Journal of Experimental Psychology: General
1995, Vol. 124, No. 4, 424-432

0096-3445/95/$3.00

Graininess of Judgment Under Uncertainty: An Accuracy–Informativeness Trade-Off

Ilan Yaniv
Hebrew University of Jerusalem

Dean P. Foster
University of Pennsylvania

This work concerns judgmental estimation of quantities under uncertainty. The authors suggest that the "graininess" or precision of uncertain judgments involves a trade-off between 2 competing objectives: accuracy and informativeness. Coarse (imprecise) judgments are less informative than finely grained judgments; however, they are likely to be more accurate. This trade-off was examined in 3 studies in which participants ranked judgmental estimates in order of preference. The patterns of preference ranking for judgments support an additive trade-off model of accuracy and informativeness. The authors suggest that this trade-off also characterizes other types of uncertain judgments, such as prediction, categorization, and diagnosis.

Efficient decision making in daily life often requires reliable information about uncertain events and future outcomes. Such information is often communicated in the form of judgments and predictions about quantities such as the arrival time of a guest, the outcome of a research project, the market value of a used car, or the success of a surgical procedure (Dawes, 1971; Griffin & Tversky, 1992; Tversky & Kahneman, 1974; Yaniv & Hogarth, 1993; Yaniv, Yates, & Smith, 1991). In making judgments under uncertainty, people often communicate the degree of confidence they accord their judgment by stating its precision or "graininess." Thus judges have some latitude in communicating either precise or vague judgments (Einhorn & Hogarth, 1985; Erev, Wallsten, & Neal, 1991; Reyna & Brainerd, 1991; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). For example, a guest might signal uncertainty about the time of her arrival by making a coarse-grained promise ("fivish") rather than a fine-grained one ("5:00 PM"). Both communications indicate the same point estimate, but they differ in their precision. Other relevant examples include, "3 weeks" versus "21 days" and "a dozen" versus "12." In a different vein, graininess can also be implied in categorical or diagnostic judgment. Compare, for instance, the precision of diagnoses such as "a virus" and "influenza type B."

The focus of this work is on the estimation of uncertain quantities. The precision of an estimate is commonly communicated in terms of the width of an interval estimate. For example, one might estimate the market value of a used car at "$2900 to $3000" or forecast an inflation rate to be "6% to 8%." The vagueness or specificity of an estimate clearly depends on the individual's confidence in his or her knowledge or the available evidence. Our main hypothesis in this article is that the vagueness or graininess of judgmental estimation under uncertainty involves a trade-off between two conflicting objectives: accuracy and informativeness. In three studies, we asked respondents to evaluate judgments made by others. The results provide support for a simple additive model of the trade-off between accuracy and informativeness. Substantively, they show that people might accept errors in the interest of securing more informative judgments. We suggest that these results have important implications for the production of judgments, because judges presumably respond to recipients' expectations. Finally, we believe that the general idea and the results also extend to other types of judgment under uncertainty, including prediction, scenario generation, categorization, and diagnosis.

Background

One of the motivations for studying the trade-off between accuracy and informativeness arises from a classic puzzling finding of Alpert and Raiffa (1969/1982) concerning the accuracy of interval judgments. A priori, one could hypothesize that in providing an interval estimate for an uncertain quantity (e.g., "air distance from Chicago to New York"), people simply set lower and upper bounds on potential judgmental errors. In a study by Alpert and Raiffa (1969/1982), students generated 98% confidence intervals for each of several uncertain quantities, for which they believed there was only a 2% chance that the true answer fell outside the bounds of the interval. If people's confidence intervals are

Ilan Yaniv, Department of Psychology, Hebrew University, Jerusalem, Israel; Dean P. Foster, Department of Statistics, Wharton School, University of Pennsylvania.

This research was supported by grants from the Graduate School of Business at the University of Chicago, the Israel Foundation Trustees, and the Israel Science Foundation. We acknowledge the helpful comments of M. Bergen, S. Hoch, R. Hogarth, J. Huttenlocher, G. Loewenstein, C. McKenzie, J. Klayman, P. Schoemaker, T. Wallsten, and E. Weber.

Correspondence concerning this article should be addressed to Ilan Yaniv, Department of Psychology, Hebrew University, Mt. Scopus, Jerusalem, Israel. Electronic mail may be sent to msilan@pluto.mscc.huji.ac.il.

calibrated with respect to their stated probabilities, then across a series of judgments, intervals stated with 98% confidence should include the correct answers 98% of the time. The proportion of intervals that actually failed to include the true answers in this study, labeled the *surprise index,* equaled 42% instead of 2%.

Subjective confidence intervals are thus too narrow and exclude the correct answers more often than expected. The difficulty of making calibrated judgments has been a puzzling phenomenon for both subjects and researchers in the numerous studies that have since replicated this finding (Lichtenstein, Fischhoff, & Phillips, 1982; Yates, 1990, chap. 4). Subsequent studies explored corrective procedures for improving calibration accuracy, such as issuing warnings, giving feedback, changing anchors, and reframing the question (Alpert & Raiffa, 1969/1982; Tversky & Kahneman, 1974; Lichtenstein et al., 1982; Sniezek & Buckley, 1991; Yates, 1990).

