Journal of Behavioral Decision Making, Vol. 10, 21–32 (1997)

Precision and Accuracy of Judgmental Estimation

ILAN YANIV
Hebrew University, Jerusalem, Israel

DEAN P. FOSTER
Wharton School, University of Pennsylvannia, USA

ABSTRACT

Whereas probabilistic calibration has been a central normative concept of accuracy in previous research on interval estimates, we suggest here that normative approaches for the evaluation of judgmental estimates should consider the communicative interaction between the individuals who *produce* the judgments and those who *receive or use* them for making decisions. We analyze precision and error in judgment and consider the role of the accuracy–informativeness trade-off (Yaniv and Foster, 1995) in the communication of estimates. The results shed light on puzzling findings reported earlier in the literature concerning the calibration of subjective confidence intervals. © 1997 by John Wiley & Sons, Ltd.

KEY WORDS judgmental estimation; subjective confidence interval; graininess of judgment; accuracy-informativeness trade-off

In the process of making our daily decisions, we commonly solicit from judges estimates and forecasts of uncertain quantities. For instance, estimates of the amount of driving are relevant for planning a trip, estimates of market prices are helpful for selling a used car, and forecasts of inflation rate are relevant in selecting a mortgage. Judges and forecasters for their part may communicate the uncertainty in their own estimates using expressions of probability (Erev and Cohen, 1990; Wallsten, 1990; Yaniv, Yates and Smith, 1991) or by intentionally varying the precision or 'graininess' of their estimates (Yaniv and Foster, 1995). For instance, in estimating the duration of a trip, one could communicate '4 to 5 hours' or '3 to 6 hours'. The present study focuses on the communication of such judgmental interval estimates under conditions of uncertainty.

A priori one could hypothesize that in setting an interval width for an uncertain quantity (e.g. 'air distance for Chicago to New York'), people simply set lower and upper bounds on potential judgmental errors. This, however, is not the case. The main impetus for this work arises from a classic finding due to Alpert and Raiffa concerning the calibration accuracy of interval estimates (1982, initially reported in 1969; Lichtenstein, Fischhoff, and Phillips, 1982). In their study, students estimated a 98% confidence interval for each of several uncertain quantities. A set of subjective 98% confidence intervals is called well calibrated if the intervals include the correct answers 98% of the time. The actual proportion of intervals that failed to include the true answer equaled 42% rather than

CCC 0894–3257/97/010021–12

Received 27 October 1994
Accepted 8 May 1996

2%, the value expected form well-calibrated judges. Subjective confidence intervals were thus too narrow as they excluded the correct answers far too often.

The difficulty of specifying intervals that contain the truth with some high probability (e.g. above 90%) is a puzzling phenomenon which has raised theoretical as well as practical questions (Lichtenstein *et al.*, 1982; Sniezek and Buckley, 1991; Tversky and Kahneman, 1974; Yates, 1990, Chap. 4; Yates, McDaniel and Brown, 1991). In classroom demonstrations of this phenomenon, students are frequently surprised that their confidence intervals so often exclude the correct answers. In principle, individuals could widen their subjective confidence intervals to increase the chances of including the correct answer, but they do not seem to do so even after receiving warnings (Alpert and Raiffa, 1982). Related methods designed to improve the calibration of confidence intervals have met relatively little success (Lichtenstein and Fischhoff, 1980; Lichtenstein *et al.*, 1982; Murphy and Winkler, 1977; Russo and Schoemaker, 1992; Seaver, von Winterfeldt, and Edwards, 1978).

In general, previous research has focused on *probabilistic calibration* as a normative standard for accuracy. One could conclude from previous work that subjective confidence intervals may be of dubious value and cannot be taken at face value as intervals that include the truth with some high probability, such as 90% or 98%.

In this work, we shift away from the focus on probabilistic calibration and suggest that a different normative approach might be applied in the evaluation of interval judgments (see also Yaniv and Foster, 1995). This approach is based on the observation that the communication of forecasts of future outcomes and judgmental estimates of unknown quantities, often takes place in the course of making a decision. Decision makers, in particular, often solicit estimates from experts and others. The norms for the evaluation of judgmental estimates in such cases are predominantly governed by the structure of the communicative interactions between the individuals who *produce* the judgements and those who *receive or use* them in making decisions.

As an illustration, consider forecasts of future events such as 'inflation rate next year'. Under uncertainty, a forecaster might contemplate guesses such as (A) '5%', (B) '4–6%', and (C) '2–20%'. Clearly, coarser estimates such as C have higher chances of being accurate. Social norms, however, preclude the communication of excessively coarse judgements and suggest that judges should be appropriately informative as well as accurate (Grice, 1975). Thus the estimate '2–20%' may not be appreciated by recipients even though it is likely to be confirmed (Tversky and Kahneman, 1983, p. 312).

