WHERE DOES THE BALANCE OF EVIDENCE LIE WITH RESPECT TO CONFIDENCE?

Douglas Vickers
Adelaide University; douglas.vickers@adelaide.edu.au

Abstract

The primary aim of most sequential sampling models of discrimination has been to explain subjects' ability to trade speed for accuracy. Less commonly, such models have attempted to account for the confidence with which responses are made. Results are reviewed from some half dozen studies, conforming to an expanded judgment paradigm, in which accuracy, response time and confidence were measured. Out of several suggested theoretical bases for confidence, the balance-of-evidence hypothesis appears to be most successful. However, the results also underline the importance of assumptions regarding the process in which internal representations of confidence are converted into overt responses.

In 1923, when Clark Trow wrote to J.B. Watson to ask what was the behaviourist position on confidence, Watson replied "I am afraid you have come to the wrong market". Despite this, and even if we exclude *self*-confidence, the topic still has a considerable market currency. We depend on assessments of confidence to make investment choices, to evaluate the credibility of eyewitness testimony, and to carry out (or undergo) surgical procedures. In psychology, confidence measures have been relied on for over a century to test hypotheses about perception, memory and decision making - both directly and through the construction of receiver operating characteristic (ROC) curves for signal detection theory analyses.

What is remarkable is that, despite its practical importance and pervasiveness, the variable of confidence seems to have played a Cinderella role in cognitive psychology – relied on for its usefulness, but overlooked as an interesting variable in its own right.

A TAXONOMY OF CONFIDENCE EFFECTS

Before looking at alternative formulations for confidence, it is useful to outline a possible taxonomy of confidence effects. First, we need to distinguish between different kinds of task. Assuming that we restrict ourselves to perceptual judgments, we might then want to treat discrimination, identification, and detection separately. We might also want to distinguish between tasks involving different numbers of response alternatives.

The next set of distinctions concerns the way a task is implemented. For example, a familiar *sensory* discrimination would be an experiment in which the subject is shown a pair of lines and has to decide which line is the longer. So-called *expanded judgments* (EJs) are less familiar. On each trial of an EJ task, the subject has the opportunity to inspect multiple

stimulus elements and to make a judgment concerning either the sample presented or the population from which the sample is drawn. The name "expanded judgment" means that the task attempts to *externalise* the hypothesised, interior process of sequential sampling in a psychophysical judgment. (In earlier versions, this meant *expanding* the process in time.)

For example, in a temporal EJ task, using *discrete* stimulus elements, the subject may be presented with a sequence of flashes on one or the other of two lamps, with instructions to decide which lamp is set to flash more frequently. By comparison, in a task using *continuously varying* elements, the subject might inspect a succession of horizontal line segments, extending to the right or the left of a central, vertical line, representing zero. The subject is told that the sequence is generated by drawing from a normal distribution, with negative numbers being represented by leftward-extending segments and positive numbers by rightward-extending segments. On a given trial, the task is to decide whether the sample (or the distribution), used to generate the segments, has a mean that is positive or negative.

In *spatial* EJ tasks, the elements are distributed spatially rather than temporally.

As with sensory judgments, EJ tasks may conform to either a *time*- or an *information-limited* paradigm. For example, in a time-limited, discrete, temporal EJ task, the subject may be presented with a sequence – of fixed length - of left or right flashes, and be asked to decide whether that sample has more right or more left flashes. In an information-limited version, subjects are allowed to continue inspecting flashes until they decide whether the population from which the flashes are drawn has more right or more left flashes.

In an information-limited EJ discrimination task, there are at least four distinct variables that affect response probability, time, and confidence. These are: (1) the *discriminability* of the two sets of stimulus elements; (2) the *speed-accuracy tradeoff* (or inferred *degree of caution* for both responses), adopted by the subject; (3) the *relative degree of caution* voluntarily exercised for one response in preference to the other; and (4) the conscious *expectation*, held by the subject, that one or the other response is more likely.