Our point of departure differs from that of earlier work. Previous studies have focused on calibration accuracy as a normative standard, whereas we take the view that interval estimates are often communicated as part of the decision-making process and thus involve "senders" and "receivers." Effective decision making requires fairly informative estimates. Judges estimating quantities thus presumably consider the informativeness of their judgments in the generation process. Preliminary evidence for this notion comes from previous work. In one study by Yaniv and Foster (in press), participants were asked to generate 95% confidence intervals, whereas in another study they were asked to provide interval estimates that they merely "felt comfortable communicating in the course of a casual social interaction." The hit rates (proportions of intervals that include the correct answers) were rather similar across the methods of elicitation (around 50%), suggesting that individuals perhaps rely on similar processes for generating estimates under both types of instruction. A more important conclusion, however, is offered by the following analyses.

In principle, individuals could increase the chances of including the correct answer by widening their estimated intervals. For each person, we calculated the factor by which his or her intervals should be symmetrically widened for them to include the truth with a probability of .95 (Yaniv & Foster, in press). The median "calibration factor" was 17. The lowest calibration factor observed was 5. For illustration, a calibration factor of 5 implies that an intuitive interval estimate of "the U.S. population in 1987" such as "200 to 240 million" would be transformed into "120 to 320 million"; similarly, an intuitive estimate of "the postal charge for an overnight express letter" such as "$8–10" would be transformed into "$4–14." Clearly, the median calibration factor would lead to far wider intervals.

Students who experience the 95% confidence-interval task (e.g., in class demonstrations) often suggest that the calibration factors found in Yaniv and Foster (in press) would frequently result in worthless estimates for the decision maker. Moreover, one might argue that a strategy of giving enormous ranges (e.g., "zero to one billion") in response to 95% of the questions would be adaptive to this task but not useful in real-life situations. Indeed, conversational norms suggest that judges should be appropriately informative as well as accurate (Grice, 1975). This implies that excessively coarse judgments are precluded in communication. This could potentially explain why judges fail to widen their intuitive estimates even when asked for 95% confidence intervals (Yaniv & Foster, in press). To examine these ideas more carefully, we investigate in this work the receivers' preferences. These are important because they may influence the judgmental estimation process that senders engage in.

Trade-Off Between Accuracy and Informativeness

We hypothesize that receivers prefer estimates that are both sufficiently informative for their current decision making and appropriately accurate. For example, the prediction that the inflation rate will be "0% to 80%" would not be appreciated by receivers, although it is likely to be confirmed (Bar-Hillel & Neter, 1993; Tversky & Kahneman, 1983, p. 312). Consider alternative judgmental forecasts of inflation rate such as "7 to 7.5%," "6 to 8%," or "4 to 12%." There is a trade-off between accuracy and informativeness such that coarser estimates are less informative, although they are likely to be more accurate. This trade-off logically implies that under uncertainty, informative predictions would be inaccurate some of the time. In other words, uncertain judgments are error-proof only if they are uninformative or vacuous. This trade-off also characterizes categorical judgments. For instance, under uncertainty, a physician making a diagnosis might choose a level of precision that is a compromise between the need to be informative and the need to be accurate.

The plan of research is as follows. First, we report preliminary results that motivate the development of a theoretical framework for the accuracy–informativeness trade-off. Then we propose a formal model of this trade-off. In three studies, we test the predictive validity of this model and compare it to that of several alternative models.

Elements of Trade-Off Model

We begin by reporting a preliminary study on the trade-off between accuracy and informativeness. The obtained evidence guides our construction of a model, which is later tested in more detail. The instructions for the preliminary study were embedded in a scenario that described a researcher preparing a presentation. The researcher solicits two judgmental estimates for some missing information about uncertain quantities from two aides called *A* and *B*. Later, on finding the correct answers, the researcher evaluates the quality of the judgments given by the aides. The respondents, taking on the role of the researcher in this scenario, were supposed to indicate in each case which of the two judgments was better in light of the correct answer. Questions from this study are given below along with the percentage of participants ($N = 20$) choosing each alternative (in parentheses):

1. Amount of money spent on education by the US federal government in 1987?
   A responds: *$20 to 40 billion* (20%)
   B responds: *$18 to 20 billion* (80%)

   The actual answer is: *$22.5 billion.* Which estimate is better?

2. Date the Sino-Japanese War began?
   A responds: *1870 to 1890* (40%)
   B responds: *1875 to 1925* (60%)

   The actual answer is: *1894.* Which estimate is better?

3. Air distance between Chicago and New York?
   A responds: *800 to 850* (15%)
   B responds: *600 to 800* (85%)

   The actual answer is: *713.* Which estimate is better?

