Perception & Psychophysics
1994, 55 (4), 412-428

The calibration and resolution of confidence in perceptual judgments

JOSEPH V. BARANSKI
Defence and Civil Institute of Environmental Medicine, North York, Ontario, Canada

and

WILLIAM M. PETRUSIC
Carleton University, Ottawa, Ontario, Canada

Confidence rating based calibration and resolution indices were obtained in two experiments requiring perceptual comparisons and in a third with visual gap detection. Four important results were obtained. First, as in the general knowledge domain, subjects were underconfident when judgments were easy and overconfident when they were difficult. Second, paralleling the clear dependence of calibration on decisional difficulty, resolution decreased with increases in decision difficulty arising either from decreases in discriminability or from increasing demands for speed at the expense of accuracy. Third, providing trial-by-trial response feedback on difficult tasks improved resolution but had no effect on calibration. Fourth, subjects can accurately report *subjective errors* (i.e., trials in which they have indicated that they made an error) with their confidence ratings. It is also shown that the properties of decision time, conditionalized on confidence category, impose a rigorous set of constraints on theories of confidence calibration.

People are often faced with the task of choosing between two alternatives. Sometimes, the correct choice is clear and can be made with a high degree of confidence. However, even when the correct choice is not immediately evident, we very often rely on our confidence to estimate the accuracy of our decisions. Indeed, because confidence judgments are so pervasive in everyday life and often play an important role in the decision-making process (see Lichtenstein, Fischhoff, & Phillips, 1982; Vickers, 1979), it is important to understand the extent to which these "metacognitive" confidence judgments reliably predict decision accuracy. The formalization of this problem is called *calibration*.

Calibration refers to the correspondence between a probability assessment, expressed as the *subjective probability*, of the occurrence of a particular event (e.g., rain tomorrow) (see, e.g., de Finetti, 1937, Phillips, 1973, and Savage, 1954, for the formal foundations of the subjectivist point of view regarding probability) and the empirical probability of the occurrence of that event. In the present context of comparative judgments with perceptual stimuli, on each trial following a response, subjects provide a subjective probability of the correctness of the response, as a confidence judgment (with .5 denoting a guess and 1.0 certainty), and over trials the *(conditional) probability of a correct response* is determined for each of various levels on the subjective probability scale. Usually, the various levels on the subjective probability-confidence scale involve forming class intervals corresponding to decades, with certainty as a special class (i.e., 0.50-0.59, 0.60-0.69, . . . , 0.90-0.99, 1.00). Ideal calibration is obtained when, over the course of many trials, a person or group is, for example, .5 correct on judgments given a .5 subjective probability of being correct, .7 correct on judgments given a .7 subjective probability of being correct, and so on.

Another important aspect of the confidence/accuracy relationship is *resolution* (Murphy, 1973), which is the extent to which a person or group can distinguish an event's occurrence or nonoccurrence. Finally, a third performance measure of considerable interest is *over/underconfidence* (Lichtenstein & Fischhoff, 1977). In line with intuition, a person or group is considered overconfident if the subjective probability (e.g., average confidence) exceeds the objective proportion correct on a given task, and underconfident if the reverse is true. These ideas will be expressed more precisely in the next section.

Following Brier (1950), the *mean probability score* (see also Yates, 1982, 1990), or *Brier score*, is used to index the accuracy of a person's assessment of probabilities. Let ψ_i denote a person's ith assessment of the probability of the occurrence of the event, E (in the present context of binary comparisons with perceptual magnitudes, as indi-

This research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada to W.M.P. A portion of the data reported in Experiment 1 was presented at the annual meeting of the Canadian Psychological Association, Montreal, June 1988. Our thanks to Lester E. Krueger, T. Gilmour Reeve, and two anonymous reviewers for their very helpful comments. Correspondence can be addressed to either J. V. Baranski, Defence and Civil Institute of Environmental Medicine, Decision Systems Group, 1133 Sheppard Ave. West, P.O. Box 2000, North York, ON, Canada M3M 3B9, or W. M. Petrusic, Department of Psychology, Carleton University, Ottawa, ON, Canada K1S 5B6.

—Accepted by previous editor, Charles W. Eriksen

cated above, the probability assessments arise in the form of postdecisional judgments of confidence), and let e_i be an indicator (Bernoulli) random variable. When the event E occurs, $e_i = 1$ (e.g., a correct response), and when it does not, $e_i = 0$ (an error). Consider n independent (Bernoulli) trials. The base rate occurrence of the event E is given by the mean, $\bar{e}$, or proportion of trials on which the event occurs, which in the present context is the proportion of correct responses, denoted $\bar{e} = p(\text{correct})$. Similarly, the variance of the random variable, e_i, is $\bar{e}(1-\bar{e})$. The mean Brier score over n trials, a squared loss function, is given by

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(\psi_i - e_i)^2. \tag{1}$$

An attractive feature of the mean Brier score is that it can be partitioned into meaningful components. As indicated above, typically, the probability assessments, ψ_i, are partitioned into J categories with n_j occurrences in category j and $n = \sum_{j=1}^{J} n_j$. Letting e_{ij} denote the occurrence indicator random variable for the jth category, and noting that $\psi_{ij} = \psi_j$ for all i in category j, the partition (see Murphy, 1972, 1973, and, more recently, Liberman & Tversky, 1993, for treatments of the partition) proceeds by first noting that the *mean probability score*,

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\psi_j - e_{ij})^2,$$

can be rewritten as

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}[(\psi_j - \bar{e}_{.j}) - (e_{ij} - \bar{e}_{.j})]^2,$$

with $e_{.j} = p_j(\text{correct})$, the mean probability of the event occurring (i.e., the probability of a correct response) *conditionally* on a probability assessment that falls into the jth category. Upon expanding the square and distributing the summation operators, this expression, algebraically equivalent to the Sanders (1963) decomposition[1] (see also Yates, 1982), is given by

$$\frac{1}{n}\left[\sum_{j=1}^{J} n_j(\psi_j - \bar{e}_{.j})^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j}(e_{ij} - \bar{e}_{.j})^2\right]. \tag{2}$$

The first term in Equation 2 is typically referred to as *calibration* (Lichtenstein & Fischhoff, 1977), but it is also known as *reliability* (Murphy, 1973) and *reliability-in-the-small* (Yates, 1982). The calibration score provides a weighted index of how closely the mean probability assessment in category j matches the obtained empirical probability of the occurrence of event E and varies between an optimal value of 0 (i.e., perfect calibration) and 1 (the worst possible calibration). As an example of the latter case, the judge would have to report absolute certainty on each trial and always be wrong. In practice, however, calibration scores above 0.10 are rarely encountered. The preceding calibration index should be distinguished from the global *calibration-in-the-large* index (Yates, 1990), which is also, more typically, referred to as *over/underconfidence* (*O/U*). As mentioned earlier, over/underconfidence is defined by the signed difference between the average, overall, confidence rating, $\bar{\psi}_{..} = \overline{\text{conf}}$, and the average, overall, proportion correct, $\bar{e}_{..} = p(\text{correct})$; that is, $O/U = \bar{\psi}_{..} - \bar{e}_{..} = \overline{\text{conf}} - p(\text{correct})$.

The second term in Equation 2 arises upon partitioning of the overall variance of the indicator random variable into between- and within-category components, as in the conventional analysis of variance:

$$\begin{aligned}\text{Var}(e_{ij}) = \bar{e}_{..}(1-\bar{e}_{..}) &= \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(e_{ij} - \bar{e}_{..})^2 \\ &= \frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_{.j} - \bar{e}_{..})^2 \\ &\quad + \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(e_{ij} - \bar{e}_{.j})^2,\end{aligned} \tag{3}$$

and, as is evident from Equation 3, this second term is the within-category component. If we rewrite Equation 3 and substitute into Equation 2, we will obtain the well-known Murphy (1973) partition of the *mean probability score*, given in Equation 4,

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\psi_j - e_{ij})^2 &= \frac{1}{n}\sum_{j=1}^{J} n_j(\psi_j - \bar{e}_{.j})^2 \\ &\quad - \frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_{.j} - \bar{e}_{..})^2 \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(e_{ij} - \bar{e}_{..})^2.\end{aligned} \tag{4}$$

Calibration, the typically reported index of the probability assessment skill of a judge, should be carefully distinguished from the second term in Equation 4, which defines *resolution* (Murphy, 1973, Yaniv, Yates, & Smith, 1991, and Yates, 1982 refer to this index as *discrimination*). The resolution index, based on the variability of conditional probabilities of event occurrence, provides a quantitative index of the ability of the judge to use the J confidence categories to effectively distinguish when the event E occurs and when it does not. As is evident from Equation 3, the resolution index is bounded above by the overall variability of the indicator variable—that is, by

$$\text{Var}(e_{ij}) = \bar{e}_{..}(1-\bar{e}_{..}) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(e_{ij} - \bar{e}_{..})^2,$$

which is also the final term in Equation 4. This index is also referred to as *knowledge* (Lichtenstein & Fischhoff, 1977). In the two-alternative, forced choice case, resolution assumes a maximal value of 0.25 when $\bar{e}_{..} = p(\text{correct}) = 0.5$ and the judge correctly assesses occurrence and nonoccurrence of the event E. On the other hand, the lower bound on the resolution score is 0, which denotes a complete inability to use the J assessment (i.e., confidence) categories to differentiate event occurrence and nonoccurrence (i.e., correct from incorrect responses in the binary choice context). As with the calibration index, resolution scores above 0.10 are rarely observed.

As noted above, the final term in Equation 4 defines the upper limit of the resolution index. Consequently, Yaniv et al., 1991 (see also Sharp, Cutler, & Penrod, 1988) have recommended that the raw resolution score be normalized by using this term—that is, $\text{Var}(e_{ij})$. This normalized resolution index is thus given by

$$\begin{aligned} \text{NRI} &= \left[\frac{1}{n}\sum_{j=1}^{J} n_j(\bar{e}_{.j}-\bar{e}_{..})^2\right] \Big/ \ \bar{e}_{..}(1-\bar{e}_{..}) \\ &= \text{Resolution}/\text{Var}(e_{ij}) = \eta^2, \end{aligned}$$

and it is interpretable as the between-category portion of the overall variance; it is directly comparable to the η^2 measure typically encountered in analyses of variance.

To date, a major focus of calibration research has been subjective probability assessment on questions of intellectual knowledge. In this domain, confidence and accuracy are almost always positively correlated and a robust finding is that people are often *overconfident* about how much they know (for reviews, see Baron, 1988; Keren, 1991; Lichtenstein et al., 1982; Yates, 1990), especially when judgments are difficult and, consequently, the probability of correct responses is low (e.g., Allwood & Montgomery, 1987; Arkes, Christensen, Lai, & Blumer, 1987; Fischhoff, Slovic, & Lichtenstein, 1977; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein & Fischhoff, 1977, 1980; Wright & Phillips, 1980). On the other hand, when accuracy is relatively high (i.e., > 80% correct), *underconfidence* is observed (e.g., Lichtenstein & Fischhoff, 1977). This interaction between judgment difficulty and over/underconfidence is quite robust and is referred to as the calibration *difficulty effect* (Griffin & Tversky, 1992) or *hard-easy effect* (Gigerenzer, Hoffrage, & Kleinbölting, 1991).

