Final Technical Report

Decision Paths in Complex Tasks

Reference Number NAGW-860

Eugene Galanter
Columbia University

Preamble

This document serves as the final report for NASA Innovative Research Program (IRP) entitled "Decision Paths in Complex Tasks," NASA reference number NAGW-860. Work under this grant has resulted in five research reports, one of which was an MA thesis in the department of psychology, Columbia University, submitted by Gloria Mark. Three reports have been published and one is in press [starred (*) in the references of this document]. This report summarizes data from our major unpublished study.

Abstract

Complex real world action and its prediction and control has escaped analysis by the classical methods of psychological research. The reason is that psychologists have no procedures to parse complex tasks into their constituents. Where such a division can be made, based say on expert judgment, there is no natural scale to measure the positive or negative values of the components. Even if we could assign numbers to task parts, we lack rules i.e., a theory, to combine them into a total task representation.

We compare here two plausible theories for the amalgamation of the value of task components. Both of these theories require a numerical representation of motivation, for motiveation is the primary variable that guides choice and action in well-learned tasks. We address this problem of motivational quantification and performance prediction by developing psychophysical scales of the desireability or aversiveness of task components based on utility scaling methods (Galanter 1990). We modify methods used originally to scale sensory magnitudes (Stevens and Galanter 1957), and that have been applied recently to the measurement of task "workload" by Gopher and Braune (1984). Our modification uses utility comparison scaling techniques which avoid the unnecessary assumptions made by Gopher and Braune (page 526). Formulæ for the utility of complex tasks based on the theoretical models are used to predict decision and choice of alternate paths to the same goal.

Introduction

Human choice and decision making has been studied from the point of view of outcome value or utility (Tversky and Kahneman, 1981), task difficulty or workload (Wickens, et al. 1983), personal qualities (Weinstein, 1972), social and managerial constraints (Helmreich 1984), and a host of literary and other scholarly disciplines. The results of the scientific part of the effort have led to very little theoretical insight or practical consequence, and no applicability at all to the real world actions that constitute chained task sequences, tasks in which each component has its own utilities and disutilities (see the short review in Gopher and Braune 1984 p. 520).

Consider an example: A pilot may wish to deviate from his planned flight path to

avoid a weather cell or frontal line. The goal is to arrive safely at the flight destination. Component goals are to minimize discomfort to himself, and to the passengers and crew, as well as to minimize flight time and its associated costs. Different paths and procedures may be available to accomplish each of these goals to varying degrees. These separate paths comprise different sequences of actions. Each of these component acts may be of greater or lesser utility. Which is chosen?

This multipath problem is not unique, but rather is the paradigm of most human decision making and choice. The route that is selected will be influenced by factors intrinsic to the different tasks, the different goals, and various aspects of the pilot's general knowledge and experience. The psychological problem is to devise a general method that lets us predict the choice and the course of action.

The prime difficulty is that we have no procedure that can parse a task into its constituents. Where we make such a division, based say on expert judgment, we have no natural scale to measure the positive or negative values of the components. Even if we could assign such numbers to task parts, we have no rules for combining them into a total task cost or benefit as distinct from the costs and benefits of the overall outcome. These limitations are further compounded by the strong interactions between the parts of a task and its outcome. Such difficulties leave us unable to predict human performance in complex environments.

Method

This experiment examines the relation between utility judgments of sub-task paths and the utility of the task as a whole. This is a convergent validation procedure (von Winterfeldt and Edwards, 1986). It is based on the assumption that measurements of the same quantity done with different methods should covary. In other studies convergent validation procedures showed high correlations (von Winterfeldt and Edwards, 1986, Hart and Bortolussi, 1984, Ogden et. al., 1979. In event related brain potential (Isreal et. al., 1980; Kramer and Wickens, 1983), these procedures also showed promise. Subjective rating techniques such as category scales (Hart et. al., 1981), magnitude estimation (Borg, 1978), and Cooper/Harper subjective ratings (Wierwille and Connor, 1983) show this validity. Finally, ratio scaling methods suggest combinatorial models with special constraints.

A significant relation between sub–task and whole task utility can have practical consequences. The experimental decomposition of a complex task into measurable components could find optimal task paths. High utility sub-task paths could be identified. Low utility paths could be discounted. A model that combines sub-task ratings also provides information on how each contributes to the variance of the total task utility. The utility measures of our sub-tasks were obtained during an "aircraft flight controller" task. The task was divided by the experimenter into two discrete sub-tasks. On successive trials, subjects use three different alternatives to reach the the first sub-task goal. The second sub-task also exposed them to three different alternatives to reach that goal. Thus, there were nine possible combinations of paths all of which lead to the task goal. During each sub-task, the subject rated the utility of each path relative to a numerical modulus. The experimenter then asked the subject to rate the utility of the combined choices relative to reaching the criterion—the task goal. The results let us decide among various models of sub-task utility combination, and indirectly, whether judgmental models need to include the equivalent of "cognitive" noise.

