COGNITIVE PSYCHOLOGY 18, 195–224 (1986)

The Psychometrics of Everyday Life

ZIVA KUNDA

Princeton University

AND

RICHARD E. NISBETT

University of Michigan

We examined people's ability to assess everyday life correlations such as the degree of agreement that exists for various kinds of evaluations and the degree of consistency that characterizes social behavior from occasion to occasion. We found substantial accuracy for correlation estimates if two conditions were met: (1) subjects were highly familiar with the data in question and (2) the data were highly "codable," that is, capable of being unitized and interpreted clearly. We generally found extreme inaccuracy if either of these conditions was not met. Subjects were particularly inaccurate about correlations involving social behavior: They severely overestimated the stability of behavior across occasions. In addition, even subjects who were statistically sophisticated showed limited appreciation of the *aggregation principle,* that is, the rule that the magnitude of a correlation increases with the number of units of evidence on which observations are based. © 1986 Academic Press, Inc.

Jane liked the movie; will you? Bill and you have served on several committees and he has always seemed very fair and very agreeable; would he make a good chairman? Our answers to such questions guide the conduct of our daily lives. Everything from the degree of pleasure to be expected from life's minor diversions to the degree of success to be expected for life's major enterprises depends on the accuracy of our answers.

Logically, answers to such questions rest on our beliefs about correlations, for example, correlations between different raters' evaluations of movies or correlations between fairness and agreeableness in different

The research reported here was supported by Grant NSF SES-8218846 from the National Science Foundation and 1 R01 MH38466-01 from the National Institute of Mental Health. We are indebted to J. E. Keith Smith for statistical advice and to David M. Buss, Nancy Cantor, Geoffrey T. Fong, Sarah Hampson, Keith Holyoak, Lee Ross, and Henri Zukier for valuable critiques of an earlier version of the manuscript. We thank Robert Vallone and Sabrina Lin for assistance in data collection and Sara Freeland for able editorial assistance. Reprint requests should be addressed to Ziva Kunda, Psychology Department, Princeton University, Princeton, NJ 08544.

0010-0285/86 $7.50

situations over time. How accurate are people's estimates of such correlations? Since the study of the correlations that underlie interrater agreement and behavioral consistency is the province of the field of psychometrics, this question may be rephrased in a way that is suggestive of the methodology that might be used to pursue it: How accurate is *lay psychometrics?*

Surprisingly little direct evidence is available on the question of lay accuracy about everday life correlations. On the other hand, a great deal of indirect evidence bears on this question. Walter Mischel (1968) and Donald Peterson (1968) set off a debate that has raged within the personality area for almost two decades by proposing that (a) the actual consistency of behavior across different situations generally presumed to tap the same trait or disposition is very low, and (b) people believe that behavioral consistency is high, and (c) people therefore suffer from what might be called an "illusion of consistency." There is little doubt that (a) is correct. Recent reviews indicate that the average correlation between any two phenotypically different behaviors generally presumed to tap the same trait (e.g., honesty, friendliness, dependency, hostility, extraversion) achieves a level of .15 or less (e.g., Mischel & Peake, 1982; Nisbett 1980).

Is (b) correct? Do people believe that the true correlation is in excess of that found in the literature? Some psychologists clearly do, at any rate. The major response by personality psychologists to the Mischel and Peterson critique was simply to deny, on methodological grounds that are in our view quite unconvincing, that the empirical evidence was very good (e.g., Block, 1977, Olweus, 1977). There is also considerable indirect evidence that laypeople overestimate behavioral consistency (Jones & Nisbett, 1972; Mischel, 1968; Nisbett & Ross, 1980; Ross, 1977). But the only direct evidence for this, to our knowledge, is a study by Jennings, Amabile, and Ross (1982).

Further evidence suggesting that people are sometimes inaccurate in perceiving correlations comes from research in the judgment and decision tradition. Two major findings about people's statistical failings seem particularly pertinent.

1. People have been shown in many laboratory studies to have difficulty in detecting covariation between complex events of a kind resembling those of daily life (Chapman, 1967; Chapman & Chapman, 1967, 1969; Golding & Rorer, 1972; Hamilton, 1979; Jennings et al., 1982; Nisbett & Ross, 1980). For example, subjects find it difficult to perceive accurately the covariation between Rorschach signs seen by clients and the clients' symptoms.

2. People often fail to apply the law of large numbers to everyday life events (Kahneman & Tversky, 1972; Tversky & Kahneman, 1971, 1974).

This bias is highly relevant to estimation of correlations because the so-called *aggregation principle,* a derivation of the law of large numbers, governs the association between reliability of evidence and correlation magnitude. The most important implication of the principle is that correlation is a function of the number of units underlying each observation. Thus, for example, IQ tests typically have test–retest (total–total) correlations of .90 or higher, but this is based on individual question (item–item) correlations of .10 or less. Similarly, as Epstein (1979, in press) has recently emphasized, the .10–.15 correlations characteristic of social behavior from situation to situation translate into substantial consistency of behavior at highly aggregated levels. Thus, applying the Spearman–Brown prophecy formula to item–item correlations of .15 gives a correlation of .78 between the average level of behavior on 20 occasions with the average level on 20 other occasions. Empirical research indicates that the Spearman–Brown formula provides a very good approximation to actual aggregation effects both for interrater agreement (Epstein, 1983; Moskowitz & Schwartz, 1982) and for behavioral consistency (Epstein, 1979; Hartshorne & May, 1928; Mischel & Peake, 1982; Moskowitz & Schwartz, 1982; Newcomb, 1929). (For a review of evidence on aggregation, see Rushton, Brainerd, & Pressley, 1983.)

The aggregation principle is relevant to estimation of everyday life correlations because it provides a means of assessing unobserved correlations. Even if one has never observed a correlation at the group level, say, for the agreement between two college classes in their evaluations of movies, one could make a good estimate of it by using an accurate estimate at the individual level and applying the aggregation principle. Similarly, an estimate of the stability of behaviors from one situation to another can be obtained by applying the aggregation principle to one's beliefs about longterm stability of behavior.

The evidence on people's statistical capabilities is mixed, however. Several studies show that people can detect correlations involving relatively barren laboratory stimuli such as columns of numbers and pairs of dial readings (e.g., Beach & Scopp, 1966; Erlick, 1966; Erlick & Mills, 1967; Jennings et al., 1982; Wright, 1962; see Alloy & Tabachnik, 1984, and Crocker, 1981, for reviews; although the work of Jennings et al. indicates that people may have difficulty detecting correlations much below .6 even with stimuli of that type).

We and our colleagues (Nisbett, Krantz, Jepson, & Kunda, 1983) recently have shown that people also have substantial ability to use at least some variants of the law of large numbers, for at least some types of problems. The factors we found to influence its use are relevant to present concerns.

1. People are more likely to use the law of large numbers for events

that are highly "codable" than for events that are less codable. Nisbett et al. defined codability as the ease with which events may be unitized and given a score characterizing them in clear and readily interpretable terms. Sports events tend to be highly codable in this sense. In principle, a machine could code most of the relevant events in a basketball game—number of baskets per player, number of baskets per ball handling, and so on. Some other events related to achievements tend to be highly codable, or at any rate to come to us in highly coded form. For example, academic performance is usually assessed by assigning numerical values to clearly defined units of performance, and accomplishments in various occupations are often similarly coded, for example, number of manufactured objects produced, sales made, or cases won. In contrast, social behavior is rarely so codable. When comparing friendliness across two occasions, for example, there is no obvious unit to use. Should we use smiles per minute or "good vibrations" per social exchange? Score assignment poses similar problems, especially for purposes of comparing different people in different situations: What coding scheme will allow you to directly compare the degree of friendliness that Jane showed at the party with the degree of friendliness that Bill showed at the meeting? Nisbett et al. (as well as Jepson, Krantz, & Nisbett, 1983, and Fong, Krantz, & Nisbett, 1986) found that subjects were much more likely to apply the law of large numbers to highly codable problems about athletics and other kinds of achievements than to less codable problems involving social behavior. They also found that manipulations designed to help people code events in such a way that the law of large numbers could be applied to them resulted in more reasoning in accordance with the law.