We examined this conjecture concerning recipients' preferences in several studies (Yaniv and Foster, 1995) where respondents were given, for instance, estimates of the 'number of United Nation (UN) member countries (in 1987)'. In one case, the following two estimates were given (A) '140–150' and (B) '50–300'. Respondents were also told that the correct answer was 159 and then asked to indicate which estimate was better. Most (90%) of the respondents preferred estimate A over B, even though only the latter included the correct answer (Yaniv and Foster, 1995, Study 3). Thus respondents were willing to accept some error in order to obtain more informative judgments.

To account for such results, we suggested that the communication of judgments under uncertainty involves a trade-off between two countervailing objectives: accuracy and informativeness (Yaniv and Foster, 1995). In particular, we assumed that the standard for assessing the accuracy of an interval estimate is the judge's own stated precision. We defined a measure of accuracy as the error-to-precision ratio $(t - m)/g$, where g is interval width, t is the true answer, m is the judge's best point estimate which is sometimes explicitly indicated by the judge or, otherwise, the midpoint of the interval. This continuous measure, which we called *normalized error*, captures our intuition that the psychological evaluation of a judgmental error $(t - m)$ depends not only on the magnitude of the error but also on the precision (g) claimed by the judge, namely, whether the judge indicates a precise or coarse estimate. For instance, an erroneous judgment stated with high precision might be disliked more than a similarly erroneous judgment stated with less precision. Consider the accuracy of two estimates of the number of UN members: (A) '120–140' and (B) '130–132'. Although both estimates miss the truth (there were 159 UN members in 1987), they might be evaluated differently. The width of the first estimate is 20, thus its normalized error is about 1.5; the second estimate ($\text{width} = 2$) has a normalized error of about 14. In terms of normalized error alone, B is less accurate, reflecting the fact that its absolute error is large relative to the claimed precision.

In our accuracy–informativeness trade-off model, accuracy is expressed in terms of the normalized error, whereas informativeness is expressed as a monotone function of the interval width (for detailed definitions, see Yaniv and Foster, 1995). The trade-off arises in this model because widening an interval estimate improves its accuracy (i.e. decreases normalized error) but impairs its informativeness. The accuracy–informativeness trade-off model accounts for recipients' evaluations of judgments such as the UN question above (Yaniv and Foster, 1995).

Recipient's preferences for judgments that are balanced in terms of their accuracy and informativeness suggests reasons to believe that producers of judgments (judges) should also be attuned to the accuracy–informativeness trade-off. According to conversational norms, judges are generally motivated to cooperate and respond to recipients' preferences (Grice, 1975). In addition, judges may gain (lose) social reputation for providing good (poor) forecasts. Interestingly, the 'timing' of the rewards that judges receive highlights the need for informativeness. Rewards for being informative are immediate, as recipients evaluate the informativeness of a forecast upon hearing it. Rewards for being accurate are typically delayed to a later point in time when the relevant feedback becomes available and the forecast's accuracy can be assessed. This timing difference may further induce judges to provide highly informative estimates.

Whereas in Yaniv and Foster (1995) we directly examined the accuracy–informativeness trade-off in recipients' preferences for judgments, in this work we examine the production of judgments. We conducted three studies. Each involved a different method of eliciting estimates for a variety of general knowledge questions, such as 'number of United Nations member countries' and 'height of Mount Everest'. In Study 1, the procedure simulated the use of interval estimates in natural situations. For each question, several scales were provided that allowed a choice among various levels of 'graininess'. In Study 2, we asked respondents to estimate 95% confidence intervals. In Study 3, we asked respondents first to give point estimates and then directly estimate their own errors.

We evaluated the accuracy of interval estimates using measures derived from the trade-off model. First, we examined the relationship between the precision (width) of interval estimates g and absolute error $|t - m|$, where t is the true answer and m is the judge's best point estimate or the midpoint of the interval. This analysis is based on the notion that precision signals to recipients the magnitude of the error they might expect. Thus, when stating a precise (coarse) interval, judges imply they expect a small (large) absolute error. This interpretation of precision differs from the notion that judges provide interval estimates that are meant to contain the correct answer with near-certainty, and therefore place less importance on hit rates.

In addition, the error-to-precision ratios are of interest as they provide indirect, comparative indices of the 'direction' of the actual trade-off between accuracy and informativeness in different conditions. Other things being equal (e.g. holding the judge and question constant), an average error-to-precision ratio of 2 represents a greater emphasis on informativeness than an average ratio of 1 (an average ratio of 1 implies that error and interval width are of the same order of magnitude whereas a ratio of 2 implies that interval width is half as large as error). With this interpretation in mind, we examine whether the error-to-precision ratio (and hence the trade-off) varies as a function of the elicitation method.

