Perception & Psychophysics
1993, 54 (1), 75-81

Realism of confidence in sensory discrimination: The underconfidence phenomenon

MATS BJÖRKMAN, PETER JUSLIN, and ANDERS WINMAN
Uppsala University, Uppsala, Sweden

This paper documents a very pervasive underconfidence bias in the area of sensory discrimination. In order to account for this phenomenon, a subjective distance theory of confidence in sensory discrimination is proposed. This theory, based on the law of comparative judgment and the assumption of confidence as an increasing function of the perceived distance between stimuli, predicts underconfidence—that is, that people should perform better than they express in their confidence assessments. Due to the fixed sensitivity of the sensory system, this underconfidence bias is practically impossible to avoid. The results of Experiment 1 confirmed the prediction of underconfidence with the help of present-day calibration methods and indicated a good quantitative fit of the theory. The results of Experiment 2 showed that prolonged experience of outcome feedback (160 trials) had no effect on underconfidence. It is concluded that the subjective distance theory provides a better explanation of the underconfidence phenomenon than do previous accounts in terms of subconscious processes.

Do people have realistic conceptions of their accuracy in judgments and decisions? This is an issue that has received considerable attention in experimental studies during the last 15-20 years. "Realism of confidence" derives from a paper by Adams and Adams (1961), in which they introduced the notion of *expected percentages*. When a subject makes an assessment of confidence, he/she should understand that "of all those decisions made with confidence p, about p% should be correct" (Adams & Adams, 1961, p. 37). This way of viewing confidence became a conceptual basis for research during the following decades (for reviews see Fischhoff & MacGregor, 1982; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; O'Connor, 1989; Wallsten & Budescu, 1983).

In order to investigate realism of confidence, the continuous variable of confidence x has to be partitioned into discrete categories x_t ($t = 1 \ldots T$)—for example, .5, .6, .7, .8, .9, and 1.0 in a forced-choice, two-alternative task, and .0, .1 . . . 1.0 in a full-range task. Let the proportion of correct responses in category t be c_t. Realism, or *calibration*, which is now a frequent term for realism, is then a question of how much x_t differs from c_t. To be more precise, calibration (C) is defined in the following way (see, e.g., Yates, 1982):

$$C = 1/N \sum n_t (x_t - c_t)^2,$$

where n_t is the number of responses in category t, and N is the total number of responses. The calibration score (C) can be decomposed into three additive components (Björkman, in press), one of which is the square of bias (i.e., the difference between mean confidence and mean proportion correct, $\bar{x} - \bar{c}$). Optimal calibration requires that $\bar{x} - \bar{c} = 0$; $\bar{x} > \bar{c}$ is *overconfidence*, and $\bar{x} < \bar{c}$ is *underconfidence*.

With very few exceptions (see below), modern research on calibration has been concerned with confidence in cognitive tasks (e.g., general knowledge questions, predictions). The *overconfidence phenomenon*, in particular, has received large interest. This phenomenon, suggesting that people overestimate the validity of their general knowledge, has been interpreted as evidence for a general cognitive bias (e.g., Kahneman, Slovic, & Tversky, 1982). Recent studies, founded on an ecological orientation (Björkman, in press; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1993, in press), however, give a different picture. According to this view, overconfidence is largely a pseudo-phenomenon created by the experimenter's biased selection of items, rather than a cognitive bias on the part of the subjects.

The present paper addresses calibration of confidence in sensory discrimination. Throughout this paper, we will be concerned with experimental settings in which data are collected by the classical *method of constant stimuli*. The subject makes a judgment as to which of two stimuli is larger (heavier, longer, etc.) and then rates his/her confidence in this judgment. More specifically, the purpose is threefold: (1) to document a very consistent underconfidence bias in the area of sensory discrimination, (2) to propose a theory that accounts for this phenomenon, and (3) to test the predictions, and the quantitative fit, of the theory to empirical data.

