Development of a Taxonomy of Human Performance:

The Task Characteristics Approach to Performance Prediction

AD736191

Alfred J. Farina

George R. Wheaton

Technical Report 7

FEBRUARY 1971

D D C

FEB 4 1972

AMERICAN INSTITUTES FOR RESEARCH

WASHINGTON OFFICE

Address 8555 Sixteenth Street Silver Spring, Maryland 20910

Telephone (301) 587 8201

R71-6

AMERICAN INSTITUTES FOR RESEARCH

WASHINGTON, D.C.

EDWIN A. FLEISHMAN, PhD, DIRECTOR

Albert S. Glickman, Ph.D., Deputy Director

INSTITUTE FOR COMMUNICATION RESEARCH

Arthur L. Korotkin, PhD, Director

Research on instructional, communication, and information systems and their effectiveness in meeting individual and social needs.

INSTITUTE FOR RESEARCH ON ORGANIZATIONAL BEHAVIOR

Clifford P. Hahn, MS, Director

Research on human resources, selection and training, management and organization, safety, and administration of justice.

INSTITUTE FOR RESEARCH IN PSYCHOBIOLOGY

Thomas I. Myers, PhD, Director (Acting)

Studies concerned with those behavioral and physiological mechanisms which determine performance and with those factors in the environment which affect these mechanisms and thereby affect performance.

ACCESSION for
NTIS WHITE SECTION
DDC BUFF SECTION
UNANNOUNCED
JUSTIFICATION
A

INTERNATIONAL RESEARCH INSTITUTE

Paul Spector, PhD, Director

Research on the development of human resources in developing countries; problems of working effectively abroad; evaluation of action programs in underdeveloped countries; role of attitudes and values in social change and economic development.

TRANSNATIONAL FAMILY RESEARCH INSTITUTE

Henry P. David, PhD, Director

Research on behavioral components of family interaction; coordination of transnational population research; and strengthening of family planning programs.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

American Institutes for Research
135 North Bellefield Avenue
Pittsburgh, Pennsylvania 15213

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: THE TASK CHARACTERISTICS APPROACH TO PERFORMANCE PREDICTION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Interim Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Alfred J. Farina and George R. Wheaton

6. REPORT DATE

February 1971

7a. TOTAL NO. OF PAGES

120

7b. NO. OF REFS

28

8a. CONTRACT OR GRANT NO.

F44620-67-C-0116 (AFOSR)
DAHC19-71-C-0004 (Army-BESRL)

b. PROJECT NO.

c. ARPA Order Number 1032
d. ARPA Order Number 1623

9a. ORIGINATOR'S REPORT NUMBER(S)

AIR-726/2035-2/71-TR7

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

BESRL Research Study 71- 7

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Advanced Research Projects Agency
Department of Defense, Washington, D.C.

13. ABSTRACT

The development and evaluation of systems for describing and classifying tasks which can improve generalization of research results about human performance is essential for organizing, communicating, and implementing these research findings. The present report describes research undertaken to develop one such system--a task characteristics approach. Basic objectives were to develop descriptive characteristics of tasks; assess the reliability of rating scales devised to measure these characteristics; and determine if these characteristics represented correlates of performance.

Major components of a task were identified and treated as categories within which to devise task characteristics or descriptors. Each characteristic was cast into a rating scale format which presented a definition of the characteristic and a seven-point scale with defined anchor- and mid-points along with examples for each point. Nineteen scales were developed and evaluated in a series of 3 reliability studies. In general, it was found that a subset of scales having adequate reliability consistently emerged in all 3.

"Post-diction" was the paradigm used to determine whether the task characteristics were correlates of performance on which predictive relationships might be established. Performance measures were abstracted from studies already existing in the literature. Two post-diction studies were conducted--the first involved 6 scales and 26 tasks, the second involved 6 scales and 20 tasks--with encouraging results.

