Journal of Experimental Psychology: Applied
2000, Vol. 6, No. 2, 130-147

1076-898X/00/$5.00 DOI: 10.1037//1076-898X.6.2.130

The Effects of Averaging Subjective Probability Estimates Between and Within Judges

Dan Ariely
Massachusetts Institute of Technology

Wing Tung Au
The Chinese University of Hong Kong

Randall H. Bender
Research Triangle Institute

David V. Budescu
University of Illinois at Urbana-Champaign

Christiane B. Dietz, Hongbin Gu, Thomas S. Wallsten
University of North Carolina

Gal Zauberman
Duke University

The average probability estimate of $J > 1$ judges is generally better than its components. Two studies test 3 predictions regarding averaging that follow from theorems based on a cognitive model of the judges and idealizations of the judgment situation. Prediction 1 is that the average of conditionally pairwise independent estimates will be highly diagnostic, and Prediction 2 is that the average of dependent estimates (differing only by independent error terms) may be well calibrated. Prediction 3 contrasts between- and within-subject averaging. Results demonstrate the predictions' robustness by showing the extent to which they hold as the information conditions depart from the ideal and as J increases. Practical consequences are that (a) substantial improvement can be obtained with as few as 2–6 judges and (b) the decision maker can estimate the nature of the expected improvement by considering the information conditions.

On many occasions, experts are required to provide decision makers or policymakers with subjective probability estimates of uncertain events (Morgan & Henrion, 1990). The extensive literature (e.g., Harvey, 1997; McClelland & Bolger, 1994) on the topic shows that in general, but with clear exceptions, subjective probability estimates are too extreme, implying overconfidence on the part of the judges. The theoretical challenge is to understand the conditions and the cognitive processes that lead to this overconfidence. The applied challenge is to figure out ways to obtain more realistic and useful estimates. The theoretical developments of Wallsten, Budescu, Erev, and Diederich (1997) provide one route to the applied goals, and they are the focus of this article.

Dan Ariely, School of Management, Massachusetts Institute of Technology, Boston, Massachusetts; Wing Tung Au, Department of Psychology, The Chinese University of Hong Kong, Hong Kong, China; Randall H. Bender, Statistics Research Division, Research Triangle Institute, Research Triangle Park, North Carolina; David V. Budescu, Department of Psychology, University of Illinois at Urbana-Champaign; Christiane B. Dietz, Hongbin Gu, and Thomas S. Wallsten, Department of Psychology, University of North Carolina; Gal Zauberman, Fuqua School of Business, Duke University.

The authorship is intentionally in alphabetical order; all authors contributed equally to this article. This research was supported by National Science Foundation Grants SBR-9632448 and SBR-9601281. We thank Peter Juslin and Anders Winman for generously sharing their data with us and Neil Bearden for comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Thomas S. Wallsten, Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27599-3270. Electronic mail may be sent to tom.wallsten@unc.edu.

Specifically, this research tests three predictions regarding the consequences of averaging multiple estimates that an event will occur or is true. The predictions follow from two theorems proposed by Wallsten et al. (1997) and proved rigorously by Wallsten and Diederich (in press). They are based on idealizations that are unlikely to hold in the real world. If, however, the conditions are approximated or if the predicted results are robust to departures from them, then the theorems are of considerable practical use.

We next provide a brief overview of background material and then develop the predictions in more detail. We test them by reanalyzing data collected for other purposes and with an original experiment. We defer discussion of the practical and theoretical consequences to the final section.

Researchers have studied subjective probability estimation in two types of tasks. In the no-choice full-scale task, respondents provide an estimate from 0 to 1 (or from 0% to 100%) that statements or forecasts are or will be true. In the other, perhaps more common task, choice half-scale, respondents select one of two answers to a question and then give confidence estimates from 0.5 to 1.0 (or 50% to 100%) that they are correct. Instructions in both the choice and nonchoice paradigms generally limit respondents to categorical probability estimates in multiples of 0.1 (or of 10). When judges are not restricted to categorical responses, the estimates generally are gathered for purposes of analysis into categories corresponding to such multiples. The graph of fraction correct in choice half-scale tasks or of statements that are true in no-choice full-scale tasks as a function of subjective probability category is called a *calibration curve*.

The most common finding in general-knowledge or forecasting domains is that probability estimates are too extreme, which is interpreted as indicating overconfidence on the part of the judge.

Specifically, in the choice half-scale paradigm, the mean subjective probability exceeds the fraction of correct answers. More generally, the fraction of correct choices within each subjective probability category is less than the category value, and the calibration curve falls below the diagonal. In the no-choice full-scale paradigm, the fraction of true statements within each response category is less extreme than the category value. That is, the calibration curve falls below the diagonal for estimates greater than 0.5 (as in the choice half-scale case) and above the diagonal for estimates less than 0.5, as illustrated in Figure 6, which portrays data that we discuss subsequently. To understand how this pattern reflects overconfidence, one must note that statements thought more likely to be true (estimates greater than 0.5) are not true as frequently as predicted, and those thought more likely to be false (estimates less than 0.5) are not false as frequently as predicted.[1]

Certain robust results argue against an unqualified generalization that probability estimates are always too extreme. One is that the judgments of some classes of well-practiced experts are relatively well calibrated. For example, calibration curves for precipitation probability estimates provided by U.S. National Weather Service forecasters fall close to the diagonal (Murphy & Winkler, 1977). The second result is the well-known hard–easy effect, whereby people tend to be overconfident in hard tasks and underconfident in easy ones (where hard and easy, of course, are relevant to an individual's level of expertise). Generally, people are overconfident in domains in which the proportion correct is below about .75 and underconfident in domains where it is above (Ferrell, 1994; McClelland & Bolger, 1994). Finally, in contrast to conceptual domains, there is some indication that underconfidence is the norm in perceptual judgments (e.g., Björkman, Juslin, & Winman, 1993). However, that conclusion is controversial (Baranski & Petrusic, 1994, 1999; Olsson & Winman, 1996) and not relevant to our concerns in this article.

Many explanations have been suggested for the observed patterns of judgment. These include the arguments by Gigerenzer, Hoffrage, and Kleinbölting (1991), Juslin (1994), Winman (1997), and others that observed patterns of overestimation are due to biases in how the items to be judged are selected. According to this perspective, subjective probability estimates are accurate when items are randomly sampled from suitably defined domains, thereby making them ecologically valid. Respondents appear to be overconfident when hard or tricky items (e.g., May, 1986) are oversampled (and presumably underconfident when they are undersampled). In contrast, Ferrell and McGoey (1980; see also Ferrell, 1994) have argued that patterns of overconfidence and underconfidence have their origin at the response selection stage of the judgment process and are due to respondents insufficiently adjusting their response criteria relative to the difficulty of the task. A third explanation, originating with the work of Koriat, Lichtenstein, and Fischhoff (1980), posits that the patterns are due to biased retrieval of information from memory. Finally, Erev, Wallsten, and Budescu (1994) suggest that trial-by-trial error in the judgment process may contribute to (or in the limit be fully responsible for) observed overconfidence (see also Pfeiffer, 1994). Considerably expanding the range of possible artifacts, Juslin, Winman, and Olsson (in press) have argued that the hard–easy effect may be driven almost exclusively by methods of measurement and analysis rather than by cognitive phenomena. The explanations are not mutually exclusive, and they are all controversial. For further discussions of them, see McClelland and Bolger (1994) and Harvey (1997).

Theory

The theory underlying this article grows from a generalization of the Erev et al. (1994) model (see also Budescu, Erev, & Wallsten, 1997). According to these authors, subjective probability estimates are perturbed by random trial-by-trial fluctuations that arise both when respondents form their covert opinions and when they translate them to overt estimates. Expressed formally (although in different notation than Erev et al., 1994, used), $R = g(X, E)$, where the judge's probability estimate, R, is an increasing monotonic function, g, of the base confidence in the item, X, perturbed (not necessarily additively) by error, E. X is a covert random variable that represents base confidence values arising from the judge's memory search for knowledge about the items in question. The randomness in this variable arises from the experimenter's (or the environment's) selection of the item to be judged and not from the memory operation itself. Thus, whatever the respondent's search strategy is, its error-free execution for item i results in a particular base confidence value, x_i. E is a random variable representing the stochastic aspects of both the memory search and the process of mapping its outcome to an overt response. This model is very general and consistent with virtually every other one in the literature. The random errors need not be additive, but their effect is one of reversion to the mean (Samuels, 1991) when calculating percentage correct or percentage true conditioned on the response category.[2] The effect may or may not cause the apparent overconfidence, but, at the very least, it contributes to its magnitude.[3]

Wallsten et al. (1997) generalized the model by allowing it to take different forms for each judge, j, $j = 1, \ldots, J$. That is,

$$R_j = g_j(X_j, E_j). \tag{1}$$

Thus, the distribution of covert confidence, X_j, may vary over judges, as may the error terms, E_j, the mapping function, g_j, to

[1] Harvey (1997), following Lichtenstein, Fischhoff, and Phillips (1982), referred to two types of overconfidence when respondents assess the probabilities that items are true. One is the type we refer to here; the other is when the calibration curve is everywhere below the diagonal, indicating a consistent overestimation in the likelihood that items are true. In our view, this is not overconfidence but rather a bias to call items true (see, for example, Gilbert, 1991, or Wallsten & González-Vallejo, 1994, for further discussion). The distinction is important to the extent that the two phenomena are mediated by different processes.

[2] Reversion to the mean is a weaker and more general condition than regression to the mean. Specifically, if X_1 and X_2 are positively (but not perfectly) related identically distributed standardized random variables with common mean μ, regression to the mean states that for all $c > \mu$, $\mu \leq E[X_2|X_1 = c] < c$, with the reverse inequality for $c < \mu$. Reversion to the mean states that for any c, $\mu \leq E[X_2|X_1 > c] < E[X_1|X_1 > c]$ and $\mu \geq E[X_2|X_1 < c] > E[X_1|X_1 < c]$. Regression to the mean implies reversion to the mean but not conversely. See Samuels (1991) and references therein for further details.

[3] Similarly, reversion to the mean also occurs when estimates are averaged conditioned on objective probability values, as is commonly done in revision of opinion research. The consequence is apparent underconfidence. This paradox was the focus of the original articles.

overt responses, R_j, and therefore the response distributions themselves.

Despite the generality of Equation 1, it does have testable predictions. One set concerns the consequences of averaging estimates per stimulus within- and between-subjects. Numerous studies (reviewed and discussed by Wallsten et al., 1997) suggest that mean estimates differ from those of individuals in one of two ways. Sometimes, the averages are better calibrated, and sometimes, they are more diagnostic (often called highly resolved, Yates, 1982). Estimates are highly resolved when, conditioned on true and false statements (correct or incorrect forecasts), they are reasonably well separated and therefore relatively predictive of the outcome.

We consider both better calibration and improved resolution to be improvements over the individual estimates, but not all authors agree. Ferrell (1994), for example, stated that such procedures "do not *improve* the calibration of individuals, they just *change* it in a systematic way. From the standpoint of decision analysis there is a serious problem if this is so" (p. 447). The systematic change to which Ferrell refers is toward underconfidence and therefore greater resolution. We disagree that this change is a problem and consider it instead as a virtue because outcome predictability is greatly enhanced. (With suitable modeling, one can recalibrate the group function if necessary.) In any case, the conditions leading to improved calibration versus improved diagnostic value are not well understood.