In addition to the ability to self-evaluate intellectual knowledge, calibration research has been concerned with issues such as the ability to forecast future events (e.g., Carlson, 1993; Fischhoff & Macgregor, 1982; Keren & Wagenaar, 1987; O'Connor & Lawrence, 1989; Vreugdenhil & Koele, 1988; Wright, 1982; Wright & Ayton, 1986; Wright & Wisudha, 1982; Yates, 1982), the ability of experts to assign appropriate subjective probabilities in their areas of expertise (e.g., Christensen-Szalanski & Bushyhead, 1981; Keren, 1987; Murphy & Winkler, 1977; Oscamp, 1965; Solomon, Ariyo, & Tomassini, 1985; Tomassini, Solomon, Romney, & Krogstad, 1982; Wagenaar & Keren, 1985), and various methods of improving poor calibration (e.g., J. K. Adams & P. A. Adams, 1961; P. A. Adams & J. K. Adams, 1958; Arkes et al., 1987; Koriat et al., 1980; Lichtenstein & Fischhoff, 1977, 1980; Sharp et al., 1988).[2]

Although there are some inconsistencies in these fields of study, the consensus of this research is that people are less likely to be overconfident in predicting future events (predictions) than in predicting past events or performance (postdictions), that experts are sometimes better calibrated than lay people, and that attempts to improve performance with feedback or some induced change in decisional strategy can be successful.

The wealth of calibration research on intellectual knowledge or personal-ability-based tasks contrasts sharply with an almost nonexistent literature on the calibration of confidence in perceptual judgments. This fact is surprising given that confidence based on perceptual information often provides us with a basis for very important decisions (e.g., whether or not to pass with oncoming traffic). The lack of calibration research on perceptual judgments is also surprising given that the pioneering studies on the confidence/accuracy relationship actually date back to some of the earliest psychophysical research (e.g., Festinger, 1943; Fullerton & Cattell, 1892; Garrett, 1922; Henmon, 1911; Johnson, 1939; Lund, 1926; Peirce & Jastrow, 1884; Seward, 1928; Trow, 1923) and that such research continues today in the context of developing and evaluating broader theories of psychophysical discrimination (e.g., Heath, 1984; Link, 1992; Petrusic, 1992; Petrusic & Baranski, 1989a, 1989b; Smith & Vickers, 1988; Vickers, 1979; Vickers & Packer, 1982; Vickers, Smith, Burt, & Brown, 1985).

In contrast to contemporary calibration research, which, as mentioned, requires confidence ratings expressed in terms of (subjective) probabilities, many of the classic psychophysical comparison studies employed alphanumeric-based confidence scales (e.g., a, b, c, d; where a = certain, b = moderate certainty, c = little certainty, and d = guess). Consequently, they permit only a limited analysis of the exact correspondence between confidence and accuracy—that is, at *guess* (confidence = 50%) and at *certain* (confidence = 100%). Interestingly, a review of the data reported in many of these studies (e.g., Fullerton & Cattell, 1892; Garrett, 1922; Henmon, 1911; Peirce & Jastrow, 1884; Trow, 1923) will show accuracy to have been very much above chance when guessing was reported, to have increased monotonically through the intermediate range of confidence levels, and to have been very near 100% correct when certainty was indicated. Hence, a "calibration-type" analysis of these data suggests that people are actually *underconfident* about perceptual judgments.

J. K. Adams (1957), in a study on perceptual word recognition, noted the limited analyses permitted by the confidence scales employed in the classic psychophysical research and thus sought to obtain a more detailed examination of the confidence/accuracy relationship by requiring his subjects to express confidence in terms of a percentage. The task involved having subjects view brief exposures of 40 different words under varying levels of illumination. After each trial, the subjects wrote down the word they thought they saw and provided a confidence rating in deciles. The results of this landmark "calibration" experiment (Adams referred to it as the "realism of confidence judgments") showed clear underconfidence.

The first direct comparison between calibration in perceptual and in intellectual knowledge tasks was provided by Dawes (1980; cf. Trow, 1923). He hypothesized that while people may tend to overestimate the power of their intellects, they may not be aware of their remarkably accurate perceptual systems and thus show underconfidence for perceptual judgments. Dawes replicated the finding of overconfidence on an intellectual knowledge task (Experiment 2), and, of the four experiments that he performed with perceptual tasks, only one revealed overconfidence (Experiment 5).

More recently, Keren (1988) readdressed the issue and found overconfidence on an intellectual knowledge task but not on a Landolt-C visual acuity task or in two other experiments involving perceptual letter identification.[3] On the basis of these findings, Keren concluded that "the assertion that overconfidence may not be present in perceptual tasks, originally proposed by Dawes (1980), is strongly supported by the present studies (p. 117)" (see Björkman, Juslin, & Winman, 1993, and Winman & Juslin, 1993, for more recent evidence of global underconfidence with perceptual comparisons). However, Keren did find less underconfidence when subjects judged the more difficult of two acuity levels in the Landolt-C task. This result is important, because it suggests a possible parallel to the previously mentioned relationship between judgment difficulty and over/underconfidence in the calibration of intellectual knowledge questions.

EXPERIMENT 1

Experiment 1 was conducted with several purposes in mind. The first was to examine in more detail the effect of perceptual judgment difficulty on the calibration of confidence ratings that follow perceptual judgments (i.e., postdictions). Because the effect of judgment difficulty on calibration in nonperceptual tasks is easily the single most robust property in the area, we reasoned that if clear over- and underconfidence could be demonstrated in the same perceptual task, mediated by only a priori perceptual judgment difficulty manipulations, then a major basis for distinguishing calibration in perceptual and nonperceptual tasks would be removed. At the least, performance in each domain could be studied in the context of a unified theory (see, e.g., Ferrell & McGoey, 1980; Gigerenzer et al., 1991).

In keeping with traditional psychophysical investigation, we studied the effects of perceptual judgment difficulty *directly*, through direct manipulation of the stimulus relations, as well as *indirectly*, by comparing calibration under conditions emphasizing either accurate or speeded responding. Consequently, relative to the accuracy condition, speeding responses would reduce accuracy for each comparison level and thus provide a further variation in difficulty levels for calibration analysis.

A second objective of the study was to provide a preliminary analysis of *subjective error* calibration. Subjects in nonperceptual calibration tasks typically provide written confidence reports between 50% and 100%, where 50% denotes a guess and 100% denotes certainty. However, in pilot studies with perceptual judgments, many subjects reported that they sometimes felt they had made an error following a two-alternative forced choice response but had no way to convey this information on a confidence scale from 50% to 100%. Consequently, in the present study, we employed a confidence scale ranging from 0% (certainty of an error) to 100% (certainty of a correct response) in order to permit an investigation of the relationship between subjective errors (i.e., confidence between 0% and 49%) and the actual error rate associated with such confidence reports (cf. Ronis & Yates, 1987).

Finally, in Experiment 1, we sought to extend Henmon's (1911) pioneering study of the relationship between *confidence* and *decision time* to the domain of contemporary calibration research. We will show that the properties of decision times, conditional on the levels of confidence employed, place a rigorous set of empirical constraints on theories of confidence calibration.

On the basis of Keren's (1988) supposition that certain perceptual tasks may be especially vulnerable to overconfidence when they involve a memory component, illusions, or "higher" mental processing (e.g., inferences, analogical reasoning, etc.), we employed a visual extent comparison task in which the two stimuli were presented simultaneously for comparison, thus precluding memory for the alternatives.

Method

Subjects. Ten Carleton University graduate students (6 female, 4 male) participated for five 90-min experimental sessions. All subjects had normal or corrected-to-normal vision and were naive with respect to the nature and aims of the experiment.

Apparatus. The stimuli were presented on an Amdek-310A video monitor. An IBM PC/XT computer controlled event sequencing, randomization, and the recording of responses and response times (RTs). On the perceptual task, subjects responded on an IBM PC mouse by using the index and middle fingers of the preferred hand. They used the nonpreferred hand to type their confidence reports on the numeric keypad of the PC keyboard. A small desk lamp, positioned behind the video monitor, provided sufficient light for responding without interfering with the visual display.

Stimuli. The stimulus display consisted of two 5-mm vertical lines presented on the left and right of a 10-mm vertical central fixation marker. The central marker served the purposes of allowing a referent from which to perform the comparative judgment and to divide the video screen (which had a horizontal resolution capability of 720 pixels) into a left and right side. All lines were 1 mm wide and appeared in amber color on a black background.

The notation (x, y) is used to denote a stimulus pair x pixels (3.13 pixels = 1 mm) to the left of the referent and y pixels to the right. Two presentation orders of eight pairs were used in the experiment: (22,20; 20,22); (40,37; 37,40); (58,55; 55,58); (76,74; 74,76); (248,241; 241,248); (266,253; 253,266); (284,263; 263,284); and

(302,274; 274,302). Henceforth, these will be referred to as Pairs 1-8, respectively. At a viewing distance of approximately 60 cm, the distance of the nearest and farthest stimulus pairs subtended visual angles of approximately 2° and 18°, respectively.

It is well known that the difficulty of visual extent comparisons can be effectively manipulated by varying the ratio (*r*) of the longer to the shorter extent of the comparison pair (see, e.g., Münsterberg, 1894; Petrusic & Jamieson, 1979). In line with intuition, judgments become progressively easier as the ratio increases. Hence, in order to investigate the effect of perceptual judgment difficulty on confidence calibration, the eight stimulus pairs were combined to form four sets of pairs varying in ratio and, consequently, judgment difficulty. Pairs 1 and 8 were combined to form the a priori lowest level of difficulty (r = 1.10, Level 1): Pairs 2 and 7, the second lowest level (r = 1.08, Level 2); Pairs 3 and 6, the second most difficult level (r = 1.05, Level 3); and Pairs 4 and 5, the most difficult level of comparison (r = 1.03, Level 4).

Procedure. Each trial began with the presentation of an instruction ("NEARER" or "FARTHER") centered near the top of the screen, followed, 1.5 sec later, by the stimulus pair. Both the instruction and the pair remained on the screen until the subject responded. The task required subjects to judge which of the two small vertical lines was, depending on the instruction, either "NEARER" or "FARTHER" from the central midline by pressing either the left or right response button on the mouse. Following the response, the screen was cleared and a visual prompt appeared (CONFIDENCE =>). Subjects then typed a confidence rating from 0 to 100, using the numeric keypad on the PC keyboard. The subjects were given detailed instructions on the use of this "full-range" confidence scale prior to Session 1. A rating of 100% confidence was to indicate absolute certainty that a correct response was made and a rating of 50% was to indicate a guess response. Ratings between 51% and 99% were to represent increasing confidence (expressed as a probability or likelihood) that a correct response had been registered. Finally, confidence ratings between 0% and 49% were to be used *only* on trials in which they felt they had made an *error*, with 0% denoting absolute certainty of an error response. There was no time limit for registering the confidence report.

