Acta Psychologica 30 *Attention and Performance II (W. G. Koster, ed.)* 1969, 16–26

THE SPEED-ACCURACY OPERATING CHARACTERISTIC

RICHARD W. PEW

Human Performance Center, University of Michigan,
330 Packard Road, Ann Arbor, Mich., U.S.A.

ABSTRACT

An analysis of the relationship between speed and accuracy of performance under a wide variety of task conditions reveals a linear relationship between log odds in favor of a correct response and reaction time. This result is consistent with the conceptual logic of the statistical decision model of choice reaction time and suggests the definition of a speed-accuracy operating characteristic analagous to the receiver operating characteristic in signal detection.

1. INTRODUCTION

Empirical research on the relationship between speed and accuracy of performance in a reaction time setting has taken on a new importance in recent years for several reasons. At the theoretical level, the research on models of choice reaction time (CRT) has found a congenial approach in the statistical decision or random-walk model. (STONE, 1960; FITTS, 1966; EDWARDS, 1965). One of the most robust predictions of this kind of model is the existence of an orderly trade-off between speed and accuracy of performance. At the empirical level CRT experiments have always been plagued with problems of intercomparison because of the difficulty of maintaining constant error rates or of comparing CRTs when they were variable. Finally, at the applied level, the assessment of performance efficiency implies understanding the relative contribution of speed and accuracy to overall efficiency. One frequently would like to ask, 'How much time is an error worth?'

R. G. Swensson of our laboratory is currently preparing a thorough review of the discrimination and CRT literature in which inferences concerning speed vs. accuracy are definable. In this paper I would like to present some previously unanalyzed data on the topic that was collected by Professor Fitts shortly before his untimely death in 1965 and to relate them to data from other experiments in which speed versus accuracy was an independent variable. Finally, I will include a preview of some data collected by Mr. Swensson as a part of his doctoral thesis under the direction of Dr. Ward Edwards that shed further light on the properties of the speed-accuracy trade-off.

GARRETT (1922) in considering the speed-accuracy trade-off in discrimination tasks suggested that, 'Judgment or perception [should] grow in accuracy with the increase in time taken to make it' (p. 1). Phrasing it differently he added, 'The factors involved would gradually coalesce, and in consequence there would be an increase in confidence with which the judgment would be made' (p. 1). Although he had no formal models in mind, this phraseology is certainly consistent with the statistical decision class of models. It is from this perspective, that of the relation between confidence of decision and speed of response, that I would like to consider the data on speed and accuracy of performance.

If one takes Bayes theorem as a model for the accumulation of evidence in a decision making context then one often finds that linear increments in the evidence in favor of a given hypothesis produce logarithmic increments in posterior odds, or relative confidence in one's decision. If the statistical decision model is relevant for studies of reaction time, then it seems a likely hypothesis that RT should be correlated with relative confidence on a logarithmic scale. Since the data seem relatively orderly in these terms, I have chosen to plot log odds vs. RT where odds refers to the ratio of correct responses to errors (more formally the ratio of probability of correct to probability of incorrect responding). Data for which a linear regression seems appropriate in these coordinates implies that logarithmic increases in confidence are related to linear increases in response time. In the two-alternative case, log odds is merely a linear transformation of d'^2, the measure that TAYLOR et al. (1967) showed to produce linear fits to the data of SCHOUTEN and BEKKER (1967) on forced reaction time.

2. THE DATA

Fig. 1a shows data from two experiments involving drastically different experimental conditions. The first is the data of SCHOUTEN and BEKKER (1967) using a method they refer to as forced reaction time in which they employ an auditory cue to constrain subjects to respond within a given interval of time. This plot shows the ratio of correct to error responses in a two-choice task conditional upon the obtained reaction times. With this technique they were able to obtain error rates ranging from chance performance, corresponding to odds of 1 to 1, up to 3% error, corresponding to odds of 35 to 1. The data have been truncated over the region in which a trade-off was observed.

