

Assessing Dependence: Some Experimental Results

Robert T. Clemen • Gregory W. Fischer • Robert L. Winkler
Fuqua School of Business, Duke University, Durham, North Carolina 27708-0120
clemen@mail.duke.edu • fischer@mail.duke.edu • rwinkler@mail.duke.edu

Constructing decision- and risk-analysis probability models often requires measures of dependence among variables. Although data are sometimes available to estimate such measures, in many applications they must be obtained by means of subjective judgment by experts. We discuss two experimental studies that compare the accuracy of six different methods for assessing dependence. Our results lead to several conclusions: First, simply asking experts to report a correlation is a reasonable approach. Direct estimation is more accurate than the other methods studied, is not prone to mathematically inconsistent responses (as are some other measures), and is judged to be less difficult than alternate methods. In addition, directly assessed correlations showed less variability than the correlations derived from other assessment methods. Our results also show that experience with the variables can improve performance somewhat, as can training in a given assessment method. Finally, if a judge uses several different assessment methods, an average of the resulting estimates can also lead to better performance.

(Correlation; Concordance Probability; Covariation Judgment; Dependence; Spearman's Rho; Subjective Assessment)

1. Introduction

Expert judgments are key inputs in important decision-making problems, and the elicitation and modeling of such judgments are formalized in methodologies such as decision analysis, risk analysis, and expert systems. As problems grow in importance and complexity (e.g., decision analysis of a major strategic move for a firm, risk analysis associated with a proposed high-level nuclear waste repository, an expert system to forecast future oil prices), with many variables of interest, assessing and modeling expert knowledge about relationships become essential. The process of building a network representation, such as an influence diagram, belief network, or Bayes net, to represent the overall structure of the uncertainty in a problem is quite well understood, as is the assessment of probability distributions for individual variables in the network. Assessing dependence among the vari-

ables has received less attention despite the fact that the ultimate probability distributions of interest may be quite sensitive to the level of dependence. When, as is often the case, "hard data" concerning relationships are limited or nonexistent, expert judgments become particularly crucial. For example, consider a product under development involving a technology that is so new that no directly relevant past data are available. The degree of dependence between variables such as time to market and availability of competing products, or advertising budget and sales, requires judgmental assessment and can have great impact on development and marketing decisions.

Typically, conditional distributions have been used to represent an expert's knowledge about a set of interrelated random variables. However, the number of required probability assessments can grow exponentially as variables are added. Also, thinking about

probabilistic relationships in terms of conditional distributions can be difficult, although not without its rewards in terms of a more complete understanding of those relationships and potentially more precise assessments overall (Ravinder et al. 1988). Practical difficulties of dealing with dependence have led some to search for rules of thumb for ignoring dependence (e.g., Smith et al. 1992). When dependence among important variables is weak, such strategies may be reasonable, but often dependence can have a strong impact as, for example, in information-aggregation problems (Clemen and Winkler 1985).

Using the divide-and-conquer strategy common in decision analysis, we prefer to separate judgments about individual variables from judgments about relationships among variables. This is a feasible modeling approach through the use of copulas (Jouini and Clemen 1996, Frees and Valdez 1998, Clemen and Reilly 1999). We can represent a joint distribution by specifying marginal distributions for the individual variables and a copula (a dependence function) that joins the variables, thereby avoiding some of the practical problems associated with the conditional-distribution approach. The assessment of marginal distributions can be accomplished readily using standard probability-assessment techniques. By considering families of copulas, we can specify the relationships among variables in terms of specific measures of dependence, such as correlations. The assessment of dependence can then be accomplished via such measures.

To our knowledge, very little prescriptive or descriptive work has been done on the assessment of dependence measures for modeling expert knowledge. Gokhale and Press (1982) studied the assessment of correlation coefficients in bivariate normal distributions. A good deal of descriptive behavioral work has been done under the rubric of covariation assessment, including illusory correlation and causation. Much of this work, beginning with Smedslund (1963), deals with relationships in a two-by-two table and looks at various heuristics that may lead people astray. These heuristics (e.g., considering only the upper-left-hand cell of the table) typically involve the failure to use all of the relevant information that is available about the

relationship (Shaklee and Mims 1981). Beyth-Marom (1982) looked more deeply into possible factors influencing the way people think about relationships: the way data are presented, the instructions subjects receive, and the types of variables being considered. For general discussions of issues involved in covariation assessment, see Nisbett and Ross (1980) and Yates (1990).

Jennings et al. (1982) dealt with continuous variables and distinguished between data-based and theory-based assessment. When making a judgment after observing data, people underestimated the strength of weak relationships and were better with stronger relationships. In a theory-based setting, with no data immediately available, this systematic bias was not observed, although the strength of correlations was overestimated in the direction of the judge's prior theories. Pechmann and Ratneshwar (1992) found that prior beliefs led to bias when they were inconsistent with the actual relationship and information was ambiguous, but had no adverse (or positive) effect when information was clear and the relationship was strong. Billman et al. (1992) showed subjects samples of data and then asked them to estimate the correlations present. They also manipulated prior beliefs about the presence and direction of the relationship between the variables. They found that theory-based estimates of correlation were superior only when the theory was correct regarding the direction of the relationship. In real-world decision-making problems, we are likely to consult experts who know relevant theories and also have a substantial experiential database.

Perhaps the most similar study from the psychology literature is Kunda and Nisbett (1986). They examined subjects' ability to make accurate judgments of dependence in a variety of situations, including course evaluations, personality traits, evaluation of scientific documents, basketball scoring ability, student GPA, and job performance. Their findings suggest that, on average, subjects can make accurate judgments in situations in which they are familiar with the domain and where the data are "codable" (i.e., a readily interpretable scale can be defined). Kunda and Nisbett's subjects used only one assessment method,

probability of concordance, which we discuss in detail in the next section. Although they do not report accuracy statistics, they do note that individual responses showed considerable variation.

In this article, we report the results of two studies designed to investigate some methods for assessing dependence. Our orientation is prescriptive in that we are interested in developing and studying procedures that lead to better decisions by improving how we elicit and model expert knowledge regarding dependence. Our intent is not to propose and test behavioral theories. On the other hand, we ultimately want methods that have a sound probabilistic foundation for modeling and also take into account current knowledge in behavioral decision theory, as well as current practice in probability elicitation. For example, training in probability assessment can include a review of some commonly encountered cognitive biases and simplifying heuristics with suggestions on how to mitigate their effects, and assessment procedures can be designed with these issues in mind (e.g., Winkler et al. 1995).

The article is structured as follows. The assessment methods are discussed in §2. Sections 3 and 4 describe the two experimental studies, respectively, including the design of the experiments, the analysis and results, and some interpretation and discussion. Section 5 summarizes the results and their implications for dependence assessment in practice.

