Image Cover Sheet

CLASSIFICATION	SYSTEM NUMBER
UNCLASSIFIED	389687

TITLE

ON THE CALIBRATION OF KNOWLEDGE AND PERCEPTION

System Number:

Patron Number:

Requester:

Notes:

DSIS Use only:

Deliver to: FF

On the Calibration of Knowledge and Perception

JOSEPH V. BARANSKI *Defence and Civil Institute of Environmental Medicine, North York*
WILLIAM M. PETRUSIC *Department of Psychology, Carleton University, Ottawa*

Abstract This study examined confidence judgements (i.e., calibration, resolution, and over/underconfidence) and response times in an intellectual knowledge task and a perceptual task requiring location comparisons. At each of four levels of judgement difficulty (i.e., Easy, Hard, Impossible and Misleading/Illusory), very similar properties were evident in the two tasks. The results are inconsistent with theories that assume a fundamentally different basis for confidence in human knowledge and perception.

Résumé Cette étude se penche sur les estimations de la confiance (c.-à-d. étalonnage, résolution et confiance exagérée ou manque de confiance) et les temps de réponse dans un test de connaissances intellectuelles et un test de perception exigeant une comparaison d'emplacement. À chacun des quatre degrés de difficulté (c.-à-d. facile, difficile, impossible, trompeur/illusoire), des caractéristiques très semblables étaient constatées dans les deux tests. Les résultats semblent contredire les théories selon lesquelles il existerait une différence fondamentale sur le plan de la confiance entre la connaissance humaine et la perception.

One of the oldest and most intriguing problems in experimental psychology concerns the relation between the accuracy of a judgement and the degree of certainty with which it is made (Peirce & Jastrow, 1884). In the classical psychophysical comparison literature (e.g., Henmon, 1911; Johnson, 1939), confidence was studied in conjunction with response accuracy and response time (RT). This tradition has continued in the contemporary development and evaluation of very general and quantitative theories of decision processing in psychophysical judgement and discrimination (e.g., Link, 1992; Vickers, 1979). Despite the long history of confidence investigation in psychophysical tasks, the preponderance of research on the confidence/accuracy relation is relatively recent, and falls under the general heading of *calibration* research (for reviews see Keren, 1991 and McClelland and Bolger, 1994).

Calibration is a normative index of the correspondence between confidence,

Canadian Journal of Experimental Psychology, 1995, 49:3, 397-407

viewed as a *subjective probability,* and overt performance, expressed in terms of a *response probability,* with the index ranging between 0 (optimal) and 1.0 (diabolically poor). An equally important index of the confidence/accuracy relation is *resolution* (Murphy, 1973), which reflects the degree to which subjects can distinguish correct from incorrect judgements. The normalized resolution index, η^2, likewise ranges between 0 (no resolution) and 1 (optimal). Finally, a third measure of considerable interest is *over/underconfidence* (Lichtenstein & Fischhoff, 1977): Judgements are considered overconfident if the mean proportion confidence exceeds the mean proportion correct, and underconfident if the reverse is true.[1]

A major focus of calibration research has been the confidence/accuracy relation for two-alternative intellectual knowledge judgements. A robust finding is that people are typically *overconfident* when judgements are difficult (e.g., < 65% correct) and *underconfident* when judgements are relatively easy (e.g., > 80% correct). This interaction between judgement difficulty and over/underconfidence is known as the calibration "difficulty effect" (Griffin & Tversky, 1992; Lichtenstein & Fischhoff, 1977) or "hard-easy effect" (Gigerenzer, Hoffrage, & Kleinbolting, 1991).