Confidence in the Perception of Small Differences: A Review

Studies of confidence in psychophysical decisions by the method of constant stimuli are as old as experimental psychology itself, the first published study presumably

This research was supported by the Swedish Council for Research in the Humanities and the Social Sciences. The authors are indebted to Thomas S. Wallsten, Claudia González-Vallejo, and two anonymous reviewers for comments on an earlier draft of this paper.

being the one by Peirce and Jastrow (1884). Judgments of confidence were originally introduced as a complement to objective measures (e.g., the probable error) in order to refute the German-inspired notion of a least perceptible difference. It is interesting to see how the early psychophysicists connected confidence to perceived difference between stimuli: "The quantity which we have called the degree of confidence was probably the secondary sensation of a difference between the primary sensations compared" (Peirce & Jastrow, 1884, p. 82), and "The clearness with which a difference is distinguished varies gradually from complete doubt to complete certainty" (Fullerton & Cattell, 1892, p. 11). These formulations express the same relationship: Confidence, "from complete doubt to complete certainty," increases with the subjective distance between stimuli.

Absence of confidence in a decision (guessing, complete doubt) was expected to result in equal proportions of right and wrong responses. But it did not. The proportion of correct responses was consistently greater than the expected 50%. This finding was taken as evidence of *subconscious processing*:

> It is interesting to note that when the decision of the observer seemed a mere guess, he was considerably more likely to be right than wrong. This bears witness to the part played by subconscious mental processes in our daily lives. (Fullerton & Cattell, 1892, pp. 132-133)

The early findings of Peirce and Jastrow (1884) and Fullerton and Cattell (1892) that subjects had a larger portion of correct responses than wrong responses under guessing (complete doubt) were repeatedly confirmed by later researchers (e.g., Garrett, 1922; Griffing, 1895), and interpreted in the same way, as an indication of subconscious processes. Hence, more than 50% correct when the subject has no confidence (guessing) is a very robust underconfidence phenomenon. This fact was emphasized by J. K. Adams (1957) in a review of laboratory studies of behavior without awareness: "The only kind of behavior without awareness which can be easily replicated is the kind reported in 1884 by Peirce and Jastrow" (Adams, 1957, p. 385).

Johnson (1939, p. 28) found that "confidence as a variable in psychophysical judgments has not received careful experimental attention." What he had in mind was that confidence had so far been treated as ordered categories rather than a continuous variable. His main concern was the function relating confidence to the physical difference between stimuli. With this aim, it was necessary to have subjects make quantitative judgments of confidence. Johnson used length of lines as stimulus variable. The standard line was 50 mm. Seven shorter and seven longer comparison stimuli ranged from 40 to 60 mm. Three subjects rated confidence by making a pencil mark on a line of 100 mm. The graphical scale values were then transformed to percentage numbers such that 0 represented complete certainty that the comparison stimulus was shorter than the standard line and 100 represented complete certainty that it was longer.

In view of the modern development in which confidence is commensurate with probability (relative frequency), the following remark is significant:

> These mean percentages ranging from 0 to 100 may be treated—for ease of computation only—as analogous to frequencies or "p" values obtained by the method of constant stimuli and a normal ogive can be fitted to the data in the usual way. (Johnson, 1939, p. 35)

Normal ogives were fitted to the plots of mean confidence against the stimulus variable. Confidence was thus described by the same function as had been used traditionally for frequency data. Johnson did not report frequencies of correct judgments (probably because the number of observations was small), but he concluded that "the confidence function resembles the usual psychometric frequency function. The difference is that the slope of the curve is much more gradual and, consequently, the range of stimuli covered is much wider" (Johnson, 1939, p. 35). The difference in slope (greater variance of confidence assessments than of frequencies of correct judgments) means that the subjects were underconfident. Their discriminative accuracy was better than they expressed in their assessments of confidence.

Within the broader framework of a quantitative decision theory, Festinger (1943) confirmed Johnson's ogival relationship between confidence and stimulus difference (length of lines). On the average, the standard deviation of the confidence function was three times larger than that of the frequency function. Again, we can conclude that the subjects made more efficient discriminations than they expressed in their confidence assessments. This conclusion rests on the assumption that the confidence assessments were not only "analogous to frequencies" for "ease of computation," but represented expected percentages in the sense defined by Adams and Adams (1961).

The studies by Johnson and Festinger seem to end a classical period of research on confidence in sensory discrimination. Though this research, which began in the 1880s, was fairly sporadic, it produced results (in particular the discovery of underconfidence) that get renewed relevance when they are viewed from the perspective of present-day theory and methodology. During the 1950s and 1960s, signal detection theory became a prominent area of psychophysics and ratings of confidence (at rank order level) became a regular element of experimentation. At this time, confidence ratings were introduced with the single purpose of getting multiple decision criteria for plotting ROC curves. Calibration has received little attention within the framework of signal detection theory.