Each of these variables can be manipulated in various ways. However, the most extensively studied variable is that of discriminability. In the case of a discrete, temporal EJ task, this would be manipulated by varying the *relative frequency* of the two binary stimulus elements. In tasks employing continuously varying stimulus elements, the situation is a little richer. Assuming that the stimulus elements are normally distributed, their discriminability can be manipulated in three main ways: (1) *varying the mean, m*, while holding the *standard deviation, s, constant*; (2) *varying both m and s*; or (3) *varying s*, while holding *m constant*.

Each manipulation of discriminability can be examined for its direct effects on response probability, time, and confidence, and on the interrelations between these variables.

HYPOTHESES REGARDING CONFIDENCE

Confidence obeys three generalisations: (1) it is a direct function of discriminability; (2) it is a direct function of accuracy; and (3) it varies inversely with response time. Peirce and Jastrow (1884) first quantified the second generalisation in the descriptive formula, $C = h \log (p/1\text{-}p)$, where C represents the measured degree of confidence, p denotes the probability of a response being correct, and h is a constant. Later, Volkmann (1934) tried to capture the third, using the equation, $C = 0.5(a/t\text{-}b) + 0.5$, where t is time and a and b are constants.

A century after Peirce's first article, three accounts appeared, each incorporating confidence into a theoretical model of the discrimination process. In one, Ratcliff (1978) proposed a *diffusion* model in which two conflicting evidence streams continuously drive a random walk towards one or the other of two thresholds. The rate of drift of the walk is determined by the discriminability of the alternatives, while the thresholds are assumed to be

set by the subject. Because the only information about discriminability is the decision time, Ratcliff proposed that confidence be an *inverse function of the actual time* taken by a subject.

A similar proposal was put forward by Link and Heath (1975). In their *random walk* model, it is differences between alternative stimulus inputs and an internal referent that are used to drive the walk towards an upper threshold, A, or a lower threshold, -A. In addition, these differences are input to the walk at discrete intervals, rather than continuously.

In Link and Heath's model, confidence is postulated to be a function of the distance (A-O) traversed by the walk, multiplied by a discriminability parameter, θ, where θ is evaluated in terms of the parameters (m, s) of the distribution of sampled differences, and O is the starting position of the walk (Heath, 1984; see also Vickers & Smith, 1985).

The third model is the *accumulator*, suggested by Vickers (1979). In this model, stimulus differences are sampled at discrete intervals, with positive and negative differences being accumulated in two separate totals until one or the other reaches a preset threshold. On this model, confidence is determined by the *balance-of-evidence* (i.e., by the difference between the two totals at the moment a decision is reached or sampling terminated.

Recently, Juslin and Olsson (1997) proposed a window-sampling model of sensory discrimination. In this model, discriminal differences are sampled, one at a time, and averaged over a moving window. Confidence is determined by the *ratio* of sampled differences, in favour of the successful response, that are present in the window when a response is made.

EMPIRICAL COMPARISONS BETWEEN ALTERNATIVE HYPOTHESES

I should like to compare these hypotheses about confidence. In particular (though not exclusively), I shall examine their accounts of the results from some half dozen EJ studies, carried out over the last three years. The experimental details are summarised in Table 1. I shall focus on specific features, moving sequentially (and selectively) through the taxonomy.

Table 1. Summary of expanded judgment tasks

Expt.	Discrete/ Continuous	Time-limited Fixed/Variable		Information-limited	Discriminability	No.Trials per Cond	No. Ss
I	Discrete	F	V	I	p=.57/.43	600	20
II	Discrete	F	V	I	p=.57/.43	600	20
III	Continuous	F	V	I	m=10, s=40	600	20
IV	Continuous	F	V	I	m/s=4/20,6/30,8/40	600	10
V	Continuous	F		I	m/s=2/12,4/24,6/36	600	20
VI	Continuous	F		I	m=4, s=12/18/30	600	3

Time-limited tasks

In our time-limited EJ experiments (I, II, III, and IV), employing variable sequence lengths, and with both discrete and continuously variable magnitudes, we found that confidence increases as a *direct* function of the time for which observations are presented. (cf Vickers, Smith, Burt, & Brown, 1985). An inverse-function-of-time (IFT) hypothesis predicts the opposite. Meanwhile, a ratio-based hypothesis would predict no increase in confidence. It is not clear how Link and Heath's hypothesis could be modified to apply to this situation.

