PSYCHOMETRIKA—VOL. 74, NO. 3, 531–553
SEPTEMBER 2009
DOI: 10.1007/S11336-009-9118-Z

WHY ARE EXPERTS CORRELATED? DECOMPOSING CORRELATIONS BETWEEN JUDGES

STEPHEN B. BROOMELL

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF ILLINOIS

DAVID V. BUDESCU

DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF ILLINOIS
AND DEPARTMENT OF PSYCHOLOGY, FORDHAM UNIVERSITY

We derive an analytic model of the inter-judge correlation as a function of five underlying parameters. Inter-cue correlation and the number of cues capture our assumptions about the environment, while differentiations between cues, the weights attached to the cues, and (un)reliability describe assumptions about the judges. We study the relative importance of, and interrelations between these five factors with respect to inter-judge correlation. Results highlight the centrality of the inter-cue correlation. We test the model's predictions with empirical data and illustrate its relevance. For example, we show that, typically, additional judges increase efficacy at a greater rate than additional cues.

Key words: information aggregation, correlation, dependence, expert advice.

1. Introduction

Decision makers (DMs) often seek advice from multiple experts[1] and aggregate the advice from the various sources to achieve more accurate decisions. Studies have shown that the accuracy of the aggregate increases monotonically as a function of the number of advisors, but at a diminishing rate that depends on the inter-judge correlation (e.g., Ariely, Au, Bender, Budescu, Dietz, Gu, Wallsten, & Zaubermann 2000; Clemen & Winkler, 1985; Hogarth, 1978; Johnson, Budescu, & Wallsten, 2001; Wallsten, Budescu, Erev, & Diederich, 1997). In other words, the relevance and usefulness of the information provided by each additional expert decreases as the correlation between experts increases. This indicates that the best new piece of information for a decision should be valid (correlated with the true state of, or value in, nature), and as uncorrelated as possible with the other forecasts.

Expert advice is based on naturally occurring cues, such as medical tests, past stock performance, weather patterns, etc., that are typically correlated. The experts often have similar education and training. These factors can account for the magnitude of the inter-expert correlations. Morris (1986, p. 143) summarizes the sources of inter-judge correlations: "*in most situations most experts have access to the same basic information and are basing their opinions on roughly the same body of data. Overlapping methodology may exist if experts in the field have similar academic and professional training. . . . The direct observation of the expert opinions, the presentation of public reports to the scientific community, and the open discussion of viewpoints and*

[1] The terms "expert," "judge," and "advisor" are used interchangeably throughout the paper.

*This work was supported by the National Science Foundation under Awards SES 02-41434 and 03-45925. The first author was supported by a grant from the National Institutes of Health under Ruth L. Kirschstein National Research Service Award PHS 2 T32 MH014257 ("Quantitative Methods for Behavioral Research") to the University of Illinois at Urbana-Champaign.

Requests for reprints should be sent to Stephen B. Broomell, Department of Psychology, University of Illinois, 603 E. Daniel Street, Champaign, IL 61820, USA. E-mail: sbroome2@uiuc.edu

hypotheses will add to the overlap among expert judgments." Surprisingly, there is no overall model of the various sources of dependence. The goal of this paper is to fill this gap by modeling the correlation between numerical estimates provided by two judges (denoted $\rho_{x_j x_{j'}}$) as a function of several basic components that characterize the nature of the decision situation. Such a model can provide new insights about various ways to analyze (and optimize) acquisition, aggregation, and use of information from multiple sources. For example, it will help us to identify conditions for perfectly correlated judges and, conversely, uncorrelated judges, and study the tradeoff between additional judges and informational cues in the environment. In addition, it will shed new light on the definition of expertise (see Einhorn's 1974 discussion).

A prototypical example of the situation which is the focus of this paper is given by Ashton (1986) and Ashton and Ashton (1985). Thirteen managers and sales personnel at *Time* magazine were asked to forecast the number of pages sold annually by *Time* from 1965 to 1978. Each judge was given 5 pieces of information: (a) the quarter to which the data applied; (b) the total number of advertising pages appearing that quarter; (c) the number of pages of liquor advertising appearing that quarter; (d) the number of pages of automobile advertising that quarter; and (e) the total number of pages committed by advertisers to date for the entire year and was asked to estimate the number of pages sold annually. The mean inter-judge correlation was 0.60. Other empirical studies (e.g., Clemen & Winkler, 1986; Winkler, 1971; Winkler & Poses, 1993) have found high correlations between experts (e.g., sports forecasting $r = .84$ to 97, economic modeling $r = .82$ to .96, and medicine $r = .76$ to .79). This suggests that in many cases it does not pay to consult many experts because their marginal contribution becomes negligible. Note, however, that there are quite large differences in the level of inter-judge agreement across domains of expertise (Shanteau, 2001; Weiss & Shanteau, 2003a).

1.1. *Aggregation of Information from Correlated Sources*

Clemen and Winkler (1985) offer an elegant analysis of effects of dependence among experts. They compute the posterior error variance, σ_D^2, of a DM who consults J experts whose opinions are equicorrelated ($\rho_{jj'} = \rho$). This value is compared to posterior error variance of a DM who consults n independent experts, σ_I^2. The authors solve for n such that $\sigma_D^2 = \sigma_I^2$ and show that the equivalent number of independent experts, denoted $n(\sigma^2, \Sigma)$, is always less than J (for $J > 1$).

$$n(\sigma^2, \Sigma) = J\left[1 + (J-1)\rho\right]^{-1} \tag{1}$$

Note that as ρ increases the equivalent number of independent judges, $n(\sigma^2, \Sigma)$, decreases.

Hogarth (1978) seeks to determine how many experts one should consult in the presence of inter-dependence. He uses a psychometric model to predict the validity of the simple mean of a group of J experts, denoted $\rho_{y\bar{x}}$. This validity is expressed as a function of the number of experts (J), the mean validity of the individual experts ($\bar{\rho}_{yx}$), and the mean correlation between experts ($\bar{\rho}_{x_i x_j}$).

$$\rho_{y\bar{x}} = J^{1/2}\bar{\rho}_{yx}\left[1 + (J-1)\bar{\rho}_{x_i x_j}\right]^{(-1/2)} \tag{2}$$

Hogarth shows that adding experts is more valuable if these additional experts lower the mean inter-correlation of the group (as opposed to simply adding the experts with the highest individual mean validity). An important insight of this analysis is that additional experts will be more valuable when $\bar{\rho}_{yx} > \bar{\rho}_{x_i x_j}$. This model shows that the (mean) inter-judge correlation has just as much influence on group validity as the (mean) individual validity. The model's main predictions were confirmed even in situations where some of its assumptions were violated (Ashton, 1986).

It is easy to see that Eqs. 1 and 2 are, essentially, equivalent. More precisely, $n(\sigma^2, \Sigma) = (\rho_{y\bar{x}}/\bar{\rho}_{yx})^2$, which can be considered a measure of efficacy of a group of correlated judges. This measure is revisited in the applications section of our paper.

Wallsten et al. (1997) propose a cognitive model of the generation of probability judgments. The judgments are a function of the judge's internal confidence, U, and random variation, E, transformed by a monotone function h (increasing in both its arguments). The overt judgment, $X_j = h_j(U_j, E_j)$, is a random variable in (0, 1). Wallsten and Diederich (2001) show that when the judgments of the experts are independent, conditional on the true state of nature, the mean of their estimates becomes increasingly diagnostic as the number of judges increases. At the other extreme, if all judges share the same "true" probability ($U_j = U$ for all j) but vary in their random components, there will be no improvement in diagnosticity. Johnson et al. (2001) show that averaging (imperfectly) correlated estimates always increases diagnosticity, but the rate of convergence to the ideal case of perfect discrimination slows as a function of the amount of inter-judge dependence. Ariely et al. (2000) confirmed these results empirically.

1.2. *The Present Paper*

Winkler (1981) describes the sources of dependence between experts: "*Experts often have some common training and experience, they may see the same data, and they may use similar aids (e.g., statistical procedures).*" The present paper incorporates this insight into a mathematical model that allows us to identify the drivers of the inter-judge correlations and analyze their relative importance. The paper is organized as follows. In Section 2, we describe a simple model of the process by which experts form estimates, and discuss the various parameters in the model. In Section 3, we derive the key results—the inter-judge correlation, analyze the effects of its various drivers, and compare their relative importance. Section 4 analyzes the effects of the constraints imposed by the cues' validity on the inter-judge correlation. In Section 5, we use the model to address various applied questions about the role of the inter-judge correlation in the advice giving and taking process and use the model to analyze a simple example. Section 6 summarizes the results and links them to the literature on expertise.

2. A Model of the Experts' Opinions

The model assumes that experts gather information (cues in nature) and aggregate it to generate an opinion or forecast, which is communicated to the DM who, in turn, aggregates the experts' opinions (e.g., Budescu, 2006). Figure 1 depicts this advice giving-and-taking process which begins at the bottom of the figure with the informational cues in the decision environment. We motivate the model with a short review of Brunswik's lens model. Then we derive the model of inter-judge correlation of linear judgment that can be analyzed as a function of several meaningful parameters that capture key features of the decision situation and the judges.