When the notions of realism of confidence and calibration began to come through, research was focused almost entirely on cognitive, rather than sensory, uncertainty. Nevertheless, the literature provides a few significant exceptions. Dawes (1980) proposed the hypothesis that the overconfidence found in general knowledge tasks would not appear in perceptual tasks. His data were not quite conclusive: "The clearest conclusion at this point is that the finding of over-confidence using intellectual content

can be reversed with some subjects making some perceptual judgments" (Dawes, 1980, p. 344). However, Dawes's hypothesis received support in a later study by Keren (1988), who argued that perceptual-like tasks would result in better calibration than would tasks requiring cognitive processing. The evidence from three experiments, the first of which was a direct comparison of a perception-like task and a general knowledge task, supported the conclusion that overconfidence gets reduced, or possibly reverses to underconfidence, in tasks that require little processing beyond sensory encoding. The stimuli in the perceptual task were black Landholt rings on a white background with a gap located either on the left or on the right side of the ring. Two types of rings were used with either a small gap or a large gap. At each trial, the subject's task was to indicate whether the gap was located to the left or to the right. The results suggested underconfidence for the perceptual tasks, a bias that reached statistical significance for the large-gap condition.

The interpretation of underconfidence favored by the early psychophysicists, as an indicator of subconscious processes, is not very convincing to a modern observer. First, the explanation seems entirely post hoc. There is no independent evidence for these subconscious processes and no account of why they should pull in the direction of underconfidence for categories of low confidence rather than, say, overconfidence for the categories of high confidence. Second, the explanation cannot be related to accepted or general principles in the area of perception or psychophysics. Third, the very robustness of the phenomenon seems to call for a similarly robust explanation in the form of a mechanism that inevitably generates underconfidence. The next section presents a theory that accounts for the underconfidence bias. This theory is an improvement over the theory of subconscious processes in all of these respects.

A Theory of Confidence in Sensory Discrimination

A useful alternative to the perception-cognition continuum approach proposed by Keren (1988) is to model the internal representations from which decisions and confidence derive. An example, in the case of general knowledge tasks, is internal cue theory (Björkman, in press), in which *internal cue validities* generate both decisions and confidence assessments. In psychophysical discrimination, on the other hand, we assume that confidence is a function of the *subjective distance* between stimuli. Let S_1 and S_2 ($S_1 > S_2$) be two stimuli close enough to be in the "uncertainty zone." Each comparison of the stimuli involves two discriminal processes (Thurstone, 1927a, 1927b) and, over trials, a resulting distribution of discriminal differences. Let this distribution (Figure 1) be a symmetric density function. Then, the proportion of correct responses is represented by the area to the right of zero, and the proportion of wrong responses is represented by the area to the left.

We now add the assumption that confidence is a monotone increasing function of the difference between discriminal processes—that is, the distance from zero in

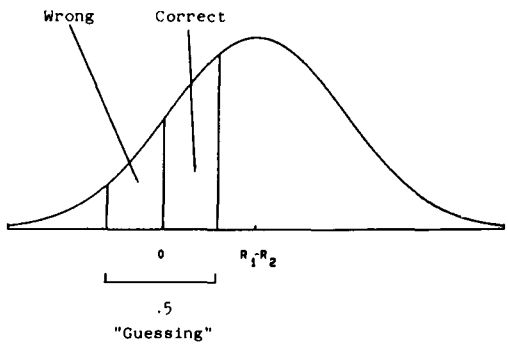

Figure 1. The distribution of differences according to the subjective distance theory: The guessing category contains more correct responses than wrong responses.

Figure 1. Hence, the categories of confidence assessments x_t (six in the following experiments) are mapped onto the continuum of sensory differences, with higher values of x_t going with larger differences. Each of the T categories is represented by an interval on the positive side (correct responses) and an interval of equal length on the negative side (wrong responses); the end category, $x_t = 1.0$, has no upper boundaries.