By making some very weak assumptions regarding Equation 1, Wallsten et al. (1997) showed that the effects of averaging the estimates depend crucially on how the covert confidence distributions of the judges relate to each other. The mathematical proofs are in Wallsten and Diederich (in press). Specifically, let J be the number of judges providing an estimate about item i. The authors showed that given four other assumptions,[4] if for all pairs of judges, j and j' ($j \neq j', j, j' = 1, \ldots, J$), X_j and $X_{j'}$ are independent, conditional on the state of the item (true or false), the mean of the J estimates becomes increasingly diagnostic of the statement's truth value as J increases. In other words, the probability that the statement is true approaches 1 or 0, according to whether the mean estimate is above or below 0.5. Expressed formally,

$$P(S_i = 1|M_J) \rightarrow \begin{cases} 1 & \text{if } M_J > 0.5 \\ 0 & \text{if } M_J < 0.5 \end{cases} \quad \text{as } J \rightarrow \infty, \qquad (2)$$

where S_i refers to the state of sentence i, which is either false (0) or true (1), J refers to the number of estimates being averaged, and M_J is that average. That is, the estimates become increasingly resolved and increasingly underconfident as the number of conditionally pairwise independent estimates increases. In fact, the theorem is much more general than stated here because it applies to any monotonic increasing transformation of the estimates.

In extreme contrast to pairwise independence, the estimates may come from the same base confidence and differ from each other only by independent random components. That is,

$$R_j = g_j(X, E_j). \qquad (3)$$

This model describes the case in which all judges have the same base confidence for a particular item, or, more realistically, the same judge gives replicated estimates at different points in time. Thus, for any item, $X = x$ for all judges (or all replications). In this case, estimates differ only due to the error component. Using only assumptions (a) and (b) of Footnote 4 and a straightforward application of the law of large numbers, Wallsten and Diederich (in press) showed that the mean of the multiple estimates converges to a mean expected value as J increases. That is,

$$M_J \rightarrow \frac{1}{J} \sum_{j=1}^{J} E(g_j(R_j)). \qquad (4)$$

Whether or not M_J, $J > 1$, is better calibrated than the estimates of $J = 1$ judges depends on the judges' collective expertise and the response functions, g_j. There is no mathematical guarantee one way or the other. Resolution will likely increase to some extent, simply because of the elimination of random scatter, but it will never become perfect as it will under the former model.

It has been long understood that the diagnostic value of multiple estimates decreases with the extent of their interdependence (e.g., Clemen, 1989; Clemen & Winkler, 1990; Ferrell, 1985; Hogarth, 1978). However, not commonly recognized has been that pooling multiple conditionally pairwise independent estimates maximizes their diagnostic value by fundamentally altering the shape of the calibration curve, whereas pooling estimates on the basis of common confidence levels may, but will not necessarily, improve resolution and calibration by reducing variability (but see Clemen & Winkler, 1990, and Ferrell, 1985, who make related points).

As Wallsten et al. (1997) indicated, the conditional pairwise independence assumption that is at the heart of the model embodied by Equation 1 is most likely to be met when judges base their estimates on distinct sources or interpretations of information. This condition, in turn, seems most applicable when judges are estimating probabilities regarding unique rather than aleatory (i.e., repeatable) events. Thus, it is in these cases that we expect the prediction implied by Equation 2 to hold. In contrast, the assumption of identical confidence per event that is the foundation of the model in Equation 3 and leads to Equation 4 is most likely to be met when judges are estimating probabilities of aleatory events on the basis of common relative-frequency information.

The assumptions leading to either model may, of course, not hold in the real world, and, therefore, their implications may not hold empirically. One may in particular question the conditional pairwise independence assumption necessary for the results in Equation 2. As a check, Wallsten et al. (1997) reanalyzed two previously published studies using full-response-scale paradigms, one by Wallsten, Budescu, and Zwick (1993), for which the model in Equation 1 was more likely to hold, and another by Erev and Wallsten (1993), for which the model in Equation 3 was more likely. Although conditional pairwise independence was violated to some extent in the first case, the mean estimates became increasingly diagnostic of a statement's truth value as the number of respondents contributing to the average increased from 1 to 21. In the second case, in contrast, the mean estimates showed im-

[4] Technically, the X_j are assumed to be discrete random variables with values, x_{jl}, $l = 1, \ldots, L_j$. That is, each judge can have a distinct number of covert confidence categories. The four other assumptions, then, are (a) the E_j are independent random variables with $E(E_j) = 0$ and σ_j^2; (b) the $f(R_j)$ are random variables with finite mean and finite variance; (c) the X_j are symmetric about their midpoints and the probabilities are equal for symmetric pairs (x_{jl}, x_{j,L_j+1}); and (d) the error distribution is such that the expected response given $X_j = x_{jl}$ is regressive, and the expected response distribution is symmetric around the midpoint of the response scale.

proved calibration but little improvement in resolution as the number of judges increased from 1 to 60.

We report reanalyses of data collected for investigation at the individual (i.e., $J = 1$) level as well as a new experiment to test three distinct predictions. The collective results paint a very coherent picture of how the degree of dependence among the judges affects the consequences of pooling multiple subjective probability estimates.

The reanalyses and the new experiment test the prediction from Equation 2, based on the model in Equation 1:

> *Prediction 1.* For the general-knowledge judgment task, indices of diagnostic value will improve substantially (to complete resolution in the limit), while calibration changes in the direction of underconfidence, as the number of estimates contributing to the group average increases.

The reanalyses, but not the new experiment, test the additional prediction based on the model in Equation 3:

> *Prediction 2.* For a task in which all respondents have the same information, indices of diagnostic value will improve to some degree, while those of calibration stabilize (perhaps, but not necessarily at better calibration), as the number of estimates contributing to the group average increases.

Finally, the new experiment, but not the reanalyses, tests the following prediction:

> *Prediction 3.* The mean estimates of two separate individuals regarding general knowledge statements are more diagnostic of an item's truth and show less overconfidence than the means of a single individual's replicated estimates collected at two different points in time.

The last prediction follows from the assumption that between-subject averaging is more likely to approach the conditional pairwise independence requirement of the model in Equation 1 than is within-subject averaging, in which the separate estimates are assumed to differ from each other only by an error component, as described in Equation 3.

Reanalyses

Prediction 1: Judgments of General-Knowledge Events

We tested Prediction 1 with the data originally published by Juslin (1994) and by Winman (1997), who used the same set of stimuli in their studies. Juslin's purpose was to compare judgment quality under a condition in which items were randomly selected versus one in which they were selected by other participants, whereas Winman's purpose was to compare the extent of the hindsight bias under the two item-selection conditions. We analyzed all the data (excluding the hindsight judgments in Winman's experiments) but present details only for the random-selection condition because the participant-selection procedure introduced considerations beyond our scope. We briefly summarize the latter results, however, in the *Discussion.* The 60 respondents in each condition include 20 from Juslin (1994), 20 from Winman's (1997) Experiment 1, and 20 from his Experiment 2; all were undergraduate university students in Uppsala, Sweden.

Method. Their stimuli consisted of 120 items in the half-range format for which respondents chose one of two alternatives as correct and gave a confidence of 50% (labeled *random*), 60%, 70%, 80%, 90%, or 100% (labeled *absolute certainty*). The items concerned six target variables: latitudes of national capitals ("Which city is further north?"), populations of national capitals ("Which city has a larger population?"), populations of countries ("Which country has a larger population?"), mean life expectancy ("In which country does the population have a higher mean life expectancy?"), area ("Which country has a larger area?"), and population density ("Which country has more inhabitants per km^2?"). The experiments were all computer controlled.

The stimuli in the random condition were constructed by randomly sampling 20 pairs of countries for each of the six target variables from the 13,366 possible pairs of 164 countries then in existence. Those in the informal condition were selected by 12 volunteer participants working in pairs. Each participant pair selected 20 pairs of countries to create questions for a single target variable. Their instructions said, in part,

> The items should be good general knowledge items. That is, the items should provide a test of the knowledge of the subjects, and in a general sense conform to your own standards for what is a good general knowledge item. (Juslin, 1994, p. 236)

Each pair of selectors was given suitable statistical tables and a world atlas. For more details, see Juslin (1994) and Winman (1997).

For our purposes, we first converted each half-scale estimate to two full-scale estimates, one each for the implied true and false statement, by assuming additivity and doing the appropriate subtraction.[5] For example, assume "A" is the correct answer to the question, "Which city is further north, A or B?" If a respondent selected "B" with confidence 70%, he or she was credited with estimates of 70% in the false statement, "City B is further north than City A," and of 30% in the true statement, "City A is further north than City B." Thus, the 120 half-scale estimates per respondent yielded 240 complementary full-scale estimates.

Results. Here, we present the results only of the random-selection condition. Prior to testing Prediction 1, we assessed conditional pairwise independence by taking all possible pairs of the respondents and, for each pair, calculating the linear correlation between their estimates of the true statements.[6] (Because the estimates of the false statements are the complements of those for the true ones, the results apply to these as well.) The correlations (i.e., the inverses of the mean Fisher Z transformations) ranged from .05 to .70. Their interquartile interval is bounded by .32 and .48, and their mean is .40. Conditional pairwise independence is clearly violated.

Nevertheless, we looked at the effect of averaging probability estimates for $J = 3$, 6, 12, 20, 30, and 60 judges. Figure 1 displays the calibration curves, including that for individual respondents ($J = 1$). First, consider the $J = 1$ curve, which shows the average proportion of statements that are true conditioned on the probabil-

[5] Justification for assuming additivity comes from Wallsten et al. (1993) and the new experiment that follows. In both cases, respondents saw true and false versions of the same items at different points in time and gave their subjective probabilities that each was true. Estimates of complementary events summed to 1.02 in both studies. This value was not significantly different from 1.00 in the case of Wallsten et al. with $N = 21$ but was in the current study with $N = 64$. Regardless, the deviation from perfect additivity is very small.

[6] It is important to note that a linear correlation of 0 (or for a sample of data, a linear correlation not significantly different from 0) is necessary but not sufficient for pairwise independence. That is, if two variables are stochastically independent, their linear correlation will be 0, but nonlinear dependencies may also yield no linear correlations. As there are an infinite number of such dependencies, one cannot look for them without a guiding theory.

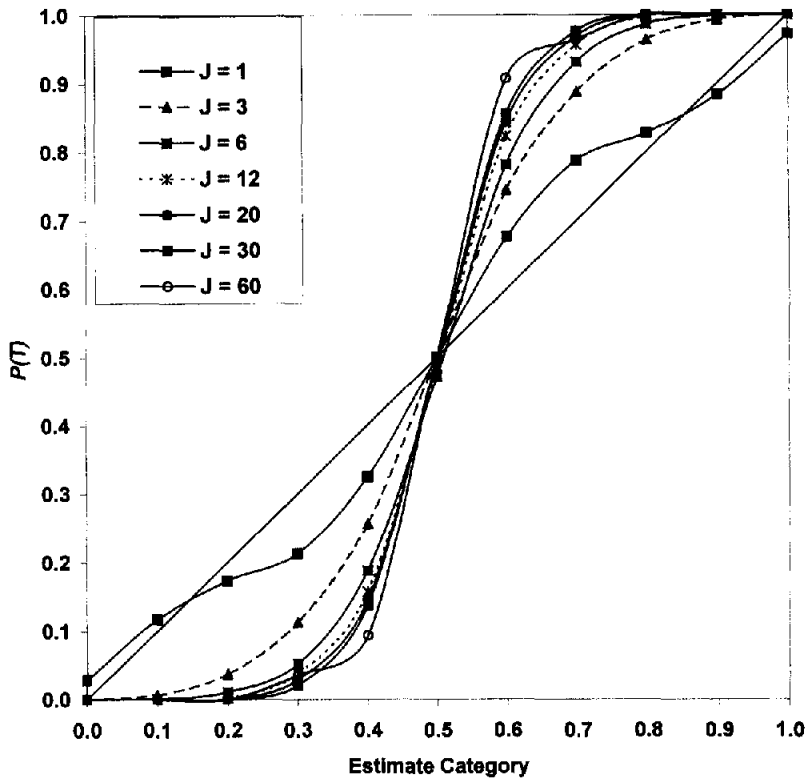

Figure 1. Calibration curves from the random condition of Juslin (1994) and Winman (1997). The abscissa is the mean estimate over J respondents per statement, with data gathered into categories that are multiples of 0.1. The ordinate is the proportion of statements per category that are true.

ity estimates assigned to them (confidence expressed as a percentage divided by 100), that is, $P(T)$ versus estimate category. This curve is the average of the 60 individual calibration plots and corresponds to the random-selection curve in Juslin (1994), except that it includes Winman's (1997) data and is reflected about 0.5. As Juslin (1994) and Winman (1997) pointed out (and expected) in their articles, respondents in this condition are reasonably well calibrated. Our analysis focuses on the calibration and diagnostic properties of the estimates as the responses from increasing numbers of judges are averaged together.