2.1. *Linear Judgments*

The cues in the decision environment are aggregated by each expert through a simple linear combination. This characterization of the information aggregation process resembles the structural formulation of the multiple lens model (e.g., Hursch, Hammond, & Hursch, 1964; Hammond, Wilkins, & Todd, 1966; Hammond & Stewart, 2001). In this research each experts' opinion, x_j $(j = 1, 2, \ldots, J)$ is a weighted average of the information cues, c_i $(i = 1, 2, \ldots, N)$ in the decision environment perturbed by random noise (e_{ij}). The cues are assumed to be jointly distributed with mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. Thus:

$$x_j = \sum_{i=1}^{N} w_{ij} c_i + e_{ij} \quad \text{where } e_{ij} \sim \left(0, \sigma^2\right) \text{ and } \boldsymbol{c} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

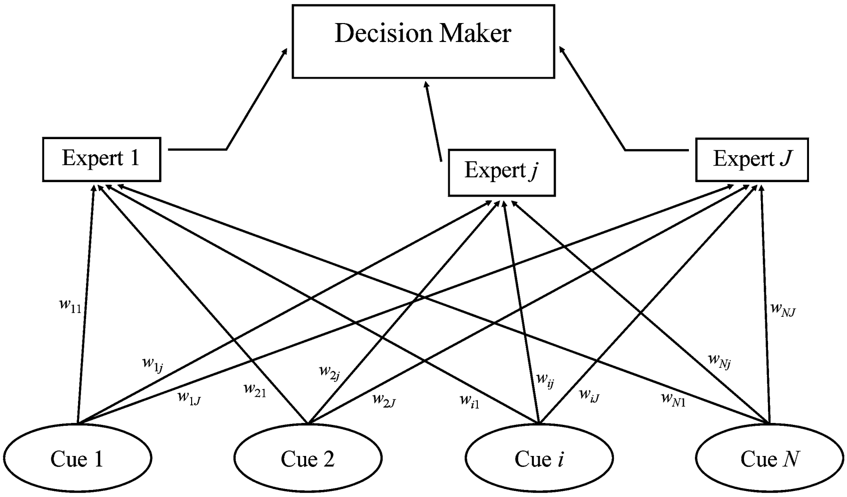


FIGURE 1.
Depiction of information aggregation scenario in seeking advice from experts.

This definition presents an expert's opinion, x_j, as a function of three variables: (a) the "natural" information cues in the environment, c_i, (b) the weights attached by each expert to each cue, w_{ij}, and (c) a random component, e_{ij}, which represents various sources of imperfection, unreliability, and error in the expert's performance.[2] These errors reflect misperception of the cues or imperfect aggregation of the cues, or both.

At this point, we deviate from Brunswik's lens model. A different set of assumptions are imposed and, more importantly, we focus on the analysis of the inter-judge agreement and not on the correspondence between the judge and the environment which is typically the focus of the lens model (see Hammond et al., 1966 for an example of interpersonal agreement). We assume that all N cues are available to all J experts and, without loss of generality, we assume that the cues are scaled in comparable units. We assume that all the weights are nonnegative and normalized to sum to one for each expert.

$$\sum_{i=1}^{N} w_{ij} = 1 \quad \text{for } j = 1, \dots, J \tag{4}$$

Our objective is to express the correlation between the opinions of two different experts (x_j and $x_{j'}$) in terms of (a) the number of cues, N, (b) the underlying correlations between the cues, $\rho_{cc'}$, (c) the variability in the weights assigned to the cues by judge j, σ_w^2, (d) the correlation of the weights used by the two experts, $\rho_{ww'}$, and (e) the amount of noise (error) of the judges, $\sigma_{e_j}^2$, and $\sigma_{e_{j'}}^2$.

The variance of an expert opinion, $\sigma_{x_j}^2$, and the covariance between two experts' opinions, $\sigma_{x_j x_{j'}}$ can be expressed in terms of the covariance matrix of the cues, the vectors of the judges' weights, $\boldsymbol{w}_j$ and $\boldsymbol{w}_{j'}$, and the random error. More specifically:

$$\sigma_{x_j}^2 = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_j + \sigma_{e_j}^2 \tag{5}$$

$$\sigma_{x_j x_{j'}} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_{j'} \tag{6}$$

[2] We assume that the judges are unbiased, but one could generalize the model to accommodate biased judges by allowing different (nonzero) values for the mean errors.

Thus, the correlation between two judges, $\rho_{x_j x_{j'}}$, is:

$$\rho_{x_j x_{j'}} = \frac{\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_{j'}}{(\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_j + \sigma_{e_j}^2)^{1/2} (\boldsymbol{w}_{j'}^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_{j'} + \sigma_{e_{j'}}^2)^{1/2}} \tag{7}$$

Note that this general formulation is not related to a particular domain or type of variable. It applies to all cases where judges aggregate cues into numerical values on an interval scale (e.g., forecasts of earnings, probabilities of rain, scores assigned to a set of contestants, etc.). Next, we add some assumptions that allow us to rewrite this equation in terms of factors that represent the conditions of the decision scenario. In the next sections, we define and discuss the factors involved in the model.

2.2. *Description of Weights*

Assumptions about the judges' weights depend on how the judges are sampled, and how the weights are determined. In many situations, judges are asked to make predictions about the target variables based on N cues, and the weights are inferred from the predictions (meaning there is no direct weight elicitation). In this case, the inferred weights are random variables, especially if the various cases are selected randomly from a certain pool. A typical example is the study by Schmidt, Johnson, and Gugel (1978) that examined the policies used by faculty members to evaluate prospective students. In this familiar scenario, each faculty member had access to several predictors (undergraduate GPA and GRE scores) and recommended to accept or reject each applicant. The authors used multiple regression to estimate the weights for each of the cues (predictors) in the various divisions of a psychology graduate program.

Recall that the weights are positive and add to one, implying that the mean weight is $1/N$. The variance of the weights is:

$$\sigma_w^2 = \frac{1}{(N-1)} \left(\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{w}_j - \frac{1}{N} \right) \tag{8}$$

The maximum and minimum of the inner product, $\boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{w}_j$, define the parameter space:

$$0 < \sigma_w^2 \le \frac{1}{N} \tag{9}$$

The lower bound is reached when all weights are equal (the judge takes a simple average of the N cues). The upper bound is reached when the judge considers only one cue (one weight = 1, and all others = 0). Appendix A presents a variety of additional reasonable weighting patterns, along with their variance formulas. This is, of course, a partial list.

The covariance between two sets of weights is also bounded:

$$\frac{-1}{N(N-1)} \le \sigma_{w_j w_{j'}} \le \frac{1}{N} \tag{10}$$

The lower bound of the covariance is reached when the two judges use (i.e., assign nonzero weights to) nonoverlapping sets of cues. Excluding the case where an expert weights all cues equally (so they have no variance), we can describe the similarity between two experts using the correlation between their respective weight vectors ($\rho_{w_j w_{j'}}$).

2.3. *Description of Cues*

The multivariate distribution of the cues has mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$. These calculations are only concerned with $\boldsymbol{\Sigma}$ (the inter-judge correlations are not affected by the values in $\boldsymbol{\mu}$). To simplify calculations, we assume that all N cues have unit variances, and are equally correlated ($\rho_{cc'} = \rho_c$ for all cues), so $\boldsymbol{\Sigma}$ can be re-expressed as a simple function of the common inter-cue correlation, ρ_c:

$$\boldsymbol{\Sigma} = (1-\rho_c)\mathbf{I}_N + \rho_c \mathbf{1}_N \mathbf{1}_N^{\mathrm{T}} \tag{11}$$

The assumption of equal variances is innocuous (the inter-judge correlations are invariant under change of scale). The assumption that all cues are equally correlated restricts somewhat the generality of our model, and it was invoked to simplify the interpretation of the results. If the assumption does not hold, the model can be viewed as a first order approximation in which ρ_c is the average correlation between cues.

2.4. *Description of Random Noise*

Finally, we define δ_j to be the "noise to signal" ratio. This is the ratio of the error variance of the jth expert to the variance of a perfect (error-free) aggregate of the N cues:

$$\delta_j = \frac{\sigma_{e_j}^2}{\sigma_{c_j}^2} \quad \text{where } c_j = \sum_{i=1}^{N} w_{ij} c_i \tag{12}$$

The variance of the expert opinion, $\sigma_{x_j}^2$, (Eq. 5) can be reexpressed as

$$\sigma_{x_j}^2 = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma} \boldsymbol{w}_j (1+\delta_j) \tag{13}$$

Table 1 summarizes the various parameters used in describing and analyzing the drivers and determinants of the inter-judge correlation.

TABLE 1.
The variables involved in the model of the inter-judge correlation ($\rho_{x_j x_{j'}}$).

	Assumptions about nature
ρ_c	The average correlation between the cues used by the judges (medical tests, stock performance, etc.)
N	The amount of total information (number of cues) available in a given decision context
	Assumptions about experts
σ_w^2	The variance of the weights—A measure of the expert's differentiation between cues
ρ_w	The correlation between the weights assigned by the experts to the various cues
δ_j	Noise-to-Signal ratio—A measure of the unreliability of the expert
	Assumptions about validity
ρ_{cy}	The average validity of the cues used by the judges to predict the criterion, y

3. The Inter-Judge Correlation

Substituting Eqs. 8, 11, and 13 into Eq. 7 expresses the inter-judge correlation in terms of the parameters listed in Table 1. (See derivation in Appendix B.)

$$\rho_{x_j x_{j'}} = \frac{\rho_c + (1-\rho_c)[(N-1)\rho_w \sigma_{w_j}\sigma_{w_{j'}} + 1/N]}{([\rho_c + (1-\rho_c)((N-1)\sigma^2_{w_j} + 1/N)][\rho_c + (1-\rho_c)((N-1)\sigma^2_{w_{j'}} + 1/N)](1+\delta_j)(1+\delta_{j'}))^{1/2}} \tag{14}$$

3.1. The Symmetric Model

The model can be simplified further by assuming symmetry among the judges in their discrimination between cues (i.e., the variance of the weights) and their unreliability (i.e., the noise-to-signal ratios). For our primary analysis, we assume $\sigma^2_w = \sigma^2_{w_j} = \sigma^2_{w_{j'}}$ and $\delta = \delta_{j'} = \delta_j$. (We provide an analysis of the sensitivity of inter-judge correlation to violations of these assumptions in the next section.) The correlation between two symmetric judges is

$$\rho_{x_j x_{j'}} = \frac{\rho_c + (1-\rho_c)[(N-1)\rho_w \sigma^2_w + 1/N]}{[\rho_c + (1-\rho_c)((N-1)\sigma^2_w + 1/N)](1+\delta)} \tag{15}$$

Taking partial derivatives with respect to each of the factors shows that ρ_c (the inter-cue correlation), N (the number of cues), and ρ_w (the inter-weight correlation), are positively associated with the inter-judge correlation, while σ^2_w (the cue differentiation) and δ (a measure of unreliability) are negatively associated with it.