Also shown in fig. 1a are the data from PACHELLA and PEW (1968),

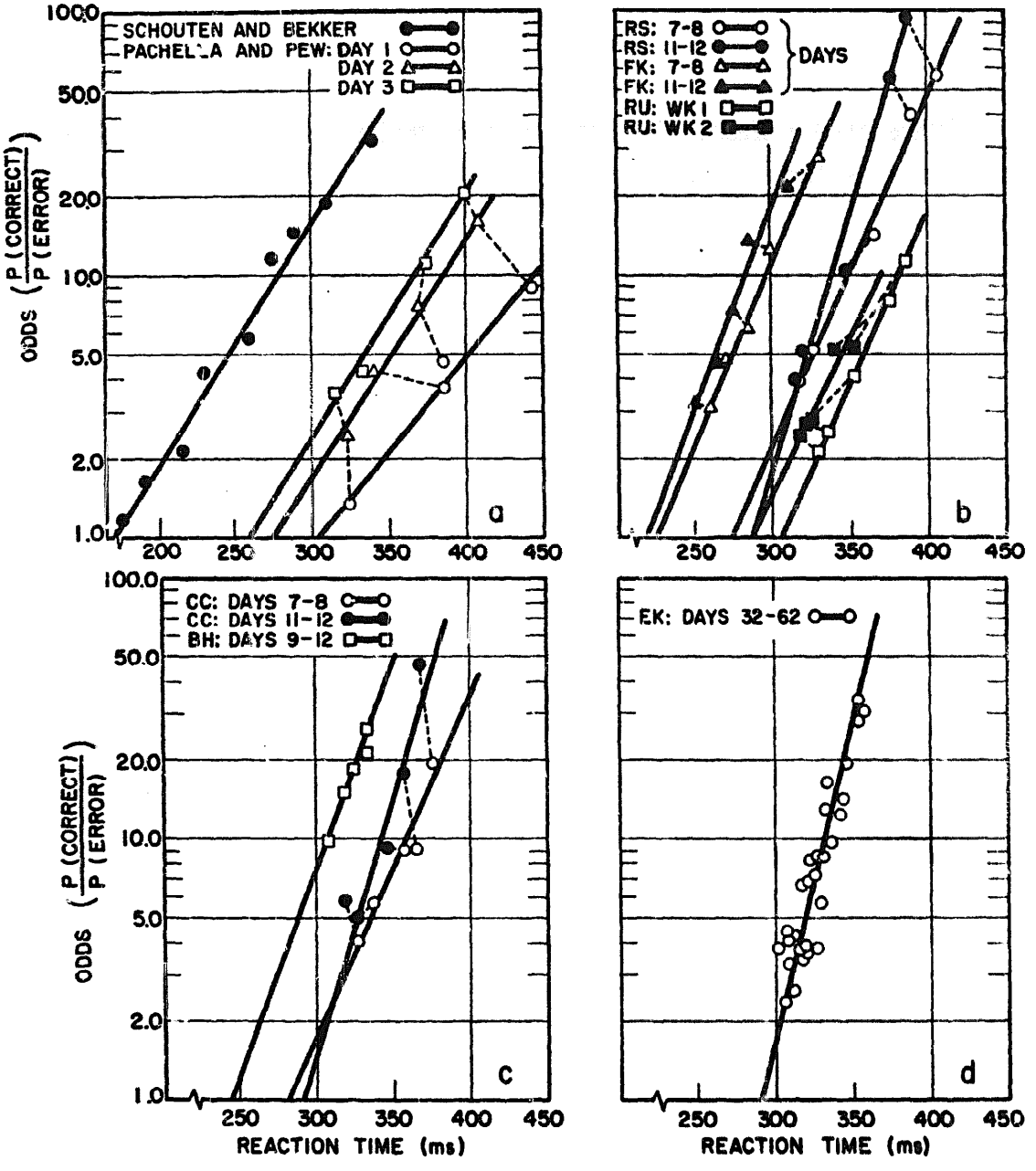

Fig. 1. The speed-accuracy operating characteristic for the data from 8 experiments. (a) Data from SCHOUTEN and BEKKER (1967) and from PACHELLA and PEW (1968). (b) Data collected by Fitts. (c) Further data from Fitts. (d) Data collected by Swensson.