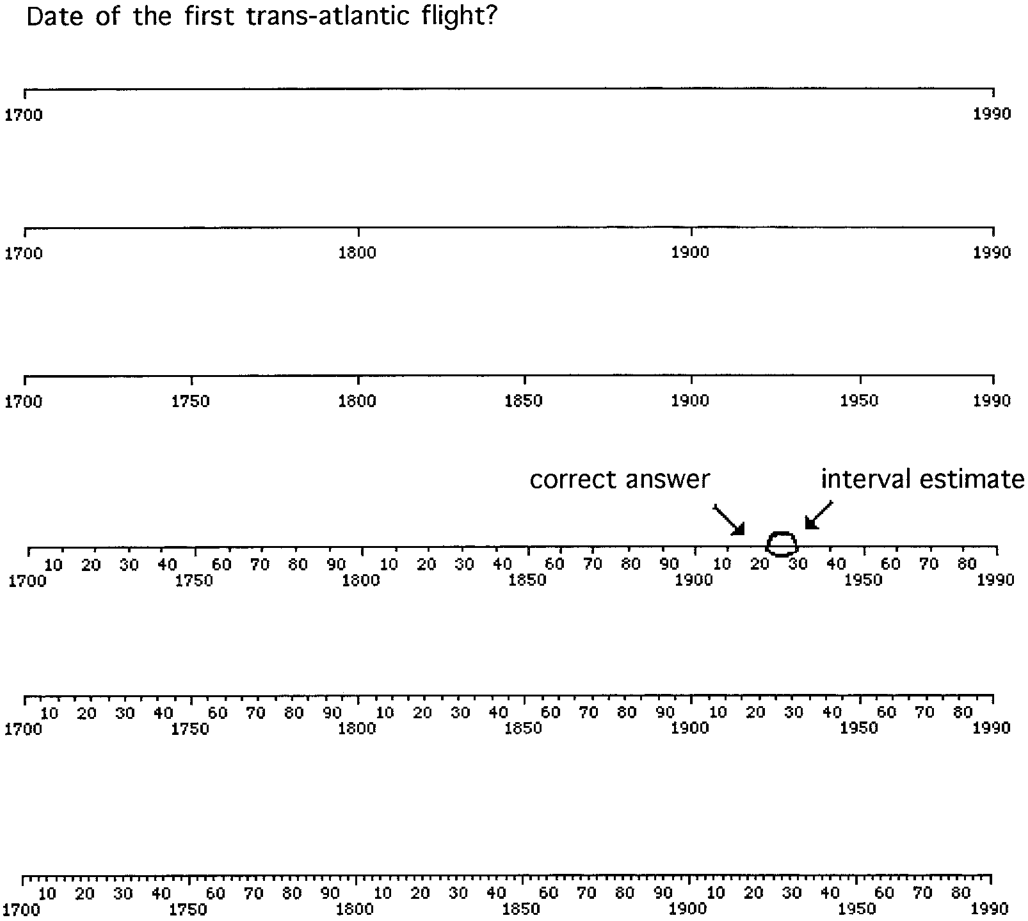

Exhibit 1. A sample question and a hypothetical respondent's answer demonstrating the grain-scale method for eliciting an interval judgment (Study 1).

STUDY 1: GRAIN SCALES

The first study involved the grain-scale method illustrated in Exhibit 1. Scales varying in precision or graininess were provided for answering the question (e.g. 'date of the first transatlantic flight'). The 'graininess' levels of these scales represented a variety of units ordinarily used in natural language communication (e.g. decades or centuries in estimating historical dates). Respondents were supposed to pick *only one of the scales* and mark *one interval* on that scale for an answer. They had to mark one *full* interval extending between two adjacent tick marks. Thus, for instance, they were not supposed to mark a 10-year period on a scale with tick marks for every 50 years. The answer illustrated in Exhibit 1 corresponds to the decade 1920–30.

The top (coarsest) scale in Exhibit 1 was provided in the event that respondents knew nothing about the topic, in which case they were supposed to circle the whole range. Otherwise, they could make a more precise judgment on *one* of the other five scales. For instance, they could circle an interval representing a century on the second scale, a decade on the fourth scale, or write down the exact year above the sixth scale (with precision of one year). The respondents were told to answer these questions as if they were posed by a friend in a casual conversation and to provide estimates that they 'felt comfortable with '. This design (in which respondents select among preset interval widths) enables us to examine the idea that respondents maintain some constant ratio of error to precision.

The questionnaire included 42 general-knowledge questions which spanned a variety of topics including history, geography, science, business, census information and sports. The format of the questions was similar to that shown in Exhibit 1. Each question was presented along with either four, five, or six scales, with the top scale for each question always consisting of a single interval that comprised the entire range. The respondents were 44 University of Chicago students. They each answered all 42 questions.

STUDY 2: 95% CONFIDENCE INTERVALS

In contrast to Study 1, where respondents were asked to indicate a level of precision that they 'felt most comfortable with', in Study 2 respondents were supposed to adhere to a well-defined probabilistic requirement. For each uncertain quantity, respondents ($N = 43$) estimated a 95% confidence interval. Specifically, they were supposed to provide low and high estimates such they were 95% confident that the correct answers fell within the intervals defined by the estimates. Their goal was to provide intervals that include the correct answer in 95% of the cases; in other words, only 5% of their answers could be in error.