To understand the joint effect of the various sources of dependence (inter-cue and inter-weight correlations), we plot several examples. Figure 2 displays the value of the predicted inter-judge correlation between judges who have access to $N = 2$ cues, as a function of ρ_w and ρ_c for two different weighting patterns—equal weights (1/2, 1/2), or weights proportional to the ranks of the cues (2/3, 1/3), and for two signal to noise ratios. For $N = 2$, both ρ_c and ρ_w range between -1 and 1. The inter-judge correlation can take on almost every possible value in $(-1, 1)$, but note that it increases very rapidly. The inter-judge correlation increases linearly with respect to ρ_w, and at a quadratic rate with respect to ρ_c (except for the case of maximal variance). Increasing the number of cues produce similar results, however, the lower bound of inter-judge correlation is higher.

We can assess which parameters will be more influential in determining the magnitude of inter-judge correlation through comparisons of the partial derivatives of the inter-judge correlation with respect to inter-cue correlation (ρ_c), the inter-weight correlation (ρ_w), and the weight variance (σ^2_w). Figure 3 has the inter-cue correlation on the vertical axis, and the inter-weight correlation on the horizontal axis. The figure shows the 'iso-influence' curves (where the ratio of the partial derivatives is 1) that divide the parameter space into two sections for various values of N. The area of the plane which is shaded (white) identifies the region under which $\rho_w(\rho_c)$ induces a higher rate of change. Clearly, the inter-cue correlation is more influential over a much larger portion of the parameter space, and the area under which ρ_w is dominant recedes as the number of cues, N, increases.

Figure 4 compares the partial derivatives of the inter-judge correlation with respect to ρ_c and σ^2_w (scaled by a factor of N to aid in comparison across various numbers of cues). This figure also has a much larger region of dominance for the inter-cue correlation. This trend is more pronounced as N increases.

This analysis shows that the cues in the environment are typically the dominant influence on the agreement between judges. This idea will be revisited in Section 5.

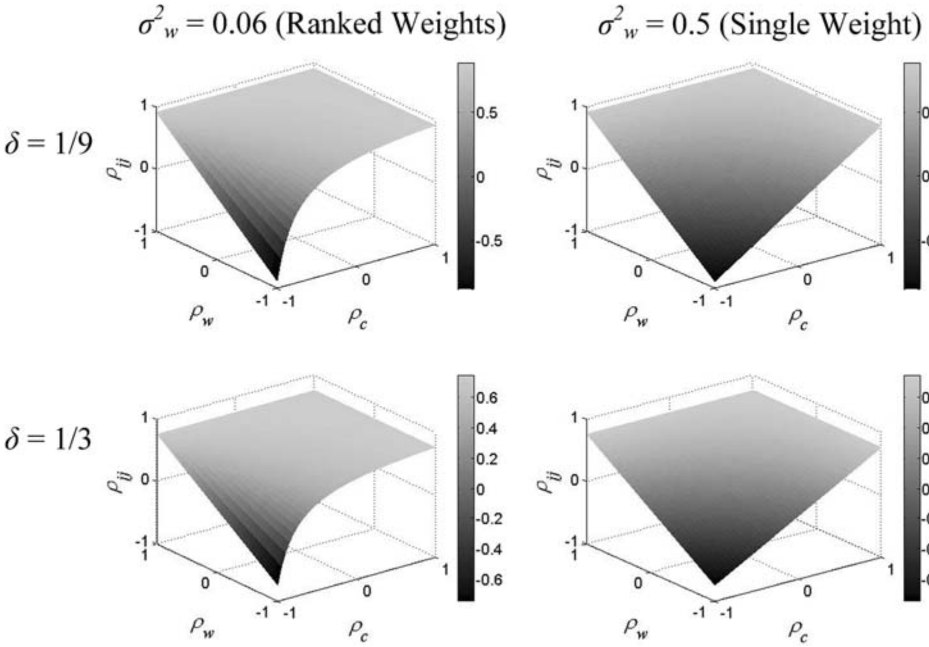


FIGURE 2.
Inter-judge correlation as a function of inter-weight and inter-cue correlations for $N = 2$ cues for two weighting schemes and two noise-to-signal ratios.

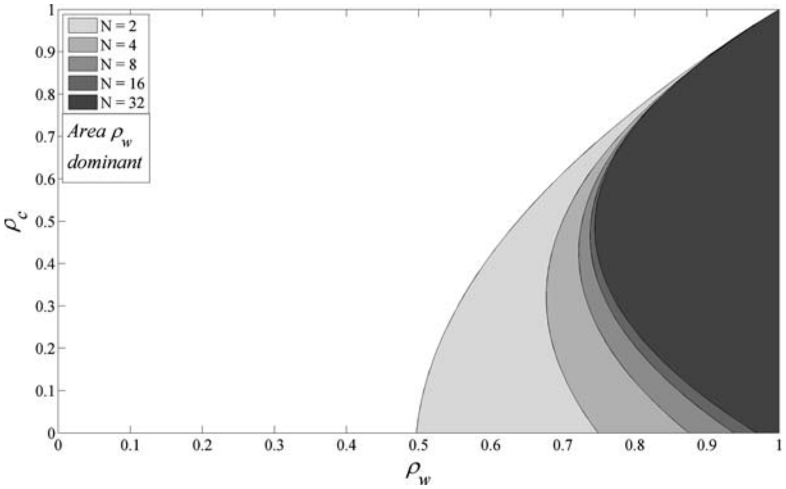


FIGURE 3.
The ratio of the partial derivatives of the inter-judge correlation with respect to inter-cue and inter-weight correlation as a function of the number of cues (σ_w^2 computed assuming the weights are proportional to ranks).

3.2. The Asymmetric Model

One can argue that in practice most experts will be similar to each other with respect to cue discrimination and reliability. For example, if one consults J physicians of similar experience and specialization, it makes sense to assume that their reliability and differentiation will not vary by much. On the other hand, if one seeks advice from judges that vary widely in experience,

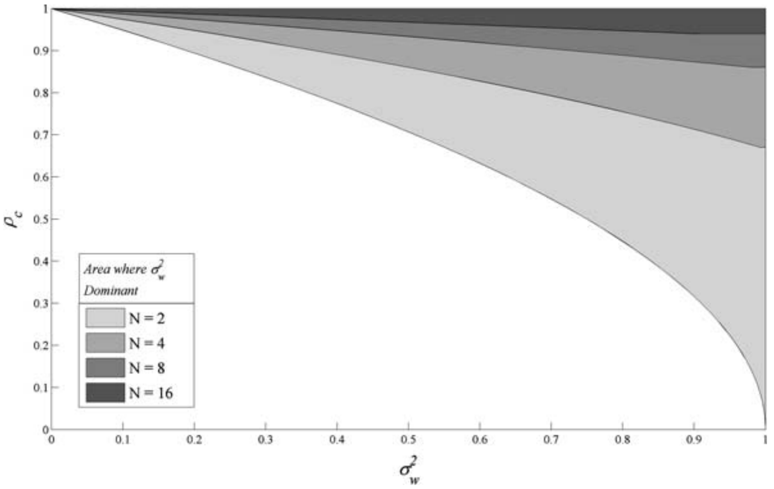


FIGURE 4.
The ratio of the partial derivatives of the inter-judge correlation with respect to inter-cue correlation and weight variance (scaled by N) as a function of the number of cues.

expertise, or specialization, the assumption of symmetry may be harder to justify. This section accommodates and compares these two sources of asymmetry between experts. Let β be the ratio of variances of the two sets of weights, so that $\sigma^2_{w_{j'}} = \beta\sigma^2_{w_j}$. Similarly, let γ represent the ratio of the two expert's reliabilities, so that $(1 + \delta_{j'}) = \gamma(1 + \delta_j)$. We can rewrite the inter-judge correlation in a way that highlights the asymmetry between two experts:

$$\rho_{x_j x_{j'}} = \frac{\rho_c + (1-\rho_c)[(N-1)\rho_w \sigma^2_{w_j}\beta^{1/2} + 1/N]}{([\rho_c + (1-\rho_c)((N-1)\sigma^2_{w_j} + 1/N)][\rho_c + (1-\rho_c)((N-1)\sigma^2_{w_j}\beta + 1/N)](1+\delta_j)^2\gamma)^{1/2}} \tag{16}$$

Figure 5 evaluates the influence of the asymmetry parameters for a case with $N = 8$ cues. We fixed the variance of one of the advisors at $\sigma^2_{w_j} = 0.012$ (corresponding to weights proportional to squared ranks of the cues), and fixed that judge's noise to signal ratio at $1/9$. We examine 4 cases produced by crossing inter-cue (ρ_c) and inter-weight (ρ_w) correlations of 0 and 0.5.