2. Dependence-Assessment Methods

What are desirable characteristics for a dependence-assessment method? For modeling purposes, the methods should have rigorous foundations that are defensible in terms of probability theory, they should be as general as possible to permit modeling of dependence in a wide variety of situations, and they should be able to be linked directly to the modeling procedure. For ease of assessment, they should have a clear intuitive interpretation, and assessors should view them as easy and credible. In terms of the resulting assessments, we would like to see coherent assessments, a reasonable amount of agreement among the assessments of individuals with similar knowledge, and accuracy (i.e., assessments that corre-

spond well with actual levels of dependence in the real world). For any given method, we can examine the underlying foundations and connections with modeling. We can also consider interpretations and ask assessors if they find the method easy to use. Assessments can be constrained to be coherent. The degree of agreement among assessors can be empirically evaluated. Finally, where data are available (as they are for the pairs of variables considered in our experiments), we can use those data to calculate accuracy statistics for the various assessment methods.

We used six different dependence assessment methods in our experiments: *S* (strength of relationship), *R* (correlation), *CF* (conditional fractile), *CNC* (concordance probability), *JP* (joint probability), and *CP* (conditional probability). These methods represent six different ways for an individual to think about dependence. For each method, we provided some guidance as to how the subject might think about the question. In the remainder of this section we present brief discussions of the six assessment methods, giving an example of the wording used for each in the context of the relationship between height and weight for male MBA students at Duke's Fuqua School of Business.

S. "How would you characterize the strength and nature of the relationship between height and weight for male MBA students?" Of the six methods, this is the most informal and the one with the least rigorous foundation in terms of probability theory. In Study 1, the response was given on a continuous line scale for which 1 = "very strong negative relationship," 4 = "no relationship," and 7 = "very strong positive relationship." In Study 2, the response was given in two parts: an indication of whether the relationship was positive or negative, and a mark on a continuous line scale with 1 = "no relationship," 4 = "moderate relationship," and 7 = "very strong relationship."

R. "What would you estimate for the correlation between height and weight for male MBA students?" This is a direct estimate of a correlation and requires that the subject understand the notion of correlation. As usual, a correlation near 0.00 implies a weak relationship and a correlation near +1.00 (−1.00) indicates a strong positive (negative) relationship. In

Study 1 the response was a mark on a continuous line scale from -1.00 to $+1.00$, whereas in Study 2 we asked the subjects to write down a correlation (same response mode as *CF*, *CNC*, *JP*, and *CP*).

CF. "A randomly chosen male MBA student is 74 inches tall, which is the 90th percentile of the height distribution. (This means that 90% of individuals are less than or equal to 74 inches tall.) For the same student, what percentile would you estimate for his weight?" The response is a number between 0 and 100. For consistency, the estimate must fall between the 10th percentile and the 90th percentile. An estimate near the 50th percentile implies little or no relationship between height and weight, an estimate near the 90th percentile means that the relationship is close to a perfect positive relationship, and an estimate near the 10th percentile means that the relationship is close to a perfect negative relationship. Asking the subject to estimate a percentile instead of the MBA student's weight means that judgments about the marginal distribution of weight should not, in principle, influence the answer, thereby maintaining the separation of judgments regarding marginal probabilities from judgments regarding dependence.

CNC. "Suppose we randomly choose two male MBA students and label them A and B. Given that A is taller than B, what is your probability that A also weighs more than B?" This asks directly for a concordance probability. (This is true as long as the variables in question—in this case height and weight—are continuous. For a rigorous definition and discussion, see Gokhale and Press 1982.) A probability close to 0.5 indicates little or no relationship, whereas a probability close to 1.00 (0.00) suggests a very strong positive (negative) relationship. The question requires the consideration of two draws from the joint distribution, which works fine for situations such as height/weight but may be difficult for unique events without a population from which to draw. Clemen and Reilly (1999) give an example of the relationship between the size of the population of an endangered species of tree frogs and a temperature index for the frogs' environment.

JP. "A male MBA student has been randomly chosen. What is your probability that this student's

height and weight are both in the lower 30% of their respective distributions (height less than or equal to 69 inches AND weight less than or equal to 160 pounds)? To put it more formally: What is $P(\text{height} \leq 69 \text{ and weight} \leq 160)$?" This requires the direct assessment of a joint probability, which experts may find relatively difficult. A probability near 0.09 would imply that the variables are not related, a probability near 0.30 indicates a very strong positive relationship, and a probability near 0.00 indicates a very strong negative relationship.

CP. "A randomly chosen male MBA student is in the lower 60% of the height distribution (height ≤ 71 inches). Given this, what is your probability that this student's weight falls in the lower 60% of the weight distribution (weight ≤ 175 pounds)?" A probability of 0.60 would indicate that height and weight are independent, whereas a response above (below) 0.60 implies a positive (negative) relationship. A potentially difficult aspect of this question is conditioning not on a specific fractile or height but on an interval of fractiles (the student's height fractile is between 0 and 60) or heights (below 71 inches). Thus, although the question has a solid foundation in probability theory, thinking about it may not be easy.

To summarize key characteristics of the assessment methods, *S* is simple but has no rigorous underpinnings in probability theory. *R* asks for correlations, a rigorous probability concept. However, decision analysts generally eschew assessments of moments, let alone cross moments (Morgan and Henrion 1990), focusing instead on the assessment of probabilities. An advantage of *R* with scientific experts is that many scientists are trained in basic statistics and hence are familiar with the concept and use of correlation. *CF* asks the assessor to provide an estimate of a fractile. Fractiles are a rigorous concept from probability, and the assessment of fractiles is common in probability assessment. However, the elicitation of conditional fractiles, given a specified condition, is not very common. *CNC*, *JP*, and *CP* all ask the assessor for probabilities and therefore are consistent with conventional decision-analysis practice. As mentioned above, however, the nature of the events and the elicitation questions may be difficult to understand. Additional

thoughts regarding the assessment methods and possible variations are given in §5.

3. Study 1

Method. In the first study, we asked 90 students in Fuqua's Weekend Executive MBA program to respond to a questionnaire. All were enrolled in the second quantitative course in the MBA curriculum, and the exercise was done at a time when the curriculum focused on Monte Carlo simulation and the problem of modeling correlated variables. All students had taken a statistics course during the previous term, during which they studied probability, regression, and correlation. The average age of the subjects was 36.5 years, 24% were female, and 76% were male.

The design is a within-subjects design. The questionnaire included the assessment questions for the six dependence measures described for five pairs of variables about which we believed the respondents would be reasonably knowledgeable. In addition, we obtained data for these variables so that we could estimate the correlation and compare the students' assessments with the estimates in order to measure the accuracy of the assessments. (Thus, our variables satisfy Kunda and Nisbett's (1986) conditions of familiarity and codability.) The five pairs of variables and the estimated correlations are:

1) *Height and Weight (HW)*. This represents the height and weight of male MBA students enrolled in Fuqua's daytime MBA program. Based on a sample of 218 individuals, the estimated Spearman correlation is 0.530.

