Operating characteristics and realism of certainty estimates [1]

Z. J. ULEHLA, K. B. LITTLE AND T. C. WEYL
UNIVERSITY OF DENVER

Nineteen Ss each estimated the probability that each of a series of conceptual discriminations would be correct. These certainty estimates were compared with the actual percentage correct at each level of certainty. Certainty estimates proved to be fairly accurate for most Ss, although consistent individual differences were found. Training sessions in certainty estimation produced no major improvement. Operating characteristics were consonant with signal detection theory.

In a number of judgment tasks Ss have shown considerable ability to rate accurately the probability that a judgment already made would prove to be correct (Little & Lintz, 1965; Adams & Adams, 1961). Such ratings are termed certainty estimates (Little, 1961). Expressed in verbal rating categories (e.g., "highly confident"), certainty estimates have been subjected to theoretical analysis, yielding operating characteristics (OCs) of a form specified by signal detection theory (SDT) (Green & Swets, 1966). When expressed as percentages (e.g., "80% certain"), certainty estimates have been subjected to analysis of their realism or accuracy, the basic measure being the discrepancy between the stated probability of being correct and the "true" probability of being correct. A "true" probability is estimated from the proportion correct of all decisions assigned that stated probability (Adams, 1957).

In the present research, percentage certainty estimates were subjected to both realism and OC analysis. In connection with realism analysis, accuracy of certainty estimation, improvement with training, transfer of this improvement to different response modes and individual differences in accuracy of certainty estimation were investigated.

Method

Certainty estimates, in percentage form, were appended to conceptual discrimination tasks involving the judgment of which source (of two) had furnished each of a large number of brief samples of English text. The stimuli were four-word passages randomly drawn from either a "male-oriented" magazine (M-mag) or from a "female-oriented" magazine (F-mag). Three forms of this conceptual discrimination task were used. A training series utilized a single stimulus (SS) condition; in each trial a single passage was presented and S judged from which source it had been taken. The transfer tasks included two alternative forced choice (2AFC) and four alternative forced choice (4AFC) conditions. On each trial of the 2AFC condition, S was presented with two passages, one drawn from each magazine, and S judged which magazine had yielded which passage. On each trial of the 4AFC condition, S was presented with one passage from one magazine combined in random order with three passages from the other magazine. The S was told which magazine had yielded only one passage as opposed to three; his task was to select this passage from the total set of four. In all of these conditions, S coupled a certainty estimate to each discrimination judgment. Estimates were permitted to range from the chance probability of being correct (50% in the SS and 2AFC conditions, 25% in the 4AFC condition) to 100% in multiples of 10% (the 25% estimate was also permitted in the 4AFC condition). The criterion for realism of certainty estimates—e.g., that about 60% of those judgments to which S assigns a certainty estimate of 60% should prove correct, etc.—was explained to S at the beginning of the experiment.

Each of the 19 Ss participated in five training sessions of 200 SS trials. Upon completion of each session, S scored his own protocol for correct responses and computed the actual percentage correct associated with each certainty estimate (the actual percentage correct associated with the 60% certainty estimate was the number of correct judgments receiving 60% certainty estimates divided by the total number of judgments receiving 60% certainty estimates). On a graph S plotted his actual percentage correct against certainty estimate, and was urged by E to correct any gross discrepancies between the two.

The measure of S's certainty estimation performance used for the realism analyses was that proposed by Adams (1957), i.e., the weighted mean of the discrepancies between each certainty estimation and the corresponding actual percentage correct, the weight for each certainty estimation category being proportional to the square root of the number of judgments assigned that certainty estimate. A progressive decrease in the Adams discrepancy measure over sessions would indicate a positive effect of certainty training.

Possible transfer of certainty training effects to different response modes was investigated by presenting S with a 200 trial 2AFC session and a 200 trial 4AFC session both before and after the training session. During these "transfer" sessions, no informative feedback was furnished S, and S was not asked to evaluate his performance.

The order of the experimental sessions, each on a different day, was as follows: 2AFC pretest; 4AFC pretest; five SS training sessions; 2AFC posttest; 4AFC posttest.