The task consisted of responding to four lights with four fingers, the middle and index finger of each hand. The four lights could occur in any of the fifteen possible combinations and the appropriate chord response was required. In their experiment speed and accuracy of performance were manipulated in two ways. Using the procedure developed

by FITTS (1966), a discrete criterion time was defined that arbitrarily dichotomized responses into fast or slow. This criterion time was used in conjunction with a payoff schedule for fast vs. slow and correct vs. incorrect responding in order to manipulate the emphasis on speed vs. accuracy in the following ways: (1) by changing the relative cost for accuracy vs. speed directly and (2) by manipulating the criterion time defining fast vs. slow. Two payoff matrices having differential emphasis on speed and accuracy and two criterion times were employed in a factorial design. The data for these four groups of four subjects each are shown in fig. 1a for each of three days of practice, one hour per day.

TABLE 1

Regression analysis of speed-accuracy trade-off data.

	No. of alt. choices	No. of data pts.	Correlation	Slope of regression line
Schouten and Bekker:	2	9	0.99	20.6
Pachella and Pew:				
Day 1	15	4	0.98	16.0
Day 2	15	4	0.99	20.6
Day 3	15	4	0.99	20.8
Fitts' data:				
RU Wk. 1	15	5	0.99	30.1
RU Wk. 2	15	5	0.96	25.6
FK days 7-8	15	5	0.99	31.1
FK days 11-12	15	5	0.98	33.3
BH days 9-12	15	5	0.97	34.3
RS days 7-8	4	5	0.99	30.8
RS days 11-12	4	5	0.98	43.6
CC days 7-8	4	5	0.96	27.7
CC days 11-12	4	5	0.93	41.3
Swensson's data:				
EK days 32-62	2	29	0.94	49.1

These curves appear to be roughly linear and there is an orderly shift in intercept with practice accompanied by a small shift in slope. (The correlations and the slope of the best fit regression lines are given in table 1.) Note that this linearity is maintained even though the various groups achieve improved efficiency in different ways. For example: The low (fast) criterion time, speed emphasis group (bottom set of points)

improved almost entirely by increasing their accuracy with practice while the low criterion, accuracy group (second from bottom set) improved almost entirely by responding faster.

Surprisingly, the curves shown for the two-alternative case of Schouten and Bekker are not inconsistent with those for this fifteen-alternative case. The difference appears to be accounted for almost entirely by a shift in intercept.

Professor Fitts was quite interested in the speed-accuracy trade-off. His last published paper (FITTS, 1966) represented his statement of the 'random walk' statistical decision model and also demonstrated the usefulness of explicit payoff matrices for reducing the individual differences that result from the usual ambiguous instructions to respond as rapidly and as accurately as you can. At the time he was also carrying out a series of experiments, almost case studies, that examined the speed-accuracy trade-off within individuals. I believe his purpose in conducting these studies was largely methodological. He was first attempting to determine whether manipulation of the relative values and costs for slow and correct versus fast and wrong responses would produce rapid adjustments in speed and accuracy within individual subjects. He was further attempting to find a set of payoff matrices that would produce a suitably wide range of error rates to make parametric studies possible. This series began with the study of subject RU. He worked with the fifteen-alternative four-key reaction time task described above using the five sets of payoff matrix values shown in the first line of table 2. Each value in the table represents the number of points (later converted to money) the subject won or lost if his response fell within a given cell of the matrix. For example, when subject RU, working with the S-1 speed emphasis condition, produced a response that was slower than the criterion time but correct, he lost eight points. The criterion time was adjusted from day to day so that approximately 50 % of his responses were faster than the critical time. His performance averaged over the first week and over the second week of practice is shown in fig. 1b. The dashed lines connect corresponding payoff matrices during each week. This subject produced a narrower range of error rates during the second week than he did in the first week.