In an influential paper, Dawes (1980) considered the possibility that, unlike our intellectual knowledge, our remarkably accurate sensory systems might not be vulnerable to the phenomenon of overconfidence. Although Dawes was able to confirm overconfidence on a difficult knowledge task, his experiments on perceptual tasks were equivocal. Subsequently, Keren (1988) provided strong empirical support for Dawes' conjecture when he found overconfidence in a general knowledge task but no overconfidence in two perceptual tasks (Landolt-C acuity and letter identification). In addition, Björkman, Juslin, and Winman (1993) and Winman and Juslin (1993) recently provided demonstrations of what they referred to as a pervasive *underconfidence bias* in perceptual judgements. They concluded that perceptual judgements will not show overconfidence or a difficulty effect and, consequently, that "the nature of confidence in sensory discriminations is different from the nature of confidence in cognitive judgements (Winman & Juslin, 1993, p. 135)." More recently, however, Baranski and Petrusic (1994) provided clear evidence of a calibration difficulty effect in three experiments involving visual comparisons and gap discrimination, and thus argued for a common and general basis

1 *Calibration* $= \frac{1}{n}\sum_{j=1}^{J} n_j(\Psi_j - e_j)^2$, *Resolution*$(\eta^2) = \left[\frac{1}{n}\sum_{j=1}^{J} n_j(e_j - e)^2\right]/e(1-e)$, and Over/underconfidence $= \Psi - e$, where Ψ = mean proportion confidence, Ψ_j = proportion confidence in category j, e = mean proportion correct, e_j = proportion correct in confidence category j, n = total number of trials, and n_j = number of trials in confidence category j. A formal development of these indices can be found in several sources (e.g., Baranski & Petrusic, 1994; Yaniv et al., 1991).

for confidence in human judgements (e.g., Baranski, 1991; Ferrell & McGoey, 1980; Vickers, 1979).

The primary goal of the present study was to provide a direct comparison – involving separate analyses for the various levels of difficulty that comprise each task – of the properties of confidence calibration for general knowledge and perceptual judgement tasks. In addition, because our knowledge questions were selected from examples provided in previous studies, we anticipated that a subset of the questions would be "misleading" (i.e., < 50% correct; see Keren, 1991). Hence we sought a perceptual task that would likewise provide instances of misleading/illusory judgements to allow a direct comparison of the effects of such judgements on perceptual and knowledge-based calibration. Accordingly, a two-dimensional location comparison task was selected because we (Baranski & Petrusic, 1993) have found that certain stimulus configurations in this task will produce illusory "space-errors" (Fechner 1966/1860).

A second goal of the study, following on the recent findings of Baranski and Petrusic (1994), was to extend the analysis of decision times in calibration research to the domain of intellectual knowledge judgements (see also Wright and Ayton, 1998). In particular, we sought to determine if the direct relation between decision time and judgement difficulty and the inverse relation between confidence and decision time, two relations that are well-known in the psychophysical comparison literature (for a review see Link, 1992), are also evident in the context of answering general knowledge questions.

METHOD

Subjects

Forty-four naive Carleton University undergraduates participated for laboratory course credits. Twenty-five participated in the knowledge task and nineteen participated in the perceptual task.

Apparatus

Stimuli were presented on an Amdek-310A video monitor, which was controlled by an IBM-PC/XT computer. Responses were made on an IBM-PC 'mouse' and confidence reports were typed on the numeric keypad of the PC keyboard.

Stimuli and Procedure

Fifty general knowledge questions were selected from examples in previous studies (e.g., Lichtenstein & Fischhoff, 1977; Wright, 1984). On each trial, a short question and two response alternatives were presented on the computer screen. For example:

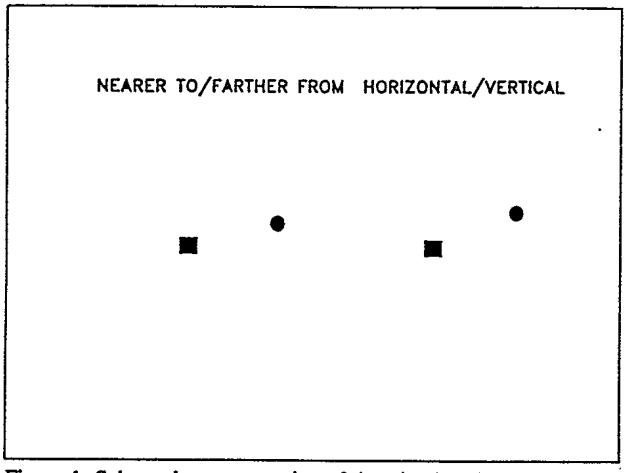

Figure 1. Schematic representation of the stimulus display in the perceptual task.