The theoretical analysis involved the cumulative response proportions obtained when the certainty estimates were combined with the stimulus judgments to yield a 12-category ordered rating scale. This scale ranged from 100% through 50% certainty that F-mag was presented and then from 50% through 100% certainty that M-mag was presented. According to SDT, the normal deviates (z transforms) of the cumulative hit rates to one stimulus source should bear a linear relation to the z transforms of the cumulative error rates to the other stimulus source. From this, two nonindependent predictions concerning the product moment correlation coefficients (r's) can be drawn: (a) The r between the z-transforms of the cumulative proportions will be higher than the r between the untransformed cumulative proportions, because curvilinearity should diminish r in the latter case. (b) The r between the z-transforms will be essentially unity. Traditional high threshold theory yields the opposite prediction that the untransformed cumulative proportions will be linearly related (see Green & Swets, 1966).

Results and Discussion

The realism analyses showed that Ss had a substantial ability to estimate the probability that a discrimination judgment would prove to be correct (see Table 1). The discrepancy measures averaged less than 10% in the SS training task, indicating that the actual percentage correct was generally within 10% of the certainty estimate. Certainty estimation appeared to be only slightly poorer in the 2AFC transfer task than in the SS training task. The discrepancy measure was greatest for the 4AFC task; however, this may simply reflect the greater range of the allowable certainty estimates (25% to 100% compared to 50% to 100% in the SS and 2AFC tasks) and thus a greater possible range for discrepancy scores.

Neither a consistent training effect nor a noteworthy transfer effect was demonstrated. The over-Ss mean of the discrepancy measures in the SS training condition decreased in training Sessions 1-3, then increased again, with considerable inter-S variability. The over-S means of the discrepancy measures in the 2AFC and 4AFC transfer conditions decreased from the pretraining sessions to the posttraining sessions, but the decrease was meager and inconsistent (see Table 1).

Table 1 **Mean Adams Discrepancy Score**

	2AFC	4AFC
Pre-test	10.52	13.91
Post-test	8.97	12.78

Trial Number	Single Stimulus
1	8.81
2	8.25
3	7.58
4	8.03
5	9.15

Table 2 **Frequency distribution of correlation coefficients**

r=	1.00	99	98	.97	.96 or less
z-transformed	14	3	0	1	1
Untransformed	0	0	1	1	17

Substantial individual differences in accuracy of certainty estimation are indicated by high correlations between discrepancy scores from session to following session (including the 2AFC and 4AFC pre- and posttests). These rs ranged from .62 to .90 with one drastic exception—an r of -.11 between the 2AFC and 4AFC pretests. From the second session on, S's certainty estimation accuracy, relative to the performance of his fellows, was fairly consistent from session to session.

For the theoretical analysis, the five 200 trial SS training conditions were pooled to form one 1000 trial condition. (Pooling was done to yield stable cumulative response proportions.) This pooling required the assumption that S retained the same location for his decision criteria from session to session, despite informative feedback. The assumption seemed reasonable in the light of the general consistency of the discrepancy scores.

All of the 19 Ss yielded higher r's between the z-transformed cumulative proportions than between the untransformed cumulative proportions, thus supporting prediction (a) above. Seventeen of the 19 Ss yielded z-transform r's of at least .99 thus supporting prediction (b) (see Table 2). Thus, ratings involving percentage certainty estimates conformed to both predictions drawn from SDT. These results suggest that the information conveyed to S by word passages varies in the continuous fashion implied by SDT (or at least in multiple steps), rather than in an all-or-nothing (threshold) fashion. In conjunction with the evidence provided by Ulehla, Canges, & Wackwitz (1967), the present results support the applicability of SDT to conceptual discrimination.

References

ADAMS, J. K. A confidence scale defined in terms of expected percentages. *Amer. J. Psychol.*, 1957, 70, 432-436.

ADAMS, J. K., & ADAMS, P. Realism of confidence judgments. *Psychol. Rev.*, 1961, 68, 33-45.

GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics.* New York: Wiley, 1966.

LITTLE, K. B. Confidence and reliability. *J. educ. psychol. Measmt.*, 1961, 21, 95-101.

LITTLE, K. B., & LINTZ, L. M. Information and certainty. *J. exp. Psychol.*, 1965, 70, 428-432.

ULEHLA, Z. J., CANGES, L. M., & WACKWITZ, F. Signal detection theory applied to conceptual discrimination. *Psychon. Sci.*, 1967, 8, 221-222.

Note

1. This research was supported in part by USPHS Grant MH-11152-01 from the National Institute of Mental Health and by Grant GB-3081 from the National Science Foundation.