Preliminary models

Based on concepts drawn from psychophysical scaling experiments, (Stevens and Galanter, 1957), a power function model of the relation of sub-task utilities to total task utility is conjectured. This model in simplest form is:

$$\log U_t >= \sum_n [w_i \log (u_i)]$$

where

U_t = the utility of the strategy used to complete the task;
u_i = the utility of the strategy used to complete sub–task i;
n = the number of sub–tasks that the task is decomposed into;
w_i = the weight assigned to sub–task i in the combination rule.

A second conjecture is an additive model using untransformed data:

$$U_t >= \sum_n [w_i u_i]$$

Both models make three assumptions. (1) The subject is interested in maximizing some criterion. (2) There is path independence, i.e., the choice made to reach the goal of sub-task X do not affect choices for other sub-tasks. This assumption is tested by the level of interaction between path choice utility ratings. (3) The combination rule should be invariant with respect to the path chosen.

Magnitude estimation methods normally require that averages be struck across a sample of subjects. Asking a subject to assess repeatedly the magnitude of the same stimulus leads to the simple repetition of his judgment. To circumvent this tendency, and to permit magnitude estimates from a single subject, we use a judgmental technique called the "shifty modulus" (Galanter, 1987).

Procedure

The experiment is a simulation of an air traffic controller's task. In this task the controller must first choose a display method for the air traffic, and then choose a procedure for conflict resolution. The overall goal is to maintain safe traffic separation for the aircraft. The specific goal in stage 1 is to choose a method to display altitude information. Stage 2 simulates some features of the decisions air traffic controller's make. These include scanning for potential collision and then taking remedial action. The task then is to choose a method to change the flight path of a target to prevent collision. The subject is told that at least one potential collision will occur on each trial.

Subjects

Five paid subjects, all students at Columbia University, participated in the experiment. Four of the subjects were male. The student ages ranged from 19 to 26. They all had vision correctable to normal.

Apparatus

The experiment was run on a Commodore Amiga microcomputer. Subjects were seated in a well lighted laboratory at a console containing a keyboard, a pointing device (mouse), and a color CRT. The subject was free to adopt a comfortable position facing the screen within reach of the keyboard. The subjects generally chose a position that placed their eyes slightly above and 45 cm distant from the CRT. For their control responses, subjects used the mouse and the keyboard.

Stimuli

The screen design was modelled after the radar displays used by air traffic controllers in 1986, in particular on the TRACON air traffic control facility in Westbury, New York. This is a terminal radar approach installation that monitors aircraft outside a five mile radius from each of four major New York airports: Kennedy, LaGuardia, Islip, and Newark. Radar at that facility is monochrome and vector drawn, but also contains digital alphanumerics associated with the radar returns from aircraft transponders.

Aircraft: Depending on the trial, 4, 8, or 16 white dots (approximately 8 mm in diameter) representing airplanes, move across the CRT display in piecewise linear paths. Alongside each dot is a smaller directional dot, approximately 2 mm in diameter, which provides information on the direction of the plane's flight vector. Tracking along with each aircraft is a two letter identification code, such as "CO." At the beginning of each trial the planes start from different positions in the display. The rate of change varies across planes from one pixel per frame to 12 pixels per frame. Each plane moves at a constant rate. The planes blink off about every six seconds and reappear about one second later in an updated position, paralleling the timing, but not the decaying appearance of a radar scope. The background color of the screen is dark grey.

Altitude information: Three choices are available to the subject for the display of altitude information: alphanumeric, voice, and digital meters. In the alphanumeric mode, altitude information appears beside the identification code of each aircraft. This altitude information is in the form of a one, two or three digit number which represents hundreds of feet. Voice interrogation is done by clicking the mouse over the plane in question. A synthetic computer voice responds with the plane identification code and a three digit altitude reading. The third altitude method, digital meters, displays columns of plane identifiers of varying hight. The altitude displays are also updated every six seconds.

Changing flight vector: Subjects could change the course of one of the planes to avoid a collision by one of three methods: altitude change, continuous lateral direction change, or limited (12°), lateral direction change. Altitude change increased or decreased the plane's altitude by one thousand feet. Continuous lateral change changed the plane's direction, left or right, by 6° per frame. 12° lateral change changed the plane's direction, left or right, by 12° only once.