2) *Math and Verbal SAT Scores (MV)*. These are scores on the two portions of the Scholastic Aptitude Test (SAT) for Duke undergraduates. Based on data for 46,278 Duke undergraduates who matriculated between 1963 and 1998, the estimated Spearman correlation is 0.377.

3) *Standard and Poor's 500 and Dow Jones Industrial Average (SD)*. This represents the monthly returns for these two indexes of overall stock market performance. Based on monthly data from February 1945 to January 1995, the estimated Spearman correlation is 0.950.

4) *Automotive Index and Chrysler Corporation (AC)*. This represents the monthly returns for Chrysler Corporation common stock and an index of stocks (defined by Compustat) from the automotive industry. Based on data from July 1976 through June 1996, the estimated Spearman correlation is 0.738.

5) *Eli Lilly and Chrysler Corporation (LC)*. These are monthly returns for Eli Lilly (a pharmaceutical firm) and Chrysler Corporation common stock. Based on data from July 1976 through June 1996, the estimated Spearman correlation is 0.173.

For each pair of variables, we provided histograms of the marginal distributions as well as estimated means, standard deviations, and deciles of the distributions. No scatterplot or other indication of the joint distribution was provided, however, because we were interested in the students' ability to assess the level of dependence between the variables based on their own subjective knowledge.

With each assessment question, we included a brief statement describing "how to think about" the assessment, including an indication of upper and lower bounds. This information was included on the basis of a previous pilot study in which such information was shown to improve accuracy and reduce the number of responses that fell outside mathematically feasible bounds.

As an example, in the case of *HW*, the assessment question for the *CF* measure is:

A randomly chosen male MBA student is 74 inches tall, which is the 90th percentile of the height distribution. (This means that 90% of individuals are less than or equal to 74 inches tall.) For the same student, what percentile would you estimate for his weight?

(How to think about this: An estimate near the 50th percentile means you think the two variables have no relationship—even though you know what the height is, your best guess for the weight is still the median value. An estimate above the 50th percentile means you think they have a positive relationship, and the greater the estimate, the stronger the relationship. Likewise, an estimate of less than the 50th percentile indicates a negative relationship. For consistency, the estimate must fall between the 10th percentile (perfect negative relationship) and 90th percentile (perfect positive relationship) of the weight distribution, or between 142 and 200 lbs.)

Your assessment: _____

The structure of each questionnaire was to group the assessment questions for each pair of variables. Thus, for example, in one condition a respondent would answer all assessment questions first for *HW*, then for *MV*, and so on. The ordering of the variable pairs was partially counterbalanced: The demographic variables (*HW* and *MV*) always appeared first, followed by the financial variables (*SD*, *AC*, *LC*), with randomized ordering within each of these groups. The ordering of the assessment questions also was partially counterbalanced. First, for a given respondent, the ordering of the assessment questions was the same for all variable pairs. Strength of relationship (*S*) was always asked first as a “conditioning” question to help respondents begin to think about the nature of the relationship between the variables. The ordering of the remaining assessment questions was varied using a Latin Square design. Complete details of the design and samples of the questionnaires are available from the authors.

The entire task took 30 minutes to complete, and no compensation was given. In a follow-up session, students received feedback regarding how their responses compared to the data-based correlation estimates and to the distribution of responses from all subjects.

Nine of the 90 students indicated that they had some experience in the financial industry, ranging from 3 to 10 years, with an average of 6.2 years. Responses from these 9 subjects regarding the financial variables have been eliminated from the following analysis because results from Study 2 indicate that experience does make a difference in terms of assessment accuracy.

Results. To analyze the dependence assessments, we first converted the responses to equivalent Spearman correlations. The responses were converted as follows:

S: We transformed *S* linearly. Thus, the equivalent correlation was calculated as $r_s = (S - 4)/3$.

R: The response was taken to be the Spearman correlation.

CF: Given random variables *X* and *Y* with distribution functions $F(x)$ and $G(y)$, the standard nonparametric regression representation is

$$E[F(X)|y] = r_{CF}[G(y) - 0.5] + 0.5.$$

The assessment question specifies $G(y)$, and we take the response to be $E[F(X)|y]$. Thus, we are able to solve for r_{CF} .

CNC: Given the probability of concordance P_C , we begin by calculating Kendall's $\tau = 2P_C - 1$. For purposes of the analysis, we assume that the bivariate normal model is a close representation of the participants' beliefs about the joint distribution. The data-based joint distributions for our variable pairs are well approximated by the bivariate normal model, and we believe that the normality assumption is reasonable and robust. Thus, we are able to use Kruskal's (1958) relationships between the Pearson correlation (r^*) and Spearman correlation (r): $r^* = 2\sin(\pi r/6)$, and between Pearson's correlation and Kendall's τ : $r^* = \sin(\pi\tau/2)$. Using these equations, we convert τ to the equivalent Spearman correlation and label it as r_C .

JP and *CP*: In both cases we explicitly assumed that the assessed probability arose from a bivariate normal distribution, and we calculated r_{JP} and r_{CP} based on that assumption. Calculations were done using Mathematica and Microsoft Excel.

In all cases, if the assessed value fell outside mathematically feasible bounds, we coded the equivalent correlation as a missing value. Across all variable pairs, such responses occurred 5 times for *CF*, 24 times for *JP*, 7 times for *CP*, and 0 for *R*, *CNC*, and *S*.

Table 1 shows summary statistics for the six assessment methods and five variable pairs after converting responses to equivalent Spearman correlations. Standard deviations in this and all subsequent tables are shown in italics. The first thing to notice in Table 1 is that, in a general sense, the problem of assessing correlations is meaningful to the subjects; higher average assessments are indeed associated with higher levels of actual correlation, and vice versa, for every assessment method. That fact notwithstanding, we also see that some of the average assessments substantially over- or underestimate the correlation (as compared to the data-based estimate). By and large, low correlations tend to be overestimated and high correlations tend to be underestimated, regardless of the assessment method used.

The standard deviations in Table 1 indicate the

Table 1 Equivalent Correlations for Study 1: Averages, Standard Deviations (in italics), and Sample Sizes

	<i>HW</i> Height & Weight	<i>MV</i> Math & Verbal SAT	<i>SD</i> Std Poor's & Dow Jones	<i>AC</i> Auto Index & Chrysler	<i>LC</i> Eli Lilly & Chrysler
<i>S</i> Strength	0.535 <i>0.276</i> 89	0.453 <i>0.300</i> 90	0.776 <i>0.212</i> 81	0.550 <i>0.245</i> 80	0.209 <i>0.266</i> 80
<i>R</i> Correlation	0.580 <i>0.246</i> 89	0.501 <i>0.259</i> 88	0.758 <i>0.174</i> 80	0.605 <i>0.229</i> 80	0.331 <i>0.273</i> 81
<i>CF</i> Conditional Fractile	0.661 <i>0.276</i> 88	0.516 <i>0.364</i> 89	0.818 <i>0.236</i> 80	0.675 <i>0.326</i> 78	0.292 <i>0.312</i> 80
<i>CNC</i> Concordance Probability	0.484 <i>0.330</i> 90	0.327 <i>0.375</i> 90	0.776 <i>0.261</i> 81	0.580 <i>0.334</i> 80	0.209 <i>0.259</i> 81
<i>JP</i> Joint Probability	0.570 <i>0.327</i> 81	0.411 <i>0.382</i> 80	0.731 <i>0.311</i> 76	0.623 <i>0.334</i> 77	0.190 <i>0.364</i> 75
<i>CP</i> Conditional Probability	0.480 <i>0.342</i> 90	0.315 <i>0.367</i> 88	0.746 <i>0.271</i> 81	0.522 <i>0.280</i> 79	0.184 <i>0.285</i> 78
Data-based Estimate	0.530	0.377	0.950	0.738	0.173
Average Error	0.021	0.044	-0.182	-0.146	0.064