STUDY 3: PLUS/MINUS ESTIMATES

In Study 3, the respondents ($N = 44$) gave their best guesses and indicated the precision of their guesses using the plus/minus method. The questionnaire listed the 42 questions (same as Studies 1 and 2 along with requests to indicate 'best guess' and 'plus-or-minus error' . For each uncertain quantity (e.g. 'date polio vaccine was discovered'), respondents first made a point estimate (under the best guess column, e.g. '1930') and indicated the error they expected for their own judgment (under the plus-or-minus column, e.g.'±15 years').

RESULTS

Studies 1, 2, and 3 were meant to examine different elicitation methods. For example, the instructions of Study 2 emphasize accuracy whereas those of Study 1 suggest an everyday communication context, and hence may highlight the need for informativeness. We found, however, few differences among the three studies, and our conclusions generalize across the different methods of elicitation.

Hit rate

In Study 1, the mean hit rate (proportion of intervals that include the truth) was 55%. Whereas these results are comparable to previous ones (Lichtenstein *et al.*, 1982), the grain-scale method of Study 1 allows us further insight into the judgment process.

One might suggest that the average hit rate of 55% simply results from averaging across scales varying in graininess (i.e. averaging of high hit rates obtained with coarse scales and low hit rates

Exhibit 2. Hit rate by grain scale in Study 1

	Hit rate (%)	
Scale	Observed	Hypothetical[b]
1st[a]	100	100
2nd	51	67
3rd	37	43
4th	46	24
5th	55	13
6th	56	5

[a]The scales are ordered from the coarsest (top scale) to the most precise scale. Answers marked on the first scale always included the correct answer (see Exhibit 1).
[b]The hypothetical hit rates are those that would have obtained had respondents used the same scale in answering all questions.

obtained with finer scales). The different patterns of observed and hypothetical hit rates in Exhibit 2 indicate otherwise, however. Hypothetically, had judges indicated their best point estimate *m* for each question on the most precise (6th) scale, they would have obtained 5% hit rate; if, in contrast, they had indicated *m* on the second coarsest scale for each question, they would have achieved 67% hit rate, with the rates for other scales varying between. However, the observed hit rates were generally similar for all grain scales, suggesting that judges systematically compensated for their uncertainty by increasing the interval width (the top scale is an exception because it always included the true answer).

In Studies 2 and 3, the intervals contained the correct answer in 43% and 45% of the cases, respectively (see Exhibit 3). Studies 2 and 3 did not differ in hit rate, $t < 1$. Hit rates in Study 1 were significantly higher than in Study 2, $t(85) = 3.10$, $p < 0.05$, and also higher than in Study 3, $t(86) = 3.26$, $p < 0.05$. The grain-scale procedure may have inflated the hit rate (55%) because the answers marked on the top scale always included the truth (see Exhibit 1). Indeed, if we eliminate the estimates marked on the top scales, then the mean hit rate in Study 1 drops to 46% — a level similar to that obtained in Studies 2 and 3.

Exhibit 3. Averages and interquartile ranges for performance measures: Studies 1–3

	Study 1: grain scale ($n = 44$)	Study 2: confidence interval ($n = 43$)	Study 3: plus/minus ($n = 44$)
Hit rate	55%	43%	45%
	41–67%[a]	29–52%	38–50%
Abs error-to-precision ratio	0.70[b]	1.08	0.71
	0.51–0.90	0.50–1.48	0.49–0.84
95% calibration factor	9[c]	17	15
	5–13	11–25	9–19
Error-precision correlation	0.82[d]	0.76	0.71
	0.78–0.85	0.71–0.82	0.65–0.76

[a]Measures of performance were calculated individually for each respondent. The first row for each measurement shows the mean of the individual averages. The second row show the interquartile range for the individual means.
[b]Mean of the individual median error-to-precision ratios (normalized errors).
[c]The factor by which intervals should be widened in order to achieve a 95% hit rate.
[d]Correlations were computed for each respondent separately across questions.

Error-to-precision ratios

The normalized errors of the estimates were computed using the transformation $(t - m)/g$, where t is the true answer, m is the midpoint of the interval, and g is the interval width. This transformation makes it possible to aggregate the data across questions. As an illustration, note that the answer shown in Exhibit 1 (1920–30) has a normalized error of -0.6, less than one interval away from the interval containing the correct date, 1919. (For further illustration, the answer '1900–1905' on the fifth scale would have a normalized error of roughly $+3.2$, while the answer '1900–1950' on the third scale would have a normalized error of -0.12.)

In Study 2, g was defined as the width of a confidence interval. For example, if a respondent's 95% confidence interval for 'height of Mount Everest' was '25,000 to 30,000 feet', then g equaled 5000 feet with m being the midpoint of that interval. In Study 3, each judgement involved a point estimate (m) and a symmetrical plus/minus error estimate (x). Thus $[m - x,\ m + x]$ was considered to be the corresponding interval judgment.