By comparison, on the balance-of-evidence (BE) hypothesis, the expected difference between the accumulated positive and negative totals should be given by $C_n = nm$, where n is the number of observations presented and C_n is the confidence after n observations. This predicts that confidence should increase as a direct function of the number of observations and that it should be higher for more discriminable stimuli. The BE hypothesis makes similar predictions for a task employing binary-valued stimulus elements with probabilities p and q ($p > q$, and $p + q = 1$). Here, $C_n = n(p - q)$, and $(p - q)$ is a measure of discriminability.

Confidence in correct and incorrect responses

When the distributions of observations favouring the two alternatives are symmetric (as in these experiments), then mean response times for correct and incorrect responses, predicted by either the diffusion or the random walk model, should be equal. Hence, these models predict that confidence in errors should be equal to that for correct reponses.

In contrast, both the ratio model and the BE hypothesis predict that confidence in incorrect responses will be lower than for correct responses. In the case of the latter, the most confident responses will be those for which all the evidence accumulated favours one alternative (and triggers a fast response). The least confident will be those where evidence totals are similar (and sampling has been continued for longer).

Results from the information-limited conditions of Experiments I-VI show that confidence in errors is uniformly lower than for correct responses (cf Vickers et al., 1985).

Effects of variations in caution: The macro-tradeoff between speed and accuracy

Vickers and Packer (1982) found that, in trials where accuracy was emphasised, subjects were more accurate, took longer and were more confident. Baranski and Petrusic (1998) found the same effect in early - but not later - sessions. They argued that subjects may have used confidence to regulate response thresholds within a trial. Be that as it may, the finding that higher thresholds result in greater accuracy, longer response times and higher confidence is inconsistent with the notion that confidence is an inverse function of time. This finding also conflicts with Juslin and Olsson's ratio hypothesis (Vickers & Pietsch, in press).

In contrast, both the BE hypothesis and Link and Heath's formulation are consistent with a covariation of confidence and response time through the macro-tradeoff.

Confidence and the micro-tradeoff

Even if subjects adopt constant threshold values, accuracy and response time will both vary whenever there is any variability in the stimulus or in its internal representation. The resulting micro-tradeoffs (or conditional accuracy functions) generally show an inverse relation between accuracy and response time (Experiments I-VI; see also Vickers et al., 1985). This conflicts with the diffusion, random walk, and window-sampling models, which all predict that the probability of making a correct response should be the same at all points on the response time distribution (except, in the last case, for responses exceeding a deadline).

Information-limited experiments with nominally constant discriminability also exhibit an inverse relation between confidence and response time (Experiments I-VI; see also Vickers et al., 1985). This is inconsistent with a ratio-based hypothesis, which predicts that confidence should be independent of response time in the micro-tradeoff (Vickers & Pietsch, in press).

This finding also contradicts the random walk formulation, since the discriminability parameter, θ, remains constant from trial to trial.

In contrast, the result is entirely consistent with a BE hypothesis.

Confidence and relative caution

Besides manipulating the subject's overall tradeoff between speed and accuracy, it is possible to influence a subject's degree of caution with respect to one response *relative* to that for the other. Heath (1984) predicts that, when there is substantial 'bias' towards one response, or when the threshold for that response is reduced, less accumulated discrepancy is required before that response occurs, and the confidence in that response should be reduced.

A similar prediction is made by the BE hypothesis. However, the prediction of a direct relation between response time and confidence is opposed to the IFT hypothesis.

As reported by Vickers (1985), when subjects are instructed, in alternate blocks of trials, to be more ready to make one or the other response more quickly, they are more likely to make that response, they make it more quickly, and they are *less* confident in making it.

These results are consistent with both Link and Heath's formulation and with the BE hypothesis. However, they are quite inconsistent with an IFT account of confidence.