Which canal is longer?
Panama Canal Suez Canal

Subjects responded by depressing the left or the right key on the mouse. Immediately following the response, the screen was cleared and subjects were visually prompted for a confidence report. A rating of 100 was to indicate complete certainty in the response (100% correct) and a rating of 50 was to indicate a guess (50% correct). Reports between 50 and 100 were to indicate increasing certainty (in terms of probability or likelihood) of a correct response.

The display for the perceptual task consisted of two 3 × 3 mm squares which were placed so as to trisect an (imaginary) horizontal line traversing the center of the screen. The squares served as the origins of unique Cartesian coordinate systems. The comparison stimuli were two small circles (radius = 1.5 mm), placed on an imaginary arc (radius = 50 mm) in the first quadrant of the coordinate system defined by each square (see Figure 1). Four pairs of circles were used in the experiment; we define the two orders of each pair by the degrees subtended by the left and right circle from the horizontal axis: (2°,3°; 3°,2°), (4°,6°; 6°,4°), (84°,86°; 86°,84°) and (87°,88°; 88°,87°).

Ten subjects selected the dot that was nearer to horizontal on half of the trials and the dot that was nearer to vertical on the other half. Another nine subjects selected dots that were farther from horizontal and farther from vertical. Subjects responded by depressing the left or the right key on the mouse. Because there was no effect of the comparative nearer/farther on accuracy or confidence, we combined the data over this instructional manipulation. In all, 16 stimulus configurations were investigated (4 pairs ×

TABLE 1
Performance Measures in the General Knowledge and Perceptual Comparison Tasks as a Function of Judgement Difficulty.

	Judgement Difficulty			
	Easy	Hard	Impossible	Misleading/ Illusory
% Correct				
Knowledge:	85.3	67.0	53.1	37.8
Perception:	85.6	66.7	53.8	29.8
% Confidence				
Knowledge:	78.5	64.6	70.3	68.4
Perception:	84.5	77.9	78.4	75.5
% Over/Underconfidence				
Knowledge:	−6.8	−2.4	+17.3	+30.6
Perception:	−1.1	+11.2	+24.6	+45.7
Calibration				
Knowledge:	.0499 (.0082)	.0752 (.0095)	.1197 (.0691)	.2261 (.1565)
Perception:	.0158 (.0026)	.0567 (.0212)	.2108 (.0908)	.3653 (.3270)
Resolution (η^2)				
Knowledge:	.0925 (.0577)	.0214 (.0240)	.0103 (.0213)	.0150 (.0272)
Perception:	.1527 (.1579)	.0188 (.0433)	.0076 (.0109)	.0222 (.0362)
RT (ms)				
Knowledge:	5918	6980	7265	6875
Perception:	1626	1914	1965	1863

Notes. Entries are based on the mean of individual subjects. Values in parentheses are based on the pooled data (Figure 2). η^2 scores were adjusted for differences in within-cell n according to Yaniv et al. (1991).

2 presentation orders × 2 instructions), and each was replicated eight times in each of four blocks for a total of 512 trials per subject. Subjects in both tasks were instructed to respond as accurately as possible without taking too much time.

RESULTS AND DISCUSSION

Stimuli in each task were grouped into four levels of judgement difficulty: Easy (≥ 75% correct), Hard (56%-74% correct), Impossible (45%-55% correct), and Misleading/Illusory (≤ 44% correct). The levels were selected so as to obtain a wide range of response probabilities and a sufficient number of observations for subsequent calibration and RT analyses.