The benchmark symmetric case is defined by the combination $\beta = \gamma = 1$ ($\log(\beta) = \log(\gamma) = 0$), and the degree of asymmetry (operationalized by the ranges of β and γ) is constrained by the fixed values of $\sigma^2_{w_j}$ and δ_j as well as by restrictions of values these parameters can take for the second expert. The effects of both facets of asymmetry are not very large and are more pronounced for the cases where the weights, and especially the cues, are uncorrelated and are quite negligible as these correlations increase (compare the top-left and bottom-right panels of the figure). In most cases studied here, the effect of asymmetry between the judges' cue differentiation ($\log(\beta)$) are more pronounced than the effects of differential reliabilities ($\log(\gamma)$).

4. The Role of Validity

So far, our analysis has ignored the validity of the judgments (the degree to which they are related to the actual value of the criterion variable, y). Of course, cues are selected in the first place solely because they are (or at least expected to be) valid (e.g., diagnostic tests, economic

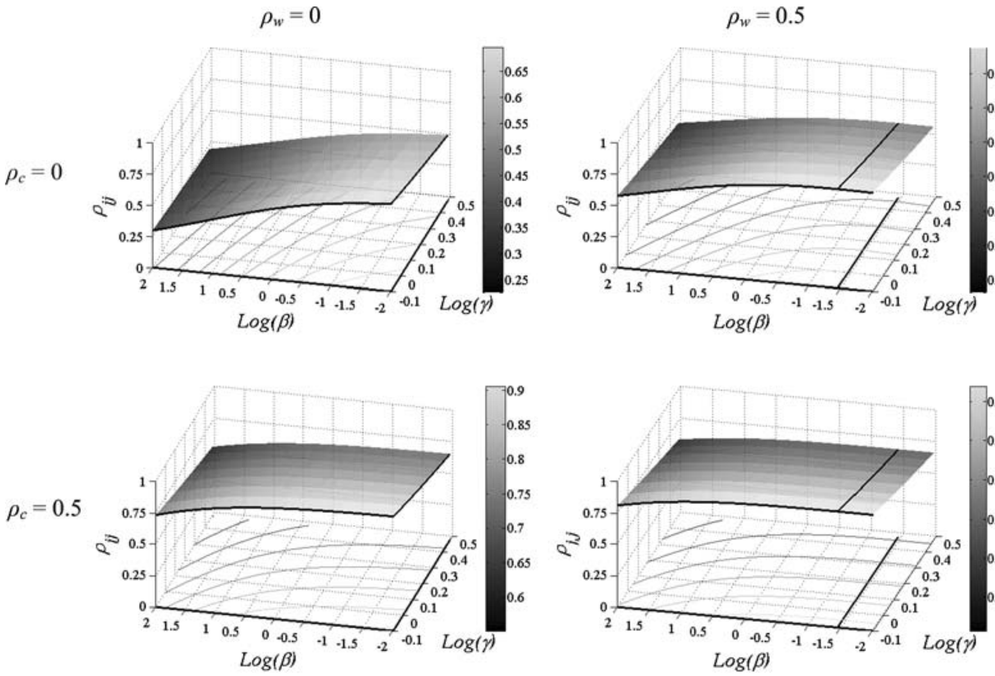


FIGURE 5.
Inter-judge correlation as a function of $\log(\beta)$, the asymmetry term for σ_w^2, and $\log(\gamma)$, the asymmetry term for δ_j, with fixed values $\delta = 0.11$, $\sigma_w^2 = 0.012$, and $N = 8$.

indicators, etc.). Is it possible that the inter-judge correlation is simply a by product of the necessary correlation between valid cues? This section explores the constraints that judge validity ($\rho_{x_j y}$) has on inter-judge correlation by “partialing out” its effects. We calculate the lower bound for inter-judge correlation, as a function of cue validity and the variance of the experts’ estimates, and use this quantity as a benchmark in our analysis. We seek to show that the inter-judge correlation goes well above the constraints due to the cues’ validities.

4.1. Modeling Validity

The predictive validity of an expert is the correlation between his/her opinion, x_j as defined in Eq. 3, and the criterion variable y. This quantity is a function of the validity of the cues the expert aggregates and the weighting scheme used. We augment the original model by adding the criterion variable to the vector of cues.

$$\boldsymbol{c}_y^{\mathrm{T}} = \left[\boldsymbol{c}^{\mathrm{T}}, y\right], \quad \text{where } \boldsymbol{c}_y \sim (\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \tag{17}$$

The covariance matrix for this new vector is a simple augmentation of the covariance matrix (Eq. 11) with an additional row (and column) that includes the validities of the cues ($\rho_{c_i y}$). To simplify the analysis, we assume that all cues are equally valid (i.e., $\rho_{c_i y} = \rho_{cy}$ for all i). This assumption can be justified in part by the notion that the essence of expertise is the ability to identify the most valid cues (e.g., Einhorn, 1974), and discard the others, reducing the variance of the cue validities. A model assuming equal validity of cues provides a good first-order approximation. The validity of the expert’s opinion is (see Appendix B):

$$\rho_{x_j y} = \frac{\rho_{yc}}{\sigma_{x_j}} \tag{18}$$

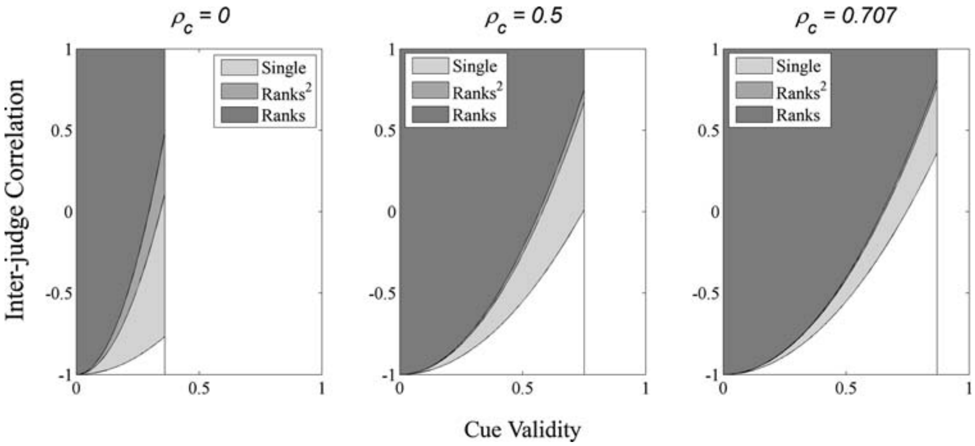


FIGURE 6.
Inter-judge correlation as a function of average cue validity for three values of inter-cue correlations and three weighting schemes. The *shaded region* shows the restricted range of inter-judge correlation.

The partial correlation between two symmetric experts (i.e., with equal variances of opinions: $\sigma^2_{x_j} = \sigma^2_{x_{j'}} = \sigma^2_x$), partialling out the criterion, y, is given by:

$$\rho_{x_j x_{j'}.y} = \frac{\rho_{x_j x_{j'}} \sigma^2_x - \rho^2_{cy}}{\sigma^2_x - \rho^2_{cy}} \tag{19}$$

Given that $(-1 \leq \rho_{x_j x_{j'}.y} \leq 1)$ it is easy to derive the bounds of the restricted range:

$$\frac{2\rho^2_{cy} - \sigma^2_x}{\sigma^2_x} \leq \rho_{x_j x_{j'}} \leq 1 \tag{20}$$

Figure 6 displays the bounds of $\rho_{x_j x_{j'}}$ as a function of the average cue validity for 9 combinations of the inter-cue correlation, ρ_c, and cue differentiation, σ^2_w, for a case with $N = 8$ cues. The shaded region in each panel represents the lower bound of the inter-judge correlation induced by a given cue validity. The range of cue validity is restricted as a function of the value of ρ_c by the requirement that $|\mathbf{\Sigma}_y| \geq 0$ (positive semidefinite). The plot shows that as (a) cue dispersion and (b) inter-cue correlation increase, the restriction of inter-judge correlation becomes more severe. Yet, the lower bounds of the inter-judge correlation are not very high (they are all less than, or equal to, 0.5 in these examples).

We calculated the difference between the inter-judge correlation predicted by the model (Eq. 15), and its lower bound (Eq. 20) and expressed it as a fraction of its possible range, to obtain a measure of Relative Improvement (RI). Formally:

$$\text{RI} = [\rho_{x_j x_{j'}} - \text{Lower Bound}]/[\text{Upper Bound} - \text{Lower Bound}] \tag{21}$$

RI is a rescaling of the inter-judge correlation that represents its increase beyond the lower bound, as a fraction of range of validities. This quantity is bounded from below at 0—when the correlation is exactly at its lower bound implied by the cues' validity—and from above at 1, when the correlation reaches its upper bound. Thus, a RI close to 0 shows little change in the inter-judge correlation beyond the values predicted from the cues' validities. Conversely, RIs closer to 1 show strong effects of the inter-cue and/or the inter-weight correlations on the inter-judge correlation, above and beyond the levels anticipated based on the cues' validities.