Significant multiple correlations of .82 and .73 were obtained between task characteristic ratings and the performance measures. It appears possible to describe tasks in terms of a task-characteristics language that is relatively free of the subjective and indirect descriptors found in many other systems. Task characteristics may represent important correlates of performance; as shown here, it was possible to describe subtle differences among tasks and relate them systematically to performance variations.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

UNCLASSIFIED
Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Task Taxonomy						
Human Performance						
Task Characteristics						
Multiple Regression						
Model of Performance						

UNCLASSIFIED
Security Classification

AIR-726/2035-2/71-TR7

AD

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE:

THE TASK CHARACTERISTICS APPROACH TO PERFORMANCE PREDICTION

Alfred J. Farina
George R. Wheaton

TECHNICAL REPORT 7

Prepared under Contract for
Advanced Research Projects Agency
Department of Defense
ARPA Orders No. 1032 and 1623

Principal Investigator: Edwin A. Fleishman

Contract Nos. F44620-67-C-0116 (AFOSR)
DAHC19-71-C-0004 (ARMY-BESRL)

American Institutes for Research
Washington Office

February 1971

Approved for public release; distribution unlimited.

PREFACE

The AIR Taxonomy Project was initiated as a basic research effort in September 1967, under a contract with the Advanced Research Projects Agency, in response to long-range and pervasive problems in a variety of research and applied areas. The effort to develop ways of describing and classifying tasks which would improve predictions about factors affecting human performance in such tasks represents one of the few attempts to find ways to bridge the gap between research on human performance and the applications of this research to the real world of personnel and human factors decisions.

The present report is one of a series which resulted from work undertaken during the first three years of project activity. In 1970, monitorship of the project was transferred from the Air Force Office of Scientific Research (AFOSR) to the U. S. Army Behavior and Systems Research Laboratory (BESRL), under a new contract. This report, completed under the new contract, is among several describing the previous developmental work. It is also being distributed separately as a BESRL Research Study.

EDWIN A. FLEISHMAN
Senior Vice President and
Director, Washington Office
American Institutes for Research

FOREWORD

The American Institutes for Research is engaged in a research program to develop and evaluate new systems for describing and classifying tasks which can improve generalization of research results about human performance and to develop a common language for researcher-decision maker communication that would help organize human performance information for maximum use in training, equipment design, and personnel selection.

The objective of this program is to develop theoretically-based language systems (taxonomies) which--when merged with appropriate sets of decision logic and appropriate sets of quantitative data--can be used to make improved predictions about human performance. Such taxonomies should be useful, for example, when future management information and decision systems are designed for Army use.

During previous project years, three different taxonomic systems were developed, each of which seemed to have maximum relevance for a different type of application: the ability-requirement approach; the task characteristics approach; and a third approach based on information theory.

The present publication reports on the development and preliminary assessment of the task characteristics approach to the prediction of human performance. The approach seeks to describe tasks in terms of a task-oriented language which, when combined with multiple regression techniques, can be used to predict task performance.

J. E. UHLANER, Director
U. S. Army Behavior and Systems
Research Laboratory

ACKNOWLEDGMENT

Conduct of a study of this type necessarily involved the efforts of many people in addition to the authors. We wish to acknowledge the overall support and guidance of the Principal Investigator, Dr. Edwin A. Fleishman. Dr. Fleishman was particularly helpful in providing full access to the data of several of his past studies which permitted both the reliability and post-diction efforts to be accomplished most efficiently.

The authors are particularly indebted to Dr. William J. Baker, formerly of AIR, who developed much of the rationale underlying the predictive model used in this study. Our thanks are also extended to Susan Emery for her many contributions to the post-diction efforts, and to Norma Lee for her able assistance during the reliability studies.