The distributions of the normalized errors (rounded to the nearest integer) are shown in Exhibit 4 for Studies 1–3. Note that an interval contains the true answer t if and only if $|(t - m)/g| \leqslant 0.5$. Thus the bar above zero represents the proportion of intervals that included the truth (hit rate). The densities of the distribution over various ranges are also shown in Exhibit 4. In Study 1, 80% of the judgment had normalized errors of -1, 0, or $+1$, whereas 88% of the judgments had normalized errors that ranged from -2 to $+2$. The results, which generalize across the three methods of elicitation, have clear implications for recipients. They suggest, for instance, that an individual who claims precision of 'one decade' in guessing a historical date is likely to make an error between 0 and two decades, but is far less likely to err by a hundred years — error-to-precision ratios greater than 10 in absolute value occurred in only 1.8% of the cases.

These conclusions are supported by individual analyses of the absolute error-to-precision ratios (normalized errors). For each individual, the median absolute error-to-precision ratio was calculated. The average and interquartile ranges of the individual medians are shown in Exhibit 3. The absolute normalized errors (Exhibit 5) were fairly close to one another in all studies, although the normalized errors in Study 2 were greater than in Study 1, $t(85) = 2.64$, $p < 0.05$, or Study 3, $t(86) = 2.64$, $p < 0.05$. Studies 1 and 3 did not differ in terms of the normalized errors. We next compared the interval widths obtained in the three studies (the means of the individual medians were 11.4, 13.2, and 12.9, for Studies 1, 2, 3, respectively) and found no significant differences. Notably, the 95% confidence-interval method did not produce significantly wider intervals than the other methods.

Ninety-five per cent calibration factors

A well-calibrated judge in Study 2 is expected to include the correct answer in 95% of her confidence intervals. None of the subjects in Study 2 was calibrated. How narrow were the intervals obtained in Study 2 relative to *calibrated* 95% confidence intervals? That is, by what factor should judges extend their intervals in order to achieve a 95% hit rate?

For each judge, we calculated the factor by which his intervals should be widened so as to include the true answers in 95% of the cases. The lowest 95% calibration factor in Study 2 was 5. Note that extension by a factor of 5 means that the interval $[m - 0.5g,\ m + 0.5g]$ is transformed into $[m - 2.5\ g,\ m + 2.5\ g]$, where $m =$ midpoint and $g =$ interval width. The median 95% calibration factor was 17. The 95% calibration factors for Studies 1 and 3 (Exhibit 3) provide a useful comparison, although the respondents in those studies were not supposed to reach a 95% hit rate. The 95% factors in Study 1 were lower in magnitude because the scales provided information that was not available to respondents in the other studies.