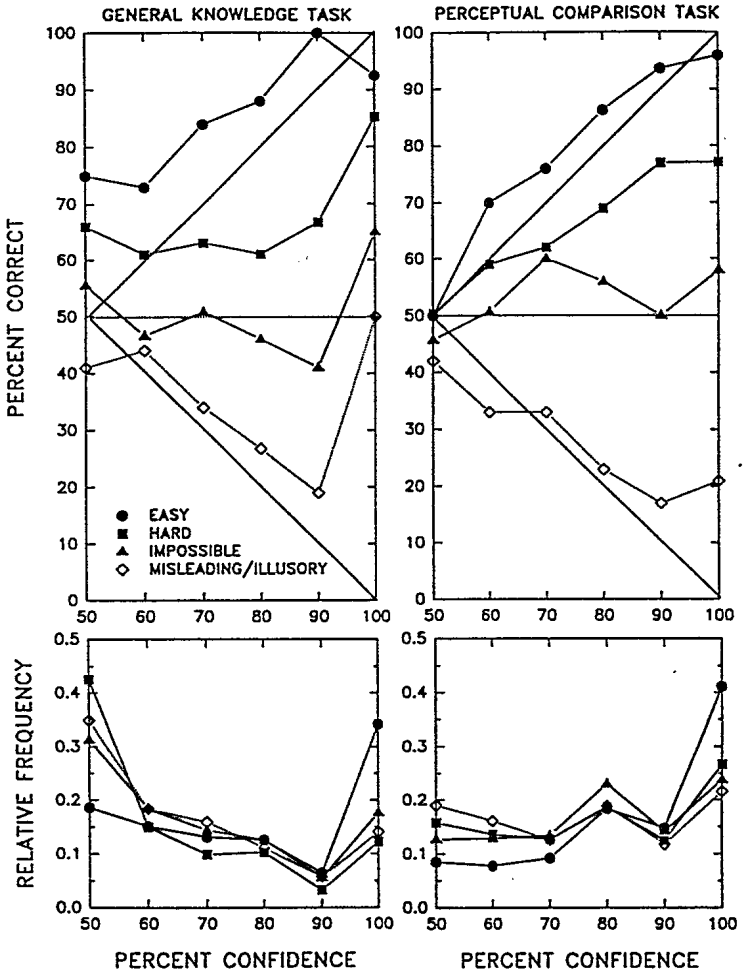

Figure 2. Calibration and response frequency curves for the four levels of judgement difficulty in the general knowledge and perceptual comparison tasks.

Calibration Analyses

The top panels of Figure 2 provide the 'calibration curves' for each level of difficulty in the two tasks. Each curve was obtained by plotting the proportion of correct judgements associated with each confidence interval (50-59%, 60-69%, 70-79%, 80-89%, 90-99%, and 100%). Hence 'perfect' calibration is denoted by points along the positive identity line, *underconfidence* by points *above* the line, and *overconfidence* by points *below* the line. The decreasing identity line denotes perfect *mis-calibration;* accuracy tends towards zero as confidence increases. The lower panels in Figure 2 provide the proportion of

trials associated with each confidence level.

Evident in Figure 2 is the overall similarity between the calibration curves in the general knowledge and perceptual comparison tasks. Specifically, there is underconfidence for the easy judgements, slight overconfidence for the hard judgements, neither calibration nor resolution and severe overconfidence for the impossible judgements and, with the exception of subjective certainty, almost perfect mis-calibration for the misleading items.

Table 1 provides summary statistics for the four levels of difficulty in the two tasks. Analyses of variance (ANOVAs) were conducted with type of task as a between-subjects factor and the four levels of judgement difficulty as a within-subjects factor (all significance levels are based on the Greenhouse-Geisser adjusted degrees of freedom).

As expected, the effect of Difficulty Level was highly reliable for both (arcsine transformed[2]) proportion correct, $F(3,126) = 78.86$, $MS_e = 11.38$, $p < .0001$, and proportion confidence, $F(3,126) = 36.26$, $MS_e = 0.56$, $p < .0001$. For proportion correct, the effect of Task and the Task × Difficulty Level interaction were not reliable, indicating that the groups were equally matched on the four levels of judgement difficulty. For proportion confidence, the effect of Task was reliable, $F(1,42) = 12.26$, $MS_e = 0.18$, $p < .005$, as was the Task × Difficulty Level interaction, $F(3,126) = 3.15$, $MS_e = 0.05$, $p < .05$. The former implies that subjects made lower confidence ratings on the knowledge task; the latter is due to a comparatively larger between-group difference in confidence for the Hard items. Despite these differences, however, it should be noted that confidence ratings follow a very similar pattern in the two tasks: appropriately high confidence for easy judgements, a decrease in confidence for the hard judgements, and inappropriately high confidence for the impossible and misleading/illusory judgements.