The foregoing qualitative discussion of the results motivates the construction of the model. On Question 1, a majority (80%) of the respondents preferred the more informative interval (B is more informative by a factor of 10 than A); we note that B seems close to the truth (in a sense that will be defined later), although it does not include it. For Question 2 there is a less clear-cut pattern of preferences between the answer of A, which is more informative (by a factor of 2.5), and the answer of B, which includes the truth. Finally, for Question 3 there is a clear preference for the more accurate answer, that of B; A is more informative than B (by a factor of 4) but seems relatively far from the truth (a definition of distance is given below).

We suggest that preference among judgments varies as a continuous function of two dimensions: informativeness, measured as the precision of the estimate, and accuracy, expressed as a continuous measure of the distance of an interval from the truth. These three examples intuitively illustrate cases along a continuum of trade-off relations between informativeness and accuracy. In the following section, we propose a simple formal model of the trade-off between these dimensions. Whereas the implementation is specific to the domain of interval estimation, we believe that the general idea of a trade-off underlying this model might be extended in future work to other types of uncertainty, such as diagnostic or categorical judgments.

Accuracy. Some traditional measures of accuracy of interval judgments (e.g., calibration) are based on a binary coding of the outcome, namely, whether or not an interval includes the truth. For instance, consider a physician's estimate that a patient's recovery will take 3 to 4 weeks. A binary all-or-none measure implies that this interval prediction is a "hit" if it includes the truth (i.e., recovery takes 3 to 4 weeks) and a "miss" if it does not.

We suggest instead a continuous measure that is based on our preliminary results. This measure assumes that receivers evaluate judgmental errors in relation to the precision claimed by the judge. Specifically, the normalized error of an interval judgment is the ratio of error to precision $(t-m)/g$, where t is the true answer and m and g are the midpoint and width (graininess) of the judgmental interval, respectively. The error-to-precision ratio implies that human evaluation of accuracy depends on (a) the estimate's distance from the truth and (b) the precision claimed by the judge. To illustrate the normalized error, imagine we ask two individuals to guess the date the University of Chicago was founded: (a) Bill's opinion is "in the 1880s" and (b) John's opinion is "1885." Suppose next we are told that the true founding date of the University was 1892, showing that both judgments miss the truth. The precision claimed by these judges seems to matter in our evaluation of their accuracy. Bill used an interval of 10 years; therefore his estimate has an error-to-precision ratio of less than 1. John implied an interval of 1 year, hence his answer has an error-to-precision ratio of about 7. Thus the normalized error of Bill's judgment is lower than that of John.

Informativeness. Coarser estimates tend to be less informative. We define informativeness in terms of ln (g), the natural logarithm of interval width. Using this transformation of g is consistent with the well-known psychophysical law (Fechner) that human responses to changes in objective magnitudes (in this case, interval width) approximate a concave function. In particular, the logarithmic transformation explains why the difference between widths (precision) of 1 and 10 years (in estimation of a historical date) is perceived to be greater than the difference between precision levels of 50 and 60 years.

Additive trade-off model. The definition of accuracy as a "ratio of error to precision" implies that coarser judgmental estimates are likely to be more accurate. However, coarser judgments are less informative. This naturally creates a trade-off between accuracy and informativeness.

The trade-off between accuracy and informativeness can be captured by a formal model of the form

$$L = f\left[\left|\frac{t-m}{g}\right|, \ln(g)\right]$$

that assigns overall evaluation scores to interval estimates. The L score is a function of one argument that corresponds to accuracy (normalized error) and a second argument that corresponds to informativeness (log of width). We assume that f is a monotonically increasing function of its arguments and propose an additive trade-off model, as shown in Equation 1:

$$L = f_1\left|\frac{t-m}{g}\right| + f_2[\ln(g)]. \qquad (1)$$

The trade-off occurs because, as g increases, accuracy improves (normalized error decreases), whereas informativeness diminishes (ln (g) increases). Note that as the accuracy or informativeness of an estimate improves, the overall L score decreases. Thus, the lower the L score the better. Suppose A and B are interval judgments estimating a given quantity. Then they can be ranked according to their L scores so that A is better than B if and only if $L(A) < L(B)$. For simplicity, we substitute f_1 and f_2 with the identity

function and the coefficient $\alpha \geq 0$, respectively, resulting in

$$L = \frac{|t - m|}{g} + \alpha \ln(g). \quad (2)$$

The coefficient α is a trade-off parameter that reflects the weights placed on the accuracy and informativeness of estimates. In particular, as α increases, the penalty for lack of informativeness also increases.

A few comments on the additive model are in order. First, we suggest that the additive form is sufficiently general. A separate consideration of multiplicative models is not necessary because, under a logarithmic transformation, a multiplicative model could be transformed into an additive model while preserving its ordinal properties. The distinction between additive and multiplicative is immaterial because only the invariant ordinal properties of L are of importance in this work.