Note. Average errors are calculated as the average of assessment minus data-based estimate. The table also shows the full names of assessment methods and variable pairs for easy reference.

extent to which the subjects agreed about the correlation in question. The first thing to note is that these standard deviations indicate a substantial amount of variability in the assessments, especially considering that almost all of the equivalent correlations are positive. Regardless, two patterns can be detected; *R* has the lowest standard deviation for all variable pairs except *LC*, and *S* consistently has the second lowest standard deviation.

Because we are interested in assessment accuracy, for each subject and each assessment we calculated the absolute error as the absolute difference between the equivalent assessed correlation and the data-based estimate. Table 2 presents average absolute errors, from which several observations can be made. First, *R* appears to be the best of the assessment methods, followed by *S*, *CNC*, *CP*, *CF*, and finally *JP*. Given the recent exposure of the students to the concept of correlation, the performance of *R* may not be surpris-

ing. Another observation, however, is that the performance of all six methods is similar; taking a weighted (by sample size) average over all of the variables, average absolute errors range from 0.213 for *R* to 0.271 for *JP*. Examining the results across the variable pairs, we see that *R* is not the best for each variable, nor is *JP* the worst.

Table 2 also includes the results of a Bayesian analysis in which we calculated the posterior probability that the expected absolute error for each method (over all variable pairs) is greater than the expected absolute error for *R*. We performed a Bayesian analysis of a standard linear model:

$$AE_i = \beta_0 + \beta_S X_S + \beta_{CF} X_{CF} + \beta_{CNC} X_{CNC} + \beta_{JP} X_{JP} + \beta_{CP} X_{CP} + \varepsilon_i$$

where AE_i is the absolute error for the i th assessment, $X_m = 1$ if the assessment used method m and 0

Table 2 Absolute Errors for Study 1: Averages and Standard Deviations

	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall	Posterior Probability
<i>S</i>	0.223 <i>0.161</i>	0.236 <i>0.199</i>	0.210 <i>0.176</i>	0.239 <i>0.195</i>	0.205 <i>0.172</i>	0.223 <i>0.181</i>	0.751
<i>R</i>	0.190 <i>0.164</i>	0.243 <i>0.151</i>	0.207 <i>0.156</i>	0.176 <i>0.196</i>	0.248 <i>0.195</i>	0.213 <i>0.174</i>	–
<i>CF</i>	0.267 <i>0.147</i>	0.334 <i>0.198</i>	0.178 <i>0.202</i>	0.221 <i>0.247</i>	0.254 <i>0.216</i>	0.253 <i>0.209</i>	0.997
<i>CNC</i>	0.253 <i>0.215</i>	0.295 <i>0.235</i>	0.186 <i>0.252</i>	0.265 <i>0.257</i>	0.196 <i>0.171</i>	0.240 <i>0.231</i>	0.970
<i>JP</i>	0.288 <i>0.158</i>	0.323 <i>0.204</i>	0.235 <i>0.300</i>	0.236 <i>0.262</i>	0.270 <i>0.243</i>	0.271 <i>0.238</i>	1.000
<i>CP</i>	0.248 <i>0.240</i>	0.255 <i>0.269</i>	0.214 <i>0.263</i>	0.281 <i>0.214</i>	0.209 <i>0.192</i>	0.242 <i>0.239</i>	0.976

Note. The posterior probability is the probability that the corresponding assessment method is less accurate (greater expected absolute error) than *R*.

otherwise, and the error terms ε_i are modeled as independent normal errors with mean 0 and variance σ^2 . Our prior distribution is the noninformative Jeffreys prior $f(\beta_0, \dots, \beta_{CP}, \sigma) \propto \sigma^{-1}$. With this specification, *R* is the baseline assessment method and β_m is interpreted as the expected change (difference) in the absolute error when the assessment method changes from *R* to *m*. The posterior probabilities shown in Table 2 are the probabilities that these differences are positive, or that the corresponding method is less accurate (in terms of expected absolute error) than *R*. The lowest posterior probability is 0.751 for *S*, and all of the other methods have posterior probabilities greater than 0.970. These calculations strengthen our conclusion that *R* is the most accurate, especially relative to *CF*, *CNC*, *JP*, and *CP*.

Table 2 also displays standard deviations for the absolute errors. In a sense, these statistics indicate the level of consensus among the subjects. To the extent that consensus among assessors is preferred, smaller standard deviations are better. *R* has the smallest overall standard deviation (0.174), but this is not uniformly true across the variable pairs.

Although details are not reported here, the rank ordering of the assessment methods is similar when average squared errors are calculated and when average absolute errors are based on equivalent concordance probabilities instead of equivalent Spearman

correlations. Moreover, our results are generally consistent with both the level of accuracy and rank ordering of methods in a related study by Clemen and Reilly (1999).

Because we have multiple assessments from each individual on each pair of variables, a reasonable question to ask is whether it would be useful to combine the equivalent correlations from the various assessment methods. Table 3 reports the results of calculating the average of the equivalent Spearman correlations for each individual and for each variable pair. (Averages were calculated in all cases, even if not all of the equivalent correlations were available.) Comparing with Table 2, we see that for *MV* and *LC*, the average performs better than any of the individual assessment methods. For *HW* and *AC*, only one of the assessment methods (*R*) is better than the average. For *SD*, both *CF* and *CNC* are better than the average. Over all variable pairs, the performance of the average improves over that of *R* by 0.015. Considering the individual variable pairs, it is interesting to note the consistency of the average absolute error for *HW*, *SD*, *AC*, and *LC*, all of which fall within 0.013 of each other. *MV* is 0.027 above the worst of the other four.

The standard deviations in Table 3 are also noteworthy. They all drop substantially from the overall standard deviations in Table 2. In fact, the worst standard deviation (*SD*, 0.176) is only 0.002 larger

Table 3 Absolute Errors for Average of Six Assessments: Averages, Standard Deviations, and Sample Sizes

	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall
Average	0.193	0.224	0.189	0.197	0.184	0.198
Standard Deviation	0.138	0.156	0.176	0.173	0.144	0.158
<i>n</i>	90	90	81	81	81	423

Note. All assessments were converted to equivalent Spearman correlations prior to combining.

than the smallest overall standard deviation (*R*) in Table 2, and the overall standard deviation from Table 3 is less than that for *R* in Table 2 by 0.016.