There are many ways to average multiple probability estimates per stimulus, and we used two. One is simply to average the estimates, r, and the other is to convert the estimates to log-odds, $\log(r/(1 - r))$, (after changing $r = 0$ and 1 to $r = .02$ and .98, respectively), average the log-odds, and then convert the result back to the original scale. Both methods lead to essentially the same results, and we present only the former. For each J greater than 1 and less than 60, the respondents were randomly gathered into subgroups of size J. Thus, for $J = 3$, there were 20 subgroups, and, in general, there were $60/J$ subgroups. We repeated the process 30 times, yielding a total of $900/J$ subgroups for each J, $1 < J < 60$. For each subgroup, we averaged the estimates of its members to each of the 120 statements and gathered the mean estimates into 11 categories corresponding to the original scale, that is, into categories with boundaries of [0, .045), [.045, .145), . . . , [.845, .945), [.945, 1]. The process was identical for $J = 60$, except that all respondents were included in a single group. Finally, we determined the proportion of true statements, $P(T)$, in each response category for each J. The results are shown as the remaining curves in Figure 1. It is important to note that all respondents contributed equally to all the curves. What differs across the conditions is how many respondents contributed to a subgroup average.

Despite the clear violation of conditional pairwise independence, the results appear to support Prediction 1 that the mean estimates become increasingly diagnostic of the true state of the item as the number of judges, J, increases. Simultaneously, as expected, calibration worsens in the direction of underconfidence. Table 1 shows four measures common in the judgment literature, which are helpful in quantifying these effects. They are as follows:

1. An overall index of quality, the well-known mean probability, or Brier, score,

Table 1
Mean Values for the Mean Probability Score ($\overline{PS}$), Calibration Index (CI), and Two Indices of Resolution (DI and DI') for the Data Summarized in Figure 1

J	$\overline{PS}$	CI	DI	DI'
1	.174	.009	.085	1.35
3	.153	.022	.119	1.78
6	.148	.034	.136	1.96
12	.146	.042	.146	2.07
20	.144	.046	.153	2.13
30	.144	.048	.154	2.15
60	.142	.056	.164	2.18

$$\overline{PS} = \frac{1}{N} \sum_{i=1}^{N} (r_i - f_i)^2,$$

where r_i is a probability estimate in [0, 1] that statement i is true, $f_i = 0$ for false statements and 1 for true ones, and N is the number of statements judged. ($N = 240$ in this study.) $\overline{PS}$ varies from 0 to 1, with lower scores being better.

2. An index of calibration,

$$CI = \frac{1}{N} \sum_{k=1}^{K} N_k(r_k - \bar{f}_k)^2,$$

where k indexes the response category, K is the number of categories ($K = 11$ here), N_k is the number of responses in category k, and $\bar{f}_k$ is the proportion of true statements in category k. CI is simply the weighted-mean squared deviation of the points in a calibration curve from the diagonal. Lower values are better, in that they indicate better calibration.

3. An index of discrimination, or resolution,

$$DI = \frac{1}{N} \sum_{k=1}^{K} N_k(\bar{f}_k - \bar{f})^2,$$

where $\bar{f}$ is the overall proportion of true statements. (In this study, $\bar{f} = 0.5$.) It is important to note that DI is the variance of the proportions of true statements conditional on response category. Higher values are better, in that they indicate greater separation among the categories, and therefore greater diagnostic value of an estimate. DI is bounded, $0 \leq DI \leq \bar{f}(1 - \bar{f}) = .25$ (Yaniv, Yates, & Smith, 1991).

4. An alternative index of discrimination,

$$DI' = \frac{\bar{r}_T - \bar{r}_F}{s_r},$$

where $\bar{r}_T$ and $\bar{r}_F$ are the means of the estimates accorded the true and false statements, respectively, and s_r is the pooled standard deviation of the two distributions of estimates. In contrast to the previous index, which is conditioned on response category, this signal-detection-like measure is conditioned on the state of the stimulus. Larger values are better.

Yates (1982) has extensively discussed Indices 1–3, including that fact that they are related through Murphy's decomposition, $\overline{PS} = \bar{f}(1 - \bar{f}) + CI - DI$. Index 4 was introduced and discussed by Wallsten et al. (1997).

The data in Table 1 quantify what is apparent in Figure 1. Calibration, CI, worsens and resolution, DI or DI', improves with J. The balance is such that the overall mean probability score, $\overline{PS}$, improves somewhat as J increases. The reason behind the increased resolution is best understood by considering DI', which grows from 1.35 at $J = 1$ to 2.18 at $J = 60$. In a manner analogous to d' in signal-detection theory, DI' indexes the extent of overlap between two distributions. The distributions in this case are of the (mean) estimates conditioned on true and false statements. Thus, resolution improves because the two response distributions become increasingly separated, as Ferrell (1994) noticed in his investigation of the effects of pooling judgments.

Figure 2 shows the conditional response distributions for se-

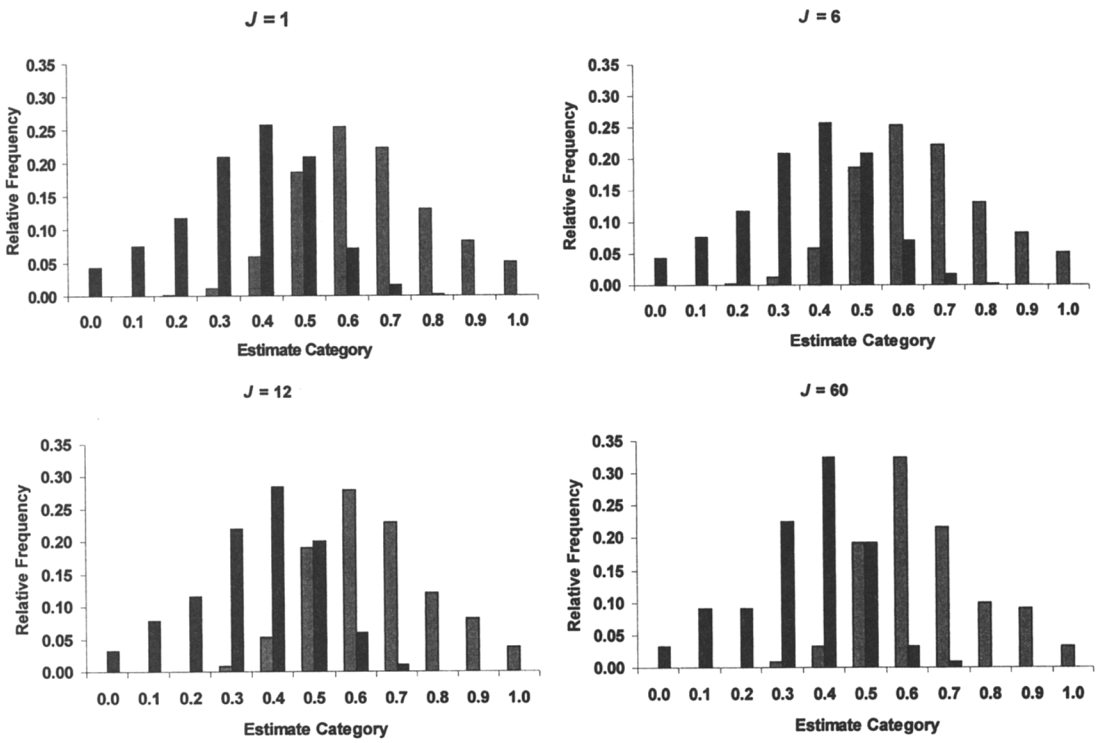

Figure 2. Response distributions conditioned on the true and false versions of the items for selected values of J in the random condition of Juslin (1994) and Winman (1997).

lected values of J from 1 to 60. It is apparent that the increase in DI' (decrease in distribution overlap) as J increases is due entirely to the decreasing variances of the two distributions. The standard deviations of the distributions—the "scatter" in Yates's (1982) terms—decrease from .218 at $J = 1$ to .138 at $J = 60$. This is the expected result under Wallsten and Diederich's (in press) Theorem 2. Their proof makes clear that under the axioms, as J increases, the expected mean estimate for true statements converges on an (unspecified) probability, r ($r > .5$), and the mean for false statements converges on $1 - r$. Convergence in the present data is not complete probably because of the violations of conditional pairwise independence and possibly also to the finite number of judges.

Discussion. The results under the random item-selection condition extend those of Wallsten et al. (1993) and provide strong confirmation of Prediction 1. Despite the violation of conditional pairwise independence, the mean subjective probability estimates of multiple judges rapidly became very diagnostic of the verity of a statement. The results are summarized and easily related to others to be discussed with the aid of Table 2, which shows the percentage of statements that were true (or false) given a probability estimate greater than (or less than) 0.5 for selected values of J. The columns are arranged according to the mean between-subject conditional pairwise correlation. Therefore, the random condition is in the fourth column. We see that the percentage is about 82% for $J = 1$. With as few as 6 judges, it increases to about 90% and with $J = 60$ to almost 95%.

For comparison purposes, Table 2 also shows selected results obtained by Johnson, Budescu, and Wallsten (in press) in Monte Carlo analyses of the effects of averaging probability estimates. Conditions in their simulations guaranteed that the axioms necessary for Wallsten and Diederich's (in press) theorem were met except where explicitly and systematically violated. Two columns in Table 2 summarize Johnson et al.'s results for the conditions in which pairwise correlations were .30 and .60, respectively. These two conditions bracket the mean pairwise correlation values of all the data to be presented in this article. It is important to note that the percentage under Johnson et al. for $J = 1$ is somewhat less than that in the random condition, suggesting that the simulation was operating at a slightly greater difficulty level. For each subsequent level of J (to $J = 32$, the upper bound of the simulation), the percentage at $r = .30$ is very close to that in the random condition, whereas for $r = .60$, it is less. Thus, the present real data mesh very well with the simulated data generated under known conditions, which gives us some confidence that the results are replicable.