Alfred J. Farina
George R. Wheaton

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: THE TASK CHARACTERISTICS APPROACH TO PERFORMANCE PREDICTION

BRIEF

Requirement:

Of the many conditions which can influence human performance, the most poorly described and least understood are those embodied in the task. As a consequence, the ability to relate performance observed in one task to that observed in other tasks is limited. The present research describes a series of studies conducted to develop an instrument in terms of which the stimulus, procedural, and response characteristics of tasks could be described. It discusses additional studies which were designed to determine whether dimensions comprising the descriptive language represented correlates of human performance.

Procedure:

The basic steps in this research were to: (a) develop descriptive characteristics of tasks; (b) assess the reliability of rating scales devised to measure these characteristics; and (c) determine if these characteristics represented correlates of performance.

The overall direction taken by the project was influenced by a heuristic model which viewed performance as a function of three sets of antecedant conditions: the operator, the environment, and the task. A decision was made to focus initial efforts on the task component of the model, holding the other components in abeyance.

Toward this end, major components of a task were identified and treated as categories within which to devise task characteristics or descriptors. Each characteristic was cast into a rating scale format which presented a definition of the characteristic and provided a seven-point scale with defined anchor - and mid-points along with examples for each point. Nineteen scales were developed and evaluated in a series of three reliability studies.

The paradigm used to determine whether the task characteristics were correlates of performance upon which predictive relationships

might be established was that of "post-diction". Post-diction referred to the situation in which performance measures were abstracted from studies already existing in the literature. Subjects rated descriptions of the tasks used in these studies on task characteristic scales and then these ratings were subjected to multiple regression analysis to establish the extent to which they were related to the performance in question. Two such post-diction studies were conducted. The first post-diction study involved six scales and 26 tasks while the second study involved six scales and 20 tasks.

Findings:

In general, it was found that a subset of scales having adequate reliability consistently emerged in all three reliability studies. The results of the two post-diction studies were encouraging in that significant multiple correlations of .82 and .73 were obtained between task characteristic ratings and the performance measures.

Utilization of Findings:

Although a final interpretation of these findings must await cross-validation efforts, it does appear possible to describe tasks in terms of a task-characteristic language which is relatively free of the subjective and indirect descriptors found in many other systems. Further, task characteristics may represent important correlates of performance; as shown here, it was possible to describe subtle differences among tasks and to relate such differences systematically to variations in performance.

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: THE TASK CHARACTERISTICS APPROACH TO PERFORMANCE PREDICTION

CONTENTS

	Page
INTRODUCTION	1
BACKGROUND	4
Heuristic Model of Performance	5
Nature and Use of the Task Descriptive System	6
Classification	7
Prediction	7
Objectives	9
SCALE DEVELOPMENT	10
Task Definition	10
Task Characteristics	10
RELIABILITY STUDIES	15
First Reliability Study	15
Second Reliability Study	17
Third Reliability Study	18
Discussion	21
POST-DICTION STUDIES	24
First Post-Diction Study	25
Second Post-Diction Study	29
Discussion	32
CONCLUSIONS AND RECOMMENDATIONS	34
REFERENCES	39
DD Form 1473 Document Control Data R&D	119
APPENDICES	43
FIGURES	
Figure 1. Relationship among the terms "task," "components," and "characteristics"	12

Page

TABLES

Table 1. Sample task characteristic rating scale 15

2. Reliability estimates for three judges using original scales to rate 37 tasks 16

3. Reliability estimates for twenty-eight judges using revised scales to rate 15 tasks 19

4. Reliability estimates for two judges using eighteen scales to rate 21 tasks 20

5. Listing of the most reliable scale within each of the three reliability studies 23

6. Basic data for the first regression analysis 27

7. Intercorrelation matrix for the first regression analysis 28

8. Basic data for the second regression analysis 30

9. Intercorrelation matrix for the second regression analysis 31

10. Comparison of post-diction studies 1 and 2 32

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: THE TASK CHARACTERISTICS APPROACH TO PERFORMANCE PREDICTION

INTRODUCTION

A major problem confronting the behavioral sciences and technologies is the lack of a structure within which to describe, interpret, and organize information about human performance. Without such a structure limits are placed on the extent to which findings from different studies can be compared, contrasted, and integrated into a systematic body of knowledge. At the root of this problem is the absence of unifying dimensions for systematically describing those antecedant conditions of which performance is a function.