Subject FK worked with another set of payoff matrices that were designed to elicit a wider range of accuracies. His performance in the 15-alternative task for days 7 and 8 for days 11 and 12 are also shown in

TABLE 2

Payoff matrices used in the Fitts studies of speed-accuracy trade-off.

	Payoff condition	Accuracy emphasis				Speed emphasis	
		A-3	A-2	A-1	N	S-1	S-2
RU	Fast-correct		+16	+16	+16	+16	+16
	Fast-wrong		—12	— 8	— 6	— 4	— 1
	Slow-correct		— 1	— 4	— 6	— 8	—12
	Slow-wrong		—16	—16	—16	—16	—16
FK	Fast-correct		+16	+16	+16	+16	+16
	Fast-wrong		—20	—12	— 6	0	+ 8
	Slow-correct		+ 8	0	— 6	—12	—20
	Slow-wrong		—24	—24	—24	—24	—24
BH	Fast-correct	+16	+16	+16		+16	+16
	Fast-wrong	—16	—12	— 8		— 4	0
	Slow-correct	+ 4	0	— 4		— 8	—12
	Slow-wrong	—20	—20	—20		—20	—20
RS and CC	Fast-correct		+ 3	+ 3	+ 3	+ 3	+ 3
	Fast-wrong		— 3	— 2	— 1	0	+ 1
	Slow-correct		— 1	0	— 1	— 2	— 3
	Slow-wrong		— 4	— 4	— 4	— 4	— 4

fig. 1b. Subject BH worked on the same 15-alternative task under still another set of payoff matrices as given in table 2. Although for clarity his data are shown separately in fig. 1c, he is responding both more rapidly and more accurately than RU but produced a relatively small range of values of accuracy. Both BH and FK were run with criterion times adjusted to produce approximately 85 to 90 % fast responses. Finally, *Ss* RS and CC performed for 12 days with a four-alternative reaction time task, using the same four fingers but singly instead of in all combinations. Their payoff matrices are also shown in table 2. Because of the similarity of their performance, their data, pooled for days 7 and 8 and for days 11 and 12, are split between fig. 1b (RS) and fig. 1c (CC), and it may be seen that both produced a substantially wider range of error rates. We don't know whether this change in odds range is attributable to the scaled down payoff matrices used with these particular subjects or simply to individual differences among subjects.

One final set of data is provided by R. G. Swensson from an as-yet unpublished experiment. In an attempt to test the EDWARDS (1966)

version of the statistical decision model Mr. Swensson used a two-alternative reaction time task with uncertain warning intervals. He introduced an error penalty and charged a unit cost for each increment in reaction time. When the stimuli were simply highly discriminable diagonal lines, and a variety of weights for speed and accuracy were tested, six out of seven subjects exhibited no evidence of a speed-accuracy trade-off. Their performance was either at chance level with a fast reaction time ($\simeq$200 msec) or somewhat slower ($\simeq$250 msec) and at a very high level of accuracy. They shifted from one strategy to the other as a function of the relative costs and payoffs, but never produced intermediate error rates.

In a follow-up study in which the stimuli were tilted rectangles having a length to width ratio that made them very difficult to discriminate, the same two strategies were observed: either the subject preprogrammed his responses, had chance accuracy and fast RTs, or he took longer and processed the stimulus information. However, in the latter case with suitable payoff manipulation is was possible to observe intermediate error rates in 3 out of 5 subjects and the linear trade between speed and accuracy shown for one subject in fig. 1d was obtained. These data are taken after a minimum of thirty days of practice and even though discriminability was low, this practice probably accounts for the steep slope of this curve. Note that while the slope is steeper than for the Schouten and Bekker data, the response time for corresponding error rates is substantially longer, especially at low levels of accuracy.