Turning, for a moment, to the participant-selection condition, the effects of averaging differed somewhat from those observed in the random condition. Specifically, the linear pairwise correlations were smaller (in fact, their mean was .30, identical to that in Johnson et al., in press, as shown in column 1 of Table 2), yet the mean probability estimates converged more slowly with J. A summary of these analyses appears in column 2 of Table 2. For a single judge, about 65% of the items were true (or false) given an estimate above (below) 0.5. With $J = 60$, this increased only to just under 75%. One might speculate that this pattern occurred simply because the selected questions were more difficult, and, therefore, the conditional response distributions overlapped to a greater extent. Therefore, more than the 60 judges available are required to achieve the theoretical asymptotic result. This explanation is unlikely to be correct because the calibration curve and associated statistics reached apparently asymptotic values at least by $J = 30$.

It is more likely that the very process of selecting items to construct a test of individuals' knowledge caused conditional pairwise independence to be violated in a systematic fashion not captured by the linear correlation coefficient. Whatever the detailed cause, we believe that the random condition is closer to the real situations to which one wants to generalize the results than is the participant-selection one. That is, real-world judgment issues arise in ways not intended to test the limits of one's knowledge but rather as the consequence of legitimate uncertainties that arise in the course of making decisions. There is nothing inherently present in these situations to trick or test people's limits, as there often is in constructed tests. This conclusion is supported by the fact that the results in the random condition are consistent with those of Wallsten et al. (1997). The study they reanalyzed (Wallsten et al., 1993) used a large number of items constructed to span a broad range of difficulty.

The general conclusion, then, is that Prediction 1 clearly holds in the face of substantial violation of conditional pairwise independence when items are randomly selected. It does not hold to the same degree when the items are selected with a view toward test construction. In either case, averaging multiple judgments yields improved forecasts, which are considerably more diagnostic than

Table 2
Percentage of Statements That Were True (or False) Given Probability Estimates (at $J = 1$) or Mean Probability Estimates (at All Other J) Greater (or Less) Than 0.5, for the Indicated Studies, Arranged in Increasing Order of Mean Pairwise Correlations

J[a]	J & W participant ($\bar{r} = .30$)	Johnson et al. (in press; $\bar{r} = .30$)	J & W random ($\bar{r} = .40$)	New experiment ($\bar{r} = .57$)	Johnson et al. (in press; $r = .60$)
1	65.1	73	81.8	64.9	72
6 or 8	70.7	88	89.9	68.7	82
10 or 12	73.1		91.8		
16		91		69.2	83
30 or 32	74.7	93	93.3	69.6	84
60 or 64	74.7		94.8	70.7	

Note. J & W refers to the data from Juslin's (1994) and Winman's (1997) experiments.
[a] When two values of J are given, one value was used in some of the studies, and the other was used in the remaining.

individual ones in the random case and somewhat more in the participant-selection one. In both cases, as few as 6 judges yield substantial improvement, and 12 are sufficient for achieving close to the maximum possible under the particular conditions.

Prediction 2: Judgments Given Identical Information Conditions Across Respondents

We evaluated Prediction 2 with data collected by Wallsten and Gu (1996). Their primary purpose was to evaluate the claims of Erev et al. (1994) that trial-by-trial stochastic error contributes to the discrepancies that arise when analyses of judgment data conditioned on responses (as is common in calibration research) are compared with those conditioned on objective probabilities (as is common in revision of opinion research). As the stochastic contributions to the probability estimates decrease, the two analyses should converge on a single conclusion (of overconfidence or underconfidence). To check this prediction, it was necessary to collect probability estimates in a situation in which objective probabilities are well defined and in which the magnitude of the stochastic component to the judgment process can be estimated. Of interest for present purposes is that all the respondents in that experiment had the same training and then saw the same stimuli. Therefore, the conditions leading to the model in Equation 3 are perfectly met.

Method. Respondents were instructed to imagine that skeletons were discovered at an archaeological site. In Session 1, they learned how to distinguish men from women based on the density of a bone substance indicated by a five-digit number. The densities were represented by two equal-variance normal distributions of five-digit numbers, separated by one standard distribution ($d' = 1$). Subsequently, in Sessions 2 and 3, the respondents saw individual density values in three replicated blocks of 150 trials each and gave probability estimates that the skeleton was of a man. They were told that although the task was hypothetical, the distributions were real, a priori half the skeletons were women and half were men, and there was a correct answer on each trial. Payment was contingent on performance. By using external distributions (Kubovy, Rapoport, & Tversky, 1971) of known discriminability in this fashion, it is possible to use Bayes's rule to calculate the objective posterior probabilities for comparison with the estimates. All respondents saw the same samples, although not in the same sequence. There were, however, two between-subject manipulations: Half the respondents selected their estimates from the 11 categorical values, .02, .10, .20, . . . , .90, and .98, whereas the other half were unrestricted, in that they could use any multiple of .01 from .02 to .98. Crossed with this manipulation, half the respondents were told their cumulative earnings every 150 trials, whereas the other half were not. With 10 participants per group, there were a total of 40 respondents in the experiment. For more details, see Wallsten and Gu, 1996. For our purposes, we averaged together up to 20 estimates per stimulus within each of the categorical and unrestricted response conditions.

Results. The model in Equation 3 implies that the mean estimate to a stimulus converges on an expected value as the number of contributing judges grows larger. Because the model is stated in terms of the mean estimate per stimulus (rather than in terms of percentage correct per mean estimate), the appropriate way to view the data is to plot the mean estimate (mean *SP*) as a function of the objective stimulus probability (*OP*). The axes of this reliability graph, thus, are reversed from those of calibration curves such as in Figure 1. Moreover, it is not necessary to aggregate observations within a probability interval, as it is with calibration curves. Rather, each stimulus is shown as a separate point. Figure 3 illustrates the pattern of results as J increases. When the data are summarized in this fashion, one can see the effects of averaging over J judges only by comparing the plots of multiple respondents with those of multiple groups of various sizes.

The panel denoted $J = 1$ in Figure 3 shows *SP* versus *OP* for 10 of the 20 respondents. Only 10 respondents are displayed so that the individual points can be discerned to some degree. Even so, the points overlap considerably. The diagonal provides a reference of perfect calibration. For the $J = 5$ panel, we formed 10 groups of 5 judges each by randomly partitioning the 20 respondents into 4 mutually exclusive and exhaustive sets of 5 each, doing so on 3 occasions (yielding 12 subgroups) and dropping the last 2 subgroups. The plot shows the means of the 5 respondents' estimates per stimulus, with a different symbol for each of the 10 groups. Again, there is much overlap among the groups. For the $J = 10$ panel, we took 5 random partitions of the full set, each time into 2 mutually exclusive and exhaustive subsets of 10 respondents each. The plot shows the mean estimate per stimulus, with a different symbol for each group and considerable overlap among the groups. Finally, the $J = 20$ panel shows the mean estimate per stimulus for all 20 respondents.

In contrast to the results of Figure 1, and consistent with Prediction 2, Figure 3 shows that the reliability curve stabilizes as J increases without fundamentally changing shape. In fact, the mean estimates stabilize on the diagonal of perfect calibration as J increases from 1 to 20.

To provide a close comparison of these results with those from the data of Juslin (1994) and Winman (1997), we calculated the indices $\overline{PS}$, *CI*, *DI*, and *DI'* exactly as we had for Table 1. That is, for $J = 1$, we partitioned the *SP* estimates of each of the 20 respondents (not just the 10 shown in the first panel of Figure 3) into categories with boundaries of [.02, .045), [.045, .145), . . . , [.845, .945), [.945, .98], took the mean *SP* and mean *OP* per category, and calculated the four indices. The mean results are shown in the first row of Table 3. Similarly, we formed many groups of $J = 5$ and $J = 10$, calculated the indices in the same fashion, and have displayed the results in the next two rows of Table 3. The last row shows the results for $J = 20$.

It is important to note the excellent *CI* values for all levels of J, despite the considerable scatter around the diagonals in the plots. This result occurs as a consequence of averaging *SP* estimates within response categories before calculating the various indices, thereby eliminating much of the noise apparent in the figure. From this perspective, therefore, calibration is excellent at the individual level and continues to be so as J increases. The more fine-grained results shown in the figure tell a different story, however, in line with Prediction 2. That is, individual estimates per stimulus vary considerably, but converge on mean values, which themselves bear an orderly relationship to the objective probabilities of the stimuli.

Prediction 2 allows resolution to improve somewhat as scatter is reduced but not to the extent expected under the model from Equation 1. The indices *DI* or *DI'* show that this is precisely what happens. As J increases, resolution comes close to the maximum allowed by the structure of the task (recall that $d' = 1$ for the external distributions), but it never achieves the level obtained with Juslin's (1994) and Winman's (1997) random condition (see Figure 1 and Table 1).

The pattern of changes in *DI* and *DI'* with J in the present data is illuminated by the conditional response distributions, which are shown in Figure 4. As before, resolution improves because the two conditional response distributions separate as J increases. The

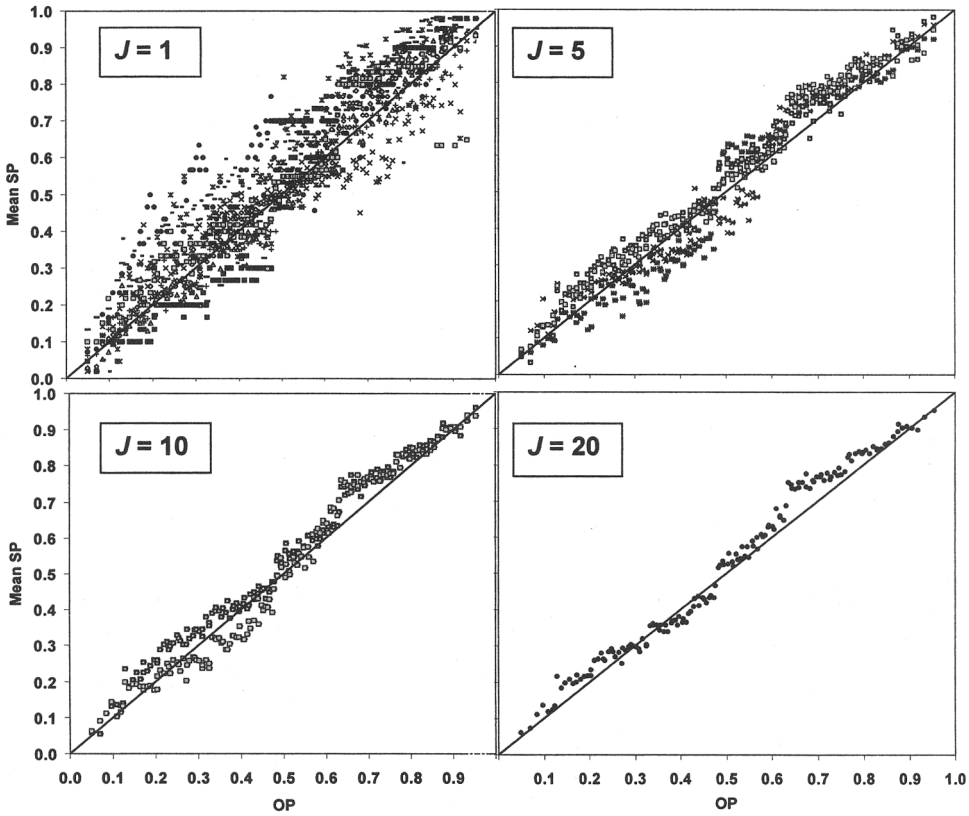

Figure 3. Calibration plots for 10 respondents ($J = 1$), for the means of 10 groups of respondents ($J = 5, 10$), and for the mean of all respondents ($J = 20$) in the unrestricted condition of Test Session 1 from Wallsten and Gu (1996). Each respondent or group of respondents in the $J = 1, 5, 10$ panels is shown by a different symbol. Points in many cases are superimposed. SP = subjective probability; OP = objective probability.

improvement in *DI'* is modest because the pooled standard deviations decrease only slightly from .269 at $J = 1$ to .224 at $J = 20$. *DI* improves more substantially with *J* because the distributions become considerably more peaked.