Second, models that have only an accuracy component, $f_1(|t - m|/g)$, are untenable because they reward judges for providing very broad ranges, contrary to social linguistic norms and experimental data showing that people normally do not use excessively broad intervals in estimation (Yaniv & Foster, in press). In a similar vein, models that include only an informativeness component, $f_2[\ln(g)]$, are untenable because they reward judges for giving the narrowest possible ranges.

Third, the additive model satisfies certain consistency requirements: (a) *Shift invariance* implies that adding a constant to the scale of measurement preserves preference order among the scores of the various judgments. (b) *Scaling invariance* implies that multiplying the scale by a constant (e.g., converting units from inches to centimeters) preserves the order of the scores assigned by the additive model to estimates. Note that this invariance is due to the logarithmic transformation of g. (In contrast, using the concave power function g^{β} with $|\beta| \leq 1$ instead would have violated scale invariance.) (c) *Distance* implies that, holding precision constant, the smaller the normalized error the better. (d) *Precision* implies that, holding the normalized error constant, the more precise the interval the better. (e) *Symmetry* implies that under- and overestimations are weighed equally.

Overview of Studies

In three studies, we tested how well a trade-off model predicts people's evaluations of judgments. We used the method described for the preliminary study with some important variations across studies. The instructions for the task were embedded in a scenario that described a researcher preparing a presentation. The researcher solicits estimates for some missing information from two aides, *A* and *B*, who are assigned to make judgmental estimates. Later, on finding the correct answers, the researcher evaluates the quality of the judgments given by the aides. Participants are supposed to take the perspective of the researcher and rank in order of quality the judgments made by the aides in light of the correct answer. The following is a sample question:

> Amount of money received by Michael Jackson in 1987 to star in a series of Pepsi commercials?
> Aide *A* responds: *$1 to 20 million*
> Aide *B* responds: *$12 to 14 million*
>
> The correct answer is: *$15 million.* Which estimate is better?

Our main thesis is that respondents' evaluations are a function of the accuracy and informativeness of these judgments. We use the additive trade-off model to generate predictions. For example, with respect to Jackson's compensation, the scores generated by the additive trade-off model with $\alpha = 1$ are $L(A) = 3.2$ and $L(B) = 1.7$; thus *B* would be ranked over *A* by the model. We fit the observed preferences to the predictions of the additive trade-off model.

Several approaches for testing the trade-off model are taken. In Study 1, we used logistic regression. In Study 2, we asked subjects to rank sets of eight estimates at a time. We calculated the correlations between the subjects' rankings and the model's ranking. This approach allowed direct comparisons between the tradeoff model and alternative models. In Study 3, we compared the tradeoff model to alternative models in more detail. The results of the last study highlight the important dimensions of respondents' preferences.

Study 1: Preference

We hypothesize that receivers' preferences among judgments depend on the scores assigned to them by the model. Specifically, let *A* and *B* be two judgmental estimates of some uncertainty. Then on the basis of the additive trade-off model, we derive the following difference score, $L(A) - L(B)$, which reflects the difference in quality between them:

$$L(A) - L(B) = \left(\frac{|m_A - t|}{g_A} - \frac{|m_B - t|}{g_B}\right) + \alpha \ln\left(\frac{g_A}{g_B}\right). \quad (3)$$

We suggest that given a choice between two estimates, the probability of choosing *B* over *A* increases as a function of the difference score. In other words, the greater the difference score, the stronger the preference for *B* over *A*. Hence, positive difference scores predict preference for *B*, whereas negative difference scores predict preference for *A*. This hypothesis is tested using a probit regression analysis with choice probability as a dependent variable and the difference score as an independent variable.

Subjects and Materials

The participants were 60 students at the University of Chicago who were assigned to one of three groups. The 20 participants in each group indicated their preferences in a different set of 21 different pairs of alternative judgmental estimates, using the method described above (e.g., the "compensation question"). The pairs of alternatives shown along with each question were selected from the set of actual confidence-interval judgments generated by

people who participated in the studies reported by Yaniv and Foster (in press). The paired alternatives had difference scores ranging from −3.0 to 3.0 on the basis of a trade-off parameter $\alpha = 1$, which appeared to predict well the participants' choices in the preliminary study.

Results

We used probit regression analysis to test whether the additive model could predict the observed preferences. This analysis evaluates whether the tendency to choose *B* increased with the difference score ($\alpha = 1$). The dependent variable was the proportion of respondents who chose alternative *B*. In the first probit analysis (logit transformation), the independent variable was the difference score $L(A) - L(B)$ (Equation 3). The resultant regression coefficient was significant ($z = 16.1$, $p < .005$). Thus, as we hypothesized, the greater the difference score, the greater the preference for *B* over *A*.

In a second probit analysis, we tested the significance of the two predictors in Equation 3: (a) the difference of normalized errors of *A* and *B*, and (b) the ratio of their interval widths. A significant regression coefficient was obtained for the normalized error, which equaled 1.03 ($z = 15.9$, $p < .005$), and for the interval width, which equaled 0.76 ($z = 15.7$, $p < .005$). Thus, both components of the additive model affect the evaluation of judgments. The trade-off parameter α of the additive model corresponds to the ratio between the second probit coefficient (on informativeness) and the first probit coefficient (on accuracy). The ratio of these coefficients, 0.76/1.03, thus corresponds to a trade-off parameter of 0.74.

