimated the total-to-item correlation to be higher than the item-to-total correlation. The planned comparison between these two estimates was significant for both abilities and traits, $t(61) = 6.92$, $p < .01$ and $t(59) = 2.57$, $p < .05$, respectively. This asymmetry is similar to that found for Tversky and Kahneman's (1980) subjects. The degree of asymmetry is greater for abilities than for traits, but there is more room for such asymmetry for abilities.

In the case of abilities, subjects recognize that item–item correlations are much lower than total–total correlations. In the case of traits, subjects recognize only a very slight difference between item–item and total–total correlations. The comparison between the two is actually not significant by itself for traits. The slight difference between item–item estimates and total–total estimates for traits constitutes a failure to replicate the pattern found in the within design study by Kunda and Nisbett (1986) where subjects' appreciation of aggregation was relatively good, and instead is closer to the pattern found with their between design, where subjects' appreciation of aggregation was poor. The most plausible explanation for this is that subjects' grasp of the aggregation principle for traits is so weak that answering all four questions, as they did in the present study, including those about total–item and item–total correlations, was sufficient to confuse them. In the Kunda and Nisbett (1986) study, where each subject only answered two questions, about item–item and total–total correlations, it may have been easier for them to focus on the important differences between questions.

We found the same pattern of results for the between-subjects version of the study as we did for the within-subjects version. It may be seen in Fig. 2 that there was less overall slope due to aggregation for both

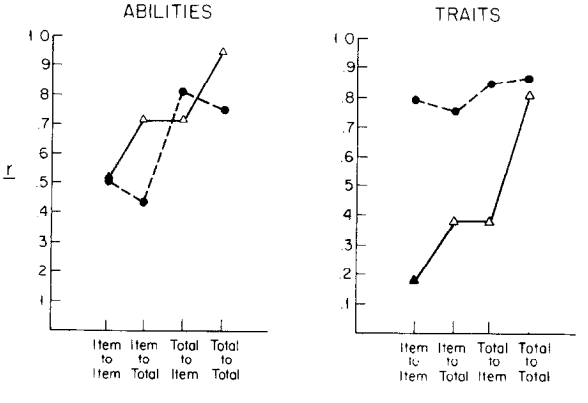

FIG. 2. Actual correlations and correlations estimated in a between-subjects design, at all four levels of aggregation, for traits and for abilities. Open triangles indicate actual *r*s predicted by Spearman-Brown from the actual item-to-item *r*s.

abilities and traits, but aside from that the results were almost identical. Subjects' estimates were again based only on the level of aggregation for the predictor variable. The $F(1, 127)$ for size of predictor for abilities was 63.05, $p < .001$, and for traits it was 7.78, $p < .01$. The Fs both for effect of the predicted variable and for the interaction were both small and nonsignificant.

Subjects' responses in the between design also showed the same pattern of inaccuracies found for the within design: for abilities, they were relatively accurate about item–item and total–total correlation, but they underestimated the item–total correlation, $t(36) = 5.56$, $p < .001$, and overestimated the total–item correlation, $t(38) = 2.83$, $p < .01$. For traits, where they grossly overestimated item–item correlations, they also overestimated both item–total and total–item correlations, both ps $< .001$.

The results are thus quite clear-cut. Subjects made predictions as if they were guided by a single, mistaken statistical principle: Aggregation is important for predictor events but not at all important for predicted events. Thus subjects appear to recognize that stability of observations is important if one wishes to make a prediction *from* them, but they appear not to recognize that stability of observations is important if one wishes to make a prediction *about* them.

STUDY 2: ARTICULATION OF RULES OF AGGREGATION

Subjects' responses in Study 1 followed the pattern expected if people were relying on a partially correct rule that recognized the importance of aggregation of the predictor sample of events but did not recognize the importance of aggregation of the predicted sample of events. But these results, in and of themselves, do not yet permit the conclusion that people's predictions are guided by such a rule, because the same pattern of responses could also be accounted for differently. In particular, it is possible that when people predict one class of events from another, their responses are determined entirely by the similarity between the two classes of events (Kahneman & Tversky, 1972). Tversky (1977) found that small objects are judged to be more similar to large objects than large objects are to small objects. People judge North Korea to be more similar to China than they judge China to be to North Korea. In the same fashion, small samples might be judged to be more similar to large samples than large samples are to small samples. Hence small samples might be judged more predictable from large samples than large samples are from small samples. Indeed, this is how Tversky & Kahneman (1980) interpreted their data.