Subjects were provided with trial-by-trial feedback ("Correct" or "Incorrect") on the accuracy of each perceptual judgment following the registration of the confidence report. A 2-sec intertrial interval, with a blank screen, separated the confidence report and the ensuing trial.

Each session began with 32 practice trials followed by 3 blocks of 160 experimental trials. The 160 trials in each block arose from the randomized factorial combination of the 8 pairs, 2 presentation orders of each pair, 2 instructions ("NEARER" or "FARTHER"), and 5 replicates.

Each subject performed two preliminary practice sessions under verbal instructions to respond as quickly and as accurately as possible (the data from these sessions are not reported). In order to study the effect of speeded responding on confidence calibration, 5 subjects performed three experimental sessions under a 450-msec response deadline. If the deadline was exceeded, subjects received the statement "Too Slow" immediately following the response. In addition, these subjects were (1) rewarded 1 cent for responding correctly and beating the deadline, (2) rewarded ½ cent for beating the deadline but responding incorrectly, (3) penalized ½ cent for being correct but missing the deadline, and (4) penalized 1 cent for being incorrect and missing the deadline. The 5 remaining subjects performed their three experimental sessions under an accuracy emphasis; they were rewarded 1 cent for each correct response and were penalized 1 cent for each incorrect response.

Results

Trials on which RTs exceeded 10 sec were excluded from the analyses (19/14,400; 0.1%). In addition, subjects reported confidence in an error response (i.e., <50%) on 1,483 of the remaining trials (10.3%). An analysis of these data is provided in the section entitled *subjective error calibration.* All other analyses reported in this Results section deal exclusively with trials on which confidence was ≥50% and, accordingly, all performance measures to be reported are conditional on this range of confidence levels.

The results are presented in three sections. The first examines the effects of perceptual judgment difficulty and speed versus accuracy instructions on confidence calibration. The second investigates the calibration of subjective errors. The third explores the properties of decision time associated with each level of confidence in the speed- and accuracy-emphasized conditions.

Effects of judgment difficulty and speed versus accuracy instructions on the calibration of perceptual judgments. The top panels in Figure 1 provide *calibration curves* for the four levels of judgment difficulty in the speed- and accuracy-emphasized conditions. These curves were obtained by plotting the percentage of correct responses associated with each confidence interval (i.e., 50%-59%, 60%-69%, 70%-79%, 80%-89%, 90%-99%, and 100%), for each level of difficulty. When viewed in this way, perfect, or ideal, calibration is obtained when the data points fall along the main diagonal, *underconfidence* is denoted by points *above* the diagonal, and *overconfidence* is denoted by points *below* the diagonal. The lower panels of Figure 1 provide the proportion of times each confidence category was used for the four levels of difficulty in the speed- and accuracy-

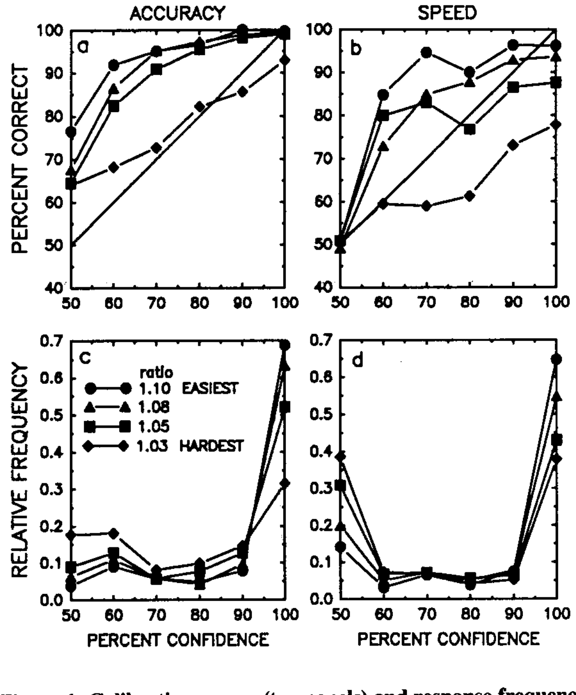

Figure 1. Calibration curves (top panels) and response frequency curves (lower panels) for the four levels of perceptual judgment difficulty in the speed and accuracy conditions in Experiment 1.

emphasized conditions. Note that a full appreciation of the confidence/accuracy relationship can be obtained only by studying the calibration and response frequency curves concurrently.

Table 1 provides, in addition to the proportion correct and mean confidence associated with each difficulty level under speed and accuracy stress, the probability assessment indices defined formally at the outset: over/underconfidence (*O*/*U*), calibration (Cal), resolution (Res), and normalized resolution (η^2).[4]

The plots in Figure 1 and the data in Table 1 provide a number of points of interest. First, in both the speed and accuracy conditions, there are clear and consistent differences in over/underconfidence with variations in comparison difficulty. Specifically, as in studies investigating the calibration of intellectual knowledge judgments, there is good calibration, good resolution, and virtually no over/underconfidence for the difficulty levels producing approximately 80% correct responses (see Lichtenstein & Fischhoff, 1977; Lichtenstein et al, 1982; Wright, 1982). On the other hand, as the level of difficulty draws the percent correct either above (accuracy condition) or below (speed condition) 80%, probability assessments become progressively poorer, with underconfidence occurring with the more accurate judgments and overconfidence with the more difficult, less accurate, speeded judgments.

Second, Figure 1 shows that resolution is much better under speed stress than under accuracy stress. The reason, evident in Figure 1, is that the precision of ''guessing'' (i.e., 50% confidence) is remarkably accurate under speed stress; for each level of difficulty, the probability of a correct response when a guess is reported is essentially .50, which is, in fact, the case for each of the 5 subjects.

Separate analyses of variance (ANOVAs) were performed on the five performance measures reported in Table 1.[5] The speed versus accuracy instructional manipulation was the between-subjects factor and difficulty level was the within-subjects factor. Overall, the proportion of correct responses was higher under accuracy stress than under speed stress [$F(1,8) = 12.13$, $p < .009$] and depended on the level of difficulty of the judgments [$F(3,24) = 95.06$, $p < .0001$]. The interaction between speed versus accuracy and difficulty level was marginally reliable [$F(3,24) = 4.27$, $p < .055$]; the difference in proportion correct between the speed and accuracy groups increased as difficulty increased.

Table 1
Proportion Correct, Mean Confidence, Over/Underconfidence, Calibration, Resolution, and η^2 for the Four Levels of Judgment Difficulty in the Speed and Accuracy Conditions in Experiment 1

Level	*p*(correct)	$\overline{\text{Conf}}$	O/U	Cal	Res	η^2
		Accuracy				
L1 ($r = 1.10$)	.977	.911	−.066	.017	.002	.089
L2 ($r = 1.08$)	.957	.891	−.066	.015	.007	.170
L3 ($r = 1.05$)	.930	.861	−.069	.013	.011	.168
L4 ($r = 1.03$)	.796	.783	−.013	.006	.013	.080
M	.915	.862	−.054	.013	.008	.127
		Speed				
L1 ($r = 1.10$)	.890	.884	−.006	.006	025	.255
L2 ($r = 1.08$)	.827	.844	.017	.004	.031	.217
L3 ($r = 1.05$)	.747	.784	.037	.010	.026	.137
L4 ($r = 1.03$)	.617	.747	.130	.023	.015	.062
M	.770	.815	.045	.011	.024	.167

Confidence did not differ between the speed and accuracy conditions ($F < 1.0$). Although counterintuitive, this result has been demonstrated many times in either between-groups or between-sessions designs (Festinger, 1943; Garrett, 1922; Johnson, 1939) and has been attributed to a context-specific (speed vs. accuracy) scaling of the confidence range (see Vickers & Packer, 1982). Apparently, only when the speed–accuracy manipulation is conducted within subjects and within sessions will confidence be higher under accuracy stress than under speed stress (Vickers & Packer, 1982; but see Baranski, 1991, for evidence that the effect may not hold over many sessions). Finally, the level of difficulty had a highly reliable effect on confidence [$F(3,24) = 25.54$, $p < .0001$].

Overall, the over/underconfidence measure was not significantly different from zero ($p > .10$) as a consequence of averaging over groups. However, there was a clear effect of difficulty level on over/underconfidence [$F(3,24) = 6.44$, $p < .027$], confirming that overconfidence increased (or underconfidence decreased) as difficulty increased.

There were neither main effects nor interactions for the calibration score. However, resolution was better under speed stress than under accuracy stress [$F(1,8) = 10.50$, $p < .02$] and this group effect interacted with difficulty level [$F(3,24) = 4.55$, $p < .039$]; the difference in resolution between the two groups decreased as difficulty level increased.

In summary, people can be overconfident on perceptual judgments if the level of difficulty of the judgments is sufficiently high. No overconfidence was observed under accuracy stress because the most difficult level of comparison was still quite easy for these subjects. In addition, calibration and resolution were generally poor under accuracy stress, but speeding responses greatly improved the resolution of confidence judgments.

Calibration of subjective errors. Figure 2 provides calibration curves for trials on which subjects thought they made an error (i.e., confidence less than 50%) in the speed- and accuracy-emphasized conditions. Because subjects rarely made such reports under accuracy stress ($N = 153$), the five confidence categories denoting subjective errors (i.e., 0%–9%, 10%–19%, 20%–29%, 30%–39%, 40%–49%) were grouped into three for the accuracy condition (i.e., 0%–9%, 10%–39%, 40%–49%) in order to provide more reliable estimates of the data. In addition, note that the *y*- and *x*-axes have been inverted in order to denote increasing error probability and increasing certainty of an error response, respectively (e.g., 0% confidence in a correct response becomes 100% certainty of an error). In this way, the plots are comparable to those

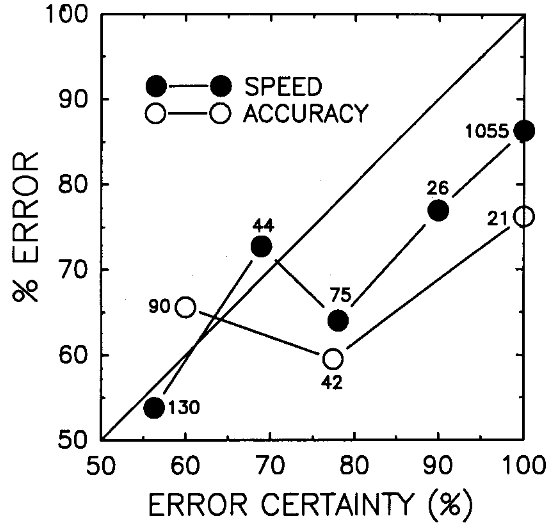

Figure 2. Subjective error calibration in the speed and accuracy emphasized conditions in Experiment 1 (number of observations provided). Percent error is plotted at the mean confidence in each confidence interval.

reported in Figure 1; that is, points above the identity line denote underconfidence and points below the identity line denote overconfidence.