Of the many conditions which can influence performance, the most poorly described and the least understood are those embodied in the task. As a consequence, the ability to relate performance observed in one task to that observed in other tasks is limited. At present, research results obtained with one task can be safely generalized only to other tasks which are so highly similar as to be almost identical. The ability to communicate research findings unambiguously is similarly hampered. Behavioral scientists, and those who must apply research findings to operational problems, are without a language for interrelating performance on different tasks.

A burgeoning research literature and a growing demand for application of findings both underscore the need for an integrative structure. A system is needed which will yield better predictions of the effects of independent variables on task performance. Similarly, a system is needed to predict more accurately the learning rates or proficiency levels associated with new tasks. These needs have been recognized by many investigators (e.g., Fleishman, 1962, 1967; Hackman, 1968; Melton & Briggs, 1960; and Miller, 1962). Fitts (1962) in particular, has called for a taxonomy which should identify important correlates of learning rate, performance level, and individual differences, and be equally applicable to laboratory tasks and to tasks encountered in industry and in military service.

The key to establishing such a taxonomy lies in developing a well-defined task descriptive language. Earlier reports under this project (e.g., Farina, 1969; Wheaton, 1968) as well as other reviewers (e.g., Ginsberg, McCullers, Merryman, Thomson, & Whitte, 1965) suggest that three general approaches are most prevalent. They differ primarily in terms of the manner in which description is accomplished.

In the first approach, description centers on the specific activities in which an operator engages while performing a task. Interest lies in specifying what the operator actually does. Those who have taken this approach (e.g., Fine, 1963; McCormick, 1968; and Reed, 1967) are more concerned with describing performance per se and less concerned with the conditions giving rise to that performance. In the second approach, description focuses on those resources of the operator which are required for performance on the task. Gagne (1962) and Miller (1966), for example, describe tasks in terms of those functions or processes which the operator is required to utilize. In a similar vein, tasks have been described in terms of the types and amounts of human abilities upon which the tasks make demands (e.g., Fleishman, 1967; Theologus, Romashko, & Fleishman, 1970). In this second general approach, emphasis is on critical aspects of the individual intervening between features of the task and consequent performance.

A third approach to developing a task descriptive language treats the task as a critical sub-set of the antecedant conditions of which performance is a function. Hackman (1968) states this position clearly:

> "...That is, if we are interested in the effects of tasks and task characteristics on behavior, it is essential that we develop a means of describing and classifying our independent variables (tasks) other than in terms of the dependent variables (behaviors) to which we ultimately wish to predict."

Investigators taking this tack (e.g., Cotterman, 1959; Fitts, 1962; Folley, 1964; and Stolurow, 1964) attempt description in terms of the characteristics of the task confronting the operator.

It is this latter approach to developing a task descriptive language which would seem appropriate for the type of taxonomy called for by Fitts. In order to eventually predict the performance which will result when a subject is exposed to a given situation, one must be able to specify and fully describe those independent variables which are in effect. Part of this specification must necessarily include that stimulus complex known as the "task" which confronts the subject. It is within this complex that many correlates of learning rate or proficiency level will be found. Knowledge of these variables would provide a basis for comparing performance on different tasks. They would also provide a basis for classifying tasks with respect to the behavioral consequences of other classes of independent variables.

The present report describes a series of studies conducted to develop an instrument in terms of which the stimulus, procedural, and response characteristics of tasks could be described. It discusses additional studies which were designed to determine whether dimensions comprising the descriptive language represented correlates of human performance.