2. People are more likely to use the law of large numbers for highly familiar domains and problem types. For example, subjects with experience in team sports were more likely to use the law of large numbers for a problem about football than subjects without experience in sports, and subjects with experience in acting were more likely to use the law of large numbers for a problem about acting than subjects without experience in acting. The fact that people are more likely to use the law of large numbers for familiar domains is undoubtedly due in large part to the fact that more familiar events are apt to be more codable, and hence the relevance of the law is more apparent.

Thus, the literature is mixed with respect to people's ability to estimate important correlations in everyday life. On the one hand, there are some conspicuous cognitive and statistical incapacities that might lead us to suspect that such estimations would pose very severe difficulties. On the other hand, the evidence is indirect, other indirect evidence suggests that accuracy may be possible at times, and several theorists have argued that the biases that produce errors in person perception in the laboratory may be muted in everyday life contexts (e.g., Hogarth, 1980; Miller & Cantor, 1982; Swann, 1984). There would appear to be no substitute for actually examining some real everyday life correlations and determining how accurate people's beliefs about them are.

Our experimental work leads us to expect that both familiarity and codability of events are important determinants of accuracy. We examine the effects of familiarity in the context of beliefs about the degree of agreement that exists for different kinds of evaluations. Evaluations do not in general pose severe coding problems. Jane's report that she liked the movie is a clear unit (one person's evaluation) with a clear code that may be compared at least on an ordinal scale to one's own evaluation and to those of other people. This is not to say that beliefs about evaluations are error free, since people may dissemble and data about evaluations may be biased in other respects as well. But the barriers to accurate perception of interrater agreement would not seem to be insurmountable. We expect people to be accurate about the degree of agreement that exists for familiar kinds of evaluations.

We examine the effects of codability on accuracy in the context of beliefs about the consistency of highly familiar behaviors. We expect people to be more accurate about correlations involving highly codable events than about correlations involving less codable events. This means we expect them to be more accurate about correlations involving ability- and achievement-related behavior than about social trait-related behavior. In both cases, we expect people to be accurate only about events at levels of aggregation that they have actually observed, since our previous work suggests that people do not have a firm understanding of the law of large numbers in the abstract and cannot be expected to be able to steer from observed to unobserved levels of aggregation.

A METRIC FOR MEASURING BELIEFS ABOUT CORRELATION

It would obviously be very useful to have a metric for measuring people's beliefs about correlations that mapped in some clear way onto the statistician's methods of measuring correlations. We propose that an appropriate metric would be one based on judgments of contingent probability, which people do with ease and, often, with substantial accuracy as well.

As it happens, one kind of probability estimate has a direct interpretation as a kind of correlation coefficient. The probability of the reversal of a pair ordering is a direct measure of Kendall's τ which is defined as the proportion of pairs of objects having the same relative order in their ranking on two variables (for example, the proportion of pairs in which observer X thinks $A > B$ and observer Y also thinks $A > B$) minus the proportion of pairs showing different relative order in the two rankings (that is, the proportion of pairs in which observer X thinks $A > B$ and observer Y thinks $A < B$). Tau yields, by derivation, an estimate of Spearman's *r*: $E(r) = \sin(\pi\tau/2)$ (Kendall, 1962, p. 124). Table 1 shows how these percentage estimates map onto correlation coefficients.

In all the studies that follow, we asked subjects to estimate the probability that two pairs of observations would have the same rank ordering, for example, the probability that two individuals or groups would agree on the ranking of objects. The general format of the questions for inter-rater agreement was, "Suppose X thought *A* was greater than *B*. What do you suppose is the probability that Y would also think that *A* was greater than *B*?"

Subjects had no difficulty in answering such questions and, in fact, were able at times to provide probability estimates that were strikingly accurate estimates of actual correlations, as will be seen. In most studies, we calibrated subjects by pointing out that an estimate of .50 is tantamount to guessing that there is no relationship between X's and Y's opinion, .60 is tantamount to a slight relationship, and so on.

In studies that paralleled those to be reported, we sometimes specified a magnitude of an evaluation or a magnitude of a comparative evaluation. For example, "Suppose X thought *A* was very good. What do you suppose is the probability that Y would also think that *A* was very good?" Or, "Suppose X thought that *A* was much greater than *B*. What do you suppose is the likelihood that Y would also think that *A* was much greater than *B*?" Answers to these questions have no clear interpretation as correlations, but they yielded results that are entirely comparable to those reported. In particular, the results for subjects' recognition of the aggregation principle were always the same whether the simple contingent probability was estimated or one of these latter two probabilities.

TABLE 1
The Conversion of Percentage Estimates into Correlation Coefficients

Percentage estimate	*r*
50	.00
55	.16
60	.31
65	.45
70	.59
75	.71
80	.81
85	.89
90	.95
95	.99

In all the studies that follow, we report results in terms of correlations, although they are based on subjects' percentage estimates for contingent probabilities, and all statistical tests are based on the percentage estimates. We do this because only correlations can be manipulated using the Spearman–Brown formula and because this is a convenient way of communicating with psychologists, who often think about association and prediction in terms of correlation coefficients. In particular, the controversy about the consistency of trait-related behaviors, for which our data have important implications, has been in terms of the magnitude of correlations.

To maintain complete comparability between estimated and actual correlations, the latter also were always derived from τ coefficients. But as a practical matter, it would have made almost no difference whether we presented standard Pearson *r*'s, Spearman *r*'s derived from τ's or *r*'s derived from τ's at the opposite level of aggregation from the target level and calculated by means of the Spearman–Brown prophecy formula. Differences among the three techniques were always trivial. Unless otherwise stated, subjects from whom estimates were obtained were University of Michigan undergraduates of both sexes who were enrolled in introductory psychology. No sex differences in estimates of correlation were found.

INTERRATER AGREEMENT

In the first series of studies to be presented, we examined actual interrater agreement among people for evaluations of different kinds of objects and we examined people's beliefs about agreement. The evaluations differed in their degree of familiarity. In the first study we examined evaluations that were familiar both at low levels of aggregation (item–item) and at high levels of aggregation (total–total). In the second study we examined evaluations that were familiar only at low levels of aggregation. In the third study we examined evaluations that were familiar at neither level of aggregation. The anticipation was that subjects would be more accurate in their estimations of correlation for types of evaluations that they had actually observed and that their estimations would be more in line with the requirements of the aggregation principle.

Study 1: Beliefs about Agreement for Course Evaluations

In the first study, we examined college students' beliefs about the degree of agreement that exists for evaluations of college courses at two levels of aggregation—the level of individuals and the level of the population of students who took the course. Students often exchange opinions about courses and thus could be familiar with the degree of agreement to be expected between any two individuals. Students are also familiar with the stability of course evaluations at the aggregate level. Some courses are known to be terrific term after term, others are perennially awful or mediocre. In addition, at some universities, including the University of Michigan, where the study was conducted, aggregate level agreement can be examined by noting the stability of summaries of course evaluations from term to term.

Method

Actual ratings. Total-to-total correlations for course evaluations were obtained by correlating the average course ratings published in the Michigan Student Assembly (MSA) course evaluation guide for 1 year with those published in the guide for the next year. These averages were based on ratings provided by students who filled out the evaluation questionnaire at the end of the term while waiting to register for the following term's classes. They evaluated the overall quality of all the courses they had attended that term on the same scale used to evaluate students' class work, which is a 13-point scale ranging from *E* to *A* +. For our calculation we included all the courses that were taught by the same professor both years and whose published ratings both years were based on at least 20 students, a total of 65 courses. Each course was rated by 71 students on the average, so the item-to-item correlations were estimated by applying Spearman–Brown to the actual total-to-total correlations using an *N* of 71. (Where actual correlations have been estimated rather than calculated directly, this is indicated on figures by an open triangle.)

Estimates. There were 63 subjects. Subjects in the item-to-item condition estimated the likelihood that they would agree with another student, identified as J.K., on the ranking of two courses. Subjects in the total-to-total condition estimated the likelihood that the MSA rankings of two courses would agree with the MSA rankings of the same courses obtained the previous year.