Under/overconfidence. ANOVA revealed reliable main effects of Task, $F(1,42) = 9.31$, $MS_e = 0.06$, $p < .004$, and Difficulty Level, $F(3,126) = 69.13$, $MS_e = 1.67$, $p < .0001$. The former was expected because the groups were matched on proportion correct but differed in their confidence judgements. The latter confirms the calibration "difficulty effect" with these perceptual and knowledge judgements; i.e., underconfidence for the Easy judgements and progressive overconfidence as difficulty increases. Importantly, the Task × Difficulty Level interaction was not reliable, indicating that judgement difficulty had the same effect in the two tasks. In addition, the 11.2% overconfidence evident for the Hard perceptual judgements is reliably different from zero (st.err. = 3.91, $t(19) = 2.85$, $p < .011$; 14/19 subjects displayed overconfidence), confirming that overconfidence can be obtained for difficult but non-illusory perceptual judgements (cf. Björkman et al., 1993).

2 An arcsine transformation was used to achieve homogeneous variance for the binomially distributed proportion correct measure.

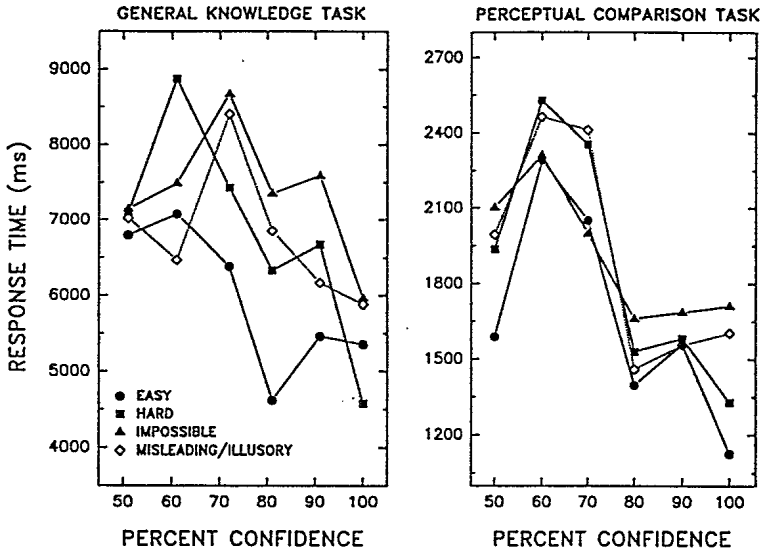

Figure 3. Confidence-response time curves for the four levels of judgment difficulty in the general knowledge and perceptual comparison tasks.

Calibration. The effect of Difficulty Level was reliable, $F(3,126) = 39.25$, $MS_e = 0.60$, $p < .0001$; for both groups, calibration became progressively worse as difficulty increased (recall that the lower the score the better). The main effect of Task was not reliable but the Task × Difficulty Level interaction was reliable, $F(3,126) = 4.99$, $p < .02$, as shown in Table 1.

Resolution (η^2). The only reliable effect for (arcsine transformed) resolution was that of Difficulty Level, $F(3,126) = 3.84$, $MS_e = 1.17$, $p < .016$, suggesting that resolution decreases as difficulty increases (recall that the higher the score the better). Interestingly, in both tasks, resolution increased for the misleading/illusory judgements. This reflects an ability to differentiate correct from incorrect judgements but, as is evident in Figure 2, the skill is actually counter to reality; i.e., error probability increases as confidence in a correct response increases!

Response Time Analyses

The Relation Between RT and Difficulty. ANOVA (performed on inverse transformed RT[3]) revealed main effects of Task, $F(1,42) = 84.74$, $MS_e = 4.10 \times 10^{-7}$, $p < .0001$, and Difficulty Level, $F(3,126) = 12.26$, $MS_e = 2.0 \times 10^{-8}$, $p < .0001$. The former was expected given the very large difference in RT between the tasks and the latter reflects the fact that, in both tasks, RTs

3 An inverse transformation (i.e., response speed) was used in order to achieve homogeneous variance.

increase as judgement difficulty increases. Also evident in Table 1 is that RTs are relatively fast for the misleading/illusory judgements, corroborating the finding that these judgements were *subjectively* easy. The Task × Difficulty Level interaction was not reliable.