In the lowest category of confidence, there will always be more correct responses than wrong responses (see Figure 1). At this end of the scale, we can safely predict underconfidence, $x_t < c_t$. Furthermore, it is not likely that the disagreement between confidence and proportion correct will disappear in the following categories. If, for example, $c_t = .65$ for $x_t = .5$, we will necessarily find $c_t > .65$ for $x_t = .6$, and so on. From a value greater than .5 at $x_t = .5$, one should expect a smooth increase of proportion correct c_t with increasing confidence. Hence, we hypothesize that the imbalance between correct and wrong responses at $x_t = .5$ extends to the higher categories of confidence with the consequence of an average underconfidence, $\bar{x} < \bar{c}$.

A second issue that will be addressed below concerns the effect on calibration of outcome feedback. Will feedback reduce or eliminate underconfidence? We believe the answer is no. The only way that underconfidence can be reduced at $x_t = .5$ is by decreasing the number of assessments in this category. This will make the interval shorter and c_t closer to .5. Expressed differently, a few of the differences between discriminal differences that were assigned to $x_t = .5$ under no feedback would be assigned to $x_t = .6$ under outcome feedback. But this is counter to the assumption that confidence is based solely on the perceived difference between stimuli. Outcome feedback does not make the sensory system more sensitive. Since this "hard-wired" sensitivity is what is reflected in confidence assessments, we do not expect that outcome feedback will reduce underconfidence.

Application of the Theory to Data

Two ways of applying the theory can be distinguished. The first concerns the general *qualitative prediction* of

underconfidence for each subject and each stimulus unit. In order to test this prediction, data were analyzed in accordance with standard practice in calibration research—that is, data were first pooled across subjects and then across stimulus units to give the proportion correct (c_t) in each confidence category. The calibration curves reported below were computed in this way.

The second way to apply the theory concerns the *quantitative fit of the theory*, once the distribution of discriminal differences has been specified. Here, this specification is made by assuming that discriminal difference is a normally distributed random variable. The unit of analysis consists of two pairs of stimuli, the standard and two of the variable stimuli, equally distant from the standard. There are good reasons to assume that R and S are linearly related in the uncertainty zone around the standard stimulus, and thus that the subjective difference is equal in the two pairs of stimuli (= stimulus unit).

The data were pooled across subjects for each stimulus unit, the idea being that the composite distribution preserves the normality of each subject's distribution (i.e., because of linear combination). For each stimulus unit, computations followed the following procedure: (1) Determine z from the overall proportion correct. (2) Use the response proportions to compute the interval boundaries. (Note that the assumption of confidence as an increasing function of the subjective difference between stimuli implies that the confidence categories occupy equal intervals on the positive side and the negative side of zero. This makes it possible to compute intervals from the proportions of responses in each confidence category.) (3) Compute the "positive" (correct responses) and the "negative" (wrong responses) portions of the response proportion. (4) Determine proportion correct as the ratio between the positive portion and the response proportion of the category.

This procedure does not give a best fit (according to some criterion) of the theory to the data, since the interval boundaries are determined exactly by the response proportions. This way of applying the theory answers the question: How well does the theory predict proportion correct (c_t) from response proportions (n_t/N)? After the predicted number of correct responses had been determined for each stimulus unit, these numbers were summed across units and divided by the total number of responses in each confidence category in order to get the predicted proportions correct.

Observed proportions correct c_t can thus be compared to (1) the confidence assessments x_t (calibration) and (2) the predicted proportions correct (fit of theory). Only in the former do we expect significant deviations (underconfidence). When the quantitative fit of the theory is tested, we make one assumption that is stronger than when testing the general prediction of underconfidence (normality is substituted for symmetry) and one that is weaker (there is no need to use the numerical assessments of confidence; rank-ordered categories would be sufficient).

EXPERIMENT 1

In Experiment 1, the prediction of underconfidence and the fit of the theory were tested in two areas of sensory discrimination: heaviness and visual judgments of length.

Method

Subjects. Sixteen undergraduate students of psychology (9 males and 7 females) participated in the experiment. Their average age was 22.3 years. They were paid 100 Swedish Crowns for their participation.