Results

Figure 1 presents actual and estimated correlations. It may be seen that subjects' estimates are highly accurate. Both item-to-item and total-to-total estimates were very close to the respective actual correlations and not significantly different from them. The estimates are also in line with those required by the aggregation principle. Tests carried out on the raw percentage estimates showed that neither of the estimated correlations was significantly different from the correlation predicted from subjects' estimates at the opposite level of aggregation (open circles in Fig. 1).[1]

These data establish that people are capable of very great accuracy about covariation at two quite different levels of aggregation. Is the accuracy due at least in part to recognition of the force of the aggregation principle or is it due solely to the fact that subjects are familiar with the (highly codable) data at various levels of aggregation? These data cannot

[1] The Spearman–Brown predicted correlations and the actual correlations were converted into percentage estimates. Both the Spearman–Brown predicted value and actual value were treated as mu when comparing estimates to them. All *p* values reported are based on two-tailed tests.

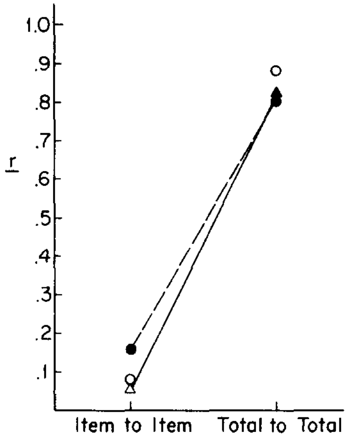


FIG. 1. Average actual (—) and estimated (- - -) correlations at both levels of aggregation for course evaluations. Open geometrical shapes indicate r predicted by Spearman–Brown from estimated r at the opposite level of aggregation.

answer this question. We pursue it in the following two studies in which the evaluations at the aggregate level are less familiar than is the case with course evaluations.

Study 2: Beliefs about Agreement for Attributes of People

Subjects proved remarkably accurate in their estimates of correlations for evaluations of courses. Would they be equally accurate in their estimates of correlations for evaluations of people? People spend a great deal of time discussing the attributes of other people, so we may presume reasonable familiarity with the relevant data on agreement. At least this is true for individual or item–item level data. People probably have little opportunity for observing aggregate level agreement about the attributes of other people, since these are rarely discussed or otherwise expressed in large group settings. If, however, people are capable of using the aggregation principle to estimate aggregate level agreement, then they might nevertheless be accurate about correlations based on aggregate data.

To obtain data concerning actual agreement, we contacted two relatively small sororities and asked all the members to rate each other on a number of personality traits and other personal characteristics such as attractiveness and degree of overweight. Beliefs about these correlations were obtained from a different group of subjects who assessed the agreement among any two individuals or groups of 20 individuals on the same personal characteristics.

Method

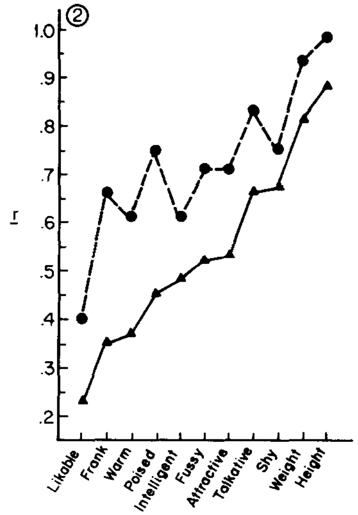
Actual ratings. Subjects were members of two small sororities who had both agreed to answer our questionnaire in exchange for a $100 honorarium. Members responded in a group session held at the sorority. The first sorority included 16 members, all of whom responded to the questionnaire. The second sorority included 33 members, 14 of whom were unable to attend the scheduled session, leaving a total of 19 respondents. For each sorority, a list of all the members was obtained in advance, and the questionnaire, which was presented as concerned with social perception, required the subjects to rate all members of the sorority on a 6-point scale on 11 characteristics—warmth, talkativeness, frankness, fussiness, poise, the extent to which the respondent liked the member, intelligence, attractiveness, degree of overweight, height, and shyness.

Estimates. A total of 55 introductory psychology subjects assessed agreement either at the item-to-item level or at the total-to-total level. Subjects were asked to imagine that a group of people who knew each other well, such as members of a fraternity or sorority, all rated each other on a series of dimensions. Some subjects in the item-to-item condition were asked to estimate the probability that they would agree with another group member on the ranking of two other members of the group for the attribute. Other subjects were asked to estimate the probability that the person on their right would agree with another group member. No differences were found between subjects making predictions about their own rankings and those making predictions about the rankings of the person on their right, so their responses were pooled. Subjects in the total-to-total condition predicted the probability that the average ranking of two group members by 20 members would agree with the average ranking given by 20 other members.

Results

Actual ratings. To obtain item-to-item correlations for each characteristic, τ coefficients were obtained in each sorority independently, converted into correlation coefficients, and then averaged across both sororities. The correlation between sororities on the coefficients for the 11 characteristics was .81. The Spearman–Brown formula was used to estimate the actual total-to-total correlations for an *n* of 20, since neither sorority had enough members to calculate *r* at this level (actual and Spearman–Brown estimated *r*'s were virtually identical at $n = 9$, however).

Estimates. Figure 2 presents actual and estimated correlations *at the item-to-item level.* It may be seen that subjects were very well calibrated indeed in their guesses about the degree of agreement between two individuals. The correlation between the estimated and the acutal item-to-item correlations is remarkably high—.93. It may also be seen that subjects systematically overestimated this agreement. The mean discrepancy between estimated and actual *r* is .20, which is statistically significant, $t(35) = 8.55$, $p < .001$. It should be noted that this does not establish that subjects overestimate the correlation in the data available to them. It may be that people mute their opinions about others and mask any disagreements. If so, then our subjects might be giving accurate estimates of the biased correlation evidence available to them.



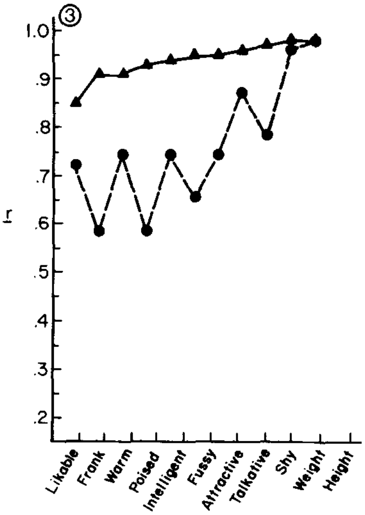


FIG. 2. Actual (—) and estimated (- - -) correlations at the item-to-item level for evaluations of attributes of people.

FIG. 3. Average actual (—) and estimated (- - -) correlations at the total-to-total level for evaluations of attributes of people.

Figure 3 presents actual and estimated correlations at the total-to-total level. It may be seen that subjects were not nearly as well calibrated about agreement at the aggregate level as they were at the item-to-item level. The correlation between estimated and actual correlations at the aggregate level was .56, which is significantly lower than the near perfect correlation obtained at the item to item level, $p < .05$.

Taken together, Fig. 2 and 3 suggest that subjects' reduced accuracy at the total-to-total level results from their failure to recognize the dramatic impact of aggregation on correlations. The actual aggregated correlations are uniformly very high. Yet subjects believe them to be as low and as variable as are the item-to-item correlations, thus markedly underestimating the actual total-to-total correlations, $t(18) = 13.03$, $p < .001$, for the mean discrepancy between actual and estimated correlation.

Subjects' estimates showed no recognition of the aggregation principle. When correlations are averaged across all 11 attributes the estimated total-to-total correlation is considerably lower than the total-to-total correlation of .98 that is expected by applying Spearman–Brown to subjects' estimated item-to-item correlation, $t(18) = 16.30$, $p < .001$. Similarly, the estimated item-to-item correlation is considerably higher than the item-to-item correlation of .16 that is expected by applying Spearman–Brown to subjects' estimated total-to-total correlation, $t(35) = 21.28$, $p < .001$.