These plots reveal that subjects are indeed likely to be in error when they think they are. In contrast to the calibration of subjective correct responses (Figure 1), however, subjects are highly *overconfident* in their assessment of errors. That is, although subjects are likely to be in error when they think they are, they overestimate the degree to which they err. Interestingly, this is a situation where overconfidence actually implies caution. Finally, note that subjects show good resolution for subjective errors under speed stress, perhaps because of their greater familiarity with making errors.

How do subjects know when they have made a mistake? It is likely that a portion of the data reflects an ability to detect errors that occur when the correct response is known but the wrong response switch is depressed (i.e., motor response errors; see Rabbitt & Vyas, 1970). These trials would be assigned a high error certainty and, intuitively, would be more likely to occur under speed demands, as is the case in the data. The remaining observations likely arise from some postdecisional, memory-based reevaluation of the primary judgment. On this view, subjects apparently have the capacity to hold the decision, response, and feeling of relative certainty in memory long enough to revise the output of the judgment if need be. Although the results are preliminary and should thus be interpreted with caution, the monotonicity and resolution exhibited in these plots suggest that confidence can provide fairly reliable information about judgment errors.

RT analyses. Figure 3 provides plots of the mean of individual subject median RTs against the levels of confidence used under speed and accuracy stress, separately for each level of perceptual judgment difficulty. Under accuracy stress we see the well-known inverse relation between RT and confidence (e.g., Festinger, 1943; Henmon, 1911; Johnson, 1939; Pierrel & Murray, 1963; Vickers & Packer, 1982; Vickers et al., 1985; Volkmann, 1934).[6] In addition, there is an ordering of the curves with respect to the level of difficulty of the judgments—an RT *difficulty effect*. Note that the occurrence of an RT difficulty effect is not trivial, because if these curves did not show a difficulty effect, then a scaling of the duration of the decision process could provide a simple and elegant basis for the judgment of confidence (e.g., Audley, 1960; Henmon, 1911; Pierrel & Murray, 1963).

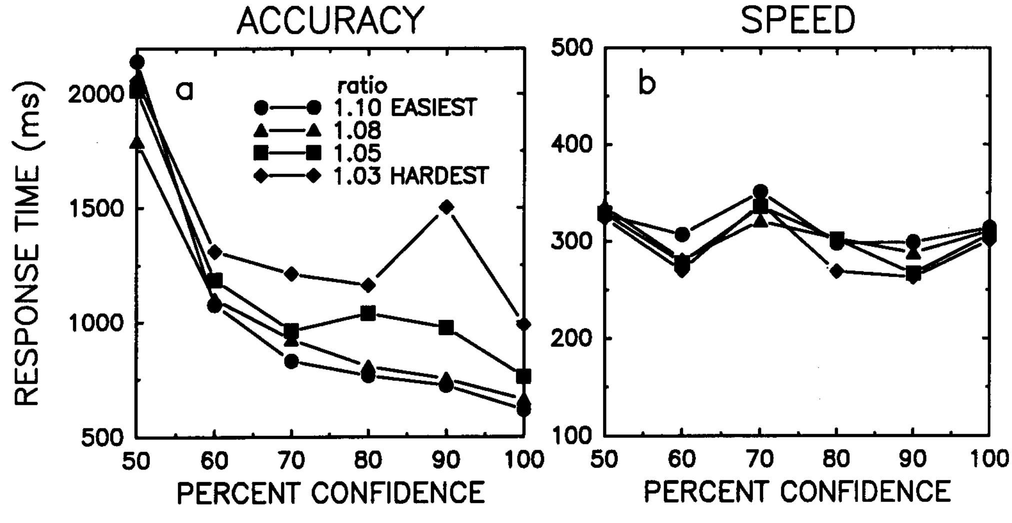

Figure 3. Mean of individual subject median response times as a function of the level of confidence for the four levels of perceptual judgment difficulty in the speed and accuracy conditions in Experiment 1.

Under speeded responding, the RT curves appear flat across the confidence levels. This is due, in part, to averaging of the data over subjects. In fact, for 3 of the 5 subjects, there is a slight inverse relationship between RT and confidence, for 1 subject RT does not vary across confidence levels, and for 1 subject the relationship is, curiously, positive. Unlike under accuracy stress, however, there is no hint of a difficulty effect for the RT measure; for each subject, the curves lie on top of one another.

Because some subjects did not use the lower confidence levels for the easier comparison levels under accuracy stress, we could not use confidence category as a within-subjects factor in an ANOVA. Thus, an ANOVA was conducted with median RT as the dependent measure (mean RTs showed the same effects), speed versus accuracy stress as a group factor, and difficulty level and error versus correct responses as within-subjects factors. As expected, subjects were considerably slower under accuracy stress than under speed stress [$F(1,8) = 16.15, p < .004$]. The main effect of difficulty level was marginally reliable [$F(3,24) = 2.97$, $.05 < p < .06$], but the interaction between difficulty level and group was reliable [$F(3,24) = 3.04, p < .05$], confirming an RT difficulty effect under accuracy stress but not under speed stress. Finally, the interaction between group and correct and error RTs was reliable [$F(1,8) = 20.17, p < .003$] because, as is typically the case in psychophysical comparison tasks, error times were longer than correct times under accuracy stress but were approximately equal to correct times under speed stress (for reviews, see Luce, 1986; Petrusic, 1992; Pike, 1968, 1971; Vickers, 1979).

In sum, decision times are inversely related to the level of confidence in the judgment and show a difficulty effect under accuracy stress but not under speed stress. Together, these results provide an important set of first-order properties for evaluating theories of confidence calibration.

Discussion

Consistent with the results of classical psychophysical comparison studies and Keren's (1988) more recent investigation of the calibration of perceptual judgments, subjects working under accuracy stress displayed no evidence of overconfidence. However, speeding responses sufficiently reduced the accuracy of each comparison level so that clear overconfidence was observed for the most difficult level. In fact, the calibration curve for the most difficult level under speed stress [$p(\text{correct}) = .617$] is strikingly similar to the calibration curves reported in several studies investigating the calibration of general knowledge questions with a similar error rate (see Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982). Nevertheless, it remains to be seen whether clear overconfidence can be obtained in a perceptual task when subjects are not speeded in their judgments. Accordingly, in Experiment 2, we employed the same task as that in Experiment 1 but investigated a range of more difficult perceptual comparison levels under conditions emphasizing accurate responding.

Another interesting finding of Experiment 1 was that subjects working under speed stress displayed high resolution in their confidence judgments. One possible explanation for this result is that the increased *global task difficulty* under speed stress alerted subjects to the fact that they were making a substantial proportion of errors. Hence, if the increase in global task difficulty was the sole factor in improving the resolution of confidence judgments, we would expect subjects to display better resolution in Experiment 2, where, as previously mentioned, we planned to increase the overall difficulty of the task. On the other hand, if the better resolution was a result of the combined effects of more difficult judgments *and* the presence of trial-by-trial response feedback, a group of subjects provided with feedback should show better resolution than a group of subjects who are not provided with feedback.

In sum, in Experiment 2 we investigated the calibration and resolution of confidence for difficult perceptual judgments under conditions emphasizing accurate responding, with and without the presence of trial-by-trial feedback on the perceptual judgments.

EXPERIMENT 2

Method

Subjects. Twenty Carleton University undergraduate students (11 female, 9 male) participated for one 90-min experimental session in order to satisfy introductory course requirements. All subjects had normal or corrected-to-normal vision and were naive concerning the nature and aims of the experiment.

Apparatus. The apparatus of Experiment 1 was used.

Stimuli. The stimuli were the same as in Experiment 1 except that six stimulus pairs were chosen so as to provide a wider range of difficulty levels (as defined by the percent correct measure) and more difficult judgments overall: (30,28; 28,30); (50,48; 48,50); (70,69; 69,70); (260,243; 243,260); (280,269; 269,280); and (300,296; 296,300). Hence, the stimulus pairs are expressed, in terms of difficulty, by the ratios 1.07, 1.04, 1.01, 1.07, 1.04, and 1.01, for Pairs 1-6, respectively. Accordingly, pairs with a ratio of 1.07 were defined, a priori, as the easiest comparisons, pairs with a ratio of 1.04 made up a moderate difficulty level, and pairs with a ratio of 1.01 defined the most difficult level of comparison.

Procedure. The procedure was the same as in Experiment 1. Here, each session began with 24 practice trials followed by 4 blocks of 96 randomized experimental trials: 6 pairs × 2 instructions ("NEARER" or "FARTHER") × 2 presentation orders × 4 replications.

The 20 subjects were randomly assigned to one of two groups. The groups were identical except that one received trial-by-trial response feedback ("Correct" or "Incorrect") on the accuracy of the perceptual judgment following each confidence report and the other did not. All subjects were instructed to respond as accurately as possible without taking too much time.

Results and Discussion

As in Experiment 1, trials on which RTs exceeded 10 sec were excluded from the analysis (84/7,680; 1%). In addition, 521 (6.6%) of the remaining trials resulted in a confidence <50% and will be analyzed in the section on subjective error calibration. The results are presented in three sections. The first provides the calibration and resolution analyses for the feedback and no-feedback conditions, the second presents a view of subjective error calibration, and the third presents the RT properties associated with the calibration analyses.

Effects of feedback and judgment difficulty on the calibration of perceptual judgments. Figure 4 provides the calibration and response frequency curves for the three levels of difficulty in the feedback and no-feedback conditions, and Table 2 provides the performance measures associated with these curves. Table 2 shows that the proportion of correct responses for each difficulty level is virtually identical in the two groups, permitting a direct comparison of the effects of feedback on the performance measures of interest, unconfounded by differences in overall discriminative accuracy.

Immediately evident in Figure 4 and Table 2 is the clear overconfidence exhibited for the most difficult judgments in the feedback and no-feedback conditions. Evidently, as in the calibration of intellectual knowledge judgments, people simply cannot avoid being overconfident when accuracy is very low (cf. Lichtenstein & Fischhoff, 1977; Lichtenstein et al., 1982).

Also evident in Figure 4 (and Table 2) is the fact that trial-by-trial feedback on the accuracy of the discriminative response has a substantial effect on the resolution of confidence judgments. The basis for this improvement becomes evident when we look at the relative response frequency curves in the two conditions. Evidently, without feedback about errors, subjects provide an unnecessarily high proportion of low confidence reports for the easier difficulty levels. This results in inflated accuracy at those confidence levels and thus in poorer resolution.

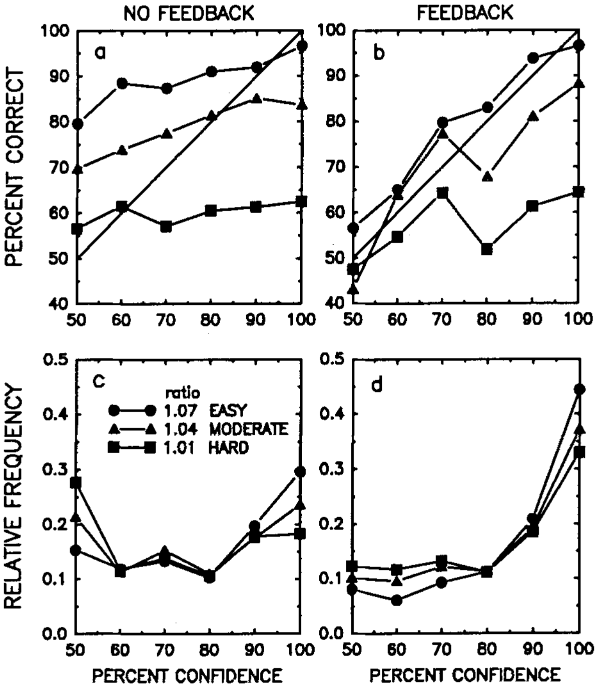

Figure 4. Calibration curves (top panels) and response frequency curves (lower panels) for the three levels of perceptual judgment difficulty in the feedback and no-feedback conditions in Experiment 2.