Confidence and expectation

In addition to both overall and relative caution, we can also manipulate the *expectation* that a subject has concerning the relative likelihood of one response rather than the other. This was examined in a study, in which subjects, in alternate blocks of trials, were told (truthfully) that the probability of one of two alternative stimuli would be greater than that of the other (Vickers, 1985). When subjects knew A stimuli would be more likely, they were more likely to make A responses (correctly and incorrectly), they made A responses more quickly, and they were more confident in A responses than when they thought B stimuli were more likely.

This finding is entirely consistent with the BE hypothesis, if it is assumed that expectation is represented by the amount by which the starting position is displaced, and that this amount is *added* to the evidence total for the corresponding response. The effect of this is to make responses for the expected alternative more likely, faster *and* more confident.

This result is also in line with an IFT hypothesis. However, it cannot be reconciled with Link and Heath's formulation because the distance traversed by the walk has been shortened (whether this accomplished by shifting the starting point or reducing the threshold).

Confidence and discriminability

Results from conventional manipulations of discriminability (varying m and holding s constant) agree qualitatively with predictions by all of the formulations for confidence.

The second two ways of manipulating discriminability yield more challenging results. Varying *both* m and s, while holding the ratio m/s constant, in a time-controlled task (Experiment V), produces higher confidence ratings for larger scalings. This agrees with predictions by the BE hypothesis, but not with those of any ratio-based or IFT hypothesis.

In the information-controlled condition, subjects are also more confident (and faster) at larger scalings. This agrees with the BE and IFT hypotheses, but contradicts the ratio-based and Link and Heath's hypotheses, according to which the scalings are equally discriminable.

In contrast, when m is held constant and s varied (Experiment VI), results are quite different. In a time-controlled task, subjects are more accurate and more confident with lower values of s. This contradicts the BE hypothesis as presently formulated.

In the information-controlled condition of Experiment VI, subjects are also more accurate, faster and more confident with more discriminable stimuli. This last result agrees

with the IFT and ratio-based hypotheses, as well as with the hypothesis of Link and Heath, but not with the current formulation of the BE hypothesis.

CONCLUSIONS

Aside from this last set of results, the balance-of-evidence hypothesis gives a good qualitative account of: (1) the contrast in the relations between time-limited and information-limited tasks; (2) the difference between correct and incorrect responses; the effects of variations in (3) overall and (4) relative caution; (5) the micro-tradeoff between confidence and response time; (6) the effects of expectation; and (7) established effects of conventional manipulations of discriminability. A concise, general summary of predicted and empirical confidence measures of these effects would say that they each provide a veridical, unbiased estimate of the probability that a judgment is correct, given all the evidence available.

Meanwhile, the balance-of-evidence hypothesis also appears to be capable of capturing the effects on performance of different scalings of stimuli of otherwise nominally equal discriminability. However, as currently formulated, the hypothesis relies for this on the absolute (unscaled) magnitudes of the accumulated evidence totals. The results of varying *s*, while holding *m* constant, suggest that, in this situation at least, we also need to take account of the way in which such magnitudes may be converted into overt confidence ratings.

REFERENCES

Baranski, J.V., & Petrusic, W.M., (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology*: *Human Perception and Performance*, *24*, 929-945.

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

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344-366.

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

Peirce, C.S., & Jastrow, J. (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73-83.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

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

Vickers, D. (1985). Antagonistic influences on performance change in detection and discrimination tasks. In: *Cognition, Information Processing and Motivation*, Proceedings of XXIII International Congress of Psychology, vol. 3, F. d'Ydewalle (Ed.), New York: North-Holland, 79-115.

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

Vickers, D., & Pietsch, A. (in press). Decision making and memory: A critique of Juslin & Olsson's (1997) sampling model of sensory discrimination. *Psychological Review*.

Vickers, D., & Smith, P. (1985). Accumulator and random-walk models of psychophysical discrimination: a counter-evaluation. *Perception*, *14*, 471-497.

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

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