There was one exception to the rule that subjects did not recognize that total-to-total correlations are higher than item-to-item correlations, however. The exception was for "likability." The $t(53)$ contrasting the two levels of aggregation for likability was 2.73, $p < .01$. It is possible that this is merely accidental, inasmuch as we would expect one or more of the contrasts to be significant at least at the .10 level by chance. On the other hand, it does seem possible that the finding is meaningful. There actually exists a concept for aggregate level likability, namely the notion of popularity. Perhaps because the likability dimension is so important, we tend to pay attention to how well liked people are in general. And, unlike for most other dimensions, we often do get opportunities to observe liking at the aggregate level. This is sometimes formal, as in voting for people for various offices, but more often informal, as when groups of people may observe the affective reactions of others.

(It is quite unlikely that subjects recognized that liking evaluations are subject to the aggregation principle simply because they estimated the item–item correlations to be very low. In follow-up studies we examined subjects' beliefs about other evaluations for which they had no opportunity to observe aggregate level agreement, for example, evaluations of black and white photographs of people and evaluations of slide photos of pictures. Even when subjects' estimates of item–item correlations were as low as .20, they failed to recognize that total–total correlations would be higher.)

Study 3: Beliefs about Agreement for Evaluations of Scientific Documents

In Study 1 subjects' estimates of correlations were in line with those required by the aggregation principle. In Study 2 they were not, despite subjects' accuracy about the relative magnitude of correlations at the item-to-item level. The explanation that we prefer for this is that people are not sufficiently aware of the aggregation principle in the abstract to allow them to apply it to domains where they have observed correlation at only one level. A relatively stringent test of this explanation would be to examine the estimates of correlation made by subjects who are knowledgeable about the aggregation principle in the abstract and see if even they are unable to apply it to relatively unfamiliar domains.

In Study 3 we examined psychologists' beliefs about the degree of agreement that exists for evaluations of manuscripts and grant proposals. Despite the importance of such documents to their professional lives, few psychologists have much familiarity with agreement about them even at the individual level. (An exception is those psychologists who review for journals, who can usually count on receiving the opinions of another reviewer and the editor. Only for a very few prolific reviewers would this amount to very much data, however.) Still fewer psychologists encounter the opinions of others about grant proposals with any regularity. And, of course, almost no psychologists ever observe the opinions of *aggregates* of colleagues, for either manuscripts or grant proposals.

We also studied the beliefs about agreement of lay subjects. Laypeople are of course even less familiar with degree of agreement for such evaluations than psychologists and hence would be expected to show little accuracy and no ability to make predictions in accordance with the aggregation principle.

Method

Actual ratings. Actual item-to-item correlations for journal manuscripts were obtained from ratings given by reviewers of *Journal of Personality and Social Psychology* (*JPSP*) manuscripts.[2] Reviewers rated manuscripts on three scales—theoretical contribution, empirical contribution, and interest value. Actual correlations for NSF proposals were estimated from data for solid state physics and economics panels obtained by Cole, Cole, and Simon (1981).

Estimates. Expert subjects were 40 members of an audience attending a symposium on statistical aspects of human judgment. Some subjects provided us with estimates of the degree of agreement to be expected, on each of the three evaluation scales, between two reviewers of manuscripts submitted to *JPSP* and between two reviewers of grant proposals submitted to either the solid state physics or the economics panel of NSF. Other subjects were asked to guess the degree of agreement for such evaluations to be expected between two panels of 8–10 reviewers each. Since the actual correlations for economics and solid state physics panels did not differ, nor did either expert or lay estimates of these correlations, results for the two disciplines were combined. Lay subjects were 120 University of Michigan students.

Results

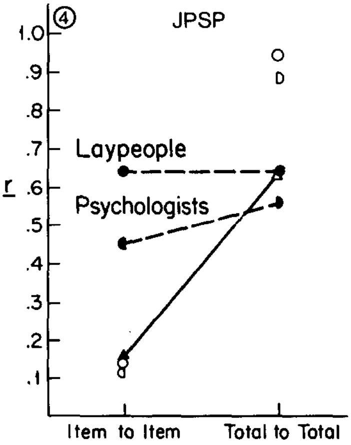
It may be seen in Fig. 4 and 5 that both psychologists and lay subjects were quite inaccurate about the degree of agreement to be expected of the ratings of manuscripts or proposals by any two individuals. In all cases the item-to-item correlations were grossly overestimated. All comparisons of estimated item-to-item correlations to actual item-to-item correlations were significant at least at the .001 level. Neither group was inaccurate at the total-to-total level for either manuscripts or proposals, but it is quite unlikely that this is because of actual observation at this level, inasmuch as few psychologists and no lay subjects have ever observed aggregation at this level. The accuracy at the total-to-total level was probably just a matter of chance, since, as we report next, it could not have been due to application of the aggregation principle.

Both psychologists and laypeople expected identical or nearly identical

[2] We are grateful to the former editor, Melvin Manis, for allowing us access to these files. It should be noted that the correlations are based on ratings by two different consultants for each of a number of articles. This method of necessity treats individual differences in rating scale usage as error and thus provides a lower bound for the actual reliability that exists.

correlations at both levels of aggregation for both *JPSP* manuscripts and NSF proposals. In no case was the estimated total-to-total correlation more than trivially higher than the estimated item-to-item correlation (all p's $> .25$). In all cases, estimated total to total correlations were significantly lower than those expected from applying the Spearman–Brown formula to the estimated item-to-item correlations (all p's $< .001$). Similarly, estimated item-to-item correlations were all considerably higher than those expected from applying Spearman–Brown to the estimated total-to-total correlations (all p's $< .001$).

In summary, laypeople substantially overestimate the degree of agreement about manuscripts and grant proposals between any 2 experts, and they do not expect agreement to be greater between two panels of 8–10 experts than between 2 experts. Experts themselves, with substantial training in statistics, show an almost identical pattern of expectations. This suggests that, even if experts understand the aggregation principle in the abstract, they are unable to apply it to important real world evaluations that they have not actually observed closely. Almost surely, then, the same thing is true of lay subjects: Any abstract appreciation of the aggregation principle they may have is probably inadequate to guarantee



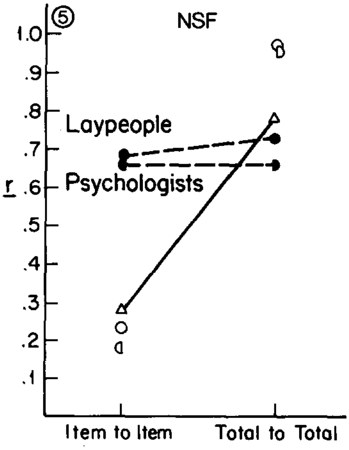


FIG. 4. Actual correlations (—) and correlations estimated by laypeople and by psychologists (- - -) for *JPSP* manuscript evaluations at both levels of aggregation. Open geometrical shapes indicate r predicted by Spearman–Brown from estimated r at the opposite level of aggregation.

FIG. 5. Actual correlations (—) and correlations estimated by laypeople and by psychologists (- - -) for evaluations of NSF grant proposals, at both levels of aggregation. Open geometrical shapes indicate r predicted by Spearman–Brown from estimated r at the opposite level of aggregation.

its use in a domain where they are not familiar with data at more than one level of aggregation.

BEHAVIORAL CONSISTENCY

We have found that when people have had little opportunity to observe other peoples' evaluations of particular objects, they can be quite inaccurate about the degree of agreement that exists for such evaluations, and they are unable to apply the aggregation principle to them. But we have found also that people can be quite accurate about other types of evaluations, so long as they have had the opportunity to observe the data at a given level of aggregation.

There is good reason to believe that the accuracy we found for evaluations is dependent on their generally high codability. Neither unitization nor interpretation of evaluations normally would be a problem for the evaluations we studied. But for behavior, where the situation or occasion is the natural item, codability can range across a variety of difficulty levels. For skill-related behavior such as academic or athletic performance the units—grades or scores—are quite clear and interpretation normally poses no problem. In addition, information about abilities typically is available at various levels of aggregation. People are given grades for individual exams and for entire courses; statistics are available on players' performance both in single games and over entire seasons. Thus, for abilities we would expect a fair degree of accuracy both at low levels of aggregation and at high levels.