Discussion. The analyses completely support Prediction 2. Increases in resolution and calibration are due solely to reduction of variability about mean values rather than to changes in the shape of the reliability functions with increasing *J*. One should not overgeneralize from the excellent calibration displayed in this task because others have obtained different results (Kubovy et al., 1971; DuCharme & Peterson, 1968) under somewhat different conditions. Our point is not to explore determinants of good calibration but simply to emphasize that the good calibration observed here is not a necessary consequence of averaging multiple estimates contingent on a common information.

Table 3
Mean Values for the Mean Probability Score ($\overline{PS}$), Calibration Index (CI), and Two Indices of Resolution (DI and DI') for the Data Summarized in Figure 3

J	$\overline{PS}$	*CI*	*DI*	*DI'*
1	.246	.000	.005	.945
5	.226	.001	.024	.978
10	.205	.001	.045	.985
20	.149	.001	.102	.988

The pattern of results from both studies is clear and informative. Predictions 1 and 2 were sustained under conditions in which the items to be judged were randomly selected from the full population, according to a uniform distribution in the first case and to the operative probability distributions in the second. The differences in the predictions and in the supporting data are seen clearly by comparing Figures 1 and 3. Prediction 1 concerns the probability that an item is true given its estimate. Accordingly, the calibration curve plots $P(T)$ as a function of the estimate category. Prediction 2 concerns the mean estimate per stimulus. Here, assuming the stimuli have relative-frequency-based posterior probabilities, the calibration curve plots *SP* as a function of *OP*.

That said, Prediction 1 dictates that the calibration curve literally should change shape as *J* increases. Regardless of the degree of calibration at $J = 1$, as long as the curve is monotonic increasing, it should demonstrate increasing underconfidence and increasing resolution as *J* increases. If the axioms leading to Equation 2 are satisfied, the curve should asymptote at perfect resolution and correspondingly extreme underconfidence. Prediction 2, in contrast, states that calibration, as indexed by reliability, increases with *J*, but that the functional relationship between mean *SP* and *OP* will not change. The necessary conditions here are much weaker than in the first case, amounting only to all judges having

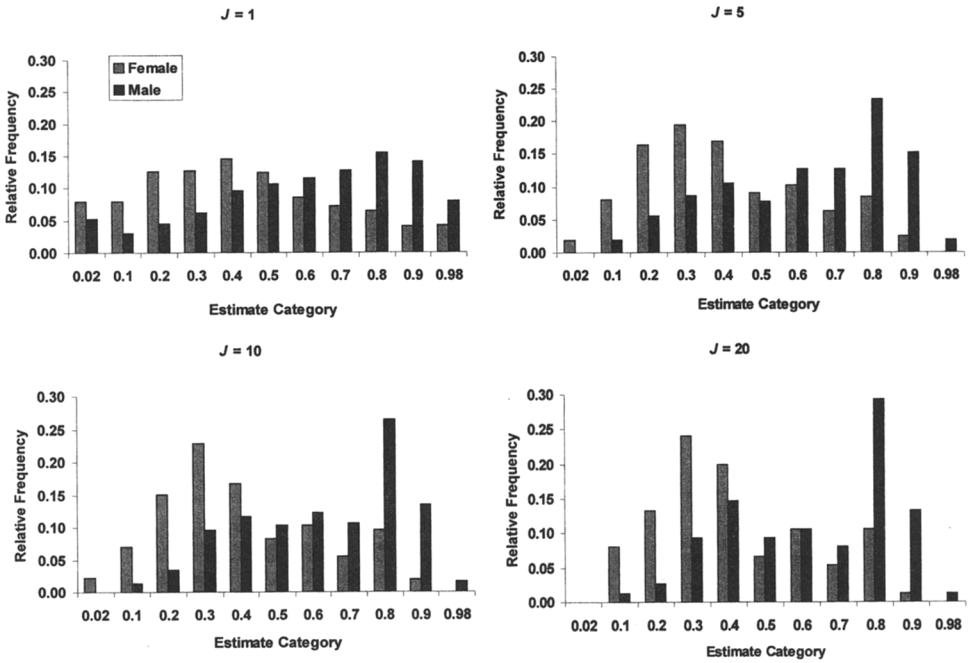

Figure 4. Response distributions conditioned on male and female samples for four levels of *J*.

the same information base, causing their estimates to differ at most by (not necessarily additive) error and unique response functions.

The conditions leading to Prediction 2 were clearly met and the prediction was perfectly sustained. Of the axioms necessary for Prediction 1 (see Footnote 4), conditional pairwise independence is the most likely to be violated, and, in fact, it was with a mean pairwise linear correlation of .40 (in the random condition). Importantly, the impact of this violation was not to nullify the prediction but to limit the asymptotic outcome to somewhat less than full resolution (95% rather than 100% of the items being true or false given a mean estimate greater than or less than 0.5). The Johnson et al. (in press) simulations confirm this result.

A New Experiment

The previous analyses relied entirely on data collected for other purposes. The present experiment was designed specifically to test Predictions 1 and 3 under varied stimulus and response conditions to assure their generality. Recall that Prediction 3 concerned the relative consequences of averaging subjective probability estimates within and between individuals under conditions of epistemic uncertainty. Assuming that conditional pairwise independence is more strongly violated within than between respondents, the diagnostic quality of the resulting means should be less in the former than the latter case.

Instead of the choice half-scale procedure of Juslin (1994) and Winman (1997), we used a no-choice full-scale procedure, in which respondents gave confidence estimates that statements of fact were true rather than false. We tested generality across response scales by having half the participants use categorical responses (0%, 10%, . . . , 90%, 100%) and half use essentially continuous responses (0%, 1%, . . . , 99%, 100%). The stimuli were either single statements or syntactically identical complementary pairs of statements, one written above the other. For example, a single statement might have been "In 1992 the population of Albuquerque was greater than that of Cleveland." A complementary pair might have been the previous example along with, "In 1992 the population of Cleveland exceeded that of Albuquerque" written below it. These manipulations were exploratory, designed to determine whether they affected within- or between-subject variability in any way and thereby affected the response distributions or the consequences of averaging.

The statements to be judged concerned the relative populations of the 50 largest cities in the United States in 1992. Consistent with Juslin (1994), Winman (1997), and our own recent work (e.g., Wallsten & González-Vallejo, 1994), we sampled stimuli randomly from the set of all possible pairs of the these cities. To anticipate one result, in contrast to Juslin and Winman, our respondents demonstrated considerable overconfidence. Budescu, Wallsten, and Au (1997) used a portion of the data from the present study to illustrate a procedure for assessing calibration following correction for trial-by-trial within-subject error. They showed most respondents to be overconfident even following this correction. Although they used a subset of the data presented here, there is virtually no overlap in the former and the present data analyses. We summarize their analyses and relate them to the present ones in the discussion.

Method

Participants. Respondents were 64 volunteers from the University of North Carolina, Chapel Hill community and were paid according to performance, with a minimum of $4 for an approximately 1-h session.

Between-subject design. Sixteen individuals served in each cell formed by crossing two stimulus with two response conditions. In Stimulus Condition S, respondents saw single true or false sentences on each trial. In Stimulus Condition P, they saw a complementary pair of sentences per trial, with the true one randomly above or below the false one. In Response Condition UR, they provided essentially unrestricted subjective probability

estimates, multiples of 1 in (0%, 100%) regarding the truth of a statement. In Condition R, their estimates were restricted to the 11 values, 0%, 10%, . . . , 90%, 100%.

Materials and procedure. Stimuli were generated by randomly sampling 100 pairs of cities from the full set (of 1,225) obtained by constructing all possible pairs of the 50 largest cities in the United States as of 1992. These 100 pairs were used to create 100 true statements (e.g., "In 1992 the population of Atlanta exceeded that of Buffalo") and their 100 false complements (e.g., "In 1992 the population of Buffalo exceeded that of Atlanta").

The 200 statements were used in 400 trials, divided into 4 blocks of 100 each, with each trial involving a distinct pair of cities. Thus, city pairs were replicated over but not within blocks. In Stimulus Condition S, Block 1 consisted of a random ordering of 50 true and 50 false statements. Block 2 used the same ordering of city pairs as Block 1, but with each sentence constructed to be the complement of the one that had appeared earlier. Blocks 3 and 4 were replications of Blocks 1 and 2, respectively. Thus, each true and each false statement appeared twice, and sentences concerning identical pairs of cities were maximally separated, with true and false versions alternating over blocks. One sequence of 400 sentences was constructed as just described, and it was used with half the respondents in Condition S. The reverse sequence was used with the other half. The ordering of city pairs was identical in Condition P, but each trial showed the pairs consisting of the true and false statements.

The experiment was computer controlled. Sentences were presented on a 14-in. (36-cm) color monitor, and responses were collected on the keyboard. Respondents studied the sentence(s) on each trial for as long as they wished. When they were ready to provide an estimate, they pressed the "Enter" key, and the stimulus disappeared. In Condition S, they typed in a subjective probability estimate that the sentence was true. In Condition P, the word *top* or *bottom* appeared to instruct them whether to provide an estimate with respect to the top or the bottom sentence in the pair. Respondents in Condition UR used any integer, and those in Condition R used any multiple of 10 from 0 to 100. The next stimulus appeared on the screen a few seconds after the response was entered.

To provide motivation, we made payoffs response contingent. Participants won or lost points on each trial according to the spherical scoring rule, $s = a + b(p/\sqrt{p^2 + (1-p)^2})$, where for true statements, $p = r$, the estimate that the statement was true, and for false statements, $p = 100 - r$. The spherical, along with all strictly proper scoring rules (Murphy & Winkler, 1970), has the property that given a subjective probability value, π, for an event, one maximizes one's expected score by setting $r = \pi$. The constants were fixed at $a = -120.71$ and $b = 170.71$, so that estimates yielded positive payments if their direction relative to 50 corresponded correctly to the statement's verity (maximum score of 50 for $p = 100$), negative payments if the direction was incorrect (minimum score of -120.71 for $p = 0$), and 0 for $p = 50$.

The instructions neither showed nor mentioned the scoring rule. Instead, they said that respondents would win or lose points (convertible to money) on each trial according to their probability estimate and whether the statement was in fact true or false. An explanation of the scoring rule principle was followed by a table with sample payoffs given selected estimates. This part of the instructions concluded by saying,

> Clearly, you will earn the maximum possible number of points by always **correctly** using 0 and 100. This will be impossible, of course, because you will not always be certain of the correct answer. The formula we are using to calculate your outcome, however, guarantees that you will maximize your **expected earnings in light of your knowledge** by always assigning that number from 0 to 100 that best reflects your actual estimate of the chances that the claim is true. No other strategy can be expected to yield better earnings.

The sample table remained available throughout the session. Respondents received feedback only at the end, and not after each trial.

Results

This section is organized as follows: First, we consider overall response distributions and response reliability. Then, we turn to the study's main predictions, considering first the properties of individual calibration curves, then the comparison of between- and within-subject averaging to test Prediction 3, and, finally, the effects of increasing the number of judges whose estimates were combined as a further test of Prediction 1. Whenever response categorization was required, we used the category boundaries [0%, 4.5%), [4.5%, 14.5%), . . . , [84.5%, 94.5%), [94.5%, 100%]. In all cases, we used analysis of variance (ANOVA) procedures to assess the effects of the stimulus and response conditions (as well as of sequential block) on the statistics of interest. Other than on the overall response distributions, the effects were close to nil. In the interest of brevity, we present these ANOVAs only for the overall response distributions.