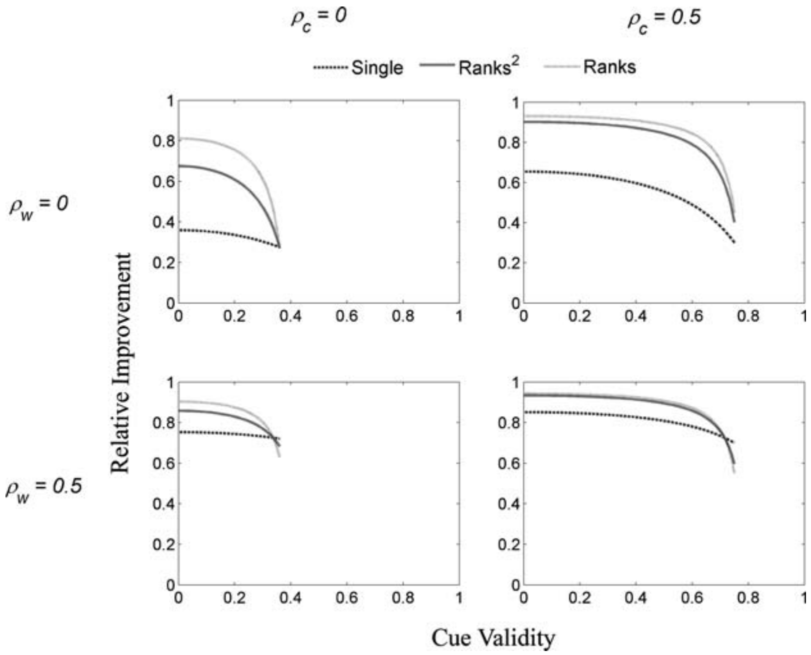


FIGURE 7.
Relative Improvement plotted as a function of cue validity for four different combinations of inter-weight and inter-cue correlation, along with three different weighting methods. Assuming symmetric judges, $N = 8$ and $\delta = 1/9$.

Figure 7 shows the relative improvement in the inter-judge correlation as a function of cue validity ranging from 0 (no validity) to the maximal possible validity for each case. The four panels pertain to 4 combinations of ρ_c and ρ_w, and each panel traces three levels of cue differentiation selected from Appendix A. Several regularities stand out. The key result is that the relative improvement is always at least 0.5, and for most cue validities it is substantially higher. Note that the case of maximal cue differentiation (using a single cue) is the least sensitive to the cue validity. This is indicative of a general pattern—decreasing cue differentiation increases the sensitivity of relative improvement to cue validity. Another interesting result is that increasing either the inter-cue correlations or inter-weight correlation (moving from the top left to the bottom right) reduces the difference between these three lines. As the inter-cue correlation increases (moving from the left column to the right column) the relative improvement becomes less sensitive to the weighting scheme. This implies that, in the presence of moderate to high inter-cue correlation, cue validity is critical only when reaching extremely high values that are higher than the validity of the cues available to experts in most domains.

5. Predictions and Applications of the Model

Our aggregation model combined with knowledge about relevant parameters can be used to formulate optimal policy guidelines, and to justify decisions in many high stakes decisions. In this section, we illustrate the use of the model to answer five questions that are of theoretical and practical importance: (a) When will judges be perfectly correlated? (b) Under what conditions will judges be uncorrelated? (c) What are the effects of different weighting methods on the inter-judge correlation? (d) Keeping everything else equal, should one recruit an additional judge, or obtain an additional cue? (e) How well does the model predict real world data?

5.1. *Perfectly Correlated Judges*

If judges are perfectly correlated, it is sufficient to retain the services of a single advisor. Thus, it is important to be able to identify such cases. A necessary condition for a perfect correlation is the absence of random noise ($\delta = 0$). In addition, if either the cues or the weighting patterns are perfectly correlated (i.e., either ρ_c or $\rho_w = 1$), one can expect perfect correlations between error-free judges.

When an expert weights all cues equally $\sigma_w^2 = 0$, and the correlation, ρ_w, is undefined. Uniform weighting of the cues is equivalent to using a single cue ($N = 1$), denoted c^* (where c^* is the mean of all N cues). Then

$$\rho_{x_j x_{j'}} = \frac{1}{(1+\delta_j)(1+\delta_{j'})} \tag{22}$$

Thus, in the absence of noise this case also induces a perfect inter-judge correlation.

Clearly, these conditions are extremely rare, so it is highly unlikely to encounter experts that are consistently in perfect agreement.

5.2. *Uncorrelated Judges*

If judges are uncorrelated, the performance of any group of advisors is a simple sum of the J individual judges' performance (see Eqs. 1 and 2). The model predicts an inter-expert correlation of 0 when the numerator of Eq. 15 vanishes, i.e., when

$$\rho_c = \frac{N(N-1)\sigma_{w_j w_{j'}} + 1}{(N-1)(N\sigma_{w_j w_{j'}} - 1)} \tag{23}$$

This occurs when the covariance of the weights is minimal ($\sigma_{w_j w_{j'}} = \frac{-1}{N(N-1)}$). All other values of $\sigma_{w_j w_{j'}}$ produce values of $\rho_c < 0$ (which is outside the parameter space of ρ_c for $N > 2$). Thus, with $N \geq 3$ cues the judges' opinions are uncorrelated only when weight covariance and inter-cue correlation take on their lowest possible values.

Such circumstances are also very rare, so it is equally unlikely to come across judges and experts whose opinions are uncorrelated.

5.3. *Controlling the Inter-Weight Correlation*

Imagine a situation where a DM has access to N cues with an inter-cue correlation of ρ_c. The DM could control, at least to some degree, the inter-judge correlation by choosing "appropriate" judges. Can a sophisticated DM exert some control over the inter-judge correlation by anticipating which weighting pattern will be invoked by the experts based on their previous performance, or by asking them about it? To answer this question, we simulated 4 groups of 1,000 experts who use distinct weighing patterns that vary with respect to the level of cue differentiation (see Appendix A): triangular (labeled TRI), proportional to ranks (RK), proportional to squared ranks (SqRK), and pick a single cue (SNG). For the purpose of this illustration, we assume that the judges have access to $N = 8$ cues, correlated at $\rho_c = .5$ and they operate with a noise to signal ratio $\delta = 1/9$. We also assume that all judges agree on the identity of the four most important (i.e., should be given the highest weights) cues. The ordering of these four best cues, as well as the ordering of the other four cues was determined randomly for each judge.

The top panel of Figure 8 shows the distribution of the inter-weight correlations for the six types of pairings of judges. Most distributions have high medians (near 0.8) and relatively homogeneous spread. The exceptions are the correlations involving judges that use a single cue. They have lower medians and higher variability. The bottom panel of Figure 8 displays the distributions

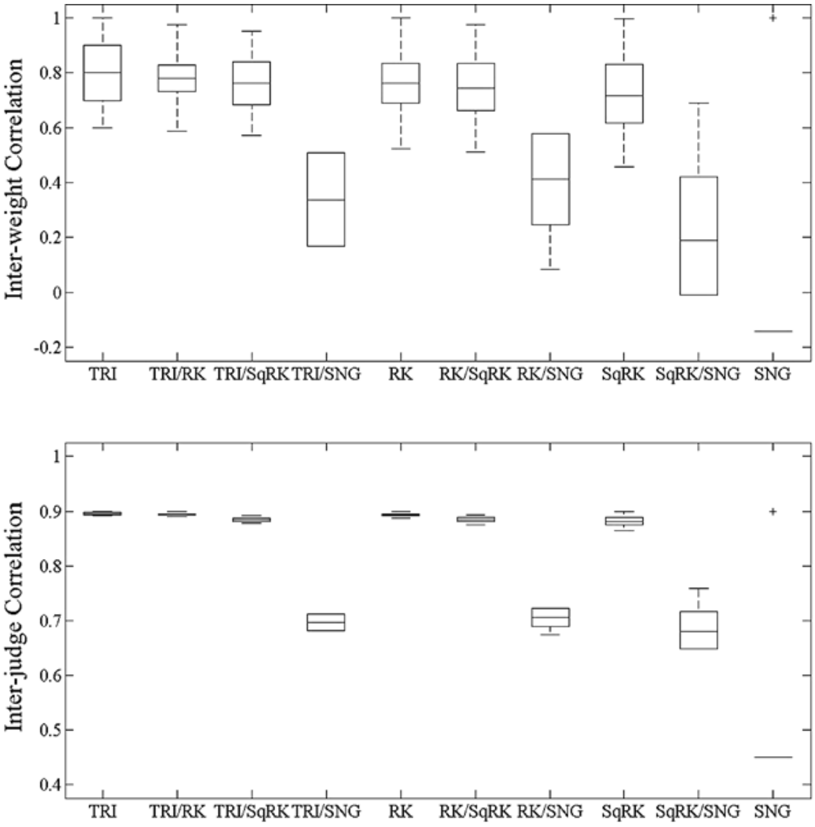


FIGURE 8.
Simulated distribution of inter-weight (*top*) and inter-judge (*bottom*) correlation between judges using various weighting schemes. Simulations assume that judges agree on best four and worst four cues using $N = 8$, $\rho_c = 0.5$, and $\delta = 1/9$. The rules compared are: a Single cue (SNG), weights proportional to Ranks (RK), weights proportional to Squared Ranks (SqRK), and Triangular distribution of weights (TRI).

of the corresponding inter-judge correlations. The inter-judge correlation is not strongly influenced by fluctuations in inter-weight correlation (confirming and complementing, Dawes, 1979). This result highlights the importance of the cues which are utilized (and their inter-correlations) over the impact of the judges' ability to distinguish between them and weigh them differentially.