Table 2
Proportion Correct, Mean Confidence, Over/Underconfidence, Calibration, Resolution, and η^2 for the Three Levels of Judgment Difficulty in the Feedback and No-Feedback Conditions in Experiment 2

Level	*p*(correct)	$\overline{\text{Conf}}$	O/U	Cal	Res	η^2
		No Feedback				
L1 (r = 1.07)	.887	.810	−.077	.024	.003	.027
L2 (r = 1.04)	.766	.779	.013	.016	.003	.017
L3 (r = 1.01)	.589	.758	.169	.052	.001	.004
M	.747	.782	.035	.031	.002	.016
		Feedback				
L1 (r = 1.07)	.869	.863	−.006	.002	.016	.141
L2 (r = 1.04)	.761	.833	.072	.010	.019	.104
L3 (r = 1.01)	.590	.815	.225	.059	.004	.016
M	.740	.837	.097	.023	.013	.087

As in Experiment 1, separate ANOVAs were conducted on the five performance measures. The feedback and no-feedback conditions provided the group factor, and difficulty level was a within-subjects factor. Difficulty level provided the only reliable effect for the proportion correct [$F(2,36) = 185.30, p < .0001$] and confidence [$F(2,36) = 9.86, p < .003$] measures. Overall, the over/underconfidence measure differed from zero in the direction of overconfidence [$F(1,18) = 4.92, p < .039$]. In addition, the main effect of difficulty level was highly reliable [$F(2,36) = 146.45, p < .0001$], confirming that overconfidence increased as difficulty increased. The effect of feedback had no effect on calibration ($p > .12$) but did improve resolution [$F(1,18) = 8.40, p < .009$). Finally, difficulty level had a large effect on both calibration [$F(2,36) = 11.09, p < .002$] and resolution [$F(2,36) = 5.54, p < .009$], confirming that performance on both measures worsened as difficulty increased.

Subjective error calibration. Figure 5 provides plots of subjective error calibration in the feedback and no-feedback conditions. As in Experiment 1, subjects in the present study are likely to be wrong when they think they are and show some resolution for subjective errors, but they are clearly overconfident in the degree to which they err. In addition, the presence of trial-by-trial feedback improved the resolution of subjective error judgments.

In summary, the present results, together with those reported in Experiment 1, demonstrate that subjects can report subjective errors fairly accurately with their confidence ratings. In addition, the provision of feedback on a difficult perceptual task (i.e., a task in which a substantial proportion of errors occurs) improves the resolution of confidence ratings for subjectively correct and incorrect decisions.

RT analyses. An ANOVA with median RT (means showed the same effects) as the dependent measure was used to evaluate the properties of RT in Experiment 2. Feedback/no-feedback was the group factor and difficulty

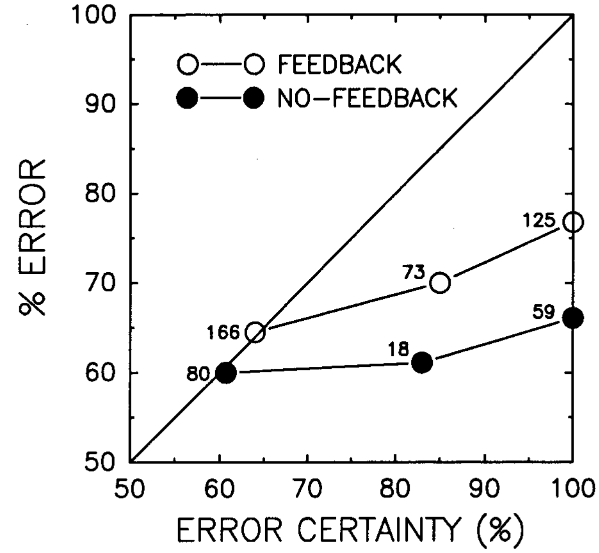

Figure 5. Subjective error calibration in the feedback and no-feedback conditions in Experiment 2 (number of observations provided). Percent error is plotted at the mean confidence in each confidence interval.

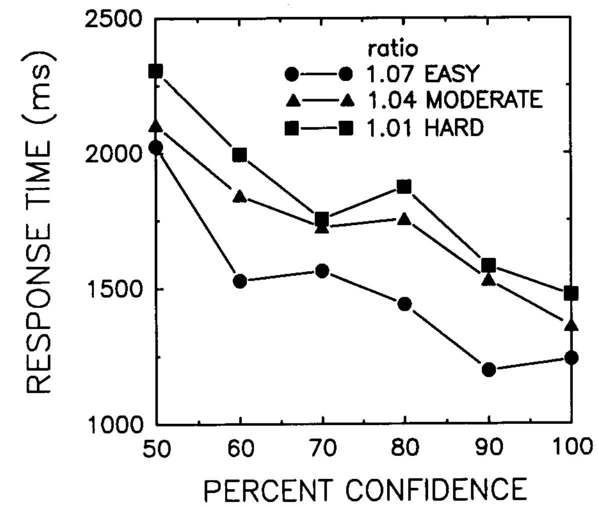

Figure 6. Mean of individual subject median response times as a function of the level of confidence for the three levels of perceptual judgment difficulty in Experiment 2 (data collapsed over the feedback and no-feedback conditions).

level and confidence level were within-subjects factors (the six confidence levels were grouped into three: 50%-69%, 70%-89%, 90%-100%) for the ANOVA because each subject did not use every confidence interval).[7] In addition, because there were neither main effects nor interactions involving feedback, we combined the data over the group factor for subsequent analyses.

Figure 6 provides a plot of the means of individual subject median RTs as a function of confidence level for the three levels of difficulty, after combining the data over subjects. As in Experiment 1 under accuracy stress, there is a strong inverse relationship between RT and confidence level [$F(2,36) = 28.21, p < .0001$] and the main effect of difficulty level was also reliable [$F(2,36) = 3.40, p < .045$], confirming the RT difficulty effect reported in Experiment 1. The interaction between confidence level and difficulty level was not reliable ($F < 1.0$).

Discussion

Two major findings were obtained in Experiment 2. The first was that trial-by-trial feedback on a difficult perceptual task improves the resolution of confidence reports. The second was that extreme overconfidence can be obtained in a perceptual task when judgments are very difficult.

Until now, overconfidence in perceptual tasks has been observed only under conditions in which illusory or misleading judgments have been presented to the subjects (e.g., Dawes, 1980; Keren, 1988). In fact, this was partly true in the present study as well.

Upon closer examination of the data, we discovered that discriminative accuracy depended on presentation order (i.e., which element in the pair was on the right or the left of the screen) for the difficult comparisons (i.e., Pairs 3 and 6). This effect was equally large in the feedback and no-feedback conditions, did not depend on the direction of the comparison (i.e., NEARER/FARTHER), and was especially large for Pair 6, the more extreme (i.e., farther) pair. This effect was not evident in the first experiment, probably because the extreme pairs were highly discriminable. A closer view of this effect is provided in Table 3.

Table 3 shows that subjects were much more accurate when the stimulus on the left was the element in the pair that was farther from the referent. This *positional-order effect* was so pronounced for Pair 6 that it led to below chance performance in the 296,300 order.

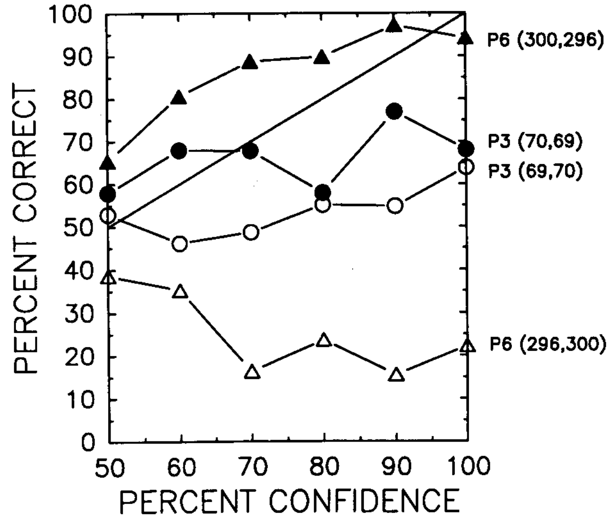

Figure 7. Calibration curves for the two presentation orders of Pairs 3 and 6 (difficult level) in Experiment 2.

Table 3
Proportion Correct, Mean Confidence, and Over/Underconfidence for the Two Presentation Orders of Pairs 3 and 6 (Difficult Comparisons) in Experiment 2

Pair	Order (pixels)	*p*(correct)	$\overline{\text{Conf}}$	O/U
3	70,69	.673	.784	.111
	69,70	.551	.780	.229
6	300,296	.871	.800	−.071
	296,300	.255	.779	.524

Figure 7 provides the calibration curves associated with the two presentation orders of Pairs 3 and 6. The most striking result is the *monotonically decreasing* calibration function for Pair 6 in the 296,300 order, which describes a decrease in accuracy with increasing confidence! This result is offset, in part, by the *underconfidence* evident in the 300,296 order, where accuracy is very high. Note, however, that subjects display extreme overconfidence for the two presentation orders of Pair 3, where accuracy is very low but clearly above chance in both orders.

In summary, subjects can indeed be overconfident on difficult perceptual judgments under accuracy stress. However, perceptual illusions or complex response biases can clearly exaggerate the degree of overconfidence obtained.

EXPERIMENT 3

In Experiments 1 and 2, we examined the effects of perceptual judgment difficulty on confidence calibration and resolution in a perceptual comparison task. Experiment 3 extends the investigation of perceptual judgment difficulty to a perceptual task requiring visual gap detection/discrimination.

Method

Subjects. Ten Carleton University graduate students (6 male, 4 female) participated for a single 45-min session in return for $7 pay. All subjects had normal or corrected-to-normal vision, had not participated in the previous experiments, and were naive about the nature and aims of the study. The data of one subject were excluded because performance was not sufficiently above a chance level [*p*(correct) = .509, $\overline{\text{conf}}$ = .680]. Another subject was run as a replacement.

Apparatus. The apparatus of Experiments 1 and 2 was used.

Stimuli. The stimulus display consisted of a 20×20 pixel (1 mm = 3.13 pixels) unfilled square which was horizontally and vertically centered on the video monitor. Each presentation of the square included a 1-pixel gap (.319 mm), which appeared equally often in one of eight possible locations (see Figure 8). At a viewing distance of approximately 2.3 m, the square subtended a visual angle of approximately .16° and the gap subtended a visual angle of approximately .008°.