Response distributions. After classifying the UR responses into the 11 categories available to the R participants, we compared response distributions across the four cells of the design in terms of the variance of each individual's frequency of category use. Taking logarithm of the variances to linearize them, we subjected the resulting values to a 2 (stimulus condition) × 2 (response condition) ANOVA. The only significant effect was that due to response condition, $F(1, 60) = 4.30$, $p < .05$.

The actual distributions for the two response conditions are shown in Figure 5. For ease of interpretation, the abscissa is scaled from 0 to 1 instead of from 0 to 100. The distributions are both markedly W-shaped, as Budescu, Weinberg, and Wallsten (1988) and Wallsten et al. (1993) also found, but the peaks in the 0, 0.5, and 1 categories are respectively higher for Condition UR than for Condition R. Substantiating that conclusion, the proportion of responses in those three categories is significantly greater in Condition UR than in Condition R, $F(1, 60) = 13.50$, $p < .05$. In Condition UR, 62% of the responses were in the end and center categories, whereas in Condition R, 43% were in the end and center categories. These W-patterns appear at the level of individuals as well.

Additivity and reliability. Respondents provided two probability estimates for each true statement and two for each complementary false one. Additivity requires that the means of these two

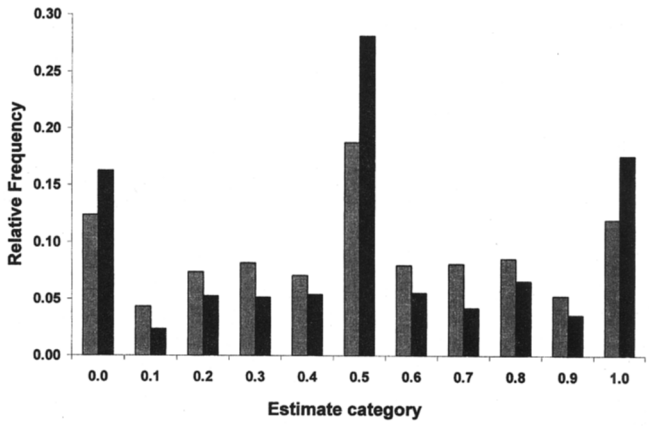

Figure 5. Response distributions for Condition R (gray) and Condition UR (black), averaged over the two stimulus conditions.

sets of judgments sum to unity. The overall mean of 1.015 (SD = 0.033) is slightly but significantly greater than 1.0, $t(63) = 3.60$, $p < .05$, based on the mean sum over all items per respondent. Nevertheless, assuming additivity, we converted the estimates of the false sentences to those of true ones by subtracting them from 1. On that basis, each respondent provided four estimates for statements concerning each pair of cities. For each respondent, we calculated the six linear correlations defined for all pairs of the four sets of estimates. The six values were similar, and we took their average following Fisher's Z transform. The overall mean correlation coefficient (i.e., the inverse of the mean Fisher's Z value) was .66.

Individual calibration. The mean of the individual calibration curves is labeled "All" in Figure 6. Separate calibration curves per group are virtually indistinguishable and are not shown here. (We discuss the remaining two curves in the figure subsequently.) Despite the random selection of items, the data show the overestimation typically found in other calibration studies rather than the pattern obtained by Juslin (1994) and Winman (1997) in their random condition. The curve rises above the diagonal for response categories less than 0.5 and falls below the diagonal for categories greater than 0.5.

To characterize each individual's performance and to look for group differences, we calculated $\overline{PS}$, CL, DI, and DI' for each participant. Overall means are given in the first row of Table 4 and serve as benchmarks for subsequent comparisons.

Prediction 3: Within- versus between-subject averaging. As before, the simple arithmetic means and the means of the estimates converted to log-odds yield the same pattern of outcomes. Thus, we present analyses based only on the first method.

Table 4
Mean Values for the Mean Probability Score ($\overline{PS}$), Calibration Index (CI), and Two Indices of Resolution (DI and DI') for Individuals and for Averages Taken Over Pairs of Estimates Both Within and Between Respondents

Level	$\overline{PS}$	CI	DI	DI'
Individual	.252	.040	.038	.666
Average within	.250	.047	.047	.657
Average between	.234	.034	.051	.732
$t(31)$[a]	21.83	11.33	−5.07	−24.67

[a] all $ps < .05$.

It is important to recall that the mean within-subject correlation coefficient is .66. Prediction 3, that between-subject averaging will be superior to within, is predicated on the assumption that there is less dependency between than within individuals. To assess whether that condition holds, we used respondents' mean estimates to the 200 statements to calculate all pairwise correlations between individuals within each of the four groups. The mean values (based on Fisher Z transformations) are .50, .57, .62, and .60 for groups P–R, P–UR, S–R, and S–UR, respectively, with a grand mean of .57. Thus, the necessary condition is met, although these values are higher than we encountered previously.

To compare the effects of averaging within- and between-subjects, we randomly paired the 16 individuals within each of the four groups, thereby forming eight dyads per group. For each dyad, we took four different means for each statement. Two were within-subject means, obtained by averaging each individual's replicated

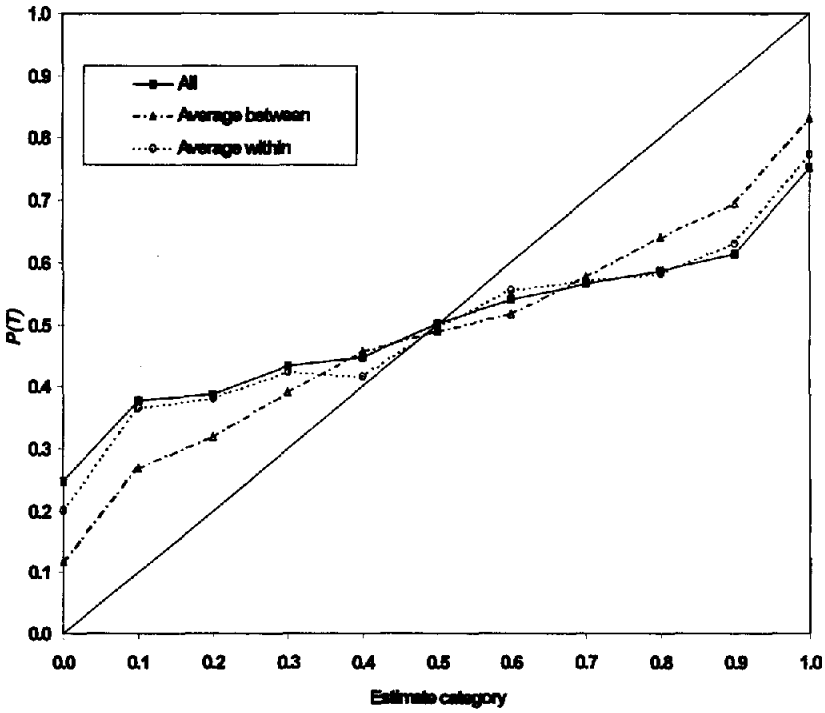

Figure 6. Calibration curves based on individuals' raw estimates (All) and on the pairwise averages of those estimates per statement within- and between-subjects. $P(T)$ = proportion of statements that are true.

estimates. Two were between-subject means, obtained by averaging each member's first estimate with the other one's second estimate. Specifically, for a given item, let r_{jk} denote respondent j's kth estimate ($k = 1, 2$; $j = 1, \ldots, 16$). Let $m_{jj'}$ denote the mean of two estimates with the first subscript indicating the person whose first estimate was included and the second subscript indicating the one whose second estimate was included. Then, the two within- and two between-subject means, respectively, are $m_{jj} = (r_{j1} + r_{j2})/2$, $m_{j'j'} = (r_{j'1} + r_{j'2})/2$, $m_{jj'} = (r_{j1} + r_{j'2})/2$, and $m_{j'j} = (r_{j'1} + r_{j2})/2$.

Calibration curves based on the within- and between-subject means, respectively, are shown in Figure 6. It is important to note that averaging pairs of replicated estimates within-subjects has virtually no effect on the quality of the estimates, whereas averaging between yields substantial improvement, as predicted.

To quantify and assess the extent of the improvement, we used the 32 sets of m_{jj} and 32 of $m_{jj'}$ to calculate the four quality indices for the within- and between-subject averaging, respectively. The means (and standard deviations) are given in rows 2 and 3, respectively, of Table 4. Consistent with the curves in Figure 6, averaging an individual's two replications per statement does very little to improve the quality of the estimates over the individual case, whereas averaging pairs of estimates between respondents provides consistent gain.

The difference between the two methods is substantiated by t tests. Because the various means are not independent of each other, we performed the t test for each index on a contrast score constructed to measure the mean within- and between-respondent difference per dyad, $D = (w_{jj} + w_{j'j'} - w_{jj'} - w_{j'j})/2$, where w denotes the index in question and the subscripts indicate the particular source of the mean estimates with which it was calculated. It is important to note that $df = 31$ for each index's test, arising from 8 independent values of D per group, one for each dyad, and therefore 32 independent values over the 4 groups. The resulting values of t are shown in the last row of Table 4. All are significantly different from 0, and all the differences favor the between-subject averaging.

Prediction 1: Increasing the number of judges. Figure 7 illustrates the consequences of averaging probability estimates over judges. All pooling across respondents used individuals' mean estimates per statement. The calibration curve based on these individual means is shown as $J = 1$ and is identical to the within-subject curve of Figure 6. For the remaining functions (those for $J \geq 2$), we ignored the stimulus condition and averaged estimates per item across participants separately within Response Conditions R and UR. Specifically, for $J = 2$, we randomly divided the 32 respondents per group into 16 pairs and, for each pair, took the average estimate per statement. The curve in the figure represents the average of the separate curves per pair. For $J = 4$, we divided the respondents into 8 sets of 4 individuals each, averaged the individual estimates per statement within each quadruple, and have shown the weighted average calibration curve. Similarly, for $J = 8$, we divided the respondents randomly into 4 sets of 8 each, and so forth through $J = 32$, where we simply averaged the estimates of all respondents per group. For $J = 64$, we simply averaged the estimates of all respondents across both conditions. Thus, all the participants contributed equivalently to pooling at each level. All that differed across levels were the numbers of estimates being averaged together per statement. The calibration curves for the separate conditions are virtually identical, and we present only the combined results in Figure 7.