Finally, participants were asked to rate the difficulty of the assessment methods on a 1–7 scale, where 1 was *very easy*, and 7 was *very difficult*. *S* was perceived as the easiest, with an average difficulty rating of 2.84, followed by *R* (3.13), *CNC* (3.34), *CF* (3.55), *CP* (3.78), and *JP* (4.16). Thus, *S* and *R* not only perform well in terms of accuracy, but they also are viewed on average as the easiest of the six techniques.

4. Study 2

Method. In the second study, we asked 289 students in Fuqua’s daytime MBA program to respond to a questionnaire. The average age was 27.5 years, 30% were female, and 70% were male. Of these students, 45 indicated that they had some experience in the financial industry, ranging from 0.5 to 6 years, with an average of 2.8 years. The timing corresponded to that of Study 1: The students were taking the second quantitative course and in particular were studying the problem of using correlated variables in Monte Carlo simulation. In the previous course on statistics, the students had been exposed to concepts of probability, regression, and correlation.

In this study, the primary objective was to determine whether some training in the assessment methods would improve performance. We created a between-subjects design, using the same pairs of variables and assessment methods as in Study 1. The questions were essentially the same, with slight changes in response modes for *S* and *R* as noted in §2.

In Study 2, subjects first answered the strength-of-relationship question for all five variable pairs, fol-

lowed by the assessment questions for all variable pairs using one particular method (the “untrained” method). Following this, the experimenter led a brief (five-minute) discussion about another of the assessment methods (the “trained” method). The training consisted of a lecture covering a sample question, advice on how to think about the method, an explanation of the response scale and how to interpret different values on this scale, and responses to any questions posed by the subjects. Finally, subjects answered the assessment questions for all five variable pairs using the trained method.

Because this study was performed in the context of a course, we were limited by the structure of the course. There were five different sections, each containing about 60 students. Each section was trained in a different randomly chosen assessment method. Thus, all students in a given section responded to the same questions for the trained method. Within that section, however, students were randomly assigned to one of the remaining four untrained methods. Also, within each section the order of the variable pairs was partially counterbalanced in the same way as Study 1: The demographic variables (*HW* and *MV*) always appeared first, followed by the financial variables (*SD*, *AC*, *LC*), with randomized ordering within these groups. The ordering of the pairs was the same for the strength-of-relationship, untrained, and trained methods.

The task took 30 minutes to complete; no compensation was given. The questionnaire was administered to the five different sections over two days, at the same point in the course curriculum in all cases.

Results. As in Study 1, we first converted the responses to equivalent Spearman correlations using

the same procedures (except for *S*, for which the linear transformation reflected the change in the question). For cases where the response was outside the mathematically feasible bounds, we coded the equivalent correlation as a missing value. For the trained assessments, this occurred 5 times for *CF*, 33 times for *JP*, 9 times for *CP*, and 0 for the other methods. For the untrained assessments, mathematically infeasible responses occurred 10 times for *JP* and 18 times for *CP*.

Table 4 presents summary statistics for the equivalent correlations in Study 2. Many of the patterns in Table 1 are repeated here. In general, higher average assessments are associated with higher actual correlations. *SD* and *AC* tend to be substantially underestimated (although the effect is somewhat reduced for the untrained assessments of *AC*). *R* has the lowest standard deviation across all variable pairs for the trained assessments, and for the untrained assessments *R* is first or second (to *S*) across all variable pairs except *HW*.

For each subject and each assessment we calculated the absolute error as the absolute difference between the equivalent assessed correlation and the data-based estimate. Because we have a relatively large sample, including 45 individuals with experience in the financial industry, we begin the analysis of these data by asking whether expertise leads to improved accuracy. For the three financial variables, we calculated average absolute errors for both experienced and inexperienced subjects; Table 5 shows the results. Comparing the overall averages, experienced subjects are uniformly more accurate than inexperienced subjects for all of the assessment methods, trained or untrained. Looking at the individual variable pairs, it is apparent that most of this effect derives from the *SD* and *AC* pairs, for which experienced subjects are always more accurate (with the exception of one cell—*CF* untrained). For *LC*, the results are mixed: In 8 of the 11 cells, experienced subjects are actually less accurate than inexperienced subjects.

In general the overall standard deviations (not presented) are smaller for experienced subjects than for inexperienced (except for *R* and *CF* untrained). The sample sizes for experienced subjects are small, however, suggesting caution in drawing conclusions.

Because the data analyzed in Table 5 involve both experience and training, it is helpful to tease apart the effects of these two variables. We performed a Bayesian analysis separately for each of the five assessment methods *R*, *CF*, *CNC*, *JP*, and *CP* using a standard linear model:

$$AE_{mj} = \alpha_m + \beta_m X_{mj} + \gamma_m Y_{mj} + \varepsilon_{mj},$$

where AE_{mj} is the absolute error for the j th observation using the m th method, X_{mj} equals 0 if AE_{mj} was based on training and 1 if not, Y_{mj} equals 0 if AE_{mj} came from an experienced individual and 1 if not. The error terms ε_{mj} are modeled as independent normal errors with mean 0 and variance σ_m^2 . For each analysis, we used the noninformative Jeffreys prior $f(\alpha_m, \beta_m, \gamma_m, \sigma_m) \propto \sigma_m^{-1}$.

Table 6 shows estimates of the expected improvement in absolute error due to training and experience (effect estimates) and posterior probabilities that the effects are positive. Effect estimates for experience are uniformly positive for all five methods, with the lowest posterior probability being 0.822 for *CNC*. We discuss the effect of training in more detail later, noting for now that the effect estimates for training are positive for each method except *CF*.

Tables 5 and 6 deliver the clear message that experience matters in terms of assessment accuracy. Averaging over all methods, the three financial variable pairs and the two levels of training, the average absolute error is 0.195 for experienced individuals and 0.254 for inexperienced individuals. It is on the strength of these results that we have chosen in our analysis of Study 1 to eliminate the responses on financial variables from experienced subjects. We have done the same for the remaining analysis of Study 2.