5.4. *An Extra Judge or an Extra Cue?*

Consider the situation where a group of J judges has access to N cues that induce a certain inter-judge correlation. If this correlation is judged to be too high, one may consider dropping a cue or reducing the number of judges in future applications. Conversely, one may decide to add an extra judge, or obtain an extra cue from the decision environment. Assuming the costs of an extra judge and an extra cue are comparable, and that all parameters are fixed, is it more effective to add (drop) a cue or an expert? We defined a measure of efficacy consistent with the analyses of Hogarth (1978) and Clemen and Winkler (1985). It is derived by dividing both sides of Eq. 2 by the mean individual validity, or by taking the square root of Eq. 1:

$$\text{Efficacy} = \frac{\rho_{y\bar{x}}}{\bar{\rho}_{yx}} = J^{1/2}\left[1 + (J-1)\bar{\rho}_{x_i x_j}\right]^{(-1/2)} \tag{24}$$

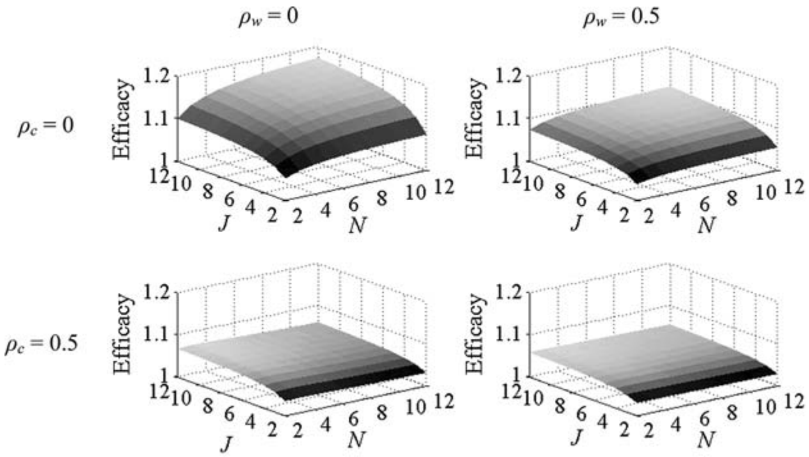


FIGURE 9.
Efficacy of a group of judges as a function of number of judges and number of cues (signal to noise ratio = 1/9 and σ_w^2 assumes weights proportional to ranks).

Using our model (Eq. 15) to predict the mean inter-judge correlation, $\bar{\rho}_{x_i x_j}$, we can compute efficacy as a function of the number of judges, J, and the number of cues, N. Figure 9 displays the efficacy under different combinations of inter-cue and inter-weight correlations (ρ_c and ρ_w, respectively). The noise to signal ratio was set to a relatively low value of $\delta = 1/9$, and σ_w^2 was computed using weights proportional to ranks (see Appendix A). Figure 9 shows that, in most cases, additional judges influence the efficacy more than additional cues. Additional cues can be slightly detrimental in the presence of high cue dependence, whereas additional judges are always monotonically increasing in efficacy. The effects of additional cues and/or additional judges are less pronounced for higher levels of inter-cue and inter-weight correlations.

The previous results assume that adding or dropping cues does not affect the other key parameters, especially the inter-cue correlation, ρ_c. In fact, one could choose new cues with the explicit intention of reducing the average inter-cue correlation. Thus, the relevant question becomes: How much must the average inter-cue correlation be lowered by the additional cue to be as efficacious as adding an extra judge? The efficacy of an additional judge can be computed and equated with the efficacy of an addition cue:

$$\text{Efficacy}\big\{(J+1), N, \bar{\rho}_c\big\} = \text{Efficacy}\big\{J, (N+1), \bar{\rho}_c^*\big\} \tag{25}$$

We treat $\bar{\rho}_c^*$ as an unknown variable and solve for the value of inter-cue correlation which makes the efficacy of an additional cue equal to an additional judge.

Table 2 presents the reduction in the inter-cue correlation ($\rho_c - \bar{\rho}_c^*$), that affects the inter-judge correlation to the same degree as the addition of a new expert. We use the same assumptions about the noise to signal ratio ($\delta = 1/9$) and cues weighted proportional to ranks (σ_w^2 computed as a function of N). In a few cases (bold typeface), the new inter-cue correlation is higher than the original one, i.e., adding a cue is more efficacious than adding a judge. Note that this can only happen when the cues are uncorrelated ($\rho_c = 0$). In the vast majority of the cases, the reduction in the inter-cue correlation must be quite extensive, especially when N and J are small and the inter-weight correlation is larger.

5.5. Modeling of Correlation of Baseball Expert Predictions

This section illustrates an application of the model to real data. We recruited 16 self-proclaimed baseball experts who participate in fantasy baseball leagues. All of them participated

TABLE 2.
The reduction in inter-cue correlation ($\rho_c - \bar{\rho}_c^*$) that equates the efficacy of an additional cue and an additional judge.

			$\rho_w = 0$						$\rho_w = 0.5$			
			N						N			
			2	3	4	5			2	3	4	5
$\rho_c = 0$	J	2	0.07	0.09	0.08	0.07	J	2	0.14	0.12	0.10	0.09
		3	**−0.05**	0.01	0.02	0.03		3	**−0.01**	0.04	0.04	0.04
		4	**−0.09**	**−0.01**	0.00	0.01		4	**−0.06**	0.00	0.01	0.02
		5	**−0.11**	**−0.03**	0.00	0.00		5	**−0.09**	**−0.02**	0.00	0.01
			N						N			
			2	3	4	5			2	3	4	5
$\rho_c = 0.5$	J	2	0.29	0.30	0.29	0.29	J	2	0.44	0.42	0.40	0.39
		3	0.11	0.15	0.15	0.15		3	0.20	0.22	0.22	0.23
		4	0.05	0.09	0.10	0.10		4	0.10	0.14	0.15	0.15
		5	0.02	0.06	0.07	0.07		5	0.05	0.10	0.10	0.11

in fantasy leagues in 2006 and 2007 (at least). Expertise was established in a previous study (Miller, 2008) where they answered a questionnaire pertaining to MLB knowledge. The average subject answered correctly 62% of the questions (SD = 18%).

Subjects were asked to predict the worth/ value of various players based on selected statistical information that we provided (cues). The experts were randomly assigned to 4 groups which differed in the number and nature of cues provided. The experts in Group 1 were provided $N = 5$ cues (statistics of baseball batters) which had an average inter-cue correlation of $\rho_c = 0.61$.[3] Groups 2 and 3 were provided only $N = 4$ of these cues. Cues were dropped such that the average inter-cue correlations would be reduced ($\rho_c = 0.49$ in group 2), or increased ($\rho_c = 0.74$ in group 3). Experts in Group 4 were given $N = 5$ different cues (statistics of baseball pitchers) which had an average inter-cue correlation of $\rho_c = 0.27$. Subjects were paid $15 for their participation.

In the first session, participants judged 40 players based on these cues. The names of the players were not provided to ensure that the subjects did not use additional private information. In the second session (a few days later), participants judged 40 players—20 new ones and 20 who were also shown in the first session. We used the 20 repeated judgments to assess the reliability and the noise to signal ratio, δ, for each expert. These ratios were quite low for most judges (Median = 0.26), especially in groups 1–3. The individual weights were inferred from their judgments using two different methods: standardized regression coefficients and general dominance statistics (Azen & Budescu, 2003). The estimated weights were highly similar and we only report results based on the regression coefficients. We used the normalized weights to estimate the variance of each expert, σ_w^2, and the inter-weight correlation, ρ_w, for each pair of judges. Table 3 presents the average inter-judge correlations observed in each group and the model's predictions. The model predicts correctly the ordering of the inter-expert correlations in the four cases (highest in Group 3 and lowest in Group 4), and the predictions are quite accurate (error in the 0.02–0.08 range) for both sets of weights.

[3] These correlations are based on the performance of all the MLB players over the last 5 seasons.

TABLE 3.
The model's prediction of the inter-judge correlation for the baseball example.

Group	Design parameters		Mean inter-judge correlation		
	$N =$ # of cues	$\rho_c =$ Inter-cue corr.	Predicted	Observed	Difference
1	5	0.61	0.72	0.66	0.05
2	4	0.49	0.85	0.77	0.08
3	4	0.74	0.92	0.88	0.05
4	5	0.27	0.60	0.61	−0.02

TABLE 4.
The model's prediction of efficacy in the baseball example.

Group	Design parameters		Mean efficacy		
	$N =$ # of cues	$J =$ # of judges	Predicted	Observed	Difference
1	5	4	1.12	1.16	0.04
2	4	5	1.07	1.11	0.04
3	4	4	1.03	1.05	0.02
4	5	3	1.17	1.08	−0.08

We compared the observed correlations between each pair of judges with the values predicted by the model. The Mean Absolute Deviation (MAD) between the correlations and the model's predictions across all 25 pairs is 0.08. The within group MADs are 0.07, 0.09, 0.04, and 0.11. The order of the MADs roughly matches the magnitude of the correlations in Table 3 indicating that the model predicted better higher correlations.

The validity was calculated for the 16 judges in predicting the players' values: The mean individual validity is an impressive 0.66 corresponding to a mean RI (Eq. 21) of 0.79 above the lower bound based on the cues' validity. Table 4 summarizes the efficacy (Eq. 25) in the four groups. The observed efficacy is the ratio of the validity of the mean forecast (across all J judges) and the average individual validity, and the predicted values are derived from the right-hand side of Eq. 24 (using Eq. 15 to calculate $\bar{\rho}_{x_i x_j}$). The predictions are very accurate, especially for groups 1–3.