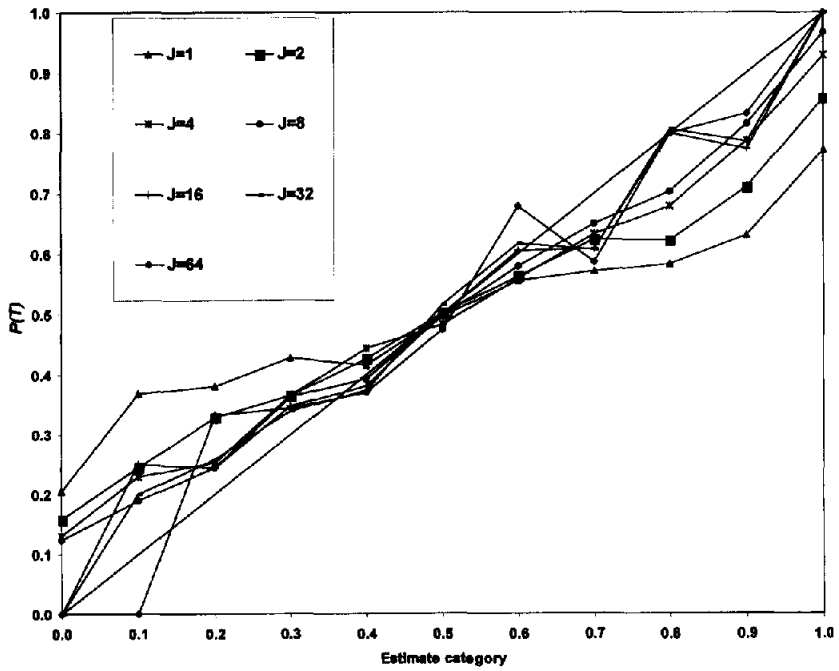

Figure 7. Calibration curves, plotting proportion of statements that are true, $P(T)$, in each response category as a function of the categorical estimate for individual respondents, $J = 1$, and for mean estimates of group sizes, $J = 2$ to 64.

It is apparent that the calibration curves change shape, consistent with Prediction 1, as the number of judges increases from 1 to 64, although not to the same extent as we saw previously (Figure 1). Table 5 and the penultimate column in Table 2 quantify the extent of the improvement. Table 5 shows mean values of the four quality indices as a function of group size and response condition, whereas the column in Table 2 shows the percentage of statements that are true (or false) given mean estimates greater (less) than 0.5 averaged over both response conditions.

The major result is that each index improves asymptotically and significantly as a function of the group size. The percentage index in Table 2 and *DI'* in Table 5 are clearly the most sensitive to number of judges. Prediction 1 called for improvements in the indices of diagnostic value and decreases in overconfidence, without making specific prediction about calibration. In fact, those indices improved as well, with the consequence that the calibration curves for $J = 32$ and 64 are very close to the diagonal. Separate linear multiple regressions for each index in Table 5 using log group size (for $J = 1, \ldots, 32$) and response condition as predictors fit the data well (with adjusted R^2 from .74 to .87, depending on the index) and show significant effects of $\ln(J)$ in all cases (at $\alpha = .05$). Response condition is significant only for *DI* and *DI'*, which are better on average in condition R than in UR.

Finally, Figure 8 shows response distributions for each group size (averaged over both response conditions) from $J = 1$ to 32 conditional on the statement being true or false. This figure tells a story similar to that of the earlier depiction of response distributions across values of J (see Figure 2). For individual judges, the conditional distributions overlap considerably. However, as J increases, these distributions separate because of their decreasing spread (pooled $SD = .392$ at $J = 1$ and .209 at $J = 32$) and increasing single peakedness. The distributions at $J = 64$ are virtually identical to those at $J = 32$, with a pooled SD of .207.

Table 5
Mean Values for the Mean Probability Score ($\overline{PS}$), Calibration Index (CI), and Two Indices of Resolution (DI and DI') for Mean Estimates Within the Indicated Group Sizes, J, Separately by Response Condition and for J = 64 Over Conditions

J	$\overline{PS}$	*CI*	*DI*	*DI'*
		Condition R		
1	.24	.03	.04	.68
2	.22	.02	.04	.78
4	.21	.01	.05	.85
8	.21	.01	.05	.93
16	.21	.01	.05	.92
32	.21	.00	.05	.91
		Condition UR		
1	.26	.04	.03	.60
2	.24	.02	.04	.70
4	.22	.01	.04	.77
8	.22	.01	.04	.79
16	.21	.01	.05	.83
32	.21	.00	.04	.88
		Over both conditions		
64	.21	.00	.05	.90

Discussion

This section discusses a few issues relevant to this experiment alone, including effects of the independent variables, response reliability, and the data with respect to Prediction 3. We defer commentary on Prediction 1 to the General Discussion because it is of interest to relate these results to those of the reanalyses. In that section, we also connect Budescu, Wallsten, et al.'s (1997) treatment of a portion of these data to the analyses presented here. Finally, the General Discussion considers the overall theoretical and practical implications of the two studies taken together.

The independent variables. The stimulus manipulation, presenting a single statement or a pair of complementary statements, had essentially no impact on any of the dependent variables. In contrast, the response manipulation did have an effect, such that individuals in the unrestricted condition used the anchor categories of 0, 0.5, and 1 (actually 0%, 50%, and 100%) more frequently than did those in the restricted. If it were the case that the anchor categories only drew from their neighbors, the effect would be relatively uninteresting. However, Figure 5 shows that use of all the nonanchor categories is depressed by roughly equivalent amounts for UR relative to R respondents, suggesting a difference in response strategy for the two groups. Nevertheless, response condition had relatively little effect on the consequences of averaging either within or between respondents.

Response reliability. The mean within-subject pairwise correlation among the responses was unaffected by the independent variables and very high, .66. This value is comforting, but we cannot rule out the possibility that, to some degree, it reflects memory of previous responses rather than simple unbiased replicability. To minimize that possibility, replicated stimuli were spaced as far apart as possible, with true and false complementary statements separated by 50 trials and repetitions of identical sentences separated by 100 trials.

Prediction 3. In accordance with this prediction, pairwise averaging of estimates between-subjects yielded better calibrated and more diagnostic results than did pairwise averaging within. We had expected the better diagnostic indices and lower overconfidence that we observed, but we had made no prediction regarding calibration, leaving open the possibility that it would turn to underconfidence. That did not happen.

The results are particularly interesting because the within- and between-subject pairwise response correlations are so close (.66 and .57, respectively), yet the outcome of averaging was so different in the two cases. Averaging estimates within individuals yielded no benefit at all, whereas averaging between yielded small, consistent, and statistically significant gains (see Figure 6 and Table 4). There are two possible explanations here, neither of which can be ruled out at the present time. One is that improvement given linear correlations in the neighborhood of .66 is simply so small that many more than two values must be averaged together to have any effect. The simulations of Johnson et al. (in press) provide some support to this notion, as they show the effects of averaging two observations at correlations of .60 to be minimal, although noticeable, with effects reaching asymptote at between 16 and 32 judges (see Table 2). The second possibility is that within- and between-subject violations of pairwise independence under these conditions differ in ways not captured by linear correlation coefficients and that the within-subject violations are such that no improvement results. For example, the within-subject correlation

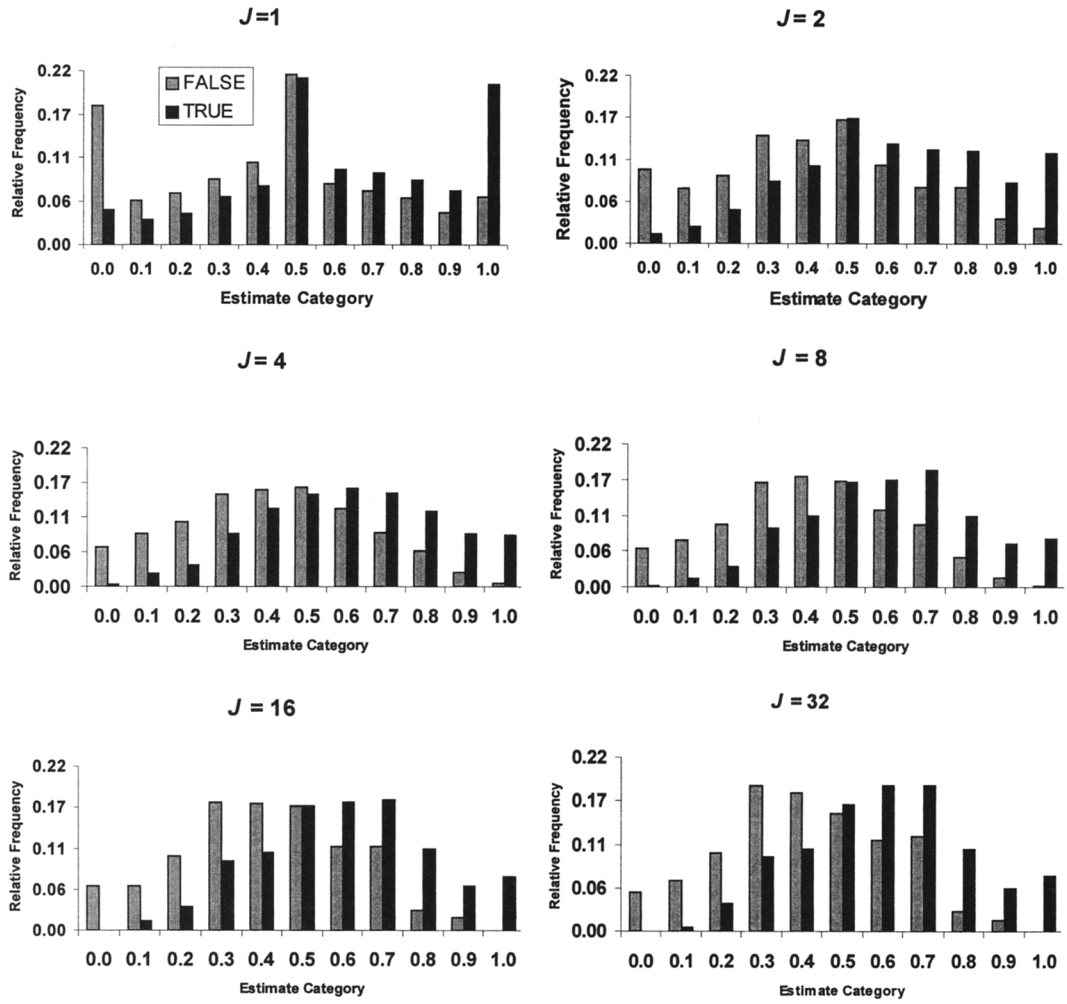

Figure 8. Distributions of mean estimates to true and false statements for group sizes, $J = 1, 2, 4, 8, 16$, and 32.

almost certainly depends solely on the degree of random error a respondent brings to his or her information base. Averaging can do no more than eliminate the perturbations caused by the error process. The between-subject correlation depends on random within-subject error but also on the extent of overlap in the judges' information bases. Averaging in this case benefits from the totality of the underlying information and thereby improves the diagnostic impact. This, of course, is exactly the distinction captured in its extreme by Equations 3 and 2, respectively. Both of these explanations support the basis of Prediction 3 that averaging within-subjects will lead to less diagnostic results than averaging between-subjects because of greater violations of pairwise independence.

General Discussion

At this point, we turn from considering the present experiment in isolation to merging discussion of it with that of the reanalyses for the purpose of drawing general conclusions and implications.

Prediction 1

Our new results join those uncovered in the reanalyses in supporting Prediction 1, but they are not so spectacular (see Figure 7 and Table 5), no doubt because of the greater violation of conditional pairwise independence. Table 2 provides a convenient comparison of all the results, including the relevant subset from Johnson et al.'s (in press) simulation. It is important to note that according to the percentage of true (or false) statements given probability estimates greater (or less) than 0.5 at $J = 1$, the mean level of accuracy in this study matches that of Juslin's (1994) and Winman's (1997) participant-selection condition. Despite the pairwise correlation being much greater in the present case, the advantages of averaging are only slightly less. This fact further suggests that the nature of the independence violation in the participant-selection case is different than in the others and not well captured by the linear correlation coefficient.