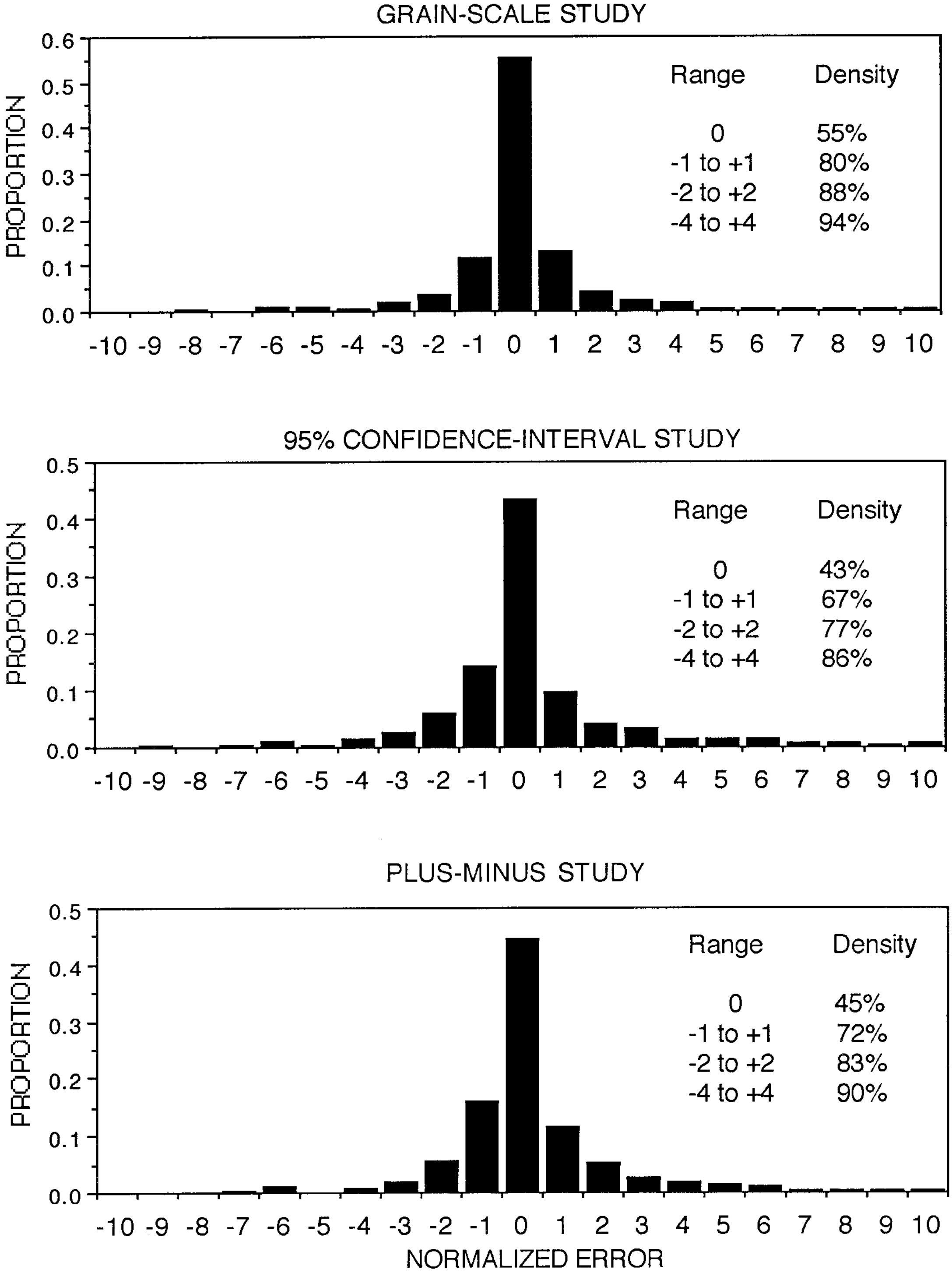

Exhibit 4. Distributions of normalized errors from Studies 1–3. Extreme normalized errors (less than −10 or greater than + 10) occurred in 1.8%, 5.5%, and 2% of the cases in Studies 1, 2, and 3, respectively.

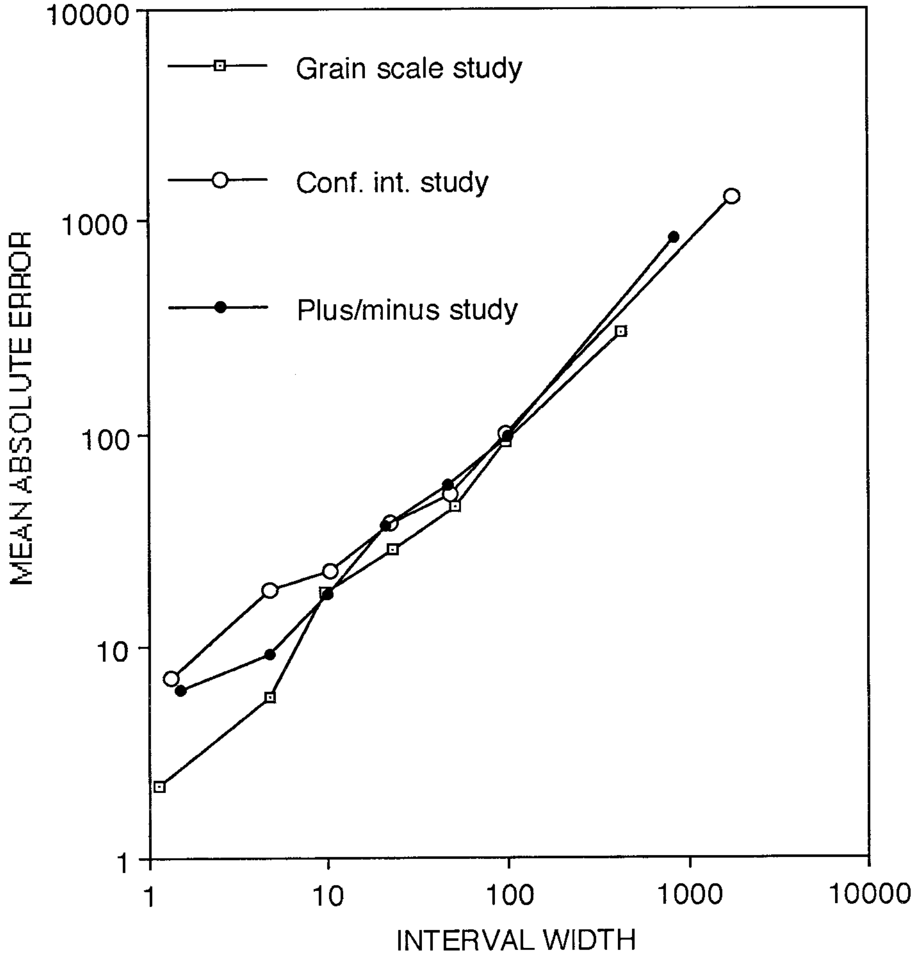

Exhibit 5. Mean absolute error plotted as a function of interval width for Studies 1–3.

Precision versus error

The relationship between the precision and absolute error was investigated at both the individual and the aggregate levels. In the aggregate analysis, the estimates were first sorted into seven categories according to the widths of their intervals. Judgments with precision levels in the range of [1–2.5) were grouped in the first category, those with precision in the range of [2.5–7) were placed in the second. The remaining categories were [7–14), [14–34), [34–70), [70–140), [140 and above). As shown in Exhibit 5, there is a monotone relationship between interval width and mean absolute error (e.g. the linear slope obtained by regressing the mean absolute error on mean interval width was 0.84 in Study 1).