To examine how the α parameter varied by subject, we performed a third set of analyses in which individual (probit) regression coefficients were fit. Separate probit analyses were performed for each person. The mean of the αs across participants ($N = 60$) was 0.76 with a standard deviation of 1.63. The interquartile range was 0.52 to 0.97. The mean α significantly differed from zero, $t(59) = 3.61$, $p < .001$. More important, the corresponding statistical 98% confidence interval around the mean trade-off parameter (0.76) ranged from 0.26 to 1.27. The method of Study 2 allowed a more systematic examination of the additive model's fit as a function of α.

Study 2: Ranking

The results from Study 1 provided evidence for a prediction derived from the additive trade-off model. The difference score (derived from that model) predicted receivers' preferences among estimates. In Study 2, we asked participants to rank larger sets of estimates. The ordinal correlation between the participants' ranks and the model scores were calculated to assess the model's goodness of fit. The advantage of using a correlational measure of fit was in allowing (for reasons that are detailed later) a straightforward comparison of the additive model and alternative models.

Subjects and Materials

The materials included 24 questions of the type shown below ("Number of United Nations member countries"). Each question involved eight judgmental estimates (*A* through *H*). In addition, the correct answer to each question was given. The instructions for the task were similar to those given before and described a researcher who had solicited estimates from aides.

Number of United Nation member countries (in 1987)?
The correct answer is: *159*

Judgmental estimates provided by aides		Subject's ranking of judgments	
Aide A	300–400	Best	1. Aide _____
Aide B	110–130		2. Aide _____
Aide C	150–160		3. Aide _____
Aide D	50–300		4. Aide _____
Aide E	100–150		5. Aide _____
Aide F	107–180		6. Aide _____
Aide G	30–50		7. Aide _____
Aide H	50–75	Worst	8. Aide _____

Taking the perspective of the researcher, participants were supposed to rank the judgments made by the aides in order from best to worst in light of the correct answer. In particular, they entered in the right-hand column the aides' letter names on the basis of the quality of their judgments. In the first row, they were supposed to indicate the interval judgment they considered best, in the second row, the interval they considered second best, and so on.

The eight alternatives shown along with the United Nation (UN) question above were selected from the set of 95% confidence-interval judgments generated by participants about this very same question in an earlier study (Yaniv & Foster, in press). As part of the selection procedure of interval estimates for this study, scores for all intervals estimating the number of UN countries ($N = 43$) were computed according to the additive trade-off model ($\alpha = 1$). Then the intervals were ranked from best to worst according to their model scores. The interval with the lowest score (i.e., the best) was ranked first, the one with the second lowest score was ranked second best, and so on. Next, 8 estimates were selected (out of the 43) whose ranks were 1st (best), 6th, 11th, 16th, 21st, 26th, 31st, and 36th. For the UN question above, the selected alternatives were C, F, E, B, D, A, H, and G, respectively.

Using this ranking and selection procedure, we created 24 different questions with corresponding sets of 8 interval judgments for each. For each question the 8 alternative judgments were presented in a randomized order. The 24 questions were divided into four sets with 6 different questions in each. The participants were 24 students from the University of Chicago. They were randomly assigned to answer one of the four sets of questions.

Results

The analyses assessed the fit of the additive model to the subjects' rankings. Each set of eight judgments was ranked from 1 through 8 according to the trade-off model. For instance, Estimates A, B, C, D, E, F, G, and H of the UN question were ranked by the trade-off model ($\alpha = 1$) as follows: 6, 4, 1, 5, 3, 2, 8, and 7, respectively. Then a Spearman correlation was calculated between each participant's ranking and the model's ranking. Correlations were thus calculated for each person and each question sepa-

rately. The mean correlation (across questions and respondents) for the additive model with $\alpha = 1$ is .84. We examined the fit of the additive model while systematically varying the α parameter across wide range. We report the mean correlation for several values across this range: with $\alpha = 0$ the mean correlation is .54; with $\alpha = 0.5$, the mean correlation is .81; with $\alpha = 2.0$ the mean correlation is .76; with $\alpha = 3.0$ the mean correlation is .71; with $\alpha = 8.0$, the mean correlation is .45; and with $\alpha = 100$, the mean correlation is .31. Whereas the maximum mean correlation (.84) is obtained with $\alpha = 1.0$, the mean correlation is close to its maximum for $0.6 \leq \alpha \leq 1.2$.

The mean correlation of the normalized error fit with subjects' rankings is .54. (We note that this is tantamount to the additive model with $\alpha = 0$.) The mean correlation of interval width with subjects' rankings is .26. These results imply that a combination of both components of the additive model is needed to account for the ranking data.