Design and Procedure. The task required subjects to judge, on each trial, whether the gap was presented on the left or the right of the bisector of the square by depressing either the left or right button on the PC mouse. The display was presented for 50 msec.

Each subject performed for 11 blocks (1 practice, 10 regular) with 32 trials (4 replicates of the 8 gap locations) in each block. All trials were completely randomized within blocks. The procedure for the reporting of confidence levels was the same as in the previous experiments. The subjects were instructed to respond as accurately as possible without taking too much time, and trial-by-trial feedback was not provided.

Results

No RTs exceeded the 10-sec cutoff and only 8 trials (out of 3,200) were made with a confidence <50% (5/8 were, in fact, incorrect). Accordingly, we present the results in two sections. The first examines the effects of gap detection difficulty on confidence calibration, and the second examines the RT properties associated with each confidence level in the task.

Effects of judgment difficulty on the calibration and resolution of confidence. Figure 9 plots the degree of under/overconfidence as a function of the probability of gap detection for each of the eight gap locations and shows that overconfidence increases as gap detection accuracy decreases—a relationship that is well described by linearity (r^2 = .884). In addition, three distinct levels of gap detection difficulty can be identified in Figure 9. Hence, for convenience in data presentation for the remaining analyses, we grouped the data as follows: Locations 3, 6, and 7 [*p*(correct) = .634] were combined to form the most difficult level of detections; Locations 4, 5, and 8 [*p*(correct) = .711] were combined to form a moderate difficulty level; and Locations 1 and 2 [*p*(correct) = .780] were combined to form the easiest level of detections.

Figure 10 provides the calibration and response frequency curves associated with these three levels of a posteriori detection difficulty, and Table 4 provides a summary of the performance measures associated with these curves. As in the previous experiments, overconfidence increases and calibration and resolution become poorer as accuracy decreases.

ANOVAs confirmed that accuracy decreased as difficulty increased [$F(2,18) = 7.34, p < .014$]. However, the decline in confidence with increasing difficulty evident in Table 4 did not attain statistical reliability [$F(2,18) = 1.92, p > .1$], providing the basis for increasing over-

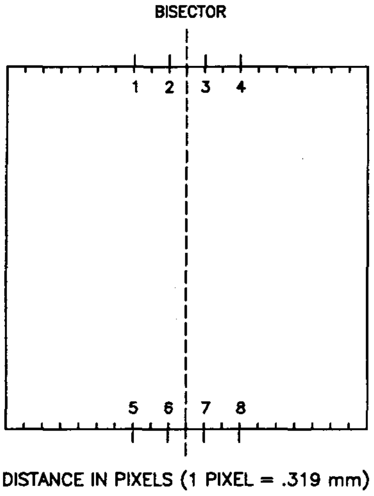

Figure 8. Positions of the eight gap locations (in pixels) in Experiment 3.

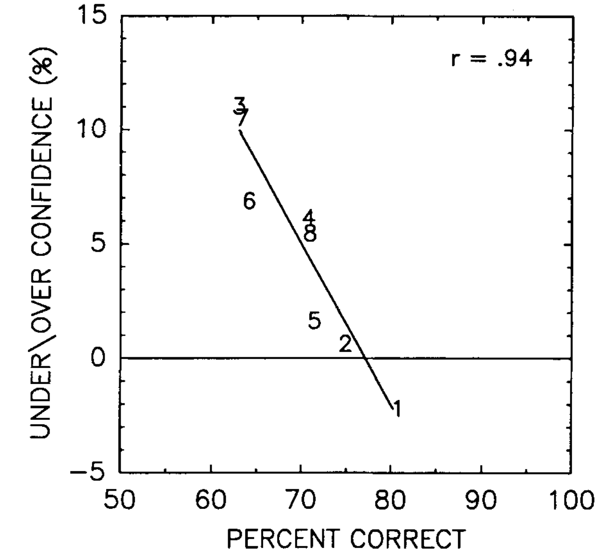

Figure 9. Under/overconfidence (in percent) for the eight gap locations in Experiment 3, plotted as a function of the percentage of correct detections.

confidence as difficulty increases [$F(2,18) = 5.36, p < .033$] and reliable overconfidence for the most difficult level of detection [$t(9) = 2.27, p < .05$]. Finally, both calibration [$F(2,18) = 4.23, p < .059; p < .031$ by a conventional test] and resolution [$F(2,18) = 5.06, p < .034$] became poorer as difficulty increased.

RT analyses. Figure 11 provides a plot of means of individual subject median RTs as a function of confidence level for the three levels of difficulty, separately for the 5 slowest and the 5 fastest responding subjects. The results provide a replication of those reported in Experiment 1. Namely, for the slower subjects, RTs show a strong inverse relation to the level of confidence and a clear RT difficulty effect. For the faster subjects, on the other hand, there is a weaker inverse relation between RT and confidence and the difficulty effect disappears. Evidently, very similar processes of judgment and confidence estimation operate in the tasks of Experiments 1 and 3. Furthermore, both instructional (Experiment 1) and a posteriori (Experiment 3) speed versus accuracy stress conditions result in a similar configuration of findings.

An ANOVA was conducted with fast and slow responding subjects as a between-subjects factor and the three difficulty levels and three confidence levels (again, the six confidence levels were combined into three because each subject did not use every confidence level) as within-subjects factors.

The main effects of group [$F(1,8) = 22.19, p < .0015$], difficulty level [$F(2,16) = 4.77, p < .038$], and confidence level [$F(2,16) = 60.85, p < .0001$] were reliable. Also reliable were the interactions between group and difficulty level [$F(2,16) = 4.52, p < .043$] and group and confidence level [$F(2,16) = 22.94, p < .0006$]. The former interaction confirms that the RT difficulty effect disappears as judgments become faster, and the latter interaction confirms that the slope of the RT versus confidence function becomes shallower as judgments become faster.

GENERAL DISCUSSION

Confidence Calibration: General Properties

Keren's (1988) claim that people are not necessarily good assessors of their performance in perceptual tasks is supported by the present findings. However, Keren's further claim that people make conservative confidence judgments on perceptual tasks and thus will not display the phenomenon of overconfidence is not supported. Specifically, clear overconfidence can be obtained in a perceptual task when judgments are difficult enough to produce a substantial decrease in accuracy, either by speeding

Table 4
Proportion Correct, Mean Confidence, Over/Underconfidence, Calibration, Resolution, and η^2 for the Three Levels of Difficulty in Experiment 3

Level	p(correct)	$\overline{\text{Conf}}$	O/U	Cal	Res	η^2
L1 (easy)	.780	.772	−.008	.003	.032	.186
L2 (moderate)	.711	.755	.044	.004	.024	.116
L3 (hard)	.634	.730	.096	.016	.012	.051

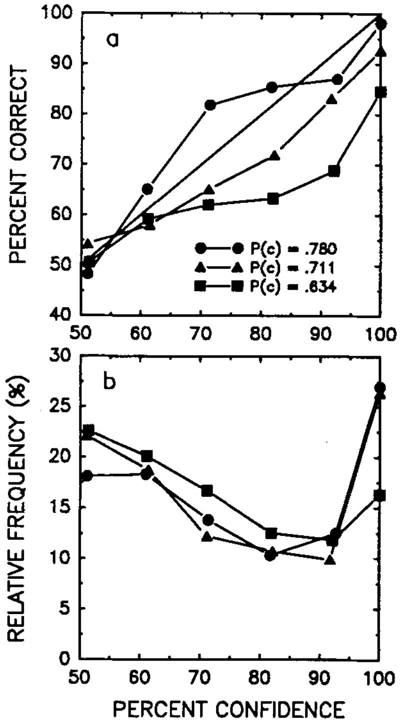

Figure 10. Calibration curves (top panel) and response frequency curves (lower panel) for the three levels of judgment difficulty in Experiment 3: Level 1 (Gap Locations 1 and 2), circles; Level 2 (Gap Locations 4, 5, and 8), triangles; and Level 3 (Gap Locations 3, 6, and 7), squares.

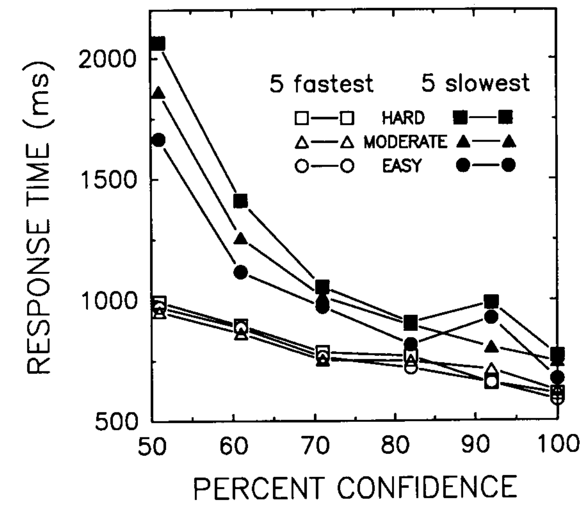

Figure 11. Mean of individual subject median response times as a function of the level of confidence for the three levels of judgment difficulty in Experiment 3: Level 1 (Gap Locations 1 and 2), circles; Level 2 (Gap Locations 4, 5, and 8), triangles; and Level 3 (Gap Locations 3, 6, and 7) squares.

responses (Experiment 1) or by appropriate stimulus discriminability manipulations under accuracy stress (Experiments 2 and 3).

The similarity between calibration in perceptual and nonperceptual tasks is not limited to the effects of judgment difficulty. For example, several authors have argued that the reason for the severe overconfidence often reported in nonperceptual tasks is that such tasks often include misleading questions, which induce lower accuracy and overconfidence through retrieval biases or memory reconstruction strategies (for discussions, see Fischhoff, 1982; Fischhoff et al., 1977; Gigerenzer et al., 1991; Keren, 1988; May, 1986; Wagenaar, 1988). With inclusion of the results reported in Experiment 2, there is now ample evidence to suggest that there are perceptual analogues to the effect of misleading general knowledge questions on confidence calibration. For example, Dawes (1980, Experiment 5) had subjects judge which of two successively presented auditory tones had the longer duration. As is typically the case with successive perceptual comparisons, Dawes obtained large *presentation order effects* or *time order errors*, as they are now more commonly known (see Allan, 1979; Baranski & Petrusic, 1992; Hellström, 1985; Jamieson & Petrusic, 1975, for reviews). Specifically, subjects were 80% correct on trials in which the second presented tone was the longer but were only 63% correct when the first presented tone was the longer. As might be expected on the basis of the present findings, Dawes's subjects were overconfident with the first-tone-longer comparisons but were well calibrated on the second-tone-longer comparisons. More recently, Keren (1988, Experiment 2) showed that the *repeated-letter inferiority effect* (Bjork & Murray, 1977; Egeth & Santee, 1981) will induce overconfidence in a perceptual letter identification task by increasing the difficulty of target letter identification.