In addition to the aggregate analyses, we performed individual analyses to examine whether interval width g predicts absolute judgmental error $|t - m|$. Following logarithmic transformation of both variables, absolute error was regressed on interval width. Regression slopes and correlations were computed individually for each respondent. Exhibit 3 shows the mean of the individual correlations between interval width and absolute error for each study. The individual analyses also suggest a monotone relationship indicating that the size of the error can be predicted from the precision of the answer. The correlations did not differ significantly across the three studies (t's < 1). The individual regression slopes averaged 0.82. 0.84, and 0.83, in Studies 1–3, respectively.

GENERAL DISCUSSION

Probabilistic calibration has been a primary concept of accuracy in previous work on judgment (Yates, 1990). Technically, a judge is well calibrated if his or her hit rate (proportion of intervals that includes the true answer) equals his or her confidence level. Studies find, however, that interval judgments fail to include the correct answers in approximately half of the cases, even when respondents are supposed to provide intervals that include the truth with a probability of 95% (Alpert and Raiffa, 1982; Lichtenstein *et al*., 1982; Russo and Schoemaker, 1992).

We have suggested that judgmental estimation are often elicited as part of social exchange and are therefore expected to obey certain social normative standards (Yaniv and Foster, 1995). These social norms may diverge from and even supersede the demands of calibration accuracy. Judges are expected to provide judgments that are not only accurate but also informative. Under uncertainty, however, accuracy and informativeness are countervailing objectives with trade-off between them. In providing interval estimates, judges must sacrifice accuracy (e.g. be willing to accept lower hit rates) in order to communicate sufficiently informative (precise) estimates.

The extent of this sacrifice is dramatically revealed by the '95% calibration factors' in Study 2. We computed the factor by which each individual judge would have to extend his or her intervals so as to achieve an overall hit rate of 95%. The minimum 95% calibration factor in Study 2 was 5 and the median factor was 17 (Exhibit 3). To illustrate this result, imagine a judge whose intuitive estimate of the US population is '200–260 millions'. A calibration factor of 5 means that the judge should report '80–370' millions instead. The median calibration factor (17) would lead to far wider intervals.

Constructed in this way, the recalibrated 95% confidence intervals are extremely broad and thus uninformative. In class demonstrations of the 95% confidence-interval task, students often claim that extending intuitive interval estimates by the 95% calibration factors observed in our studies would yield worthless estimates. Some point out that a strategy of giving an enormous range (e.g. 'zero to one billion') in response to 95% of the questions would be adaptive to this task, but often maladaptive in real-life situations. Such reactions indicate that respondents dislike making excessively coarse judgments (Yaniv and Foster, 1990, 1995).

Indeed, we did not see a preponderance of very coarse judgments in Study 1, where respondent had a choice among grain scales varying in coarseness (see Exhibit 1). They used the coarsest (top) scale in only 16% of the cases and the second coarsest scale in 20% of the cases; the more finely grained scales were used in 64% of the cases. This pattern is consistent with the notion that informativeness is a powerful motive. The constancy of the hit rates across the varying graininess (precision) levels (Study 1, Exhibit 2) is consistent with the idea that judges generally preserved a balance between informativeness and accuracy. In recent work Bar-Hillel and Neter (1993) asked their respondents to predict whether a given individual (whose description was given) was a member of some basic-level category (e.g. department of physics) or the encompassing superordinate category (natural sciences). Their respondents often predicted the narrower category even though predicting the wider (superordinate) category would have guaranteed greater accuracy, and in one of the studies would have resulted in a higher expected payoff. Bar-Hillel and Neter's results are consistent with the notion that the need for informativeness exerts an influence that is difficult to overcome even when the instructions suggest otherwise.

One primary focus of this work has been on the possibility that the precision of a judgment (interval width) signals to recipients the magnitude of the error to be expected. The positive monotonic relationship that was found between precision and absolute error substantiates this idea. One might be better off interpreting intuitive interval judgments as predictors of absolute error rather than ranges that include the truth with some high probability, such as 90% or 99%. Moreover, the individual median error-to-precision ratios were mostly between 0.5 and 1.5 in our three studies (see the inter-quartile ranges in Exhibit 3). Roughly speaking, then, precision and error were on the same order of magnitude.

Note that the absolute error of a judgment reflects the judge's knowledge and is not subject to strategic behavior. In contrast, the error-to-precision ratio reflects not only knowledge but also strategic behavior. For instance, obtaining a relatively high average error-to-precision ratio in a study might suggest that greater importance was placed on informativeness. Therefore, we suggested earlier that, other things being equal (e.g. same questions and same level of knowledge), one could compare the average absolute error-to-precision ratios obtained in different studies to infer differences in the trade-off being made between accuracy and informativeness. We found that the average absolute ratios in the three studies were generally similar, thus offering little reason to believe that the different elicitation methods induced different trade-off levels.