Another interesting feature of Swensson's low discriminability data is that the intercept with chance performance (odds of 1.0) occurs at a substantially higher RT than is represented by the preprogramming strategy (for the subject shown the intercept is 80 msec longer than the average obtained RT with chance performance). The implication is that there is a residual decision time that is involved in the information seeking strategy over and above the incremental increases in RT associated with improved accuracy. It is as if a constant decision time is added to the overall latency of response when *S* chooses to make use of the stimulus information. Although this extra time may be taken in visual information processing or response selection activities, it is in addition to the residual delays associated with response execution or stimulus detection since these activities are required even with the preprogramming strategy.

3. Discussion

The observation that over a wide variety of tasks and a wide variety of techniques for manipulating the relative emphasis on speed or accuracy produces curves best described as linear on a plot of log odds versus reaction time provides reassurance that the negative correlation between speed and accuracy is indeed a performance limitation of individual subjects and not a task specific effect.

That this generally linear trend holds regardless of the manner in which speed or accuracy is manipulated argues strongly for the basic concepts embodied in the statistical decision model. The curves shown may be thought of as the average path of the random walk process, a kind of Speed-Accuracy Operating Characteristic (S-A OC) that represents the average rate at which evidence accumulates.

Schouten and Bekker (1967) argue: (1) that speed-accuracy trade-off is a function of what *S* actually does and not what he is trying to do; (2) that such a finding is inconsistent with fixed boundaries in a statistical decision model; and (3) that a perceptual focussing model is an appropriate representation of the obtained relationships. The results presented here tend to support the notion that the trade-off depends on what *S* actually does, but these studies, employing a range of explicit payoffs, do suggest that *S* has a substantial degree of control over the level of speed or accuracy at which he chooses to operate.

Fitts' (1966) data argued for fixed criterion boundaries although he referred to the possibility of adjusting those boundaries continuously on the basis of error feedback. Schouten and Bekker's own data, and the finding in many of the cases described here and elsewhere that the average RT for error responses is faster than that for correct responses favors the diffuse rather than strict control of the boundary. However, this certainty does not refute the basic statistical decision concepts.

Finally, it is true that the perceptual focussing hypothesis is tenable, but it incorporates no mechanism for defining the effect of explicit payoff structures on performance. These data show that error constraints and time constraints are both effective in influencing *S*'s operating point along the Speed-Accuracy Operating Characteristic.

Although the data discussed here are relatively unsystematic and drawn from several different sources, it may be interesting to speculate about the variables influencing the two parameters: slope and intercept of the (S-A OC). The slope is interpretable as the time required for a unit increase in confidence. The intercept is more difficult to interpret

since the lower bound on odds is set by chance performance, and chance performance is at a different level of odds depending on the number of alternative stimuli. Nevertheless, an operating characteristic lying everywhere to the northwest of another one must represent better performance in terms of some decrease in response time for all points along the operating characteristic.

Of those independent variables that may be examined here, practice appears to have the largest effect on the slope and intercept. In all but one case practice increases the slope and moves the intercept to the left. Both these results seem intuitively appropriate. In statistical decision terms the rate of accumulation of evidence increases and in addition the residual processing time is reduced as learning progresses.

Although not shown here, Swensson has obtained evidence suggesting that the slope of the operating characteristic may be increased by decreasing the difficulty of the required visual discrimination (increased length to width ratio of two tilted rectangles). Since the statistical decision model postulates the need for discrimination of the proper alternative even when the stimuli are visually distinct, Swensson's result is also a confirmation of the basic model.

The slope of the Schouten and Bekker data (2 stimuli) is essentially the same as the Pachella and Pew data (15 stimuli). The data for FK (15 stimuli) have a slope after 11-12 days of practice very similar to that of the RS data and CC data (4 stimuli). This consistency of slope independent of the number of alternative stimuli may very well be a fortuitous result of the confounding of practice and individual differences, but taking it at face value it is curious. A frequently suggested way to generalize the statistical decision model from the two-alternative case has been to think of a test of the most likely alternative against all others. But this results suggests that these comparisons may be made in parallel since the rate of build up in relative confidence appears to be independent of the number of possible stimuli that must be tested. Clearly an experiment is needed that examines the role of these variables in a factorial design.