Experimental Design.

The nine choice combinations were assigned to each flight scenario in a Latin Square design. Twelve flight scenarios were randomized within each cell. A trial consists of one flight scenario. The 12 flight scenarios were all different and consisted of four scenarios each of 4 planes, 8 planes, and 16 planes. The data from this experiment consisted of nine cells (108 trials per subject). The experiment yielded a data matrix as

follows:

		Stage 1 choice		
		1	2	3
	1	12	12	12
Stage 2 choice	2	12	12	12
	3	12	12	12

Each cell contains observations from the 12 different flight scenarios. Each observation consists of the utility estimates reported for subtask 1, subtask 2, and the overall task.

[Descriptions of the practice session, the method of utility estimation, the modulus formats and the procedures and data collection may be obtained from the author.]

Results

After demonstrating that modulus invariance holds, we converted the utility estimates into relative utility estimates for the sake of easy comparison across trials with differing modulus values. This simple conversion consists of dividing the reported utility value by the modulus utility value. A preliminary analysis showed that two of the subjects, C and D, used the ratio method. Subjects B and E appear to have used a category type judgment in their reports, indicating that they did not understand the verbal and written instructions provided at the beginning of the experiment. Subject A seemed unable to master the task.

To determine which rating system a subject used we assumethat the noise, or "scatter," in the reports is symmetrically distributed about the mean under the appropriate transformation. Because ratio and category judgments are both modulus comparisons, the sources of noise in both judgmental modes are presumed similar. This technique becomes clear when the data are viewed graphically. Figure 1 shows the idealized result of the inferred category or ratio judgments transformed either linearly or logarithmically.