Stimuli. Two kinds of stimuli were used: rectangles (width, 8 mm) varying in lengths and weights. The standard length of the rectangles was 190 mm. In the *easy* set, the comparison stimuli had lengths of 181, 184, 186, 188, 190, 192, 194, 196, and 199 mm. In the *hard* set, the lengths were 184, 186, 188, 189, 190, 191, 192, 194, and 196 mm. The standard weight was 200 g. In the *easy* set, the comparison stimuli were 130, 160, 180, 190, 200, 210, 220, 240, and 270 g. In the *hard* set, the stimuli were 180, 185, 190, 195, 200, 205, 210, 215, and 220 g.

Procedure. The subjects viewed the rectangles from a distance of 7 m. The weights were lifted with one hand, and the subjects were unable to see the weights. Subjects 1-8 had the easy set of weights and the hard set of rectangles; Subjects 9-16 judged the hard set of weights and the easy set of rectangles. Each subject made 14 judgments for each comparison stimulus, making $9 \times 14 = 126$ judgments per subject and stimulus set, and a total of 1,008 (8×126) judgments for each of the four sets of stimuli. In order to get the predicted proportions correct, the theory was applied to the 224 responses (14 judgments $\times$ 2 stimuli $\times$ 8 subjects) that resulted from each stimulus unit (when the comparison stimulus equals the standard, the theory, of course, predicts $c_t = .5$, for all confidence categories).

After each decision as to which stimulus was longer (heavier), the subject assessed his/her confidence that the decision was correct. This was done on a percentage scale with six values: .5, .6, .7, .8, .9, and 1.0. The end points were described in the instructions as *guessing* (flipping a coin) and *absolute certainty*. The use of a confidence scale with six predetermined alternatives has been common practice in calibration research, with the rationale being that people tend to use two-digit rounded numbers ending in zero, even if presented with a continuous scale (see, e.g., Winkler, 1971). The subjects were briefly introduced to the concept of calibration. In the data analysis, the subject's choice of stimulus was scored with 1 when correct, with 0 when wrong, and with .5 when the comparison stimulus had the same weight (length) as the standard stimulus. The subjects' work was divided into two sessions.

Results

The results are reported in Table 1. Notice first that there is a negative bias, underconfidence, of considerable size in all four conditions (−.12, −.12, −.15, and −.12). In all four conditions, the chi-square goodness-of-fit statistic indicates significant deviations between observed proportions correct c_t and the proportions correct required for perfect calibration (i.e., $c_t = x_t$). Of 32 cases (16 subjects $\times$ 2 conditions), only one was found in which the bias was positive (+.020).

The four conditions in Table 1 represent a considerable variation of proportion correct ($\bar{c}$), from .667 to .885. Mean confidence ($\bar{x}$) follows proportion correct with a fairly constant negative bias. This is in contrast to the

Table 1
Experiment 1: Number of Responses n_t and Proportion Correct c_t in Different Confidence Categories x_t

Condition	Subjects	Confidence (x_t)	n_t	c_t	Theory
Easy-weights*	1-8	.5	200	.655	.619
		.6	184	.845	.838
		.7	118	.903	.948
		.8	89	.949	.967
		.9	85	.994	.992
		1.0	332	.994	1.0
Hard-weights†	9-16	.5	264	.672	.574
		.6	267	.757	.719
		.7	217	.802	.830
		.8	117	.850	.879
		.9	89	.938	.944
		1.0	54	.954	.970
Easy-rectangles‡	1-8	.5	386	.650	.606
		.6	354	.778	.789
		.7	182	.821	.871
		.8	44	.886	.930
		.9	28	.857	.890
		1.0	14	1.0	1.0
Hard-rectangles§	9-16	.5	612	.618	.600
		.6	317	.710	.756
		.7	70	.864	.830
		.8	9	.944	.855

*In the easy-weights condition, mean confidence $\bar{x} = .767$, proportion correct $\bar{c} = .885$, calibration = .0233, bias $(\bar{x} - \bar{c}) = -.118$, chi-square: perfect calibration = 35.2 ($df = 5, p < .005$), fit of model = .73 ($df = 5, p = .98$).

†In the hard-weights condition, mean confidence $\bar{x} = .677$, proportion correct $\bar{c} = .782$, calibration = .0170, bias $(\bar{x} - \bar{c}) = -.115$, chi-square: perfect calibration = 20.0 ($df = 5, p < .005$), fit of model = 4.4 ($df = 5, p = .49$).