The overall picture from our analyses along with the Johnson et al. (in press) simulations is one of substantial support for Prediction 1, with good but still incomplete indication of its degree of robustness. In all cases, the calibration curve plotting $P(T)$ as a function of the mean estimate becomes less regressive, even anti-regressive, as the number of judges increases, in accordance with Prediction 1. The comparisons in Table 2 and the relevant figures suggest that the degree to which that occurs depends on both the difficulty of the task and the nature of the conditional pairwise dependency. When the task is relatively easy, items are not se-

lected with a view to testing the judge, and violations of independence are moderate (Juslin's, 1994, and Winman's, 1997, random condition and the Johnson et al. simulation with $r = .30$), averaging yields considerable underconfidence and good resolution. Under the same conditions, but more severe violations of independence (our experiment and Johnson et al., $r = .60$), averaging still improves the diagnostic value to a considerable degree. Good resolution given the degree of conditional dependence was achieved with 6 judges and close to the maximum possible with 12 or 16.

Calibration and Difficulty Differences Among the Various Conditions

According to the ecological view (e.g., Gigerenzer et al., 1991; Juslin, 1994), overconfidence in probability estimation is an artifact of biased sampling of events. Juslin's (1994) and Winman's (1997) data, which we have looked at so closely, support this view. Accordingly, then, respondents in our Study 2 should have been well calibrated because their stimuli also were randomly sampled from a well-defined domain. In contrast, they were overconfident.

Associated with this difference in calibration is a difference in difficulty, as indexed by the percentage of true or false items given mean estimates respectively above or below 0.5. (This index is a direct function of the percentage of correct answers in the choice half-scale task.) In fact, others (see Brenner, Koehler, Liberman, & Tversky, 1996, and references therein) have also shown overconfidence given random sampling in cases with low proportions of correct answers. Thus, we might attribute the calibration differences simply to the hard–easy effect. However, this effect is not an explanation, it is simply a description of a widely observed pattern of data. In a recent article, Juslin et al. (in press) argued that the hard–easy effect is primarily, if not entirely, due to statistical and measurement artifacts. When these are properly corrected for, they argued, most of the effect goes away in tasks that involved random item selection (see also Klayman, Soll, González-Vallejo, & Barlas, 1999). As part of their argument, Juslin et al. showed that the means and standard deviations of the subjective probability estimates were identical in their random- and participant-selection conditions. What differed in the two cases were the probabilities of correct choices (and, as a consequence, the level of calibration). Indeed, comparing the response distributions in Juslin's (1994) and Winman's (1997) random condition and our Study 2 (which can be done by averaging over the conditional distributions in Figure 2 and Figure 8, respectively), we see that they barely differ. Thus, the groups of respondents are identical at the level of response distributions and differ in calibration only to the extent that the tasks vary in difficulty. (See Wallsten, 1996, for additional discussion and illustration of the role $P[T]$ plays in determining calibration.)

Budescu, Wallsten, et al.'s (1997) Analyses

The ecological approach may explain why the response distributions do not change for informal versus random sampling of items from a given domain, but it does not explain why the distribution is fixed across domains. Exploration of that issue is beyond the scope of this article. However, Budescu, Wallsten, et al. (1997) analyzed a portion of the present data to determine whether the observed overconfidence remained after correcting for trial-by-trial random perturbations in judgment and response processes. The analyses, which used the data of the respondents in Condition R and were done only at the level of $J = 1$, were distinct from those presented here. Rather than investigate the effects of averaging, they focused on Wallsten and González-Vallejo's (1994) stochastic judgment model, according to which respondents' probability estimates depend on the location of their covert confidence for an item relative to response criteria, with both the confidence and the criteria subject to trial-by-trial fluctuation. Omitting all details, the conclusion was that the vast majority of the respondents displayed overconfidence, even after correcting for stochastic perturbations. Moreover, the overconfidence was directly due to respondents setting insufficiently extreme response criteria for the task at hand.

Practical Implications

The present data add to a growing body of results that strongly points to the importance of considering within- and between-subject variability in probability estimates (e.g., Soll, 1996). Trial-by-trial error within respondents may add to the apparent degree of overconfidence and miscalibration within an individual (Erev et al., 1994), although it did not do so to any appreciable degree in the experiment. We continue to expect on theoretical grounds that it will in conditions of lower response reliability, and that in those cases, averaging individual judgments may yield some or even substantial improvement.

However, the story is different with between-subject variability. In this case, even when pairwise correlations among estimates (conditional on the state of the item—true or false) are high, averaging over multiple judges distinctly improves the quality of the forecast. The degree of improvement varies with the extent of the dependency but can be quite substantial.

These results yield three very clear practical conclusions: One is that averaging works. The mean estimate of only 2 judges generally is more diagnostic than that of either one alone. Additionally, substantial improvement can be obtained by averaging the estimates of as few as 6 judges. The second conclusion is that the judges should operate from as distinct information bases as possible, thereby reducing the conditional pairwise correlations among their estimates and maximizing the diagnostic result. These two points echo Hogarth's (1989) and Sorkin and Dai's (1994) recommendations (see also Ashton, 1986). Thus, a fixed budget is better spent on diversifying across experts with access to different information or with different perspectives than on increasing the number of experts looking at the same information in the same way. Third, there seems to be some advantage in having respondents use categorical rather than unrestricted estimates. Interestingly, most tend to adopt this strategy spontaneously. (Almost 85% of the estimates in Condition UR were multiples of 10 and an additional 9% were multiples of 5 but not of 10. This pattern is similar to that observed by Budescu et al., 1988, and Wallsten et al., 1993).

Can we more specifically advise the decision maker on how to treat the average estimate of J (say 6) forecasters? Before addressing this issue, it is appropriate to point out that the optimal way to use multiple subjective probability estimates is through Bayes's rule (Morris, 1977). That is, the decision maker should treat the estimates (say of Event A) as data, estimate the likelihood of the particular data pattern (combination of estimates) given A and

given not-A, then use the likelihood ratio to revise his or her own subjective prior odds about the event. Although this approach sounds straightforward, it is very difficult to apply in practice because the decision maker rarely has sufficient experience with the particular experts in the particular context to reliably estimate likelihood ratios. To make matters worse, the greater the conditional dependency among judges, the more history is required to get good estimates. In the end, the decision maker must rely on his or her intuition on the basis of whatever information or history is available. Researchers have suggested a host of formal models to aid the decision maker in incorporating this intuition into Bayes's rule (see, for example, Clemen & Winkler, 1993; Genest & Schervish, 1985; Winkler, 1989).

The advantage of taking averages, as we are suggesting, is that one does not have to focus on different combinations of estimates but only on their central value. Thus, perhaps (we have not studied the issue) global and informal estimates of conditional dependency will do, based on general knowledge of the experts' backgrounds, biases, and data sources. Of course, the more this estimate is informed by explicit past experience, the better. If the decision maker considers the experts to have mild and unsystematic overlap in their information sources, then he or she should treat the probability of the event as considerably more extreme than the mean estimate. In contrast, if the overlap is large (perhaps complete), then the decision maker may wish to treat the probability as closer or roughly equal to the mean estimate.

Finally, although not specifically addressed in this study, we should briefly say something about the type of average to take. Wallsten and Diederich's (in press) result is that given the assumptions, the mean of any monotonic transformation of the probability estimates will converge in the manner of Equation 2. Thus, arithmetic averages, such as used in this study, are not required. One can use mean log-odds, weighted means, trimmed means (including medians), or any other central statistic and should select the one that converges fastest. Empirical comparisons thus far suggest little practical difference between arithmetic means and means of the log-odds transformation. Nevertheless, one form of central tendency may be substantially better than others under particular conditions. For example, it may turn out to be better to average log-odds than to take arithmetic means when only rare events (e.g., with probabilities under .01) are being judged. These points, among others, remain to be more thoroughly investigated.

References

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

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55,* 412–428.

Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception & Psychophysics, 61,* 1369–1383.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & Psychophysics, 54,* 75–81.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65,* 212–219.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of a probability judgment. Part 1: New theoretical developments. *Journal of Behavioral Decision Making, 10,* 157–171.

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment. Part II: Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making, 10,* 173–188.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 281–294.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5,* 559–583.

Clemen, R. T., & Winkler, R. (1990). Unanimity and compromise among probability forecasters. *Management Science, 36,* 767–779.

Clemen, R. T., & Winkler, R. (1993). Aggregating point estimates: A flexible modeling approach. *Management Science, 39,* 501–515.

DuCharme, W. M., & Peterson, C. R. (1968). Intuitive inference about normally distributed populations. *Journal of Experimental Psychology, 78,* 269–275.

Erev, I., & Wallsten, T. S. (1993). The effect of explicit probabilities on decision weights and the reflection effect. *Journal of Behavioral Decision Making, 6,* 221–241.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

Ferrell, W. R. (1985). Combining individual judgments. In A. Wright (Ed.), *Behavioral decision making* (pp. 111–145). New York: Plenum Press.

Ferrell, W. R. (1994). Discrete subjective probabilities and decision analysis: Elicitation, calibration and combination. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 411–451). Chichester, England: Wiley.

Ferrell, W. R., & McGoey, P. J. (1980). A model of calibration for subjective probabilities. *Organizational Behavior and Human Performance, 26,* 32–53.

Genest, C., & Schervish, M. J. (1985). Modeling expert judgments for Bayesian updating. *Annals of Statistics, 13,* 1198–1212.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46,* 107–119.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Science, 1,* 78–82.

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

Hogarth, R. (1989). On combining diagnostic "forecasts": Thoughts and some evidence. *International Journal of Forecasting, 5,* 593–597.

Johnson, T., Budescu, D. V., & Wallsten, T. S. (in press). Averaging probability judgments: Monte-Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making.*

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226–246.

Juslin, P., Winman, A., & Olsson, H. (in press). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review.*

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79,* 216–247.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107–118.

Kubovy, M., Rapoport, A., & Tversky, A. (1971). Deterministic vs. prob-

abilistic strategies in detection. *Perception & Psychophysics, 9,* 427–429.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

May, R. S. (1986). Overconfidence as a result of incomplete and wrong knowledge. In R. W. Scholz (Ed.), *Current issues in West German decision research* (pp. 13–30). Frankfurt, Germany: P. Lang.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.* Cambridge, England: Cambridge University Press.

Morris, P. (1977). Combining expert judgments: A Bayesian approach. *Management Science, 23,* 679–693.

Murphy, A. H., & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica, 34,* 273–286.

Murphy, A. H., & Winkler, R. L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society: Series C, 26,* 41–47.

Olsson, H., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental settings and response strategies. *Perception & Psychophysics, 58,* 374–382.

Pfeiffer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes, 58,* 203–213.

Samuels, M. L. (1991). Statistical reversion toward the mean: More universal than regression toward the mean. *The American Statistician, 45,* 344–346.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65,* 117–137.

Sorkin, R. D., & Dai, H. (1994). Signal detection analysis of the ideal group. *Organizational Behavior and Human Decision Processes, 60,* 1–13.

Wallsten, T. S. (1996). An analysis of judgment research analyses. *Organizational Behavior and Human Decision Processes, 65,* 220–226.

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., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39,* 176–190.

Wallsten, T. S., & Diederich, A. (in press). Understanding pooled subjective probability estimates. *Mathematical Social Sciences.*

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504.

Wallsten, T. S., & Gu, H. (1996). *Effects of criterion variance on judgment: Model and data.* Paper presented at the 37th annual meeting of the Psychonomic Society, Chicago.

Winkler, R. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting, 5,* 605–609.

Winman, A. (1997). The importance of item selection in "knew-it-all-along" studies of general knowledge. *Scandinavian Journal of Psychology, 38,* 63–72.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110,* 611–617.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132–156.

Received October 19, 1998
Revision received November 22, 1999
Accepted December 2, 1999 ■