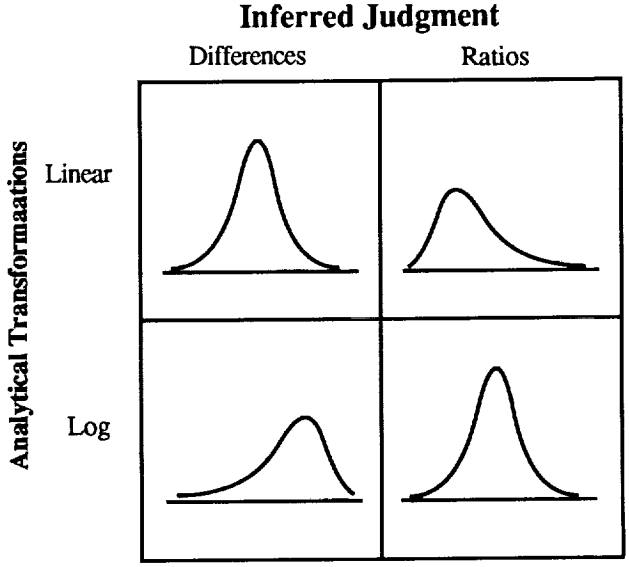

Figure 1

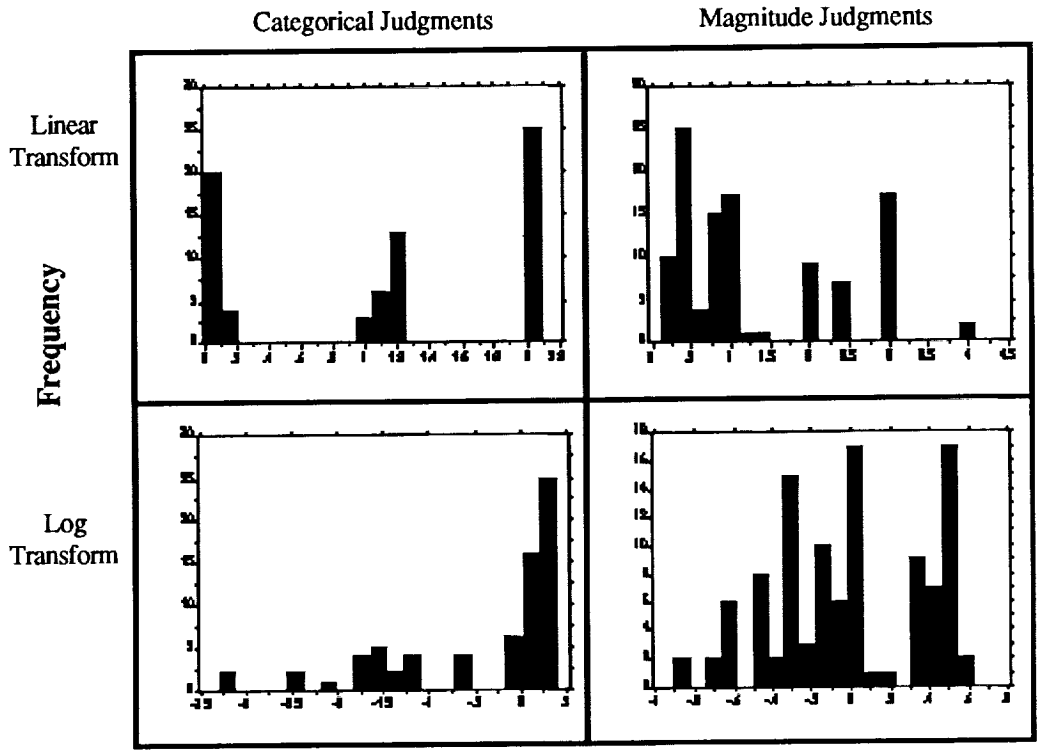

Figure 2

In Figure 2, data from two subjects who arguably used different judgmental modes are displayed. These graphical representations support the assumptions outlined above. Although the data from subjects A, B, and E must be ignored when determining the relationship between the sub-task utilities and the overall utility, their data can still be used to test the assumptions of path and modulus independence.

The mean utility estimates for subjects are shown in table I. Subjects C, D, and E usually gave utility estimates that were significantly different between the paths in sub-task 1 and 2. Subject A did not differentiate between the choices based on the ratings shown. Subjects were generally consistent in rating voice interrogation with the lowest utility and alphanumerics as having the highest utility. In sub-task 2, subjects were generally consistent in rating 12° lateral change with the lowest utility and altitude change with the highest utility.

Table I

(Entries Represent modulus ratio)

Subject	Sub-Task 1			Sub-Task 2			Total Task
	A-N	V-R	D-M	A-C	12°	LC	
A	2.72	2.28	2.42	2.46	2.20	2.29	3.02
B	1.26	0.72	0.74	1.30	1.19	1.29	1.52
C	1.44	0.90	1.30	1.36	0.97	1.20	1.21
D	2.58	0.54	0.79	3.32	1.23	2.03	1.87
E	2.00	0.05	1.16	1.82	0.54	1.11	1.18

Sub-Task 1: A-N=Alphanumerics; V-R=Voice Report; D-M=Digital Meters
Sub-Task 2: A-C=Altitude Change; 12° = 12° Lateral Change; LC=Continuous Lat Chg

A multiple regression analysis was performed on each subject's data. The dependent variable is the overall utility estimate of the combined path choice for both sub-tasks, and the independent variables are the sub-task 1 utility estimate and the sub-task 2 utility estimate. The model that was tested is of the form:

$$\log (Y_i) = a + b_1 \log (X_{1i}) + b_2 \log (X_{2i}) + e_i$$

The results of the regression analysis are shown in Table II. The asterisks refer to whether the beta coefficients of the intercept and sub-task utilities are significant.

Table II

		Beta Estimates			
Subject	N	Intercept	Sub-Task 1	Sub-Task 2	p—Value	
A	95	0.2281	0.2829	0.7483	0.0001	0.7971
B	92	0.4118*	0.2790*	0.0324	0.0197	0.0845
C	105	0.1164*	0.0906*	0.3406*	0.0001	0.3343
D	99	0.1174*	0.5176*	0.9215*	0.0001	0.8211
E	59	0.1675*	0.1707	0.5379*	0.0001	0.7751

*Significant at $p < 0.05$ Mean $R^2 =$ 0.562

Four out of the five subjects gave subtask utility estimates that were predictive of the overall utility according to the model specified in [3]. Subject B's utility estimate for the paths used in sub-task 2 was not significant. The percentage of variance explained in the total task judgment, R^2, ranged from values of .08 to .82. Models fit for three subjects, A, D, and E explained more than 77% of the variance of the holistic task judgment. The residuals plotted against the predicted values for each subject show the residuals to be evenly distributed around zero. They also indicate that the fit of the log log model is appropriate.

Next, a linear additive model was tested to compare the fit with the fit obtained in the log transformed model. The level of significance of beta parameters for the stage utility estimates are shown in Table III.