The proposition that people do in fact possess and use the partially correct rule of aggregation would be strengthened if it could be shown that people are capable of articulating such a rule, and that they spontaneously explain their responses in terms of this rule. Study 2 assessed

this possibility. Each subject was asked to make two predictions, which differed from each other either in the size of the predictor sample, or in the size of the predicted sample, or in both. If subjects' responses to the two questions were not identical, they were asked to explain why.

This within-subject design provides an even stronger test of rule appreciation than the within design employed in Study 1, because subjects were not susceptible to being confused by the presence of more than two questions. The strongest test is provided by the conditions where only the predictor or only the predicted sample size are varied. It seems fair to assume that if subjects do not appreciate aggregation under such advantageous circumstances, they simply do not understand the rule.

It was expected that, once again, subjects' responses would reflect appreciation of the effects of aggregation of the predictor sample and no appreciation of the effects of aggregation of the predicted sample. And the explanations subjects gave to account for differences in their responses were expected to reflect the same partial appreciation. Subjects were expected to articulate good approximations of the rule concerning the impact of aggregation of the predictor, but to be unable to articulate the rule concerning the impact of aggregation of the predicted sample.

Method

Subjects were 96 University of Michigan undergraduates of both sexes who were enrolled in introductory psychology.

The predictions subjects were asked to make were identical to those employed in Study 1: predictions about basketball, spelling, honesty, and friendliness. Each subject responded to a pair of questions about one of the traits or abilities, at two different levels of aggregation. The pairs were obtained by pairing each level of aggregation with all the remaining ones, which led to the following six kinds of pairs: (1) item–item and total–item, (2) item–total and total–total, (3) item–item and item–total, (4) total–item and total–total, (5) item–item and total–total, (6) item–total and total–item. It may be seen that in the first two pairs only the size of the predictor sample is varied, in the next two pairs only the size of the predicted sample is varied, and in the last two pairs the sizes of both predictor and predicted samples are varied. Order of predictions in each pair type was varied orthogonally to content (basketball, spelling, honesty, and friendliness) and pair type. The full design was therefore 4(content) × 6(pair type) × 2(order) with two subjects in each of the resulting 48 cells. Subjects were randomly assigned to condition.

Each prediction appeared on a separate page. On the last, third, page subjects were asked: "If your answers to the questions on the two previous pages were not identical, please explain why."

Results

Predictions. It may be seen in Fig. 3 that the pattern of results obtained for subjects' predictions at each level of aggregation was similar to that obtained in Study 1. These predictions follow, once again, the pattern that could be expected if subjects recognized the effects of aggregation on the predictor sample without recognizing the effects of aggregation on the predicted sample.

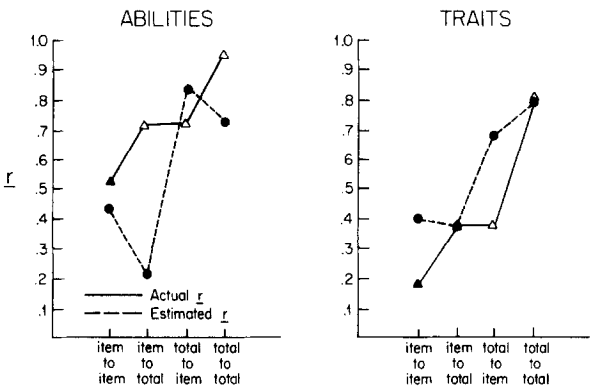

FIG. 3. Actual and estimated correlations, at all four levels of aggregation, for traits and for abilities. Open triangles indicate actual *r*s predicted by Spearman-Brown from the actual item-to-item *r*s.

ANOVAs for traits and for abilities yielded results similar to those obtained in Study 1. Separate ANOVAs were carried out for subjects' first responses and for their second responses for traits and for abilities. All but one of the four Fs(1, 44) for size of predictor were significant in the correct direction, all at $p < .001$. The exception, $F(1, 44) < 1$, was for first responses to trait questions, where the rule is expected to be least salient. All but one of the four Fs(1, 44) for the predicted variable were nonsignificant, with the exception of second responses to ability questions, which yielded a significant effect in the counter-normative direction, $F(1, 44) = 9.16$, $p < .01$. All but one of the four Fs(1, 44) for the interaction terms were nonsignificant, with the exception of first responses to the ability question, $F(1, 44) = 4.67$, $p < .05$. Since this is the only case where a significant interaction was found, this finding is most probably due to chance.