Finally, we discuss the implications of the present results for related research. First, analyses of the relationship between precision and error could clarify findings on the effects of expertise on judgmental estimation. Consider a study of experts' judgmental estimates by Russo and Schoemaker (1992), in which business managers estimated 90% confidence-intervals for uncertain quantities in their areas of expertise (e.g. petroleum, banking, etc). The hit rates obtained in various samples of managers ranged from 38% to 58%, performance levels similar to those typically found in studies of lay people (cf. Yates, McDaniel and Brown, 1991). While these results are surprising, we suggest that they do not necessarily imply that expertise fails to improve estimation. It is possible that experts make relatively smaller errors *and* also provide more precise estimates than do lay people. Hit rate, which is a joint function of both precision and error, might mask these beneficial effects of knowledge. Clearly, further work would be needed to examine this conjecture.

Similarly, our work may have implications for the aggregation of opinions (Yaniv and Hogarth, 1993). A recurring practical and theoretical problem for decision makers is the need to form a composite estimate when presented with the opinions of different individuals who disagree with one another. Whereas averaging has been a common method for the aggregation of estimates (Ashton and Ashton, 1985), an interesting question for further work is whether weighting estimates by their precision could further improve the accuracy of the composite estimate (Yaniv, 1996).

ACKNOWLEDGMENTS

This research was supported by grants from the Graduate School of Business at the University of Chicago, the Israel Foundations Trustees, and the Israel Science Foundation (administered by the Israel Academy of Sciences and Humanities).

REFERENCES

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

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

Bar-Hillel, M. and Neter, E. 'How alike is it versus how likely is it: A disjunction fallacy in probability judgments', *Journal of Personality and Social Psychology*, **65** (1993), 1119–31.

Erev, I. and Cohen, B. L. 'Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox', *Organizational Behavior and Human Decision Processes*, **45** (1990), 1–18.

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

Lichtenstein, S. and Fischhoff, B. 'Training for calibration', *Organizational Behavior and Human Performance*, **26** (1980), 149–171.

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

Murphy, A. H. and Winkler, R. L. 'The use of credible intervals in temperature forecasting: Some experimental results', in H. Jungermann and G. deZeekuw (eds), *Decision Making and Change in Human Affairs*, Amsterdam: Reidel, 1977.

Russo, J. E. and Schoemaker, P. J. H. 'Managing overconfidence', *Sloan Management Review*, **33** (1992), 7–17.

Seaver, D. A., von Winterfeldt, D. and Edwards, W. 'Eliciting subjective probability distributions on continuous variables', *Organizational Behavior and Human Performance*, **21** (1978), 379–91.

Sniezek, J. A. and Buckley, T. 'Confidence depends on levels of aggregation', *Journal of Behavioral Decision Making*, **4** (1991), 263–72.

Tversky, A. and Kahneman, D. 'Judgment and uncertainty: Heuristics and biases', *Science*, **185** (1974), 1124–31.

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

Wallsten, T. S. 'The costs and benefits of vague information', in R. H. Hogarth (ed.) *Insights in Decision Making: A tribute to Hillel J. Einhorn* (pp. 28–43), Chicago, IL: University of Chicago Press, 1990.

Yaniv, I. 'Weighting and trimming: Heuristics for aggregating judgments under uncertainty', Working paper, Hebrew University, 1996.

Yaniv,I. and Foster, D. P. 'Judgment, graininess, and categories', in *Cognitive Science Proceedings* (pp. 133–140), Hillsdale, NJ: Lawrence Erlbaum, 1990.

Yaniv, I. and Foster, D. P. 'Graininess of judgment under uncertainty: An accuracy–informativeness tradeoff', *Journal of Experimental Psychology: General*, **124** (1995), 424–432.

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

Yaniv, I. Yates, J. F. and Smith, J. E. K. 'Measures of discrimination skill in probabilistic judgment', *Psychological Bulletin*, **110** (1991), 611–17.

Yates, J. F. *Judgment and Decision Making* (pp. 75–111), Englewood Cliffs, NJ: Prentice Hall, 1990.

Yates, J. F., McDaniel, L. S., and Brown, E. S. 'Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise', *Organizational Behavior and Human Decision Processes*, **42** (1991), 145–71.

Authors' biographies:

Ilan Yaniv received his PhD in psychology from the University of Michigan in 1988. He now teaches at the Hebrew University.

Dean P. Foster received his PhD in statistics from the University of Maryland in 1988. He now teaches at the University of Pennsylvannia.

Authors' addresses:

Ilan Yaniv, Department of Psychology, Hebrew University, Mt Scopus, Jerusalem, Israel. E-mail: msilan@pluto.mscc.huji.ac.il.

Dean P. Foster, Department of Statistics, Wharton School, University of Pennsylvannia, Philadelphia, PA 19104, USA. E-mail: foster@hellspark.wharton.upenn.edu.