The Relation Between Confidence and RT. Figure 3 provides plots of the RT associated with each confidence interval for the four levels of difficulty in the two tasks. In each case, we see the inverse relation between confidence and RT that is well known in the psychophysical comparison literature (and see Wright and Ayton, 1998), and an ordering of the curves with respect to the level of difficulty of the judgements: an "RT difficulty effect". Importantly, the RT difficulty effect denies the possibility that confidence is *determined* by scaling the duration of the decision process (Audley, 1960), because a specific decision time is not associated with a specific confidence level.

SUMMARY AND CONCLUSIONS

The present findings demonstrate that if general knowledge and perceptual comparison tasks are conducted under a common judgement paradigm (i.e., one in which the subject controls the duration of processing), and if the levels of judgement difficulty are appropriately matched, then very similar properties of confidence judgements will be evident. Moreover, many of the well-established RT properties in perceptual comparison tasks are also evident in the context of answering intellectual knowledge questions. Hence theories that predict that perceptual judgements will exhibit only underconfidence, and thus assume a fundamentally different basis for the judgement of confidence in perceptual and non-perceptual tasks (Björkman et al., 1993; Winman & Juslin, 1993), are falsified by the present data. Conversely, these data provide a valuable replication and extension of the Baranski and Petrusic (1994) findings, and in doing so encourage the development of more general theories of confidence calibration. Specifically, the dependence of RT on decision difficulty and the inverse relation between RT and confidence taken together dramatically constrain the classes of admissible models of confidence calibration. Currently, the only models that provide a quantitative account of these relations are those based on stochastic-evidence-accrual decision processes. Several models of this class, developed recently by Baranski (1991) as extensions of Vickers' (1979) "balance-of-evidence" view (i.e., confidence is a scaling of the difference in accumulated evidence in favor of each alternative), provide an excellent first-order account of the present data, and thus provide a potentially powerful and general theoretical framework for research on the confidence/accuracy relation.

This research was funded by an NSERC grant to William Petrusic. Address correspondence to J.V. Baranski, DCIEM, 1133 Sheppard Ave. West, P.O. Box 2000, North York, Ontario, M3M 3B9, or to W.M. Petrusic, Department of Psychology,

Carleton University, Ottawa, Ontario, K1S 5B6. E-mail addresses are, respectively, joe.baranski@dciem.dnd.ca and bill_petrusic@carleton.ca.

References

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

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

Baranski, J.V., & Petrusic, W.M. (1993). *Comparing locations in the plane.* Presented at the 3rd Annual Meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, Toronto, Ontario.

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

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

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.

Fechner, G.T. (1966/1860). *Elements of psychophysics.* New York: Holt.

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

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

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

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

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

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? The calibration of probability judgments. *Organizational Behavior and Human Performance, 20,* 159- 183.

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

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

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

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

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

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

Wright, G. (1984). *Behavioural decision theory: An introduction.* Middlesex: Penguin.

Wright, G., & Ayton, P. (1988). Decision time, subjective probability and task difficulty. *Memory and Cognition, 16,* 176-185.

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

Date of acceptance: September 16, 1994

#389687

NO. OF COPIES NOMBRE DE COPIES	COPY NO. COPIE N°	INFORMATION SCIENTIST'S INITIALS INITIALES DE L'AGENT D'INFORMATION SCIENTIFIQUE
1	1	JC
AQUISITION ROUTE FOURNI PAR ▶	DCIEM	
DATE ▶	20 MARCH 1996	
DSIS ACCESSION NO. NUMÉRO DSIS ▶		

DND 1158 (6-87)

National Defence / Défense nationale

PLEASE RETURN THIS DOCUMENT TO THE FOLLOWING ADDRESS:

DIRECTOR
SCIENTIFIC INFORMATION SERVICES
NATIONAL DEFENCE
HEADQUARTERS
OTTAWA, ONT. - CANADA K1A 0K2

PRIÈRE DE RETOURNER CE DOCUMENT À L'ADRESSE SUIVANTE:

DIRECTEUR
SERVICES D'INFORMATION SCIENTIFIQUES
QUARTIER GÉNÉRAL
DE LA DÉFENSE NATIONALE
OTTAWA, ONT. - CANADA K1A 0K2