In the analyses described above, a single trade-off parameter was fit for all respondents and questions. Next, individual trade-off parameters were calculated. In particular, for each of the 24 respondents we computed the α parameter with the highest Spearman correlation between the additive model and his or her subjective ranks. The resulting 24 optimal individual α parameters had a mean of 0.90, a standard deviation of 0.46, an interquartile range between 0.5 and 1.1, and a range between 0.2 and 2.4. The 98% confidence interval for the mean trade-off parameter (using the critical value for t_{23}) ranged from 0.66 to 1.14. It should be noted that the derivation of the trade-off parameters in Studies 1 and 2 involved entirely different analytical tools and also different experimental methods. In Study 1, the optimal α values were derived from the ratios of probit coefficients, whereas in Study 2 they were computed such as to maximize the ordinal (Spearman) correlations. The individual parameter values were nevertheless roughly similar in magnitude across both studies, which attests to the robustness of the findings.

Alternative Models

We next compare the fit of the additive trade-off model to the fit of several other choice models (Table 1). These other models provide useful baselines for assessing the importance of including the dimensions of accuracy and informativeness. In illustrating the various models, we refer to the example above on Jackson's compensation and the two estimates: *(A)* "$1 to 20 million" and *(B)* "$12 to 14 million" (correct answer is $15 million).

First, we consider lexicographic (semiorder) models. They merit attention because they have been suggested as behavioral heuristic choice rules in deciding among multiattribute alternatives (Payne, Bettman, & Johnson, 1988; Tversky, 1969). With a lexicographic procedure, the dimensions of accuracy and informativeness are not numerically added. Instead, these two dimensions are used serially. Estimates are first considered with respect to accuracy. If the alternatives are tied on the first dimension, then the informativeness dimension is invoked. In particular, two alternatives are tied if their values are either identical or "close," that is, if the difference on that dimension is less than some threshold k (Tversky, 1969). Assume, for instance, $k = 1$, and consider Jackson's compensation question. The normalized errors of intervals A and B are 0.24 and 0, respectively. The difference between these values is less than 1; hence the two estimates are tied on accuracy. The choice, which is thus based on informativeness, results in a preference for B because it is narrower. We evaluated the lexicographic semiorder model with different threshold levels (e.g., $k = 0.5$, 1, and 4).

Table 1
Fit of Various Models for Respondents' Rankings in Study 2

Model	Mean correlation
Additive trade-off	.84
Absolute error plus half width[a]	.77
Nearest boundary	.73
Lexicographic semiorder	.82
Absolute error	.82
Normalized error	.54
Interval width	.27
Inclusion	.61

[a] Formally equivalent to a "farthest boundary model" (see text).

We also assessed a simplified version of the additive trade-off model called "absolute error plus half width," defined in terms of the function $L = |t - m| + \frac{1}{2} g$. Considering the estimates of Jackson's compensation, the model assigns a lower score to B than to A, resulting in a preference for B. Note that absolute error plus half width is algebraically equivalent to the "farthest boundary" model, namely, the model that says that in choosing between two intervals, people are less likely to prefer the interval whose remote boundary is farther away. In addition, we evaluated the "nearest boundary" model, which says that respondents' preferences are determined by the distance of the truth from the nearest boundary of the interval estimates. According to it, A is preferred to B if the truth is closer to the near boundary of A than it is to the near boundary of B. The other models in Table 1—absolute error, normalized error, and interval width—were defined in earlier sections.

Additional Analyses of Study 2

In calculating the fits of the models, the eight judgments in each set were ranked from 1 through 8 according to each model. Then, for each question separately, correlations were calculated between each respondent's ranking and the model's ranking. The mean correlation (across questions and respondents) is reported in Table 1 for each model. Although the additive trade-off model (Equation 2) provided the highest fit, several other models achieved relatively high correlations as well. Perhaps this is not completely surprising, as other models include components that correlate with those of the additive trade-off model. The lexicographic semiorder model also takes both accuracy and informative-

ness into account, as does "absolute error plus half width." The rankings produced by these models are thus partially correlated with those of the additive trade-off model. For illustration, estimates C and F of the UN question in Study 2 are ranked 1st and 2nd by several models, including additive trade-off, lexicographic semiorder, and absolute error plus half width. Therefore, a more powerful comparison of the models requires sets of interval estimates for which the models make different predictions.

Study 3: Comparing Models

In this study, we presented to respondents pairs of alternative interval estimates and asked them to indicate their preferences among them. The method was similar to that used in Study 1, except that pairs of interval estimates were specifically constructed to distinguish among the models.

Subjects and Materials

The 30 individuals in this study were recruited from the same population as in Studies 1 and 2. They were given a total of 24 questions. Sample questions are shown in Table 2, along with the predictions of the alternative models.