The pattern of results for traits is much closer to the pattern found for within design subjects in Kunda and Nisbett (1986) than to the pattern found for within design subjects in Study 1. This suggests that the four judgments required of within design subjects in Study 1 did in fact serve to confuse them.

Examination of the pattern of responses provided by each subject yielded similar results. Table 1 presents the number of subjects in each pair type condition (collapsed over content and order) whose predictions were in the normatively correct rank order, and the number whose predictions were in the remaining, incorrect, rank orders. It may be seen that in the pairs where only the size of the predictor was varied, the majority of subjects correctly expected greater predictability for the larger predictor (69%). But in the pairs where only the size of the predicted sample was varied, only a minority of the subjects correctly expected

TABLE 1
NUMBER OF SUBJECTS IN EACH PAIR TYPE WHOSE PREDICTIONS FOLLOWED EACH PATTERN[a]

	N		N
	Only predictor varied		
Item–item < total–item	*12*	Item–total < total– total	*10*
Item–item = total–item	3	Item–total = total–total	5
Item–item > total–item	1	Item–total > total–total	1
	Only predicted varied		
Item–item < item–total	*1*	Total–item < total–total	*5*
Item–item = item–total	10	Total–item = total–total	8
Item–item > item–total	5	Total–item > total–total	3
	Both predictor and predicted varied		
Item–item < total–total	*13*	Item–total < total–item	15
Item–item = total–total	1	Item–total = total–item	*0*
Item–item > total–total	3	Item–total > total–item	1

[a] Numbers in italics refer to correct patterns of responses.

greater predictability for the larger predicted sample (19%). When the predictor and predicted sample sizes were varied simultaneously, 81% of subjects correctly recognized that total–total predictions are greater than item–item predictions, but 94% of subjects erroneously believed that total–item predictions were greater than item–total predictions. These results suggest, once again, that subjects appear to recognize that increasing the size of the predictor sample increases predictability, but do not recognize that increasing the size of the predicted sample increases predictability.

Explanations. It appears, therefore, that subjects' predictions are compatible with a partially correct rule of aggregation. But do they actually possess such a rule? If they do, we would expect them to articulate it when explaining why their responses to the two questions were not identical. Subjects' explanations were coded as to whether they reflected correct or incorrect beliefs about the impact of aggregation on the predictor events and about the impact of aggregation on the predicted events. Any expression of the idea that increasing sample size increases predictability were coded as correct, whereas expressions of the idea that increasing sample size decreases predictability were coded as incorrect. A small number of explanations did not fit either of these categories, and these were coded as nonanswers. Interrater agreement on coding was 92%.

It may be seen in Table 2 that subjects clearly understand and are able to articulate the rule that increasing the size of the predictor increases predictability. In those conditions where only the size of the predictor sample was varied, the majority of subjects explained their responses in ways that reflected correct understanding of the impact of aggregation

TABLE 2
NUMBER[a] OF SUBJECTS IN EACH PAIR TYPE WHO GAVE CORRECT AND INCORRECT EXPLANATIONS OF THE EFFECT OF SAMPLE SIZE

	Predictor		Predicted	
	Correct	Incorrect	Correct	Incorrect
Only predictor varied				
Item–item & total–item	12	0	0	0
Item–total & total–total	8	0	0	0
Only predicted varied				
Item–item & item–total	0	0	0	3
Total–item & total–total	0	0	4	2
Predictor and predicted varied				
Item–item & total–total	12	1	1	1
Item–total & total–item	14	0	0	0

[a] N for each cell is 16.

(63%), and none expressed incorrect beliefs. Two examples will help to convey subjects' appreciation of the importance of sample size for the predictor sample. One subject, asked to explain why she estimated the total–item association for spelling test scores to be higher than the item–item association, said the following: "There is more data to use in the first question than in the second. On one test, Johnny's results could be a fluke. After 20 the results are more conclusive." Another subject, asked to explain why he said that the total–total association for friendliness was higher than the item–total association, said, "The first was higher because I observed Tom and Joe in 20 associations so I have more evidence to support that Tom is a nicer guy."