Social behavior, however, is harder to unitize and more subject to interpretive vagaries. Thus we would expect people to be less accurate about the correlations that exist for social behavior. We expect particularly poor accuracy at the item-to-item level, where psychologists at any rate appear to have been surprised by the lack of consistency from one situation to another.

In the next series of studies we examined people's beliefs about the degree of consistency to be expected for ability-related behaviors and for trait-related behaviors, both at the level of individual occasions and at highly aggregated levels.

Study 4: Lay and Expert Perceptions of the Consistency of Traits and Abilities

Method

Each of 55 University of Michigan students provided estimates of correlations at either the item-to-item or the total-to-total level for two traits, namely, honesty and friendliness, and for two abilities, namely, basketball scoring ability and spelling ability as measured by spelling tests. In addition, the same experts as in Study 3 also provided estimates. Subjects were asked to estimate the probability that for a given trait or ability two individuals would maintain their relative ranking from one situation to another (for the item-to-item correlation) or from the average of 20 situations to the average of 20 other situations (for the total-to-total correlation). The item-to-item question for honesty read as follows: "Suppose you observed Jane and Jill in a particular situation and found that Jane was more honest than Jill. What do you suppose is the probability that in the next situation in which you observe them you would also find Jane to be more honest than Jill?"

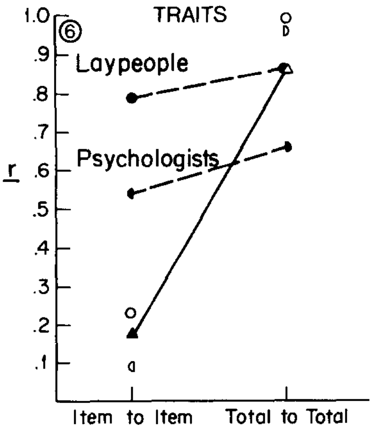
The total-to-total version of the question substituted "20 different situations" for "a particular situation" and asked the subjects to suppose that Jane had been found to be more honest "on the average." The item-to-item level of aggregation for basketball was the number of points scored in a particular game and the total-to-total level was the number scored over the first 20 games of the season vs the last 20. The item-to-item level for spelling was one test. The total-to-total level was the average for the 20 tests of the first term vs the 20 tests of the second term.

Actual correlations for basketball were obtained by correlating the scores of University of Michigan players for the previous season. Actual correlations for spelling tests were assessed by examining spelling scores in 2 fifth-grade classes in two different schools. Actual correlations for honesty are available from the landmark work by Hartshorne and May (1928) who conducted a study in which they measured the behaviors of thousands of children in situations contrived to measure honesty behavior. The average correlation that they obtained across situations was .23, though it should be noted that this should be regarded as an upper bound, because the .23 figure is based on values that are themselves aggregations in some cases. Actual correlations for friendliness are based on an average from three studies that examined people's friendliness in two or more situations and obtained ratings from observers (Bem & Allen, 1974; Chaplin & Goldberg, 1985; Mischel & Peake, 1982). The average correlation for these studies was .13, but it should be noted that this correlation also is based on aggregated measures for the most part and that the correlation at the level of one situation with one other situation would be lower.

Results

Subjects' estimates of the consistency of the traits of honesty and friendliness did not differ at either level of aggregation nor did their estimates of consistency of the abilities of basketball and spelling. The actual correlations for the two traits and for the two abilities were also similar. So both trait estimates and ability estimates were pooled at each level of aggregation and so were the actual correlations. It may be seen in Fig. 6 that subjects' estimates were very seriously in error for traits at the item-to-item level. This was true both for lay subjects and for expert subjects, $p < .001$ and $< .01$, respectively. In addition, the experts were also mistaken about the correlation at the total-to-total level, $p < .05$. We suspect that the experts yielded a curve that was lower overall because some of them, at least, were aware that traits are not very good predictors. Their memories may have been jogged by the presence of Walter Mischel, seated prominently in front of the room!

Unfortunately, neither Mischel's presence nor the statistical training of the expert subjects was sufficient to enable them to recognize the relevance of the aggregation principle for trait data. Neither they nor the lay



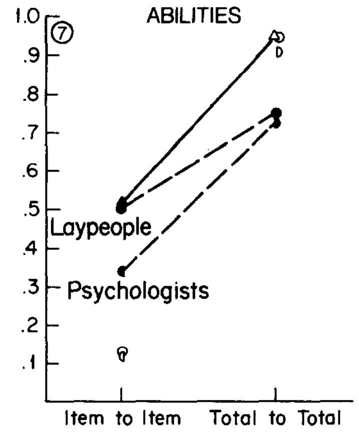


FIG. 6. Actual correlations (—) and correlations estimated by laypeople and by psychologists (- - -) for trait-related behaviors, at both levels of aggregation. Open geometrical shapes indicate *r* predicted by Spearman–Brown from estimated *r* at the opposite level of aggregation.

FIG. 7. Actual correlations (—) and correlations estimated by laypeople and by psychologists (- - -) for ability-related behaviors, at both levels of aggregation. Open geometrical shapes indicate *r* predicted by Spearman–Brown from estimated *r* at the opposite level of aggregation.

subjects provided estimates for the total-to-total level that were significantly different from their estimates for the item-to-item level. In addition, both group's estimates of the item-to-item level correlations were very far from the item-to-item correlation predicted from subjects' estimates of the correlations at the total-to-total level; p for lay subjects $<.001$, p for experts $<.01$. Finally, both groups' estimates of the total-to-total correlations were very far from the total-to-total predicted from subjects' estimates of the item-to-item correlations; both p's $< .001$. This failure of subjects to recognize that their beliefs about total-to-total correlations *entail* very low item-to-item correlations is most probably at the heart of many psychologists' disbelieving reactions to the Mischel and Peterson critiques.[3]

[3] Readers who have been following the literature on the trait controversy will recall that Mischel and Peake (1982) argued that people's tendency to overestimate correlations between any two situations is due to their mistaken assumption that (cross-situational) consistency is as high as (within-situation) temporal stability. In a follow-up study, we modified our questions so as to create two clearly different conditions—one in which subjects were to make a prediction about the same kind of situation as the one for which they already had observations of behavior and one in which they were to make a prediction about a different kind of situation. Subjects dramatically overestimated both temporal stability and cross-situational consistency. Thus, a failure to distinguish between temporal stability and cross-situational consistency is not the only source of overestimation of the latter.

Matters are quite different for both lay and expert estimates for abilities, as may be seen in Fig. 7. Lay estimates at the item-to-item level are not different from the actual values, and psychologists' estimates differ from the actual values only at the .10 level. Both lay and expert subjects significantly underestimated correlations at the total-to-total level; both p's $< .001$. But it may be seen that both groups recognized that correlations at the total-to-total level should be greater than at the item-to-item level; both p's $< .01$.

The results for abilities do not, however, show a full recognition of the force of the aggregation principle. Both experts and lay subjects were fairly far off the predictions made from estimates by subjects for the opposite level; p for total-to-total level for both groups of subjects $<.001$, p for item-to-item level for lay subjects $<.001$, for psychologists $<.05$. And the experts show no better recognition of the differences at the two levels than do lay subjects; $F(1,70)$ for difference between slopes <1.

The major anticipations thus were supported fully. Both lay and expert subjects severely overestimated the consistency of trait-related behavior at the level of the situation. Both groups' estimates of correlation for traits failed to show any influence of the aggregation principle. Estimates made by both groups for abilities were relatively accurate at the item-to-item level and showed some congruence with the aggregation principle. It therefore appears that greater codability contributes to greater accuracy about correlations and to greater appreciation of the aggregation principle.

Study 5: Predicting Trait-Based and Ability-Based Outcomes

An important suggestion of Study 4 is that people actually may believe that social trait-related behaviors are more consistent, and therefore more predictable, than ability-related behaviors. Lay subjects estimated that trait-related behaviors were more highly correlated than ability-related behaviors both at the item-to-item level, $p < .001$, and at the total-to-total level, $p < .01$. This, of course, reverses the true state of affairs, since the abilities we examined were in fact more consistent than the traits. In Study 5 we explored the implications of this finding. We studied an outcome that we expected that most people would regard as primarily ability based, namely grade point average (GPA), and an outcome that we expected that most people would regard as trait based, namely success as a community action organizer in the Peace Corps. We wanted to determine whether subjects would think that success in the Peace Corps was more predictable than GPA.