Results

The percentage of respondents choosing each alternative are shown in parentheses in Table 2 for the sample questions. The mean percentage agreement with the additive trade-off model ($\alpha = 1$) was 87% across the 24 questions. For 23 of the 24 questions, respondents strongly preferred the estimate predicted by the additive trade-off model; for one question, the votes were tied. In contrast, each of the alternative models listed in Table 1 is rejected by data from at least 5 questions out of the 24.

The other models seem less well supported by the data. It is interesting to note why they fared less well than did the additive trade-off model in this study. Some of the alternative models predict preferences that are not supported by social norms. For instance, the absolute error plus half width model implies that when error is held constant, individuals always prefer the more precise interval; thus it does not permit a trade-off between accuracy and informativeness, contrary to the results for Question 3 in Table 2. The absolute error model implies that individuals are indifferent with respect to the interval width of an estimate. Question 5 contrasts two estimates that have the same absolute errors but different graininess. Whereas the trade-off model predicts the direction of preference, the absolute error model

Table 2
Sample Questions, Model Predictions, and Results From Study 3

Question[a]	Truth	Model predictions: Additive ($\alpha = 1$) trade-off	Absolute error	Absolute error + half width	Nearest boundary	Inclusion	Lexicographic semiorder with $k = 1$
1. Number of United Nation member countries? A. 140–150 (90%) B. 50–300 (10%)	159	A	A	A	A	B	B
2. Air distance between Chicago and New York? A. 730–780 miles (90%) B. 700–1500 miles (10%)	713 miles	A	A	A	B	B	A
3. Average number of rainy days in Chicago? A. 160–165 (13%) B. 140–180 (87%)	130	B	B	A	B	B	B
4. Amount of money received by Michael Jackson in 1987 to star in a series of Pepsi commercials? A. $1–20 million (7%) B. $12–14 million (93%)	$15 million	B	B	B	B	A	B
5. Total number of points scored by Kareem Abdul-Jabaar in 19 years of playing basketball? (as of 1987–1988 season). A. 30,000–45,000 (3%) B. 37,000–40,000 (97%)	37,639	B	A	B	B	B	B
Average percentage correct prediction across all 24 questions (Chance performance = 50%)		87%	59%	64%	68%	44%	64%

[a] Percentages of respondents who chose each estimate are shown in parentheses.

ignores the trade-off. Similarly, the inclusion model always gives priority to an interval that contains the truth, regardless of interval width (Question 1).

General Discussion

The theory and results of this research highlight the importance of the accuracy–informativeness trade-off in judgmental processes under uncertainty. Subjects' evaluations of judgments could be predicted from an additive model that weighs accuracy and informativeness. Consistent with this trade-off, participants were willing to accept some error in order to obtain more informative judgments on uncertain quantities. For example, when given a question concerning the "money spent on education by the US" along with the correct answer, $22.5 billion, 80% of respondents said that the judgmental estimate, "$18 to 20 billion," was better than the estimate "$20 to 40 billion," even though only the latter interval included the correct answer. This preference is predicted by the additive trade-off model but not, for instance, by a model of choice that places primary importance on including the true answer in the interval.

This trade-off appears systematic in the sense that a relatively simple model could account for the data. Whereas the additive trade-off model provides an adequate fit of individual's preferences, we do not claim that it is the best descriptive model. Future work could provide refinements of this model with additional dimensions of evaluation, more parameters, and better fit. It is likely, however, that accuracy and informativeness will play a major role in any such model. Moreover, we suggest that the accuracy–informativeness trade-off also affects the production of judgments and the level of precision used. In the following sections, we turn to these issues.

Extensions of the Model

The informativeness and accuracy of judgments are typically assessed at different points in time. The "payoff" for making an informative judgment under uncertainty is immediate, whereas the reward for an accurate (or penalty for an inaccurate) judgment is delayed until the true answer or outcome is observed. Various social and cognitive factors could affect the structure and timing of the payoffs. For example, election candidates may make bold predictions if they believe public memory is short and malleable. The benefit of making a definitive, highly informative statement about expected achievements could outweigh the cost of being proven incorrect at a remote point in the future. In other situations, however, the reverse could be true. Scientists whose assessments are regularly archived and eventually tested might be inclined to hedge their assessments in the interest of protecting their accuracy level.

The trade-off parameter of the trade-off model controls the relative weights on accuracy and informativeness. Suppose, for example, a decision-maker places a high premium on getting accurate estimates before planning a hike through a desert area. Other things being equal, the decision-maker should effectively be willing to accept less informative judgments, a preference that is tantamount to setting a lower trade-off parameter α.

Our model assumed equal penalty for over- and underestimation of the true answers because we had no reason to expect a bias in people's preferences about general-knowledge estimates. This assumption is consistent with our previous results (Yaniv & Foster, in press), which indicated no strong bias one way or another in the production of general-knowledge estimates. However, in some situations the costs of over- and underestimation errors are asymmetric (Einhorn & Hogarth, 1985). For instance, arriving at a movie theater 10 min late is far worse than arriving 10 min early. The additive trade-off model could be revised to accommodate asymmetric costs by applying differential weights for over- and underestimation.