Table III

		Beta Estimates				
Subject	N	Intercept	Sub-Task 1	Sub-Task 2	p—Value	R^2
A	95	0.0841	0.2096*	1.0516*	0.0001	0.7770
B	92	0.6953*	0.4399*	0.3343*	0.0015	0.1368
C	105	0.5868*	0.1665*	0.3559*	0.0001	0.2872
D	99	0.5286*	0.6868*	0.7156*	0.0001	0.8013
E	59	0.0248	0.5146*	0.5483*	0.0001	0.8456

*Significant at $p < 0.05$ Mean $R^2 =$ 0.5695

Subjects A, D, and E show that the model explains more than 77% of the variance of the total task judgment. All beta estimates for the slopes are significant for both variables. In order to explain certain anomalies in these data it will be useful to have the range of utility estimates available, as contained in Table IV.

Table IV

Subject	Subtask 1			Subtask 2			Whole Task		
	Min	Max	Range	Min	Max	Range	Min	Max	Range
A	0.75	6.00	5.25	0.38	4.50	4.19	0.67	6.50	5.83
B	0.50	1.70	1.20	0.02	1.93	1.50	0.75	2.25	1.50
C	0.03	2.50	2.47	0.40	2.50	2.10	0.71	2.50	1.79
D	0.13	4.00	3.88	0.50	4.00	3.50	0.25	4.44	4.19
E	.001	2.00	2.00	0.10	2.00	1.90	0.20	2.40	2.20
All Subjects	0.001	6.00		0.02	4.50		0.20	6.50	

[Details of the analysis including statistical tests and measures of multicollinearity, and further discussion of these results may be obtained from the author unpon request.]

References

Borg, Gunnar. Subjective aspects of physical and mental load. *Ergonomics*, 1978, 21 (3), 215220.

Galanter, E. The direct measurement of utility and subjective probability. *American Journal of Psychology,* 1962, 75, 208-220.

Galanter, E., & Holman, G. L. Some invariances of the isosensitivity function and their implications for the utility function of money. *Journal of Experimental Psychology,* 1967, 73, 333-339.

Galanter, E., & Pliner, P. Cross-modality matching of money against other continua. In H. Moskowitz, B. Sharf, & J. C. Stevens (Eds.), *Sensation and measurement: Papers in honor of S. S. Stevens*.Dordrecht, Netherlands: Reidel, 1974.

Galanter, E., Popper, R., & Perera, T., Annoyance scales for simulated VTOL and CTOL overflights., *Journal of the Acoustical Society of America,* 1977, 62, S8A.

Galanter, E. Timing of motor programs and temporal patterns. (in Timing and time perception). *Proceedings of the NY Academy of Science,* May 1983.

*Galanter, E. The shifty modulus: Psychophysical scales for individual subjects, Psychophysics Laboratory, NY:Psychophysics Laboratory Report 87/3, 1987.

*Galanter, E. Utility functions for non-monetary events. *American Journal of Psychology,* 1990, **103**, 4, 449-470.

*Galanter, E. and Wiegand, T. E., Multiple moduli and payoff functions in psychophysical scaling. *Ratio Scaling of Psychological Magnitude,* (Ed. S. J. Bolanowski & G. A. Gescheider), Hillsdale NJ:LAE Associates, 1991.

*Galanter, E. Modulus estimation quantification of single events. *American Journal of Psychology,* (in press)

Gopher D. & Braune, R., On the psychophysics of workload: Why bother with subjective measures? *Human Factors,* 1984, 26, 519-532.

Hochberg, J. & Galanter, E. Behavioral indicators of pilot workload., *Proceedings of Second Symposium on Aviation Psychology,* (Ed. Jensen, R. S.) Columbus Ohio, April 1983.

Isreal, Jack B., Wickens, Christopher D, Chesney, Gregory L., Donchin, Emanuel. The eventrelated brain potential as an index of displaymonitoring workload. *Human Factors*, 1980, 22 (2), 211224.

Kornbrot, D. E., Donnelly, M. & Galanter, E. Estimates of utility function parameters from signal detection experiments. *Journal of Experimental Psychology: Human Perception and Performance,* 1981, 7, 441-458.

Stevens, S. S., & Galanter, E. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology,* 1957, 54, 377-411.

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *Science,* 1981, 211, 453-458.

von Winterfeldt, Detlof, and Edwards, Ward. *Decision Analysis and Behavioral Research.* Cambridge: Cambridge University Press, 1986.

Weinstein, A. G., Predicting behavior from attitudes. *Public Opinion Quart.,* 1972, 36, 355-360.

Wierwille, Walter W; Connor, Sidney A. Evaluation of 20 workload measures using a psychomotor task in a movingbase aircraft simulator. *Human Factors*, 1983, 25 (1), 116.

Wickens, E. D., Sandry, D., & Vidulich, M. Compatibility and resource competition between modalities of input, output, and central processing: Testing a model of complex task performance. *Human Factors,* 1983, 25, 227-248.