Finally, individual differences among subjects appear to have their primary effect in the intercept of the operating characteristic which also absorbs most of the variance in task characteristics, as might be expected. Individual differences appear to have an even larger effect than the number of stimuli in these experiments.

Since none of these results is in conflict with the statistical decision

model logic, although some fall outside its domain, it seems safe to conclude that one may argue about the nature of and accuracy of the decision criterion imposed by *S*, but the concept of *S* making a sequential, statistical decision seems well supported in the context of speed-accuracy adjustment.

From a methodological point of view these data suggest that if one wishes to compare information processing efficiency across a set of tasks, at the very least it is appropriate to attempt to fix the error rate or to fix the response time and examine the error rate. A better procedure would be to regard the trade between speed and accuracy to be a parameter and compare the slopes and the intercepts of the resulting Speed-Accuracy Operating Characteristics.

Returning to the practical level, it appears that in tasks which must be performed under time pressure, it is not meaningful to ask the question, 'What is the cost in time of making an error?' because the cost depends on the level of accuracy at which one is operating at the moment. An error saves more time at high levels of accuracy than at low levels. Rather one should ask, 'How long does it take to increase the confidence of one's judgment by a factor of 10 or 100?'

Acknowledgement

I wish to acknowledge the assistance of Mrs. Linda K. Fensch. This research was supported in part by the National Aeronautics and Space Administration under Contract NASr 54(06) and in part by the Advanced Research Projects Agency of the Department of Defense under Contract AF 49(638)-1235 between the U.S. Air Force and the Human Performance Center.

References

Edwards, W., 1965. Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. J. math. Psychol. **2**, 312–329.

Fitts, P. M., 1966. Cognitive aspects of information processing: III. Set for speed versus accuracy. J. exp. Psychol. **71**, 849—857.

Garrett, H. E., 1922. A study of the relation of accuracy to speed. Arch. Psychol, N. Y., No. 56.

Pachella, R. C. and R. W. Pew, 1968. Speed-accuracy tradeoff in reaction time: effect of discrete criterion times. J. exp. Psychol. **76**, 19—24.

SCHOUTEN, J. F. and J. A. M. BEKKER, 1967. Reaction time and accuracy. In: Attention and performance, A. F. Sanders (ed.), Acta Psychol. **27**, 143—153.

STONE, M., 1960. Models for choice-reaction time. Psychometrika, **25**, 251—260.

TAYLOR, M. M., P. H. LINDSAY and S. M. FORBES, 1967. Quantification of shared capacity processing in auditory and visual discrimination. In: Attention and performance, A. F. Sanders (ed.), Acta Psychol. **27**, 233—229.

DISCUSSION

Schouten: Have you found any evidence of a shifting criterion during the serial experiments?

Pew: In these experiments we did not make any sequential analysis. In his tasks, Swensson indeed found a shift in strategy within a block of trials.

Rabbitt: In a two-choice task two classes of errors may be distinguished: (1) a class of errors which occurs because the subject repeats his response that he should not; (2) a class of errors when he makes a new response when he should in fact repeat it. The difference between the mean RTs of correct responses and of errors is very much greater in the second case (new response) than in the first case (repetition).

Pew: If the decomposition of error distributions in these two ways is made, it might be found that the kind of functional relationship discussed here does apply to non-repeats but does not apply to the case of repeats.

Rabbitt: A repeat might very well be an impulsive guess, whereas a new response is a failure in accumulating evidence.

Sanders: Would you expect considerable changes in the slope of the curve when the stimulus-response compatibility is varied?

Pew: In a study of Fitts' in which different tasks were used (2 alternative cases and 4 alternative cases) one of the variables was compatibility. It was found that the slope was steeper for the more compatible cases.