For both outcomes, we examined subjects' beliefs about predictability based on an interview. Interviews have been shown to have little predictive power for either intellectual performance or for various kinds of job performance. Most validity coefficients are correlations of less than .10 (cf. Hunter & Hunter, 1984; Nisbett & Ross, 1980). But, because of the power of the consistency illusion, it could be anticipated that subjects would overestimate the predictability attainable on the basis of an interview, especially for the outcome based on social traits. We also examined subjects' beliefs about the predictability of both outcomes based on highly aggregated forms of evidence—the reports of acquaintances in the case of the Peace Corps outcome and high school GPA in the case of the college GPA outcome.

Method

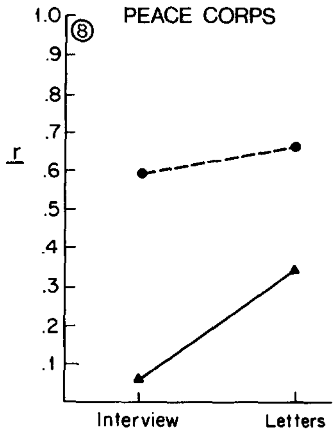
One hundred thirty-two Michigan students assessed the predictability of yearlong performance in the Peace Corps from either a single event, namely an interview, or from an aggregate of events, namely the average rating given to letters of recommendation by teachers, ministers, and community leaders who knew the applicants well. Other subjects assessed the predictability of overall University of Michigan GPA, either from an interview or from a different kind of aggregate—high school GPA. The actual predictability of Peace Corps performance from an interview and from letters of recommendation were obtained from Stein (1966). The actual predictability of GPA at the University of Michigan from high school GPA was provided by Michigan's admissions office. Our estimate for the actual predictability of GPA from an interview is somewhat arbitrary. The .07 estimate reflects the fact that all coefficients for interview validity with which we are familiar are nonnegative, while cross-validated coefficient rarely exceeds .10 (see, for example, Klitgaard, 1985; Mayfield, 1964; Ulrich & Trumbo, 1965).

Results

Figures 8 and 9 present predicted and actual correlations for both types of outcomes. The most striking finding is that people appear to believe that Peace Corps performance is far more predictable than it actually is, for both kinds of evidence. The validity coefficient for job interviews, in general, and for the Peace Corps interview, in particular, is less than .10, yet subjects estimated that it was .59, $t(33) = 7.75, p < .001$. The validity coefficient for the letters of recommendation was .35, yet subjects estimated that it was .66, $t(32) = 4.47, p < .01$.

Subjects do not overestimate the predictability of the ability-based GPA outcome to anything like the same extent, though they do significantly overestimate the predictability of GPA from the interview, $t(32) = 4.08, p < .001$. Ironically, subjects tended to *underestimate* the predictability of GPA from the modestly valid predictor of high school GPA, $t(31) = 2.60, p < .05$.

Finally, subjects believed that Peace Corps performance is considerably more predictable from an interview than is GPA, $t(65) = 2.99, p <$



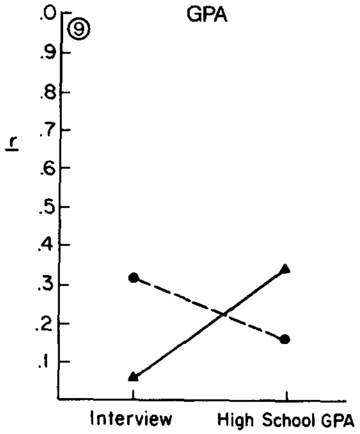


FIG. 8. Actual (—) and estimated (- - -) correlations for the prediction of Peace Corps performance from an interview and from an aggregate (letters of recommendation).

FIG. 9. Actual (—) and estimated (- - -) correlations for the prediction of college GPA from an interview and from an aggregate (high school GPA).

.01. Similarly, when predicting from aggregates of events, subjects believe Peace Corps performance to be far more predictable than GPA, $t(63) = 4.70, p < .001$ (though here it should be noted that we are comparing aggregated apples with aggregated oranges).

Thus it appears that subjects do not realize that chance plays at least as big a role in affecting a trait-based outcome as it does in affecting an ability-based outcome. They seem to believe that in a single interview one can figure out how people will behave in novel situations over a long period of time. The "interview illusion" exists for both the ability-based outcome and the trait-based outcome, but it seems to be more acute for the latter.

The data are also suggestive of a lack of appreciation of the aggregation principle in the domain of traits. Subjects do not expect the predictability of Peace Corps performance from an aggregated measure of acquaintance to be any greater than that obtained from a single exposure: They believe that one observer who has interacted with a person for a single hour can predict the person's behavior just as well as the aggregated assessment of several people who know the person well ($t < 1$). (This observation must be treated only as a tentative suggestion, however, because the units that are aggregated by acquaintances include information of a different type than is conveyed in an interview. And of course the study does not permit evaluation of people's appreciation of the aggregation principle for abilities because the aggregated measure, namely high school GPA, was composed of completely different types of information than the single-item measure, namely the interview.)

APPRECIATION OF THE AGGREGATION PRINCIPLE

Study 6: Cuing Recognition of the Aggregation Principle

The data so far suggest that neither laypeople nor psychologists have a sufficiently robust appreciation of the aggregation principle in the abstract to ensure that they will apply it to unfamiliar data or to data that are difficult to code. In Study 3, both expert and lay subjects failed to show any recognition of the aggregation principle when estimating correlations for relatively unfamiliar data concerning manuscripts and grant proposals. In Study 4 both groups failed to show any recognition of the principle when making estimates about the consistency of difficult-to-code, trait-related behavior. The data do suggest, however, that subjects *may* make use of the aggregation principle if the data are codable and are familiar at more than one level of aggregation, as they are for abilities (Study 4) and for course evaluations (Study 1). This would imply a highly domain-specific ability to apply the aggregation principle and would be consistent with the domain specificity found for various other versions of the law of large numbers (Fong et al., 1986; Jepson et al., 1983; Nisbett et al., 1983). It is also possible, though, that neither laypeople nor experts have any ability to apply the aggregation principle to the data of everyday life: Both groups seem to be accurate if the data are easily codable and have been observed at a given level of aggregation and neither is accurate if either of these requirements is violated. And psychologists show no *more* recognition of the principle when making judgments about abilities than do lay subjects. Such limited accuracy could be due entirely to the correct detection of, and memory for, the covariations at each level of aggregation.

A better way to test whether people appreciate statistical rules is to use studies employing within-subject designs (cf. Fischhoff, Slovic, & Lichtenstein, 1979). In Study 6, we examined whether subjects could be cued to recognize the aggregation principle by requiring them to answer for two levels of aggregation. In previous studies we required subjects to answer for only one level of aggregation. This may have made it harder for subjects to recognize the relevance of the principle even if they had some understanding of it in the abstract and some ability to apply it under optimal circumstances. We examined estimates of the consistency of trait- and ability-related behaviors.

The design was identical to that of Study 4 except that instead of having subjects make estimates for both traits and abilities either at the item-to-item or the total-to-total level of aggregation, they made estimates for only one of the traits or abilities at *both* the item-to-item and the total-to-total level. In addition, half of the 144 subjects were required to *justify* their answers: "If your answers to the above two questions were not identical, please indicate why."

The results were dramatic and clear-cut. The within design caused subjects to give estimates for both traits and abilities that were more in line with the aggregation principle. In both cases the interaction between level of aggregation and design was significant, $F(1,199) = 10.08$, $p < .001$ for traits, $F(1,191) = 4.53$, $p < .05$ for abilities.[4] In both cases the effect was produced entirely by lowering the estimate of the item-to-item level correlation—from .77 to .51 for traits, and from .51 to .28 for abilities.

Requiring subjects to justify different answers at the two levels had no effect over and above simply requiring them to answer for both levels, but the open-ended justification data make it very clear that subjects do indeed appreciate the aggregation principle when the problem context cues them to its relevance. Two examples of answers are given below. The first is a justification for giving a higher estimate for the total-to-total level for traits. The second is a justification for giving a higher estimate for the total-to-total level for abilities.