‡In the easy-rectangles condition, mean confidence $\bar{x} = .602$, proportion correct $\bar{c} = .747$, calibration = .0228, bias $(\bar{x} - \bar{c}) = -.145$, chi-square: perfect calibration = 19.2 ($df = 5, p < .005$), fit of model = 1.3 ($df = 5, p = .72$).

§In the hard-rectangles condition, mean confidence $\bar{x} = .548$, proportion correct $\bar{c} = .667$, calibration = .0143, bias $(\bar{x} - \bar{c}) = -.119$, chi-square: perfect calibration = 34.4 ($df = 3, p < .005$), fit of model = 2.1 ($df = 3, p = .83$).

"hard-easy" effect in general knowledge tasks (Lichtenstein & Fischhoff, 1977; see also elaborate discussions in Gigerenzer et al., 1991, and Juslin, 1993). This effect means that hard samples of items (low $\bar{c}$) result in overconfidence and easy items (high $\bar{c}$) result in underconfidence. Psychophysical discrimination does not exhibit any hard-easy effect, and it shouldn't. The theory predicts underconfidence for all levels of $\bar{c}$ (with .5 and 1.0 as trivial exceptions).

In a calibration diagram, the proportions correct c_t are plotted against the corresponding confidence levels x_t. Figure 2 shows the predicted and the observed calibration curves for the four conditions.[1] In Table 1, we see that the chi-square statistics for the fit of theory to the observed proportions correct c_t suggest that deviations are well within expectations of sampling errors.

Finally, in the hard-rectangles condition (Table 1), the subjects used only the four lowest confidence categories. This suggests that the subjects did not feel "forced" to distribute their assessments across all six scale values on the confidence scale. The subjects were able to use the scale in a discriminative way. For instance, in the easy-weights condition the average confidence assessment for the stimulus unit with the largest objective difference (200 g ± 70 g) was .97 (SD = .08), whereas the average confidence for the stimulus unit with the smallest difference (200 g ± 10 g) was .62 (SD = .14).

EXPERIMENT 2

Experiment 2 addressed the issue of whether outcome feedback would eliminate the underconfidence bias observed in Experiment 1. Keren (1988) compared a group that received outcome feedback during the experimental session with a group that received no outcome feedback. He found no differences in performance, neither for the cognitive nor for the perceptual tasks, but he reported no actual data for pretest and posttest performance. The subjective distance theory suggests that the underconfidence bias is very robust and thus not likely to be much affected by experience. We therefore expected feedback to have no effect on underconfidence.

Method

Subjects. Twelve undergraduate students of psychology (5 males and 7 females) participated in the experiment. Their average age was 23.4 years. They were paid 120 Swedish Crowns for their participation.

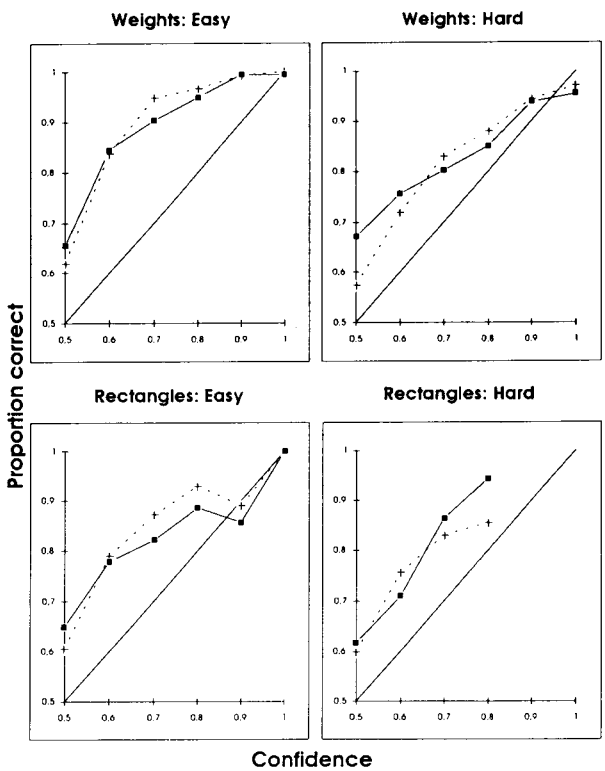

Figure 2. Predicted (+) and observed (■) calibration curves for the four conditions of Experiment 1.