The next step is to make a more complete comparison of the results of the untrained assessments from Study 2 with results from Study 1 (also untrained). Table 7 shows average absolute errors and sample sizes for the assessment methods and variable pairs (excluding assessments from experienced subjects on the financial variables). The overall average absolute errors for the untrained methods in Study 2 are uniformly greater than in Study 1, with the differences ranging from a low of 0.008 (*CF*) to a high of 0.066

Table 4 Equivalent Correlations for Study 2: Averages, Standard Deviations, and Sample Sizes

Untrained	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>
<i>S</i>	0.639	0.533	0.772	0.680	0.312
	<i>0.291</i>	<i>0.336</i>	<i>0.190</i>	<i>0.263</i>	<i>0.292</i>
	281	285	239	241	231
<i>R</i>	0.574	0.559	0.741	0.661	0.290
	<i>0.324</i>	<i>0.290</i>	<i>0.213</i>	<i>0.244</i>	<i>0.298</i>
	55	55	40	40	40
<i>CF</i>	0.607	0.518	0.827	0.673	0.249
	<i>0.275</i>	<i>0.374</i>	<i>0.228</i>	<i>0.326</i>	<i>0.416</i>
	57	58	49	49	47
<i>CNC</i>	0.545	0.392	0.691	0.542	0.145
	<i>0.345</i>	<i>0.314</i>	<i>0.282</i>	<i>0.349</i>	<i>0.390</i>
	60	60	53	52	53
<i>JP</i>	0.640	0.405	0.632	0.629	0.091
	<i>0.359</i>	<i>0.450</i>	<i>0.417</i>	<i>0.322</i>	<i>0.418</i>
	47	47	39	41	43
<i>CP</i>	0.557	0.304	0.605	0.441	0.116
	<i>0.309</i>	<i>0.438</i>	<i>0.422</i>	<i>0.450</i>	<i>0.397</i>
	47	54	45	34	39
Average Error	0.082	0.109	-0.212	-0.098	0.073

Trained	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>
<i>R</i>	0.669	0.572	0.775	0.668	0.269
	<i>0.170</i>	<i>0.214</i>	<i>0.165</i>	<i>0.223</i>	<i>0.272</i>
	57	57	47	47	47
<i>CF</i>	0.612	0.580	0.511	0.485	0.532
	<i>0.295</i>	<i>0.361</i>	<i>0.356</i>	<i>0.433</i>	<i>0.376</i>
	55	55	48	48	48
<i>CNC</i>	0.604	0.476	0.772	0.628	0.122
	<i>0.288</i>	<i>0.381</i>	<i>0.341</i>	<i>0.327</i>	<i>0.456</i>
	53	53	43	43	43
<i>JP</i>	0.676	0.530	0.707	0.655	0.131
	<i>0.223</i>	<i>0.338</i>	<i>0.349</i>	<i>0.339</i>	<i>0.364</i>
	60	60	48	48	48
<i>CP</i>	0.540	0.311	0.697	0.574	0.176
	<i>0.365</i>	<i>0.512</i>	<i>0.342</i>	<i>0.379</i>	<i>0.333</i>
	55	56	53	58	49
Average Error	0.092	0.118	-0.259	-0.138	0.075

Note. Average errors are calculated as the average of assessment minus data-based estimate. Responses on financial variables from subjects with financial industry experience have been excluded; see Table 5 and the discussion thereof.

(*CP*). Subjects in Study 1 may experience a slight learning effect from making so many assessments, although the differences in study designs make a meaningful comparison impossible. Despite the slight increase

in average absolute errors, we note that the rank order of the untrained methods is similar in the two studies; in both, the top two performers are *S* and *R*, the next three are *CF*, *CNC*, and *CP*, and sixth is *JP*.

Table 5 Averages of Absolute Errors and Sample Sizes for Financial Variable Pairs, Segregated by Experienced and Inexperienced Subjects and by Training Level

	Experienced					Inexperienced			
	<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall		<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall
Untrained					Untrained				
<i>S</i>	0.130 44	0.140 44	0.257 44	0.176 132	<i>S</i>	0.200 239	0.182 241	0.252 231	0.211 711
<i>R</i>	0.135 15	0.130 15	0.276 15	0.180 45	<i>R</i>	0.213 40	0.186 40	0.254 40	0.218 120
<i>CF</i>	0.144 8	0.321 9	0.168 8	0.215 25	<i>CF</i>	0.174 49	0.233 49	0.335 47	0.247 145
<i>CNC</i>	0.131 7	0.236 7	0.339 7	0.235 21	<i>CNC</i>	0.266 53	0.294 52	0.275 53	0.278 158
<i>JP</i>	0.045 6	0.199 6	0.240 6	0.161 18	<i>JP</i>	0.334 39	0.249 41	0.327 43	0.303 123
<i>CP</i>	0.097 8	0.143 6	0.417 8	0.226 22	<i>CP</i>	0.351 45	0.380 34	0.289 39	0.339 118
Trained					Trained				
<i>R</i>	0.065 10	0.087 10	0.246 10	0.132 30	<i>R</i>	0.177 47	0.172 47	0.228 47	0.192 141
<i>CF</i>	0.229 7	0.152 7	0.476 7	0.285 21	<i>CF</i>	0.447 48	0.360 48	0.444 48	0.417 144
<i>CNC</i>	0.140 10	0.151 10	0.400 10	0.230 30	<i>CNC</i>	0.190 43	0.255 43	0.331 43	0.259 129
<i>JP</i>	0.180 12	0.150 12	0.248 12	0.192 36	<i>JP</i>	0.253 48	0.238 48	0.270 48	0.254 144
<i>CP</i>	0.215 5	0.182 5	0.349 4	0.241 14	<i>CP</i>	0.260 53	0.287 53	0.247 45	0.265 151

The main question addressed by Study 2 is the effect of the specialized training that we performed for the five assessment methods. Judging from Tables 6 and 7, training appears to help. Examining Table 7, for both *R* and *JP* the overall average absolute error improves (by 0.029 and 0.061, respectively), and the improvement is uniform across all variable pairs. For *CP*, average absolute error decreases by 0.018, but the decrease occurs only for the financial variables. For *CNC*, performance improves very slightly (by 0.003). For *CF*, training appears to have a very deleterious effect, increasing the average absolute error by 0.101. We have no explanation for this anomalous result.

The Bayesian analysis in Table 6, although based only on the financial variables, tells a similar story.

The training effect estimates are all positive except for the anomaly with *CF*. The effect estimate for *CP* is the largest, but the posterior probability of a positive effect is greater than 0.90 for both *R* and *CP*. For *JP*, the training effect estimate is actually greater than for *R*, but the posterior probability is slightly lower. The training effect size and the corresponding posterior probability drop somewhat for *CNC*.

Finally, looking back at Table 5, the lowest overall average absolute error occurs for trained and experienced subjects using method *R*, for which the statistic is 0.132. This is about 60% of the average absolute error for untrained and inexperienced subjects using *R*, although the result must be used with caution, based as it is on only 30 assessments.

Examining the standard deviations in Table 7, we note

Table 6 Effect Estimates for Experience and Training and Posterior Probabilities that Effects Are Positive

Method	Experience	Training
<i>R</i>	0.047 0.976	0.030 0.937
<i>CF</i>	0.120 0.996	-0.155 0.000
<i>CNC</i>	0.038 0.822	0.018 0.721
<i>JP</i>	0.067 0.947	0.037 0.880
<i>CP</i>	0.077 0.913	0.063 0.958

Note. Effects are interpreted as the expected improvement in absolute error due to experience or training. Posterior probabilities are in bold face.

that training does not necessarily lead to decreased standard deviations. However, for *R* in particular, the standard deviation drops from 0.203 (untrained) to 0.148 (trained), a drop of 27%.