It may be seen just as clearly that most subjects do not understand that increasing the size of the predicted sample also increases predictability. In those conditions where only the size of the predicted sample was varied, only very few subjects expressed correct understanding of the impact of aggregation (12%). Indeed, subjects were more likely to express incorrect beliefs in these conditions (16%) than they were to express correct beliefs. The same pattern of rule comprehension emerges from subjects' explanations in those conditions where both the sizes of the predictor and predicted samples were varied simultaneously. Here, a majority of subjects expressed correct understanding of the impact of aggregation of the predictor (81%), whereas almost none expressed correct understanding of the impact of aggregation of the predicted sample (3%). Here, subjects were just as likely to be incorrect about aggregation of the predicted sample (3%) as they were to be correct.

DISCUSSION

It is important to note that the law of large numbers is probably best viewed not as a simple, unitary rule but, rather, as a complex set of interrelated rules. In its most general form the law of large numbers states that the larger a sample is the more likely it is to be representative of the population from which it is drawn. This broad statement has diverse implications. Some of the notions encompassed within its scope are that we may generalize with greater confidence from large samples than from small ones, that small samples are more likely than large ones to deviate from the population mean, that extreme scores are likely to regress to the mean, and that increasing the size of predictor and predicted samples increases predictability.

The major thrust of most of the early work on statistical reasoning was to demonstrate people's many failures to appreciate these different statistical rules, and the resulting view of people's capabilities was almost uniformly bleak (Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980). More recently, a more optimistic view of people's capabilities has emerged from research attempting to delineate the conditions under which people do appreciate statistical principles. Some of the notions that initially appeared to be missing from people's intuitions (Kahneman & Tversky, 1972, 1973) were later shown to have intuitive equivalents which could be used in familiar domains that lend themselves to statistical reasoning, and with particular problem structures. This was true for the impact of sample size on generalization and prediction and for the phenomenon of regression to the mean (Kunda & Nisbett, 1986; Nisbett et al., 1983).

While we do not yet have a complete mapping of people's understanding of the many implications of the law of large numbers, the present data suggest that although the initial extreme pessimism about people's inferential abilities was exaggerated, extreme optimism is also unwarranted. It now appears that some of the rules derived from the law of large numbers do have intuitive counterparts, whereas others remain extremely counterintuitive.

Using the same problem structures, in the same domains, at times even the same subjects, we found that people understand, apply, and articulate the notion that increasing the size of the predictor sample increases predictability. But we found no evidence indicating that people understand the notion that increasing the size of the predicted sample increases predictability, even in the within-subject designs that are most likely to elicit appropriate responses (Fischhoff et al., 1979; Kunda & Nisbett, 1986), and even for highly codable events such as athletic and academic performance which lend themselves easily to other types of statistical reasoning (Kunda & Nisbett, 1986; Nisbett et al., 1983).

Why do people develop the particular partial version of the law of large numbers that they do? Why do they learn that predictability increases

with the size of the predictor sample and at the same time fail to learn that predictability increases with the size of the predicted sample?

Our answer is by necessity speculative. We begin with the obvious fact that people induce rules from regularities found in the pattern of data available to them. It seems likely that this data pattern may encourage recognition of the importance of N on the predictor side far more than it does on the predicted side, for several reasons.

First, we suspect that people normally treat the size of the to-be-predicted event as a given, and look for evidence of varying types that may serve as predictors. The to-be-predicted event might be, for example, Jane's likely performance on the job, or John's likely enjoyment of the party. One of the ways in which predictor evidence may differ is with respect to its quantity or level of aggregation. People discover, through a process of trial and error, that more evidence of a given type on the predictor side generally does a better job of predicting than less evidence. The reciprocal process, that of fixing the quantity of evidence on the predictor side and allowing the quantity of evidence on the to-be-predicted side to vary, does not normally occur. People are far more likely to search for various types of evidence needed to make a given prediction than they are to search for various types of prediction that may be based on a given piece of evidence. As a consequence, people learn about the importance of aggregation on the predictor side without learning about the importance of aggregation on the to-be-predicted side.