Stimuli. In Experiment 2, only weights were used as stimuli. As in Experiment 1, the standard weight was 200 g. This standard was compared with variable weights of 160, 179, 189, 194, 197, 203, 206, 211, 221, and 240 g. In each comparison, the subject selected the weight judged to be the heavier one and assessed confidence on the same scale as in Experiment 1. To guard against fatigue, the subjects were asked to alternate between the left hand and the right hand on every 40 trials.

Procedure. Six blocks of 40 trials each were divided into two sessions separated by a rest pause. After the first (pretest) block, the subjects were given verbal outcome feedback in the form: "The answer is right" or "The answer is wrong." This feedback was maintained until the last (posttest) block. The subjects were instructed to try to make their confidence assessments more realistic on the basis of the feedback.

Results

Table 2 shows the measures for pretest (the first block) and posttest (the sixth block). As is obvious, the 160 trials with outcome feedback had little or no effect on these measures. None of the differences even approached statistical significance ($p > .2$; Wilcoxon signed-rank tests).

Figure 3 shows confidence and proportion correct for the six blocks. Note that the curve of confidence assessments is perfectly horizontal across all 160 trials of outcome feedback.

The subjects in Experiment 1 also made a large number of weight discriminations separated by a rest pause (2×63), but received no feedback. As a control condition for Experiment 2, the data from the first occasion and the second occasion of Experiment 1 were analyzed separately. Again, there was a slight but statistically insignificant increase in underconfidence between the first occasion ($\bar{x} = .781$, $\bar{c} = .877$, $\bar{x} - \bar{c} = -0.096$) and the second occasion ($\bar{x} = .755$, $\bar{c} = .893$, $\bar{x} - \bar{c} = -0.138$) of Experiment 1. It seems safe to conclude that outcome feedback does not eliminate underconfidence. Finally, we computed calibration on data collapsed across all six blocks. We again found underconfidence [$n = 2{,}880$, $\bar{x} = .739$, $\bar{c} = .822$, calibration $= .0107$, bias $(\bar{x} - \bar{c}) = -0.083$], with a calibration curve with proportions correct c_t higher than those required for perfect calibration (see Figure 4).

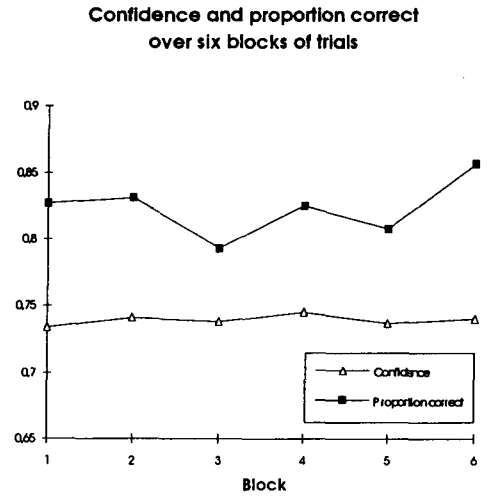

Figure 3. Mean confidence $\bar{x}$ and proportion correct $\bar{c}$ for the six blocks of Experiment 2 (see text).

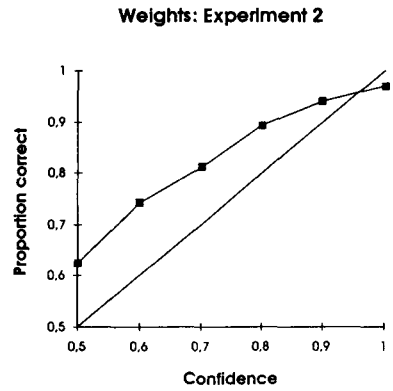

Figure 4. The calibration curve for data aggregated across all six blocks of Experiment 2.

DISCUSSION

The underconfidence observed in these two experiments is a very robust phenomenon. The amount of bias is substantial (-0.08 to -0.15), and it occurred both when proportion correct was fairly low (.667) and when it was high (.885). Furthermore, as we have seen, underconfidence was also a common observation among the early psychophysicists. The subjective distance theory explains underconfidence as an immediate consequence of the (symmetrically distributed) random fluctuations of the sensory system. Due to the fixed sensitivity of the nervous system, prolonged experience of outcome feedback does not affect the subjective experience of uncertainty.