5. Discussion

Our studies suggest that the most accurate way to obtain a subjective dependence measure is simply to ask the expert to estimate the correlation between the two variables in question. The direct assessment of correlation consistently performed better than any of the other assessment methods in terms of average absolute error. *R* also was judged to be one of the two easiest assessments to make. In addition, the standard deviation (of both the equivalent correlations themselves as well as the absolute errors) was consistently less, indicating a greater level of consensus among the subjects. Finally, in comparison with *CF*, *JP*, and

Table 7 Averages and Standard Deviation of Absolute Errors for Trained and Untrained Assessment Methods and Variables in Study 2

Untrained	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall
<i>S</i>	0.239 <i>0.198</i>	0.296 <i>0.222</i>	0.200 <i>0.167</i>	0.182 <i>0.198</i>	0.252 <i>0.203</i>	0.236 <i>0.203</i>
<i>R</i>	0.231 <i>0.230</i>	0.279 <i>0.196</i>	0.213 <i>0.209</i>	0.186 <i>0.173</i>	0.254 <i>0.191</i>	0.236 <i>0.203</i>
<i>CF</i>	0.230 <i>0.167</i>	0.327 <i>0.225</i>	0.174 <i>0.191</i>	0.233 <i>0.234</i>	0.335 <i>0.254</i>	0.261 <i>0.222</i>
<i>CNC</i>	0.250 <i>0.236</i>	0.258 <i>0.178</i>	0.266 <i>0.275</i>	0.294 <i>0.270</i>	0.275 <i>0.275</i>	0.268 <i>0.246</i>
<i>JP</i>	0.330 <i>0.172</i>	0.376 <i>0.242</i>	0.334 <i>0.404</i>	0.249 <i>0.228</i>	0.327 <i>0.268</i>	0.325 <i>0.270</i>
<i>CP</i>	0.217 <i>0.220</i>	0.319 <i>0.306</i>	0.351 <i>0.417</i>	0.380 <i>0.380</i>	0.289 <i>0.274</i>	0.308 <i>0.326</i>
Trained	<i>HW</i>	<i>MV</i>	<i>SD</i>	<i>AC</i>	<i>LC</i>	Overall
<i>R</i>	0.199 <i>0.091</i>	0.252 <i>0.142</i>	0.177 <i>0.163</i>	0.172 <i>0.156</i>	0.228 <i>0.175</i>	0.207 <i>0.148</i>
<i>CF</i>	0.240 <i>0.187</i>	0.339 <i>0.234</i>	0.447 <i>0.346</i>	0.360 <i>0.346</i>	0.444 <i>0.268</i>	0.362 <i>0.288</i>
<i>CNC</i>	0.229 <i>0.186</i>	0.316 <i>0.231</i>	0.190 <i>0.334</i>	0.255 <i>0.230</i>	0.331 <i>0.313</i>	0.265 <i>0.263</i>
<i>JP</i>	0.233 <i>0.126</i>	0.317 <i>0.190</i>	0.253 <i>0.343</i>	0.238 <i>0.253</i>	0.270 <i>0.244</i>	0.264 <i>0.236</i>
<i>CP</i>	0.278 <i>0.234</i>	0.372 <i>0.355</i>	0.260 <i>0.337</i>	0.278 <i>0.304</i>	0.255 <i>0.211</i>	0.290 <i>0.296</i>

CNC, which can (and sometimes do) lead to assessments outside of the mathematically feasible range, responses for R fall naturally between $+1.00$ and -1.00 .

Accuracy can be improved slightly in two ways. One is to provide some training to the expert regarding the measure of dependence being assessed, as we did in Study 2. The training we did was minimal; for R we discussed the meaning of a correlation as a measure of dependence ranging from $+1.00$ to -1.00 , and we showed several scatterplots with different levels of correlation to help the subject visualize what different correlations would imply graphically. The increase in performance for R especially was accompanied by a marked (more than 25%) decrease in the standard deviation of absolute errors. An open question is whether more extensive training and practice would lead to further improvements.

The second way to improve an expert's performance is to ask for several different dependence measures and average them. The improvement is slight, about 0.015 in terms of average absolute error, and is about the same order of magnitude as the effect of training. However, the averaging also led to a substantial decrease in the standard deviation of absolute errors, which implies a lower risk of getting a very large absolute error when multiple measures are assessed and averaged. These results are consistent with the findings of Makridakis and Winkler (1983) that averaging forecasts from different forecasting techniques led to some improvements in accuracy and dramatic reductions in the variability of accuracy measures. Thus, although R performs better than the other methods, it may be worthwhile to use other methods in addition to (not in place of) R . Standard practice in probability assessment in decision analysis is to ask questions in a variety of ways to ferret out inconsistencies and stimulate careful thought. It is also common to have the same probabilities assessed by more than one expert and to average the resulting probabilities, with the same types of gains noted above for averages of assessments from different methods. If multiple experts are available to assess a dependence measure for a given relationship, averaging their assessments is an attractive strategy.

Another conclusion of our studies is that the ability of our subjects to assess correlations accurately is somewhat limited. Even the most accurate methods have overall average absolute errors around 0.20, and for particular variable pair/assessment method combinations average absolute errors between 0.30 and 0.40 are not unusual. Our results for experienced subjects, however, suggest that experience can improve accuracy, and the low average absolute error on financial variables (0.132) for experienced subjects trained to use R is encouraging if not conclusive. The effects of experience and training in correlation assessment accuracy parallel their roles in probability assessment, where the evidence suggests that expertise, training, and feedback can reduce the effect of overconfidence and improve calibration. (For a review, see Yates 1990.)

One possible explanation for the generally good performance of R is simply that our subjects had been exposed to correlation fairly recently. Although this is true, they also had been exposed to probability concepts. Regardless, if familiarity with correlation is of value, this bodes well for the use of R in real-world applications, where many experts will have been trained in statistical methods and hence should be familiar with correlation. Also, we note that the Excel add-in Crystal Ball presents an example scatterplot for a specified correlation, thereby essentially mimicking the training procedure we used for R . For an expert who is knowledgeable about the variables in question and familiar with correlation, using Crystal Ball's correlation tool appears to be a suitable approach for correlation assessment.

Although the results from our study are sound, we have considered only five variable pairs, albeit with a wide range of positive correlations. Given constraints on the availability of time for the subjects and appropriate variables about which the subjects had some knowledge, we did not include any variables with negative relationships. We speculate that results for negative relationships would parallel those for positive relationships, with the strength of the relationship being more crucial than the direction. However, a more general study, including negative as well as positive correlations, would be of value. Perhaps this

would reveal that some method performs well for some range of correlation values but not for others. Even the best-performing method, *R*, tended to overestimate low correlations; of course, we can always adjust for such miscalibration, and *R* did well in terms of average absolute error, which is our primary criterion, for low as well as high correlations.