> It is very possible for one to misjudge a person in a given situation. However, after observing many more situations the average reaction to one's actions becomes more accurate.
>
> It can simply be a fluke (i.e., Johnny could have had a good day), while an average grade is reflective of his whole attitude and study capabilities in that course.

Altogether, 80% of the answers given by subjects in the no-justification condition and 78% of the answers given by subjects in the justification condition indicated that subjects believed that total-to-total correlations were higher than item-to-item correlations. Eighty-seven percent of the justifications for answers indicating that total-to-total correlations were higher gave as the reason at least some crude version of the aggregation principle as the justification for the answer.

Undoubtedly real problems sometimes do provide reminders of the sort given in our study, and so people probably are capable of using the aggregation principle occasionally even in domains that are unfamiliar or hard to code. But it would be a mistake to assume that correlations at a level different from the focal one are normally salient to people. The structure of most inference tasks is of the form "What prediction do I make given the evidence?" rather than of the form "What prediction do I make given the evidence and given what I would guess the evidence to

[4] For the purpose of this analysis the "within" data were treated as though they were "between." The resulting p levels thus are very conservative.

look like if there were more (or less) of it?" Life, as Nisbett and Ross (1980) put it, has a between design, and we all too rarely conduct thought experiments having a within design. In addition, though subjects in within conditions showed some appreciation of the aggregation principle, they did not show enough: Their judgments showed very insufficient influence of the principle for both traits and abilities, and they still markedly overestimated trait consistency at the level of the situation, $p < .001$.

Domain Specificity of Recognition of the Aggregation Principle

The aggregation principle appears to resemble other versions of the law of large numbers in that spontaneous appreciation of the principle is extremely domain and problem specific. Although people can be cued to apply the principle even to difficult domains, spontaneous recognition of the aggregation principle seems to occur only in domains where people are able to detect covariations at more than one level. Even when subjects accurately perceive item-to-item correlations to be very low, they do not use the aggregation principle to extrapolate to total-to-total correlations unless they have also had the opportunity to perceive that the total-to-total correlations are high. Accurate covariation detection in a given domain, at more than one level of aggregation, may be required if the rule is to be induced in that domain.

Thus, ironically, people are probably able to apply the aggregation principle best for domains for which they already have had substantial opportunity to observe covariation. They may not benefit much from its use in domains where it would be most beneficial—domains where they are familiar with covariation at only one level of aggregation, or at no level. As a consequence, people make very serious errors when assessing covariation. When they are familiar only with covariation at the item-to-item level, as in several of the domains we examined, they tend to grossly *underestimate* covariation at the total-to-total level. It seems likely that when they are familiar only with covariation at the total-to-total level, they would tend to grossly *overestimate* covariation at the item-to-item level. And when they have no familiarity with a domain, they seem sure to make at least one of these errors.

DISCUSSION

How accurate, then, is lay psychometrics? When people make estimates of the correlations that guide their predictions and choices in everyday life, how likely are they to be correct?

The correlations that we examined do not constitute a random sample of everyday life correlations, nor anything approximating such a sample. Nevertheless, the types of correlations are sufficiently representative and diverse, and the degree of accuracy found is sufficiently broad, that three important generalizations may be proffered.

1. Notwithstanding people's demonstrated difficulties in assessing covariation and their lack of abstract appreciation of the law of large numbers, they are capable of impressive accuracy when making estimates of some important kinds of everyday life correlations.
2. The accuracy that we found was limited to cases where several important factors were all favorable to correct estimation.
3. We found serious inaccuracy where these factors were less favorable. Such inaccuracy was found even where the events in question are both common and important and even when the judges were expert in both psychology and statistics. We shall now amplify each of these points.

Factors Influencing Accuracy about Predictions for Social Events

We have found that people can be remarkably accurate about correlations in the social world if each of three conditions obtain. Two of these have been discussed at length already. They are (a) familiarity with the data and (b) codability of the data.

A third factor that undoubtedly influences the accuracy of perceiving correlations was not salient to us before we began the research, but is clear in retrospect. This is whether or not the data to be correlated are drawn from *distributions of the same kind of events*. There is a two-decade-old literature showing that people can be reasonably accurate about covariation when estimating correlations among two sets of numbers or among two sets of readings for pointers on identical dials (e.g., Beach & Scopp, 1966; Erlick, 1966; Erlick & Mills, 1967; Wright, 1962). Nisbett and Ross (1980) were inclined to attribute accuracy in these cases to the impoverishment of the stimuli and a corresponding lack of a priori theories that might serve to bias judgments about covariation. But an interpretation in terms of common versus disparate event distributions seems more likely in view of present results. Most previous research examining people's perception of covariation in social domains has examined events drawn from qualitatively different distributions—for example, between "Draw-a-Person" test responses such as treatment of the eyes (normal eyes vs large, small or otherwise distorted eyes) and psychiatric diagnoses (of paranoia vs some other pathology). While it is true that such judgments are rife with opportunities for interference from prior theories that people hold, they also present *cross-category* coding problems of a kind that parallel columns of numbers, or concomitant dial readings, or most of the events we studied, do not.

Correlations among variables coming from distributions of the same type are much easier to assess because in this case each pair of observations in and of itself contains information, namely, the distance between the two observations, that can be used to assess the correlation. For example, when assessing the correlation between people's opinions about courses, we may ask each of two people who have attended a given course for their evaluations. A comparison of the distance between the evaluations provides us with a rough idea about the correlation. Compare this to cases where the two variables of interest come from different distributions. For example, imagine trying to estimate the correlation between evaluations of a course and performance in the course. We cannot, of course, directly compare a person's evaluation of the course and the person's performance in the course to each other—that would be like comparing apples and oranges. Instead, we need to locate the person's evaluation on the distribution of evaluations and do the mental equivalent of calculating the person's percentile score for evaluations, then locating the person's performance on the distribution of performance scores, and calculating the person's percentile score for performance. Only then can we compare the two percentile scores to each other to obtain a distance estimate. The process is more complicated, and requires knowledge about the two distributions, knowledge that is not necessary when assessing correlation between two identical variables.

Many important correlations in everyday life are characterized by all three of the factors that our research suggests are important, namely familiarity, codability, and common distribution of events. Many kinds of evaluations, in particular, would seem to meet all three of these criteria. In addition, many ability-related behaviors, at least if they are coded on a common distribution, would seem to meet our criteria.

Accuracy about such matters is almost surely of great utility to people. They are probably well prepared to take appropriate action on the basis of information about the evaluations of others concerning, for example, the personal attributes of other people and the desirability of college courses. Similarly, they can probably take effective advantage of information about the abilities of others in many athletic, academic, and professional domains.

Consequences of Inaccuracy about Correlations among Social Events

But the present results also suggest that the inferential failings that have been demonstrated by judgment researchers in laboratory settings are sometimes manifested in full force in judgments about everyday events. Even for domains where subjects, on average, show substantial accuracy, many individuals do not: Not everyone is accurate just because the mean is on target. Thus, for example, the mean estimate for the item-to-item correlation for the abilities we examined was .51, which was almost exactly correct. However, a third of the subjects made guesses about the correlation that were either over .75 or under .31. Similarly, the mean estimate for total-to-total correlations for course evaluations was .79, which again was almost exactly right. But a third of these subjects guessed the correlation to be either as high as .99 or lower than .45. Thus, even where the mean was very close to the actual value, many individual subjects were quite inaccurate, and their errors would be of the sort that could sometimes produce unhappy consequences in their lives.

Untoward consequences would seem to be the norm for decisions and behaviors based on judgments about covariation for which the *majority* of people are badly mistaken, as in the case of judgments about the stability of social behavior and judgments about the reliability at the individual level for judgments about documents such as manuscripts and grant proposals. At the very least, such errors mean that we will be constantly surprised at outcomes. We will be surprised when the woman who seemed so nice when the realtor introduced her turns out to be such an undesirable neighbor. We will be surprised when the man who made such a poor impression in his job interview turns out to be a rising star at the institution that (uproariously, we thought at the time) hired him. We will be astonished that two such eminent scientists could have such different views of the same manuscript. And we will be dubious when psychological research shows low cross-situational consistency for trait-related behaviors.