The two tables illustrate nicely the interplay between adding judges and cues discussed in Section 5.4. Note that group 3 has $N = 4$ cues (with $\rho_c = 0.74$) and $J = 4$ judges. In group 2, we also have $N = 4$ cues (with a lower $\rho_c = 0.49$) and more judges ($J = 5$). On the other hand, in group 1, we have the same number of judges ($J = 4$), but more cues ($N = 5$ with $\rho_c = 0.61$). Adding either a judge (group 3) or a cue (group 1) reduced the mean inter-judge correlation (Table 3) and increased the efficacy of the process (Table 4). In this case, adding a cue had a slightly stronger effect. Overall, these results show that the model predicts quite well empirical results assessed in a way that matches closely the assumptions of the model. The discrepancies are due, in part, to the fact that the cues are not necessarily equicorrelated, especially in group 4.

6. Summary and Discussion

We have presented a quantification of Morris' (1986) and Winkler's (1981) analysis of the sources of inter-judge correlation. The model invokes a set of reasonable simplifying assumptions to describe the overt agreement between experts as a function of several easy to interpret parameters. We examined how these parameters interact to drive the observed correlations between experts. The model shows that ρ_c (inter-cue correlation), N (the number of cues), and ρ_w (inter-weight correlation), are positively associated with the inter-judge correlation, while σ_w^2 (a measure of cue differentiation) and δ (a measure of unreliability) are negatively associated with the inter-judge correlation. The model enabled us to evaluate analytically the impact of each parameter and to compare their relative influence on the inter-judge correlation.

6.1. *Summary of Analysis*

Relative Importance of the Factors The iso-influence curves in Figs. 3 and 4 show that in most cases the inter-cue correlation, ρ_c, dominates both ρ_w and σ_w^2. The region of the parameter space under which inter-cue correlation is the most important driver of inter-judge correlation includes most cases that are likely to be observed in real-life applications. The relative importance of ρ_c, compared to ρ_w and σ_w^2, increases as the numbers of cues increases.

Asymmetry Figure 5 shows that the overall impact of asymmetry between judges who vary in their qualifications and/or performance is not as strong as the other parameters. The effects of asymmetry are more pronounced for the case of uncorrelated cues and weights.

Validity The cues' (and consequently, the experts') validities restrict the range of the inter-judge correlations. Accounting for the contribution of validity reveals weak support for the argument that experts are highly correlated simply because they are highly valid. We show that validities which mimic the magnitude of empirical results exhibit the potential for a high RI (see Eq. 21). This confirms the importance of the model's parameters that capture our assumptions about the decision environment and the judges.

6.2. *On Expertise*

Throughout the paper, we have used the term "expert" without defining it formally. We are not alone in taking this approach since expertise is an elusive concept that defies simple definition. It is not our intention to offer a definitive characterization of expertise, nor to take sides in the debate on this topic. We simply seek to illustrate how our model of the environment and the judgment process, and the decomposition of the inter-judge correlation can contribute to the analysis of expertise.

Einhorn's (1974) analysis listed a set of necessary conditions for expertise. His framework and his assumptions about the sources of the inter-judge dependence (similarity in cues and weights) are quite similar to those in our model. Among the necessary conditions for expertise, Einhorn (1974) lists (a) high reliability/consistency that corresponds to a low noise to signal ratio (δ) in our model, (b) inter-expert agreement in the identification and grouping of cues, and (c) inter-judge similarity in the way cues are weighted and combined. Thus, he equates expertise (especially in cases where there is no external criterion) with inter-judge agreement.[4] Interestingly, recent empirical work (Budescu & Yu, 2007; Yaniv, Choshen-Hillel, & Milyavsky, 2009) suggests that most people share this intuition and they are highly confident in consensus opinions. The results of the current analysis (especially Section 5.3) suggest that not all these factors are equally important. While high reliability (low δ) is crucial, the similarity of the weighting schemes (ρ_w) play a much smaller role than the identification of the key cues and their correlations (ρ_c). Based on our analysis and on Dawes' (1979) results, we would argue that in most cases similarity in weighting is not a necessary condition for high inter-judge agreement and for expertise.

[4] He found large differences between the weighting schemes of expert pathologists reviewing biopsy slides.

Einhorn's emphasis on consensus is not universally accepted. Weiss and Shanteau (2003a) make a compelling case against this view, and argue that agreement with other experts is neither necessary, nor sufficient, for expertise. Weiss and Shanteau (2003b) advocate a definition of expertise that requires high discrimination between different cases and high consistency (reliability). Their empirical index of expertise, CWS, is calculated as the ratio of measures of discriminative ability and inconsistency, so experts, who have high discriminative ability and are highly consistent, have high CWS scores. Our analysis confirms that high reliability (low δ) is critical, but suggests that the CWS definition of discrimination—"as the stimulus changes, the evaluation changes accordingly"—may be too narrow. It certainly applies to cases where there are few and/or uncorrelated cues, but not cases with many highly correlated cues where the same final outcome, y, can be obtained for multiple values of a certain cue, say c_1, under various combinations of the other cues, $c_2, \ldots, c_N$. Moreover, the CWS index ignores the role of the external validity of the expert judges that was illustrated in Section 4.1.

6.3. *Final Remarks on Choosing Advisors*

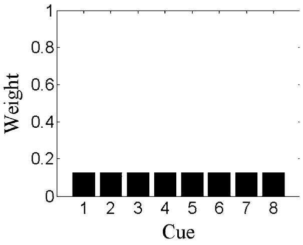
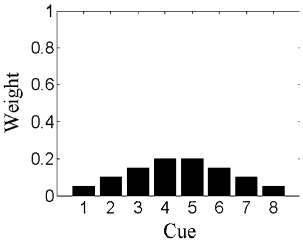
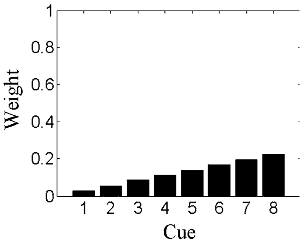
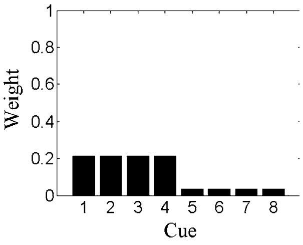
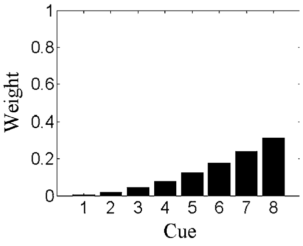
We conclude with a short discussion of the two critical questions in the area of advice giving and taking—which and how many, advisors to employ—in light of the results of our analysis. It is well understood that, everything else being equal, to maximize efficacy one should seek advisors whose opinions are valid but as uncorrelated with each other as possible. Low inter-judge correlations are associated with low inter-cue correlations, low inter-weight correlations, and unreliable judges with high levels of differentiation between the cues. Some of circumstances that reduce inter-judge correlations are clearly undesirable—no reasonable DM would seek unreliable experts (if the term can be even used in such cases) simply because they appear to be uncorrelated. Instead, we focus on the level of the dependence between the cues in nature, and more importantly, whether experts are using overlapping sets of cues, or different sets of cues.

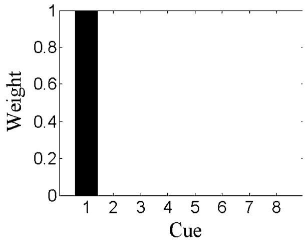
The predictability of the target variables varies considerably across domains (e.g., Shanteau, 2001; Weiss & Shanteau, 2003a). As we showed in Section 4.1, high levels of validity induce high lower bounds of the correlations between expert opinions and reduce the likelihood of finding uncorrelated experts. The analysis in Section 4.1 provides clear benchmarks for inter-judge correlations in various environments that are characterized by various levels of cue validity and inter-cue correlations. This can help a DM determine to what degree the inter-judge agreement is driven by the highly valid cues, or due to some of the other parameters. Consider, for example, the middle panel of Fig. 6, where the inter-cue correlation is $\rho_c = 0.5$, and imagine that two judges are moderately correlated, say around $\rho_{xx'} = 0.4$. This is an excellent situation in an environment with highly valid cues (e.g., $\rho_{yc} \geq 0.6$) that imposes high constraints on the judges. On the other hand, if the cues are not very valid (e.g., $\rho_{yc} \leq 0.2$), this suggests that the two judges are too similar, and that one should be able to find other, more informative advisors.

The question regarding the number of advisors is best answered through analysis of relative measures, such as our index of efficacy (Eq. 24), rather than absolute ones. Such indices ask how to achieve a desired level of efficacy (say twice, or thrice, as efficacious as a single "typical' advisor in the domain) at minimal cost, i.e., with the smallest number of advisors. The efficacy measure is a negative monotone function of the inter-judge correlations so, keeping everything else fixed, any factor that reduces the inter-judge correlation increases the efficacy, and reduces the number of advisors required to achieve a particular level of improvement in the performance of the group of advisors.

The analysis of efficacy presented in Section 5.4 also reveals new insight into how cues can be valued in comparison to additional advisors. The overarching theme of our analyses is that the underlying dependence in cues (the environment) plays the largest role in undermining the value of additional opinions. Using the current model in combination with the formulation of efficacy in Eq. 24 it is possible to assess the reduction of the average inter-cue correlation which is necessary for an additional cue to contribute more than an additional judge in terms of efficacy.