In preparing for our empirical studies, we searched the literature and worked diligently to develop dependence-assessment methods. We chose to include methods that seemed reasonable for experts to assess. Other measures of dependence are indeed available in the literature (e.g., see Schweizer and Wolff 1981, Genest and Rivest 1993), but they are difficult to interpret intuitively and do not readily lend themselves to subjective assessment. Regardless, further work might consider other dependence measures and ways to assess dependence. Along these lines, a useful study would compare direct assessment of measures of dependence (as we have done here) with the assessment of dependence via conditional probability distributions, the conventional approach in decision-analysis modeling.

Assessing dependence is not an easy judgmental task, and cognitive problems can influence the assessed measures. For instance, difficulties in judging a joint probability, often labeled the "conjunction fallacy," are well documented (Tversky and Kahneman 1983). *JP* yielded more correlations outside the feasible bounds than any other technique and performed poorly in general. *CNC* also requires thinking about joint events and may cause assessments to be based on special pairs or limited (nonrandom) samples of pairs. *CF* is susceptible to the nonregressive prediction effect (Kahneman and Tversky 1973), as suggested here by the high correlations obtained from *CF*. In Table 1, for example, *CF* had the highest average correlation for four of the five variable pairs and the second highest for the remaining pair. As noted in §2, we believe that conditioning on an interval of values, as is done in *CP*, is difficult. Cognitive problems such as these may have contributed to the fact that *JP*, *CF*, and *CP* had the highest average absolute errors in Table 2 and the worst average difficulty ratings, with *CNC* next in line.

For several of the assessment methods we studied, variations in how the question is framed or in other details could be considered. For instance, the question in *CF* could be asked in terms of a value (e.g., a weight) instead of a percentile. We chose to use a percentile to disentangle judgments about dependence from judgments about marginal probabilities. In this study, the marginal distributions were given to the subjects, but in a real-world decision-making problem that is unlikely to be the case. However, conditioning on fractiles and asking for fractiles may be difficult cognitively. Similarly, the conditional probability in *CP* could be conditioned on a specific value or fractile instead of a range of values. For the methods involving the assessment of probabilities (*CNC*, *JP*, and *CP*), questions can be framed in different ways (e.g., in terms of frequencies), and an alternative to the direct assessment of probabilities is the use of lotteries or other indirect procedures for assessing probabilities. Such procedures are commonly used in decision analysis. Finer details can also be modified. For instance, a percentile other than the 90th percentile could be used in *CF*. We chose the 90th percentile in order to have a wide range of feasible responses (from the 10th to the 90th percentile). Similarly, different percentiles could be used in *JP* and *CP*. However, even with modifications, methods such as *CF*, *CNC*, *JP*, and *CP* might not perform as well as *R*.

Finally, we have considered here methods for assessing dependence among two variables only. Even if we consider only bivariate measures of dependence, when more than two variables must be modeled, the matrix of such bivariate measures must obey certain constraints (e.g., a correlation matrix must be positive definite). Methods to ensure that pairwise assessments obey such constraints must be developed. In addition, in some cases pairwise dependence may be inadequate for modeling the pattern of dependence among a set of variables.¹

¹This research was supported in part by the Fuqua School of Business and by the National Science Foundation under grants SBR 94-22527, SBR 95-96176, and SBR 98-18855. The authors are grateful to the associate editor and referees for helpful comments.

References

- Beyth-Marom, R. 1982. Perception of correlation reexamined. *Memory and Cogn.* **10** 511–519.
- Billman, D., B. Bornstein, J. Richards. 1992. Effects of expectancy on assessing covariation in data: "Prior belief" versus "meaning." *Organ. Behavior Human Decision Process.* **53** 74–88.
- Clemen, R. T., T. Reilly. 1999. Correlations and copulas for decision and risk analysis. *Management Sci.* **45** 208–224.
- , R. T., R. L. Winkler. 1985. Limits for the precision and value of information from dependent sources. *Oper. Res.* **33** 427–442.
- Frees, E., E. Valdez. 1998. Understanding relationships using copulas. *North Amer. Actuarial J.* **2** 1–25.
- Genest, C., L.-P. Rivest. 1993. Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.* **88** 1034–1043.
- Gokhale, D. V., S. J. Press. 1982. Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J. Royal Statist. Soc. Ser. A.* **145** 237–249.
- Jennings, D. L., T. M. Amabile, L. Ross. 1982. Information covariation assessment: Data-based versus theory-based judgments. D. Kahneman, P. Slovic, A. Tversky, eds. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, England 211–230.
- Jouini, M., R. T. Clemen. 1996. Copula models for aggregating expert opinions. *Oper. Res.* **44** 444–457.
- Kahneman, D., A. Tversky. 1973. On the psychology of prediction. *Psych. Rev.* **80** 237–251.
- Kruskal, W. 1958. Ordinal measures of association. *J. Amer. Statist. Assoc.* **53** 814–861.
- Kunda, Z., R. E. Nisbett. 1986. The psychometrics of everyday life. *Cogn. Psych.* **18** 195–224.
- Makridakis, S., R. L. Winkler. 1983. Averages of forecasts: Some empirical results. *Management Sci.* **29** 987–996.
- Morgan, M. G., M. Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, Cambridge, England.
- Nisbett, R., L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. Prentice-Hall, Englewood Cliffs, NJ.
- Pechmann, C., S. Ratneshwar. 1992. Consumer covariation judgments: Theory or data driven. *J. Consumer Res.* **19** 373–386.
- Ravinder, H. V., D. N. Kleinmuntz, J. S. Dyer. 1988. The reliability of subjective probabilities obtained through decomposition. *Management Sci.* **34** 186–199.
- Schweizer, B., E. F. Wolff. 1981. On nonparametric measures of dependence for random variables. *Ann. Statist.* **9** 879–885.
- Shaklee, H., M. Mims. 1981. Development of rule use in judgments of covariation between events. *Child Development* **52** 317–325.
- Smedslund, J. 1963. The concept of correlation in adults. *Scand. J. Psych.* **4** 165–173.
- Smith, A. E., P. B. Ryan, J. S. Evans. 1992. The effect of neglecting correlations when propagating uncertainty and estimating the population distribution of risk. *Risk Anal.* **12** 467–474.
- Tversky, A., D. Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psych. Rev.* **90** 293–315.
- Winkler, R. L., T. S. Wallsten, R. G. Whitfield, H. M. Richmond, S. R. Hayes, A. S. Rosenbaum. 1995. An assessment of the risk of chronic lung injury attributable to long-term ozone exposure. *Oper. Res.* **43** 19–28.
- Yates, J. F. 1990. *Judgment and Decision Making*. Prentice-Hall, Englewood Cliffs, NJ.

Accepted by Martin Weber; received on November 3, 1998. This paper was with the authors 2 months and 20 days for 2 revisions.