The trial-by-trial feedback employed in Experiment 2 improved the resolution but not the calibration of confidence judgments. The latter result is consistent with those of Keren (1988, Experiment 1), who found no effect of trial-by-trial feedback for any of the performance measures investigated in his Landolt-C visual acuity task. Unfortunately, Keren did not provide resolution measures in his analyses, and thus a full comparison between his results and those reported here cannot be made. A main finding of the present study is that trial-by-trial feedback has its most pronounced effects when the *global difficulty* of the task is high. Presumably, calibration and resolution were poor in Experiment 1 under accuracy stress, because subjects received a preponderance of "Correct" feedback reports and thus felt no need to alter their confidence judgments. However, under speed stress in Experiment 1 and in the globally more difficult context in Experiment 2, subjects made more errors. Apparently, feedback about making errors leads to better resolution of the difference between correct and incorrect judgments.

The positive effect of feedback on resolution obtained in Experiment 2 is consistent with the findings of Sharp et al. (1988; cf. Lichtenstein & Fischhoff, 1980), who found that feedback improved the resolution of confidence for intellectual knowledge questions. However, it is important to note that the latter studies did not provide trial-by-trial response feedback as was done in the present studies. Rather, subjects in those studies were given detailed feedback about their performance and calibration after each session. It remains to be seen whether trial-by-trial feedback can improve the calibration or resolution of confidence judgments in nonperceptual tasks.

In sum, there are a number of similarities between the properties of confidence calibration in perceptual and nonperceptual judgment tasks. Although the type of information used to arrive at a feeling of certainty is clearly different in the two domains, there appears to be a common and general process for evaluating probabilistic information, obtaining a feeling of relative certainty in a judgment, and transforming that feeling into a numerical probability estimate.

Decision Times and Calibration Theory

The present study confirms the inverse relationship between confidence and decision time in the context of calibration analyses, and it has demonstrated an RT difficulty effect at each level of confidence. As previously mentioned, the RT difficulty effect is important in ruling out the possibility that confidence is *determined* by scaling the duration of the decision process—that is, that a specific decision time is not associated with a specific level of confidence. The RT difficulty effect is most likely to be evident under conditions of slower, more cautious, responding and when the range of difficulty levels is wide.

On the other hand, when caution is sacrificed for speed, the slope of the curve relating RT to confidence diminishes and the RT difficulty effect disappears.

The simultaneous study of calibration and RT data is important because it directs our theoretical consideration to a very specific class of potential theories of confidence calibration. Evidently, a successful theory of confidence calibration requires the simultaneous modeling of at least two aspects of human performance. The first is a *general* model of the decision process—one that is applicable to perceptual and nonperceptual judgment situations and one that is sufficiently articulated to permit quantitative relations among decision difficulty, decision time, and decision accuracy. The second requirement of a comprehensive theory of confidence calibration, as assumed by May (1986), is a model of how to evaluate the result of the decision process and then translate that evaluation into a numerical probability estimate.

Recently, Gigerenzer et al. (1991) have proposed a theory of confidence in the context of answering general knowledge questions which they claim is also applicable to perceptual judgments (pp. 521-523). According to this view, the subject generates and tests a sequence of potentially relevant cues until one of these cues becomes "activated" and permits the selection of one alternative over the other. Gigerenzer et al. hypothesize (see p. 524) that the process of cue testing might be temporally organized according to "cue validity"; that is, cues that are high in validity are tested sooner, will have higher accuracy, and will be associated with higher confidence levels. If all cues have been tested but none have been activated, the subject "guesses." In fact, temporal prioritization according to cue validity is necessary if the theory is to predict the inverse relationship between confidence and decision time, although it is not immediately evident how the theory might account for the RT "difficulty effect" evident in the present studies. More importantly, with direct reference to perceptual judgments, Gigerenzer et al. predict that overconfidence will be observed in a perceptual task "if perceptual tasks are selected for perceptual illusions—that is, for being misleading—whereas zero overconfidence is to be expected if tasks are not selected" (p. 522). With respect to the present findings, the theory cannot account for either the underconfidence obtained for easy perceptual judgments or the overconfidence obtained for difficult but nonillusory perceptual judgments (see Griffin & Tversky, 1992, for a discussion of other problems with this theory).

On the other hand, Björkman et al. (1993) and Winman and Juslin (1993) recently have developed a "subjective distance" theory of confidence calibration in perceptual judgments that predicts, *exclusively*, underconfidence. This theory, based on a Thurstonian scaling of stimulus differences, and a variant of the conventional strength theories (see below), posits cut-points along the decisional axis such that confidence increases monotonically with distance from the decisional criterion. The theory is able to provide an impressive account of the global underconfidence evident in their data and provides excellent quantitative theoretical fits to their calibration curves.

However, this theory is clearly unable to account for several aspects of the present data. First, both the stimulus-discriminability-based calibration difficulty effect evident in each of the three experiments reported in the present paper and the speed-accuracy tradeoff-induced difficulty effect obtained in Experiment 1 are not permitted, because as Björkman et al. (1993, p. 79) state, "The theory predicts underconfidence for all levels of $\bar{c}$ (with .5 and 1.0 as trivial exceptions)." (Björkman et al. use $\bar{c}$ to denote the proportion of correct responses). Thus, the clear occurrence of overconfidence in each of the three experiments reported here cannot be accounted for by the subjective distance model. Second, their subjective distance theory predicts that underconfidence must always occur with the "guessing" confidence category (Björkman et al., 1993, p. 77). The data from Experiments 1 and 3 of the present paper clearly contradict this prediction (Figures 1 [speed] and 10). Third, Björkman et al. (1993, Experiment 2) and Winman and Juslin (1993) failed to find any effects of trial-by-trial feedback, as was predicted by their theory, and this agrees with the failures to find any effects of feedback on calibration in Experiment 2 of the present paper. However, the present studies have established that a globally very difficult task is necessary in order for one to observe the effects of feedback and that these effects are evident with the resolution measure. Finally, the subjective distance theory, as will become evident below, is unable to account for the full configuration of RT, confidence, and response probability interrelations established in the present study.

To our knowledge, the first quantitative theory of confidence calibration fitted to empirical data is the signal-detection-based calibration model proposed by Ferrell and McGoey (1980). This theory comprises the two requisite components for the successful modeling of calibration data: (1) a primary decision process, characterized by signal detection theory (e.g., Green & Swets, 1966; Tanner & Swets, 1954), and (2) a process for obtaining a numerical probability (confidence) estimate, based on the precision of the detection and the width of confidence intervals, which are denoted by cut-points on the primary decision axis. Importantly, the signal-detection-based conceptualization and the basis for confidence are independent of the nature of the stimulus representation; the model assumes a common basis for the primary judgment and the judgment of confidence in perceptual and nonperceptual tasks.

However, as currently formulated, the Ferrell and McGoey (1980) model cannot account for the inverse relation between confidence and decision time. Specifically, because the model is theoretically conceived in the context of "single-sample" signal detection theory, and thus assumes only one "ideal observation" per trial (Tanner & Swets, 1954), the model predicts, contrary to the present findings, that decision times will be the same for all confidence categories (see Pike, 1973; Vickers, 1979).

This general limitation of signal detection theory can be overcome by assuming the existence of a *latency function* (Pike, 1973; Thomas, 1971) for the detection process in which RTs decrease as a function of the distance of the sampled observation from the decisional criterion axis. This view, commonly referred to as *strength theory* (e.g., Norman & Wickelgren, 1969; but see Coombs, 1964, p. 530 for an earlier statement), when applied in the context of the Ferrell and McGoey (1980) model, predicts that RTs will monotonically decrease as discriminability, and thus confidence, increases.

However, strength theory is likely to fail in the long run. First, Murdock and Dufty (1972) showed that, contrary to empirical results, strength theory predicts smaller RT variability for errors than for correct responses. Second, strength theory predicts that error RTs will *always* be longer than correct RTs (e.g., Coombs, 1964; Petrusic & Jamieson, 1978; Pike, 1973; Vickers, 1979), a prediction which is easily disconfirmed under conditions emphasizing speeded responding (e.g., Experiment 1) where error RTs are typically either the same as or faster than correct RTs (see Luce, 1986; Petrusic, 1992; Pike, 1971; Vickers, 1979).

Note that the inability of a signal-detection-based view to account for the RT data reflects a limitation of the assumed decisional model and not necessarily a limitation of the basis for confidence assumed by such a model. In fact, an appealing avenue for theoretical consideration would be to maintain (some variant of) Ferrell and McGoey's (1980) notion of confidence being a scaling of decision strength but to consider such a view in the context of a decision model that permits the full range of RT-response probability relations. For example, Vickers (1979) and his associates (Smith & Vickers, 1988; Vickers & Packer, 1982; Vickers et al., 1985) have developed, in the context of his accumulator model of psychophysical discrimination (Vickers, 1970, 1979), the *balance of evidence* hypothesis for the basis of confidence in two-choice situations. On this view, rated confidence will be proportional to the difference between the amount of information accrued favoring the dominant and nondominant responses at the completion of the judgment. To date, Vickers and his associates have provided impressive empirical support for the balance of evidence hypothesis in terms of its ability to account for the major properties of confidence in two-choice situations. The applicability of the balance of evidence hypothesis to confidence calibration and RT data, and the consideration of alternative models of confidence based on alternative evidence accumulation theories (e.g., Link, 1975, 1992; Link & Heath, 1975; Petrusic, 1992), provides the focus of our current work on the calibration problem (Baranski & Petrusic, 1991).

REFERENCES

Adams, J. K. (1957). A confidence scale defined in terms of expected percentages. *American Journal of Psychology*, **70**, 432-436.

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review*, **68**, 33-45.

Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *American Journal of Psychology*, **71**, 747-751.

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, **26**, 340-354.

Allwood, C. M., & Montgomery, H. (1987). Response selection strategies and realism of confidence judgments. *Organizational Behavior & Human Decision Processes*, **39**, 365-383.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior & Human Decision Processes*, **39**, 133-144.

Audley, R. J. (1960). A stochastic model for individual choice behavior. *Psychological Review*, **67**, 1-15.

Baranski, J. V. (1991). *Theories of confidence calibration and experiments on the time to determine confidence*. Unpublished doctoral dissertation, Carleton University, Ottawa, ON.

Baranski, J. V., & Petrusic, W. M. (1991). *Confidence calibration and the scaling of doubt*. Paper presented at the 52nd annual meeting of the Canadian Psychological Association, Calgary, AB.

Baranski, J. V., & Petrusic, W. M. (1992). The discriminability of remembered magnitudes. *Memory & Cognition*, **20**, 254-270.

Baron, J. (1988). *Thinking and deciding*. Cambridge: Cambridge University Press.

Bjork, E. L., & Murray, J. T. (1977). On the nature of input channels in visual processing. *Psychological Review*, **84**, 472-484.

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

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78** 1-3.

Carlson, B. W. (1993). The accuracy of future forecasts and past judgments. *Organizational Behavior & Human Decision Processes*, **54**, 245-276.