When the theory is fitted to data, the interval boundaries are determined exactly by the response distribution, and all measurement error in regard to n_t/N is retained in the predicted c_t. Furthermore, the choice of a normal distribution of discriminal differences followed traditional practice. A different distribution may provide even better numerical fit. In view of these facts, the fit of the theory is highly satisfactory. This suggests that reference to more elaborate unconscious processes is unnecessary.

It should, of course, be noted that the subjective distance theory presented here is derived for the case of the method of constant stimuli. Another method may require

Table 2
Pretest and Posttest Measures From Experiment 2

Measure	Pretest	Posttest
Mean confidence: $\bar{x}$	.734	.740
Standard deviation (s_x)	.184	.172
Proportion correct ($\bar{c}$)	.827	.856
Bias ($\bar{x} - \bar{c}$)	−0.093	−0.116
Calibration (C)	.0183	.0183

a different representation. For instance, the method of single stimulus requires a representation in which both the individual (the decision criterion) and the stimulus are mapped into points on the subjective continuum. Confidence will be a monotone increasing function of the distance between the decision criterion and the stimulus.

The results presented here along with recent research on confidence in one's general knowledge (Björkman, in press; Gigerenzer et al., 1991; Juslin, in press) suggests a different nature of confidence in sensory judgments, relative to cognitive judgments. Confidence in cognitive judgments is adaptive, and it reflects knowledge structures (cues) formed by experience of natural environments. The notorious overconfidence phenomenon observed in previous studies with general knowledge items seems to a large extent to be a pseudo-phenomenon caused by the experimenter's biased selection of items. Indeed, when general knowledge questions are selected in order to be more representative of natural environments, people show good calibration (Gigerenzer et al., 1991; Juslin, in press) or even excellent calibration (Juslin, 1993). Confidence in sensory discriminations, on the other hand, reflects hard-wired features of the nervous system, features that lead to a systematic and robust underconfidence bias.

REFERENCES

ADAMS, J. K. (1957). Laboratory studies of behavior without awareness. *Psychological Bulletin*, **54**, 383-405.

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

BJÖRKMAN, M. (in press). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior & Human Decision Processes*.

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

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

FULLERTON, G. S., & CATTELL, J. M. (1892). On the perception of small differences. *Publications of the University of Pennsylvania* (No. 2).

GARRETT, H. E. (1922). A study of the relation of accuracy to speed. *Archives of Psychology*, No. 56.

GIGERENZER, G., HOFFRAGE, U., & KLEINBÖLTING, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, **4**, 506-528.

GRIFFING, J. H. (1895). On sensations from pressure and impact. *Psychological Review Monographs*, **1**, 1-88.

JOHNSON, D. M. (1939). Confidence and speed in two-category judgment. *Archives of Psychology*, No. 241.

JUSLIN, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in general knowledge. *European Journal of Cognitive Psychology*, **5**, 55-71.

JUSLIN, P. (in press). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior & Human Decision Processes*.

KAHNEMAN, D., SLOVIC, P., & TVERSKY, A. (Eds.) (1982). *Judgments under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

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.

LICHTENSTEIN, S., & FISCHHOFF, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance*, **20**, 159-183.

LICHTENSTEIN, S., FISCHHOFF, B., & PHILLIPS, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 306-334). Cambridge: Cambridge University Press.

O'CONNOR, M. (1989). Models of human behavior and confidence in judgment: A review. *International Journal of Forecasting*, **5**, 159-169.

PEIRCE, C. S., & JASTROW, J. (1884). On small differences of sensation. *Memoirs National Academy of Sciences*, **3**, 73-83.

THURSTONE, L. L. (1927a). A law of comparative judgment. *Psychological Review*, **34**, 273-286.

THURSTONE, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, **38**, 368-369.

WALLSTEN, T. S., & BUDESCU, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, **29**, 2, 152-173.

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

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

NOTE

1. In the condition of easy rectangles (presented in Figure 2), the proportion correct for Confidence Category .9 is lower than the proportion correct for Category .8. This is a consequence of the aggregation of data from several stimulus units into one calibration curve. At the level of each individual stimulus unit, proportion correct c_i is, of course, a monotone positive function of confidence x_i.

(Manuscript received February 12, 1992;
revision accepted for publication January 12, 1993.)