Appendix A

Weighting method	Description	Formula for variance	Weights	Value for $N = 8$
Uniform		0	$w_i = 1/N$	0
Triangular		$(N/2 - 1)/(3N(N - 1)(N/2 + 1))$	$w_i = 2i/(N(N + 1))$ $i = 1, \dots, i/2$	0.004
Ranked		$1/(3N(N + 1))$	$w_i = 2i/(N(N + 1))$	0.005
Two equal sets		$(1/(N - 1))((n_1 + n_2r^2)/((n_1 + n_2r)^2) - 1/N)$	$w_i = c, w_j = cr,$ $i \neq j$	0.009 $(r = 1/6)$
Squared ranked		$(8N + 11)/(5N(N + 1)(2N + 1))$	$w_i = [2i/(N(N + 1))]^2$	0.012

Weighting method	Description	Formula for variance	Weights	Value for $N = 8$
Single		$1/N$	$w_i = 1, w_j = 0,$ $j = 1, \ldots, i-1,$ $i+1, \ldots, N$	0.125

Note: Computational formulas for several weighting methods.

Appendix B

Derivation of the model of inter-judge correlation, Eq. 14 By the definition of the vector $\boldsymbol{w}_j$ in Eq. 4, the mean weight is

$$\bar{w} = \frac{1}{N}\boldsymbol{w}_j^{\mathrm{T}}\mathbb{1}_N = \frac{1}{N}$$

Using the definition of the unbiased variance estimator we obtain the following equality.

$$\sigma_w^2 = \frac{1}{(N-1)}\left(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_j - N(\bar{w})^2\right) = \frac{1}{(N-1)}\left(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_j - \frac{1}{N}\right)$$

Solving for $\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_j$, we can rewrite the above as

$$\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_j = (N-1)\sigma_w^2 + \frac{1}{N}$$

Using the same method, we can rewrite the covariance:

$$\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_{j'} = (N-1)\sigma_{ww'} + \frac{1}{N} = (N-1)\rho_w\sigma_w\sigma_{w'} + \frac{1}{N}$$

Note from Eqs. 11 and 13 that

$$\boldsymbol{\Sigma} = (1-\rho_c)\mathbf{I}_N + \rho_c\mathbb{1}_N\mathbb{1}_N^{\mathrm{T}}$$
$$\sigma_{x_j}^2 = \boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}_j(1+\delta_j)$$

Proof: Substituting the above into Eq. 7, we get

$$\begin{aligned}
\rho_{x_jx_{j'}} &= \frac{\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}_{j'}}{(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}_j + \sigma_{e_j}^2)^{1/2}(\boldsymbol{w}_{j'}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}_{j'} + \sigma_{e_{j'}}^2)^{1/2}} \\
&= \frac{\boldsymbol{w}_j^{\mathrm{T}}[(1-\rho_c)\mathbf{I}_N + \rho_c\mathbb{1}_N\mathbb{1}_N^{\mathrm{T}}]\boldsymbol{w}_{j'}}{(\boldsymbol{w}_j^{\mathrm{T}}[(1-\rho_c)\mathbf{I}_N + \rho_c\mathbb{1}_N\mathbb{1}_N^{\mathrm{T}}]\boldsymbol{w}_j(1+\delta_j))^{1/2}(\boldsymbol{w}_{j'}^{\mathrm{T}}[(1-\rho_c)\mathbf{I}_N + \rho_c\mathbb{1}_N\mathbb{1}_N^{\mathrm{T}}]\boldsymbol{w}_{j'}(1+\delta_{j'}))^{1/2}} \\
&= \frac{[(1-\rho_c)\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_{j'} + \rho_c]}{([(1-\rho_c)\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{w}_j + \rho_c](1+\delta_j))^{1/2}([(1-\rho_c)\boldsymbol{w}_{j'}^{\mathrm{T}}\boldsymbol{w}_{j'} + \rho_c](1+\delta_{j'}))^{1/2}}
\end{aligned}$$

$$= \frac{[(1-\rho_c)((N-1)\rho_w\sigma_w\sigma_{w'}+1/N)+\rho_c]}{([(1-\rho_c)((N-1)\sigma_w^2+1/N)](1+\delta_j))^{1/2}([(1-\rho_c)((N-1)\sigma_{w'}^2+1/N)+\rho_c](1+\delta_{j'}))^{1/2}}$$

This equation defines inter-judge correlation with the following constraints.

$$\rho_{x_j x_{j'}} = f\left(\rho_c, \rho_w, \sigma_w^2, N, \delta\right) \quad \text{for } 0 \le \rho_c \le 1,\ -1 \le \rho_w \le 1$$

$$0 \le \sigma_w^2 \le \frac{1}{N},\ 0 < N,\ \text{and } 0 \le \delta \qquad \square$$

Derivation of expert validity, Eq. 18 Assume the vectors of cues has included at the end the criterion variable y, Eq. 17,

$$\boldsymbol{c}_y^{\mathrm{T}} = \left[\boldsymbol{c}^{\mathrm{T}}, y\right], \quad \text{where } \boldsymbol{c}_y \sim (\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

The matrix $\boldsymbol{\Sigma}_y$ is of the same form as $\boldsymbol{\Sigma}$ in Eq. 11 with an added $(N+1)$th row and column which have the value ρ_{yc}. The weight vectors have an extra $(N+1)$th element added to the end as follows,

$$\boldsymbol{w}_j^{\mathrm{T}} = [w_{1j} \quad w_{2j} \quad \ldots \quad w_{Nj} \quad 0] \quad \text{and} \quad \boldsymbol{w}_y^{\mathrm{T}} = [0 \quad 0 \quad \ldots \quad 0 \quad 1]$$

Proof: This allows the expression of the covariance of the opinion, x_j, and the criterion, y,

$$\sigma_{x_j y} = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma}_y \boldsymbol{w}_y = \rho_{yc}$$

The variance of the criterion, y,

$$\sigma_y^2 = \boldsymbol{w}_y^{\mathrm{T}} \boldsymbol{\Sigma}_y \boldsymbol{w}_y = 1$$

The variance of the opinion, x_j,

$$\sigma_{x_j}^2 = \boldsymbol{w}_j^{\mathrm{T}} \boldsymbol{\Sigma}_y \boldsymbol{w}_j + \sigma_{e_j}^2 = \left[\rho_c + (1-\rho_c)\big((N-1)\sigma_w^2 + 1/N\big)\right](1+\delta_j)$$

Therefore, we get the correlation between x_j and y,

$$\rho_{x_j y} = \frac{\rho_{yc}}{\sigma_{x_j}} \qquad \square$$

References

Ariely, D., Au, W.T., Bender, R.H., Budescu, D.V., Dietz, C.B., Gu, H., Wallsten, T.S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*, 130–147.

Ashton, R.H. (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes*, *38*, 405–414.

Ashton, A.H., & Ashton, R.H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, *31*, 1499–1508.

Azen, R., & Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*, 129–148.

Budescu, D.V. (2006). Confidence in aggregation of opinions from multiple sources. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 327–354). Cambridge: Cambridge University Press.

Budescu, D.V., & Yu, H.T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*, 153–177.

Clemen, R.T., & Winkler, R.L. (1985). Limits for precision and value of information from dependent sources. *Operations Research*, *33*, 427–442.

Clemen, R.T., & Winkler, R.L. (1986). Combining economic forecasts. *Journal of Business and Economic Statistics*, *4*, 39–46.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*, 571–582.

Einhorn, H.J. (1974). Expert Judgment: Some necessary conditions and an example. *Journal of Applied Psychology*, *59*, 562–571.

Hammond, K.R., & Stewart, T.R. (2001). *The essential Brunswik: beginnings, explications, application*. London: Oxford University Press.

Hammond, K.R., Wilkins, M.M., & Todd, F.J. (1966). A research paradigm for the study of interpersonal learning. *Psychological Bulletin*, *65*, 221–232.

Hogarth, R.M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance*, *21*, 40–46.

Hursch, C.J., Hammond, K.R., & Hursch, J.L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review*, *71*, 42–60.

Johnson, T.R., Budescu, D.V., & Wallsten, T.S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making*, *14*, 123–140.

Miller, S. (2008). *Supporting joint human-computer judgment under uncertainty*. Unpublished Dissertation at the University of Illinois at Urbana-Champaign.

Morris, P.A. (1986). Comment on Genest and Zideck's "Combining probability distributions: A critique and annotated bibliography". *Statistical Science*, *1*, 141–144.

Shanteau, J. (2001). What does it mean when experts disagree? In E. Salas & G. Klein (Eds.), *Linking expertise and naturalistic decision making*. Earlbaum: Mahwa.

Schmidt, F.L., Johnson, R.H., & Gugel, J.F. (1978). Utility of policy capturing as an approach to graduate admissions decision making. *Applied Psychological Measurement*, *2*, 345–357.

Wallsten, T.S., Budescu, D.V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*, 243–268.

Wallsten, T.S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *18*, 1–18.

Weiss, D.J., & Shanteau, J. (2003a). The vice of consensus and the virtue of consistency. In J. Shanteau, P. Johnson, & C. Smith (Eds.), *Psychological explorations of competent decision making*. Cambridge: Cambridge University Press.

Weiss, D.J., & Shanteau, J. (2003b). Empirical assessment of expertise. *Human Factors*, *45*, 104–116.

Winkler, R.L. (1971). Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*, *66*, 675–685.

Winkler, R.L. (1981). Combining probability distributions from dependent information sources. *Management Science*, *27*, 479–488.

Winkler, R.L., & Poses, R.M. (1993). Evaluating and combining physician's probabilities of survival in an intensive care unit. *Management Science*, *39*, 1526–1543.

Yaniv, I., Choshen-Hillel, S., & Milyavsky, M. (2009). Spurious consensus and opinion revision: Why might people be more confident in their less accurate judgments? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 558–563.

Manuscript Received: 11 SEP 2008
Final Version Received: 10 DEC 2008
Published Online Date: 29 APR 2009