Production of Judgmental Estimates

Perhaps the most important implications of the present results concern the production of judgments under uncertainty (e.g., Huttenlocher, Hedges, & Bradburn, 1990). Receivers' preferences are central, as they shape the process by which judges generate estimates. Judges presumably consider the risk of losing credibility if they claim greater precision than is warranted by their ultimate accuracy. Judges also consider the disutility of excessively coarse (i.e., uninformative) judgments. Clearly, under uncertainty, judges do not know the true answer for sure. However, they can subjectively assess the expected accuracy of their estimation and then, on the basis of that assessment, choose a level of graininess of judgment that is also sufficiently informative. In future research, the choice of graininess could be modeled by replacing the normalized error (in the additive model) with the expected normalized absolute error, for which the expectation is calculated over some subjective distribution of the true estimate.

One of the studies in Yaniv and Foster (in press) actually provides some empirical evidence consistent with the notion that judges provide balanced judgments in terms of their expected accuracy and informativeness. Judges made interval estimates "that they felt comfortable with" concerning questions such as "the date the University of Chicago was founded" or "the date the polio vaccine was discovered." In particular, they were supposed to use one of several "grain scales" that differed in their precision. For instance, one scale allowed them to indicate a 100-year period as their estimate. Another scale allowed them to indicate a 50-year period as their estimate. A third scale allowed them to indicate a 10-year period as their estimate. Other scales allowed them to indicate 5- and 1-year periods. The overall hit rate (proportion of intervals that included the true answers) in this study was 55%. The hit rates for the five scales (from the coarsest to the most precise) were 51%, 37%, 46%, 55%, and 56%, respectively. This constancy in the accuracy levels across scales is striking. It is consistent with the notion that people choose a level of precision that preserves some stable trade-off between expected accuracy

and informativeness. Future work could examine more directly the accuracy–informativeness trade-off in the production of judgmental estimates.

Finally, we wish to emphasize that despite our present focus on the accuracy–informativeness trade-off in judgmental estimation, the issues studied here have close parallels in a variety of other situations. The conflict between being precise and being right occurs, for instance, in science, where one often faces a tension between commitment to precise theoretical statements and the need for moderate conclusions to account for variability in the data. Whereas many subscribe to the notion that "I'd rather be precisely wrong than vaguely right," others have shown preference for being "vaguely right." Our main point is that being "precisely right" is rarely an available option, hence the accuracy–informativeness trade-off implicated in such quotes.

In a similar vein, the issues studied here have some close analogs in categorization research. Within a hierarchical category system, a particular object may be classified at various levels of specificity. Superordinate categories are coarsely grained, whereas subordinate categories are finely grained (Bar-Hillel & Neter, 1993; Smith & Medin, 1981). Under conditions of certainty, the considerations in choosing a label, such as the name *dog* rather than *animal* or *collie,* are based on pragmatic aspects of the communication, for instance, the speaker's goal in the conversation and the listener's ability to comprehend the message (Brown, 1958; Grice, 1975; Teigen, 1990). We suggest that uncertainty breeds additional considerations. Consider an antiques expert who is uncertain about the origin of a piece of furniture and thus contemplates one of two potential descriptions, a specific description such as "French dining table from the 1800s" or the more general one, "antique dining table." The present research on judgment under conditions of uncertainty implies that the expert's choice of a category label would depend on his or her trade-off between accuracy and informativeness. A full-fledged model of categorical judgment under uncertainty would, therefore, require appropriate definitions for graininess, accuracy, and informativeness.

References

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky, *Judgment under uncertainty: Heuristics and biases* (pp. 294–305). New York: Cambridge University Press. (Original work published 1969)

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology, 65,* 1119–1131.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65,* 14–21.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188.

Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review, 92,* 433–461.

Erev, I., Wallsten, T., & Neal, M. (1991). Vagueness, ambiguity, and the cost of mutual understanding. *Psychological Science, 2,* 321–324.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). New York: Academic Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Huttenlocher, J., Hedges, L., & Bradburn, N. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 196–213.

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

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4,* 249–262.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Sniezek, J. A., & Buckley, T. (1991). Confidence depends on level of aggregation. *Journal of Behavioral Decision Making, 4,* 263–272.

Teigen, K. H. (1990). To be convincing or to be right: A question of preciseness. In K. J. Gilhooly, M. T. G. Keane, R. H. Logie, & G. Erdos (Eds.), *Line of thinking* (pp. 299–313). New York: Wiley.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76,* 31–48.

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

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meaning of probability terms. *Journal of Experimental Psychology: General, 115,* 348–365.

Yaniv, I., & Foster, D. P. (in press). Precision and accuracy in judgmental estimation. *Journal of Behavioral Decision Making.*

Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological Science, 4,* 58–62.

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. (1990). *Judgment and decision making* (pp. 75–111). Englewood Cliffs, NJ: Prentice-Hall.

Received February 15, 1995
Revision received May 22, 1995
Accepted June 15, 1995 ■