But of course our predictions often have consequences beyond mere surprise. Our predictions, and the choices they engender, often will *produce* outcomes that are undesirable and that could have been avoided, in principle and on the average. We do not hire the candidate who made a rather poor personal impression, even though the folder provided clear evidence of superiority. We turn to only one or two consultants for help in a decision when the outcome is of some real moment either to ourselves or to institutions that we value and when there is generally low agreement for the relevant judgments. We avoid contact with people who strike us as dull, silly, or obnoxious on a brief encounter, even though a fair fraction of such people would have been regarded as pleasant or even delightful on longer acquaintance.

Most of the above consequences are not new ones to social psychologists. They have for some time been asserted to be the consequence of the fundamental attribution error—known to Kurt Lewin, described by Fritz Heider, established empirically by Edward E. Jones, named by Lee Ross, and documented at length by Nisbett and Ross. We believe, however, that the present data provide the best evidence to date for the reality of the phenomenon. We have little doubt that our method of measuring people's beliefs about correlations maps well onto whatever representation people actually use for such judgments. It seems quite unlikely that the estimates in, for example, Fig. 1 presenting estimated correlations for course evaluations, or Fig. 2 presenting estimated correlations for evaluations of the attributes of people, are as close to the actual correlations as they are simply as a matter of chance. The accuracy that we found for many types of correlations, especially in ability domains, indicates that this representation may be adequate for some of the important purposes of daily life. The striking inaccuracy we found for beliefs about the consistency of social behavior thus seems all the more real and serious. If we take the data for lay subjects' estimates of the stability of social behavior at their face value, and we feel justified in doing so, they indicate that people are enormously more confident of the expected nature of a person's social behavior, given knowledge of the nature of their behavior on one occasion, than reality affords them any right to be. This is true both for predictions for a single occasion given observation of actual behavior in a situation tapping a particular trait (Fig. 6) and for predictions for complex behavior over a long period given observation in an interview (Fig. 8).

The implications of these results for the trait controversy should be spelled out explicitly. In our view, the debate has lasted as long as it has because psychologists' intuitions, like those of laypeople, tell them that there is very substantial predictability at the level of individual acts, as much predictability in fact as at highly aggregated levels. The error here is a very basic one, amounting not merely to an empirical mistake, but to literal incoherence. What both psychologists and laypeople do not realize is that their beliefs about predictability at the aggregate level actually preclude a belief in comparable predictability at the individual level. This is powerful testimony to the strength of the illusions underlying perception of personal consistency.

Statistical Expertise and Intuitive Psychometrics

Can anything be done about the fundamental attribution error and about related errors in perceiving covariation in the social domain? The present data have important implications for the possibility of improving lay psychometrics. It is clear that abstract training in statistical principles will not suffice to alleviate all of people's difficulties, or perhaps any of them. Even statistically knowledgeable people were unable to recognize the aggregation principle for data with which they were unfamiliar or for data that are hard to code, and their judgments were not more guided by the principle than those of laypeople (cf. Tversky & Kahneman, 1971). On the other hand, even laypeople were able to recognize the aggregation principle for the highly problematic trait domain when strongly cued to do so in a within-subject design. This suggests that the key to improving lay psychometrics lies not so much in teaching people abstract principles as in teaching people to map the elements of unfamiliar domains onto such principles. This may be done even for domains that are difficult to unitize if people are prompted to assess the *relative* numbers of units in different types of information. For example, even if one does not know exactly what units to use to measure friendliness, it is useful to realize that, whatever the unit, there will be far more units in a yearlong acquaintance than in an hourlong interview. Our within-subject study suggests that the simple thought experiment of considering the predictability of behavior over the long haul when making predictions about behavior over the "short haul" is sufficient to drive down people's estimates of the predictability of one behavior from one other behavior, and thus to improve accuracy.

The value of teaching people to map events onto statistical principles has been demonstrated by Fong et al. (1986). They studied several principles derivable from the law of large numbers and showed that abstract training in the law and training in mapping everyday events onto the law each contributed independently to improving statistical reasoning.

Thus it would be premature to be pessimistic about the possibility that training might improve people's ability to recognize the applicability of the aggregation principle across a wide domain of events and problem types. We already know that even modest amounts of statistical training can have a big impact on some types of judgment, and we have little basis for predicting how much more improvement is feasible.

REFERENCES

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, 91, 112–149.

Beach, L. R., & Scopp, T. S. (1966). Inferences about correlations. *Psychonomic Science*, 6, 253–254.

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.

Block, J. (1977). Advancing the psychology of personality: Paradigmatic shift or improving the quality of research. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Erlbaum.

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.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, 6, 151–155.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous diagnostic observations. *Journal of Abnormal Psychology*, 72, 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271–280.

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science (Washington, D.C.)*, 214, 881–886.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin,* **90,** 279–292.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology,* **37,** 1097–1126.

Epstein, S. (1983). Aggregation and beyond: Some basic issues in the prediction of behavior. *Journal of Personality,* **51,** 360–391.

Epstein, S. (in press). The stability of behavior across time and situations. In A. I. Rabin, J. Aronoff, A. M. Barclay, & R. Zucker (Eds.), *Further explorations in personality* (Vol. 2). New York: Wiley.

Erlick, D. E. (1966). Human estimates of statistical relatedness. *Psychonomic Science,* **5,** 365–366.

Erlick, D. E., & Mills, R. G. (1967). Perceptual quantification of conditional dependency. *Journal of Experimental Psychology,* **73,** 9–14.

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,* in press.

Golding, S. L., & Rorer, L. G. (1972). Illusory correlation and subjective judgment. *Journal of Abnormal Psychology,* **80,** 249–260.

Hamilton, D. L. (1979). A cognitive attributional analysis of stereotyping. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12). New York: Academic Press.

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

Hogarth, R. M. (1980). *Judgment and choice.* New York: Wiley.

Holyoak, K. J., & Gordon, P. C. (1983). Social reference points. *Journal of Personality and Social Psychology,* **44,** 881–887.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin,* **96,** 72–98.

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge Univ. Press.

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

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones et al. (Eds.), *Attribution: Perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

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

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

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

Klitgaard, R. (1985). *Choosing elites.* New York: Basic Books.

Mayfield, E. C. (1964). The selection interview: A re-evaluation of published research. *Personnel Psychology,* 17, 239–260.

Miller, G. A., & Cantor, N. (1982). Book review of R. Nisbett & L. Ross, *Human inference: Strategies and shortcomings of social judgment. Social Cognition,* **1,** 83–93.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

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

Moskowitz, D. S., & Schwartz, J. C. (1982). Validity comparison of behavior counts and ratings by knowledgeable informants. *Journal of Personality and Social Psychology,* **42,** 518–528.

Newcomb, T. M. (1929). *Consistency of certain extrovert–introvert behavior patterns in 51 problem boys.* New York: Columbia University, Teachers College, Bureau of Publications.

Nisbett, R. E. (1980). The trait construct in lay and professional psychology. In L. Festinger (Ed.), *Retrospections on social psychology.* New York: Oxford Univ. Press.

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.

Olweus, D. (1977). A critical analysis of the "modern" interactionist position. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology.* Hillsdale, NJ: Erlbaum.

Peterson, D. R. (1968). *The clinical study of social behavior.* New York: Appleton-Century-Crofts.

Ross, L. (1977). The intuitive psychologist and his shortcomings. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10). New York: Academic Press.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin,* **94,** 18–38.

Stein, M. I. (1966). *Volunteers for peace.* New York: Wiley.

Swann, W. B., Jr. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review,* **91,** 457–477.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin,* **76,** 105–110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (Washington, D.C.),* **185,** 1124–1131.

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

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin,* **63,** 100–116.

Wright, J. C. (1962). Consistency and complexity of response sequences as a function of schedules of noncontingent reward. *Journal of Experimental Psychology,* **63,** 601–609.

(Accepted December 6, 1985)