Christensen-Szalanski, J., & Bushyhead, J. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 928-935.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Dawes, R. M. (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice: Papers in honor of Clyde Coombs* (pp. 327-345). Bern: Hans Huber.

de Finetti, B. (1937). La prevision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, **7**, 1-68. English translation in H. E. Kyburg, Jr., & H. E. Smokler (Eds.) (1964), *Studies in subjective probability* (pp. 93-158). New York: Wiley.

Egeth, H. E., & Santee, J. L. (1981). Conceptual and perceptual components of interletter inhibition. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 506-517.

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

Festinger, L. (1943). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, **32**, 291-306.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 422-444). Cambridge: Cambridge University Press.

Fischhoff, B., & MacGregor, D. (1982). Subjective confidence in forecasts. *Journal of Forecasting*, **1**, 155-172.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception & Performance*, **3**, 552-564.

Fullerton, G. S., & Cattell, J. M. (1892). *On the perception of small differences*. Philadelphia: University of Pennsylvania Press.

Garrett, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, **56**, 3-104.

Geller, E. S., & Pitz, G. F. (1968). Confidence and decision speed in the revision of opinion. *Organizational Behavior & Human Performance*, **3**, 190-201.

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

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **11**, 702-718.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, **15**, 84-93.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, **116**, 119-136.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

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

Heath, R. A. (1984). Random-walk and accumulator models of psychophysical discrimination: A critical evaluation. *Perception*, **13**, 57-65.

Hellström, A. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, **97**, 35-61.

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, **18**, 186-201.

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology*, **51**, 261-268.

Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics*, **17**, 197-202.

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. *Archives of Psychology*, **34**, 1-53.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior & Human Decision Processes*, **39**, 98-114.

Keren, G. (1988). On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. *Acta Psychologica*, **67**, 95-119.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, **77**, 217-273.

Keren, G., & Wagenaar, W. A. (1987). Temporal aspects of probabilistic predictions. *Bulletin of the Psychonomic Society*, **25**, 61-64.

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

Liberman, V., & Tversky, A. (1993). On the evaluation of probability judgments: Calibration, resolution and monotonicity. *Psychological Bulletin*, **114**, 162-173.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? The calibration of probability judgments. *Organizational Behavior & Human Performance*, **20**, 159-183.

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

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

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, **12**, 114-135.

Link, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: Erlbaum.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, **40**, 77-105.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Lund, F. H. (1926). The criteria of confidence. *American Journal of Psychology*, **37**, 372-381.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 663-679.

Maki, R. H., & Swett, S. (1987). Metamemory for narrative text. *Memory & Cognition*, **15**, 72-83.

May, R. S. (1986). Inferences, subjective probability and frequency of correct answers: A cognitive approach to the overconfidence phenomenon. In B. Brehmer, H. Jungermann, & G. Sevon (Eds.), *New directions in research on decision making* (pp. 175-189). Amsterdam: Elsevier.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 223-232.

Münsterberg, H. (1894). Studies from the Harvard psychological laboratory: A psychometric investigation of the psychophysic law. *Psychological Review*, **1**, 45-51.

Murdock, B. B., Jr., & Dufty, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology*, **94**, 284-290.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: Part 1. Two-state situation. *Journal of Applied Meteorology*, **11**, 273-282.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595-600.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, **2**, 2-9.

Myers, J. L., & Well, A. D. (1991). *Research design & statistical analysis*. New York: Harper Collins.

Nickerson, R. S., & McGoldrick, C. C. (1963). Confidence, correctness, and difficulty with non-psychophysical comparative judgments. *Perceptual & Motor Skills*, **17**, 159-167.

Nickerson, R. S., & McGoldrick, C. C. (1965). Confidence ratings and level of performance on a judgmental task. *Perceptual & Motor Skills*, **20**, 311-316.

Norman, D. A., & Wickelgren, W. A. (1969). Strength theory of decision rules and latency in short-term memory. *Journal of Mathematical Psychology*, **6**, 192-208.

O'Connor, M., & Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting*, **8**, 141-155.

Oscamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, **29**, 261-265.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences*, **3**, 75-83.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 962-986.

Petrusic, W. M., & Baranski, J. V. (1989a). Context, context shifts, and semantic congruity effects in comparative judgments. In D. Vickers & P. Smith (Eds.), *Human information processing: Measures, mechanisms, and models* (pp. 231-251). Amsterdam: Elsevier.

Petrusic, W. M., & Baranski, J. V. (1989b). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics*, **45**, 439-452.

Petrusic, W. M., & Jamieson, D. G. (1978). The relation between probability of preferential choice and the time to choose changes with practice. *Journal of Experimental Psychology: Human Perception & Performance*, **4**, 471-482.

Petrusic, W. M., & Jamieson, D. G. (1979). Resolution time and the coding of arithmetic relations on supraliminally different visual extents. *Journal of Mathematical Psychology*, **19**, 89-107.

Phillips, L. D. (1973). *Bayesian statistics for social scientists*. London: Nelson.

Pierrel, R., & Murray, C. S. (1963). Some relationships between comparative judgment, confidence, and decision-time in weight-lifting. *American Journal of Psychology*, **76**, 28-38.

Pike, A. R. (1968). Latency and relative frequency of response in psychophysical discrimination. *British Journal of Mathematical and Statistical Psychology*, **21**, 161-182.

Pike, A. R. (1971). The latencies of correct and incorrect responses in discrimination and detection tasks: Their interpretation in terms of a model based on simple counting. *Perception & Psychophysics*, **9**, 455-460.

Pike, A. R. (1973). Response latency mechanisms for signal detection. *Psychological Review*, **80**, 53-68.

Rabbitt, P. M. A., & Vyas, S. M. (1970). An elementary preliminary taxonomy for some errors in laboratory choice RT tasks. *Acta Psychologica*, **33**, 56-76.

Ronis, D. L., & Yates, J. F. (1987). Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior & Human Decision Processes*, **40**, 193-218.

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, **2**, 191-201.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Seward, G. H. (1928). Recognition time as a measure of confidence. *Archives of Psychology*, **99**, 2-54.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior & Human Decision Processes*, **42**, 271-283.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two choice discrimination. *Journal of Mathematical Psychology*, **32**, 135-168.

Solomon, I., Ariyo, A., & Tomassini, L. A. (1985). Contextual effects on the calibration of probabilistic judgments. *Journal of Applied Psychology*, **70**, 528-532.

Tanner, W. P., & Swets, J. A. (1954). A decision making theory of visual detection. *Psychological Review*, **61**, 401-409.

Thomas, E. A. C. (1971). Sufficient conditions for monotone hazard rate: An application to latency-probability curves. *Journal of Mathematical Psychology*, **8**, 303-332.

Tomassini, L. A., Solomon, I., Romney, M. B., & Krogstad, J. L. (1982). Calibration of auditors' probabilistic judgments: Some empirical evidence. *Organizational Behavior & Human Performance*, **30**, 391-406.

Trow, W. C. (1923). The psychology of confidence. *Archives of Psychology*, **67**, 1-47.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, **13**, 37-58.

Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed versus accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica*, **50**, 179-197.

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental paradigms emphasizing state or process limitations: II. Effects on confidence. *Acta Psychologica*, **59**, 163-193.

Volkmann, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin*, **31**, 672-673.

Vreugdenhil, H., & Koele, P. (1988). Underconfidence in predicting future events. *Bulletin of the Psychonomic Society*, **26**, 236-237.

Wagenaar, W. A. (1988). Calibration and the effects of knowledge and reconstruction in retrieval from memory. *Cognition*, **28**, 277-296.

Wagenaar, W. A., & Keren, G. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. *Organizational Behavior & Human Decision Processes*, **36**, 406-416.

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **16**, 214-222.

Winman, A., & Juslin, P. (1993). Calibration of sensory and cognitive judgments: Two different accounts. *Scandinavian Journal of Psychology*, **34**, 135-148.

Wright, G. (1982). Changes in the realism and distribution of probability assessments as a function of question type. *Acta Psychologica*, **52**, 165-174.

Wright, G., & Ayton, P. (1986). Subjective confidence in forecasts: A response to Fischhoff and MacGregor. *Journal of Forecasting*, **5**, 117-123.

Wright, G., & Phillips, L. D. (1980). Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. *International Journal of Psychology*, **15**, 239-257.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessment for almanac and future event questions. *Scandinavian Journal of Psychology*, **23**, 219-224.

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 & Human Decision Processes*, **30**, 132-156.

Yates, J. F. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice-Hall.

NOTES

[1] The expression in Sanders's partition follows from that given in Equation 2 upon expanding

$$\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(e_{ij}-\bar{e}_{.j})^2,$$

distributing the summation operators, and noting that

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}e_{ij}^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j}e_{ij} = \sum_{j=1}^{J}n_j\bar{e}_{.j}.$$

(Note: $e_{ij}^2 = e_{ij}$ because $e_{ij} = 0$ or 1.) After some algebra, the expression reduces to

$$\frac{1}{n}\left(\sum_{j=1}^{J}n_j\bar{e}_{.j}-\sum_{j=1}^{J}n_j\bar{e}^2_{.j}\right) = \frac{1}{n}\sum_{j=1}^{J}n_j\bar{e}_{.j}(1-\bar{e}_{.j}).$$

[2] Recently, an interesting and related research area has emerged—the investigation of the relationship between confidence and the accuracy of reading comprehension (e.g., Glenberg & Epstein, 1985, 1987; Glenberg, Sanocki, Epstein, & Morris, 1987; Maki & Berry, 1984; Maki & Swett, 1987; Morris, 1990; Weaver, 1990). Although this research area is generally referred to as the "calibration of comprehension," the primary index of performance is the correlation between confidence (rated on ordinal scales) and comprehension accuracy, rather than the more formal index developed in the context of subjective probability analyses (see Nickerson & McGoldrick, 1963, 1965, for similar analyses in the context of intellectual knowledge judgments).

[3] Keren (1988) found overconfidence in one condition of the letter identification task (Experiment 2). This result, and Dawes's (1980, Experiment 5) overconfidence result, will be discussed in the General Discussion.

[4] Yaniv et al. (1991) demonstrated that an adjustment to η^2 is required in order to remove an inherent bias due to the J/n ratio, where J is the number of confidence categories and n is the total number of observations. In the present studies, n is always very large and thus bias is negligible (see Yaniv et al., 1991, p. 615).

[5] All analyses reported in this paper are based on the Greenhouse-Geisser epsilon adjusted degrees of freedom. However, the degrees of freedom shown in the text are defined by the design.

[6] It should be noted that the inverse relationship between confidence and RT is common to paradigms such as the present one, in which the subject controls the duration of the trial (termed *S-controlled* studies; see Vickers et al., 1985). Interestingly, when the duration of the trial is regulated by the experimenter (termed *E-controlled* studies; see Vickers et al., 1985), confidence is found to increase as the duration of the trial increases (e.g., Geller & Pitz, 1968; Irwin, Smith, & Mayfield, 1956; Vickers et al., 1985).

[7] One missing data point was estimated for the ANOVA according to the procedure recommended by Myers and Well (1991).

(Manuscript received November 24, 1991;
revision accepted for publication September 3, 1993.)