Second, the available data patterns may be biased because of asymmetrical similarity judgments. This is particularly true for item–total and total–item correlations. Small samples may be judged as more similar to large samples than large samples are to small ones (Tversky, 1977), and, if predictability is estimated on the basis of such similarity judgments, small samples will be judged to be better predicted by large samples than large samples are by small ones. Repeated reliance on the representativeness heuristic in this way will lead to a pattern of predictions from which the partial rule subjects appear to possess may be induced. Thus we suspect that similarity judgments, or the representativeness heuristic, play an important role in shaping predictions. But we also believe that the pattern of predictions that results from repeated use of this heuristic leads to the induction of a broad inferential rule which eventually gains independence from the underlying similarity judgments, and may then guide predictions in its own right.

Not only do subjects fail to induce the correct impact of aggregation of the predicted sample on predictability, but they actually tend to believe that increasing the size of the predicted sample *decreases* predictability. Ironically, they use this counternormative heuristic precisely in those circumstances where statistical reasoning is likely to be most salient: ability questions in the within-subjects design in Study 1 and second

responses to ability questions in Study 2. This suggests that this counternormative heuristic, like other more correct ones, reflects recognition of the role of chance in predictions. But instead of realizing that chance influences cancel each other out with aggregation, subjects seem to think that the uncertainty aggregates with the number of events to be predicted.

Whatever the reason for people having the partially correct rule of aggregation, we believe our studies show that people do indeed have it and often use it as a basis for predictions. The correct half of the rule is likely to serve them well. The ability to base confidence in predictions on the amount of evidence available can save people from overreliance on unreliable evidence as well as from underreliance on reliable evidence. But the missing half of the rule is likely to hurt their ability to assess confidence in predictions accurately, and may at times undercut and distort any benefits obtained through the application of the correct half. This is especially true in the case of item–total and total–item predictions, about which subjects were particularly unable to reason correctly.

It is also important to note that although our studies show that people are capable of appreciating the impact of aggregation of predictors on prediction, it does not follow that they will always do so spontaneously. We have reason to believe that their typical performance may be considerably lower than their competence. Our within-subjects studies were designed to maximize the likelihood that people will use any rules that they possess. Other studies have shown that when the relevance of statistical rules is not cued by the use of within designs or by other means, statistical reasoning is less likely (Fischhoff et al., 1979; Kunda & Nisbett, 1986; Nisbett et al., 1983).

It is hoped that, eventually, we will arrive at a complete mapping of people's understanding of the many implications of the law of large numbers, and will also have more detailed knowledge about the factors that determine when and how people access and apply inferential rules. Such a mapping is essential in order to maximize the effects of statistical training on reasoning about everyday life problems (cf. Fong et al., 1986).

REFERENCES

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, **81**, 506–520.

Chaplin, W. F., & Goldberg, L. R. (1985). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology*, **47**, 1074–1090.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, **23**, 339–359.

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.

Hartshorne, H., & May, M. A. (1928). *Studies in deceit*. New York: MacMillan.

Jepson, C., Krantz, D. H., & Nisbett, R. E. (1983). Inductive reasoning: Competence or skill? *Behavioral and Brain Sciences*, **6**, 494–501.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge Univ. Press.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–251.

Kendall, M. G. (1962). *Rank correlation methods*. London: Griffin.

Kunda, Z., & Nisbett, R. E. (1986) The psychometrics of everyday life. *Cognitive Psychology*, **18**, 195–224.

Mischel, W., & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review*, **89**, 730–755.

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.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice–Hall.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327–352.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments about uncertainty. In M. Fishbein (Ed.), *Progress in social psychology*, Hillsdale, NJ: Erlbaum.

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY

Published bimonthly by Academic Press, Inc., 1 East First Street, Duluth, MN 55802. Number of issues published annually: 6. Editor: Dr. Thomas M. Ostrom, Department of Psychology, Ohio State University, 404C West 17th Avenue, Columbus, OH 43210.

Owned by Academic Press, Inc., 1250 Sixth Avenue, San Diego, CA 92101. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.

Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securities of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.

Total no. copies printed: average no. copies each issue during preceding 12 months: 1994; single issue nearest to filing date: 1953. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 1638; single issue nearest to filing date: 1690. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 0; single issue nearest to filing date: 0. Free distribution by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 72; single issue nearest to filing date: 72. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 1730; single issue nearest to filing date: 1712.

(Signed) Roselle Coviello, Senior Vice President