BACKGROUND

The research described in the present report was part of a larger programmatic effort concerned with development of a taxonomy of human performance (Fleishman, 1967; Fleishman, Kinkade, & Chambers, 1968; Fleishman, Teichner, & Stephenson, 1970; Fleishman & Stephenson, 1970). In support of this general program of research, several alternative task descriptive systems were developed. The general purpose of each of these systems was to provide a basis for classifying tasks in order to permit better organization and increased generalization of performance data within and between task categories.

Studies described in the present report were concerned with the development and initial use of one such system. Known as the task characteristics approach, it attempted to provide for the description of tasks in terms of a variety of task-intrinsic properties including goals, stimuli, procedures, response modes, etc. The decision to describe tasks in these rather morphological terms, instead of using more behavioral-, process- or ability-oriented descriptors, stemmed from the conviction that tasks, in their own right, represented a potent class of independent variables. Accordingly, if the variables comprising a task were manipulated singly or in combination (e.g., creating a number of different tasks), the resultant effects on performance could be mapped systematically. Knowledge of how performance varied, as a result of manipulating the characteristics of tasks, would provide a basis for estimating performance on other tasks whose characteristics could be described.

The consequences of the foregoing rationale for development and use of a task descriptive system were explored by constructing an heuristic model of performance. In turn, this model helped specify what was to be described, how description was to be accomplished, and how the task descriptive indices were to be related to performance.

Heuristic Model of Performance

During early stages of the project an heuristic model of performance was entertained. The model, known as POET, simply stated that any obtained performance score (P) was necessarily the function of at least three major classes of independent variables. These included the particular task (T) on which performance was measured, the specific operator (O) whose performance was monitored, and the environmental conditions (E) under which performance took place. Included in the latter class were all variables (e.g., ambient noise, drug dosages, conditions of practice, etc.) which were extrinsic to either the task or the operator and primarily impinged on the latter.

The POET model, therefore, suggested that the difference in performance which might be observed when comparing two experiments could be due to variations within any one or all three of the major classes of independent variables. Observed differences in performance could arise from the use of different samples of operators, or from different tasks, or from the application of different treatments (extrinsic variables). Consequently, it seemed obvious that any system which was developed to permit increased generalization of performance data would have to take all three classes of variables into consideration. This in turn meant that descriptive systems would eventually be required for each of the major components within the model.

Instead of attacking the problem at this general level, however, the decision was made to develop descriptive systems sequentially. The issue, therefore, was to decide upon which descriptive system to place initial emphasis. There appeared to be a variety of ways in which to describe different operators based on such variables as age, intelligence, abilities, interests, etc. Indeed, many studies have been conducted in which individual differences on these and similar "personal" variables were systematically related to variations in performance. By the same token there seemed to be farily adequate description and specification of what were termed the "environmental" variables. In most cases

descriptive systems dealing with this component have been sufficient to permit investigation of the effects of different levels of treatment upon performance of a large number and variety of variables.

While description of the operator and of the environment seemed adequate, description of the task component was not. Most of the available descriptive systems were inadequate because they failed to emphasize the task as an antecedant condition of performance, a condition which could be subjected to systematic and specifiable manipulation. Such systems prevented one from readily talking about type or, more significantly, level of treatment in the sense that he could for the operator and environment components. Yet the ability to make such statements seemed essential if one were to investigate the effects of variations in tasks on subsequent performance. Therefore, while recognizing the importance of descriptive systems for all three components, the decision was made to focus initial efforts on a task descriptive system. As explained in a later section of this report, description was based on a variety of task characteristics.

Nature and Use of the Task Descriptive System

During early stages of the project consideration was also given to the manner in which the descriptive data provided by the system were to be used in organizing tasks and consequent performance data. This issue was of importance for it was felt that specification of the intended use(s) of the descriptive data would culminate in a set of requirements for the language itself. Two major uses were identified: classification and prediction. Task characteristics data would provide a basis for classifying tasks in terms of their observed similarities and dissimilarities. The descriptive data could also be utilized within a multiple regression context to relate variations in the characteristics of tasks to variations in performance.

Classification - Although several alternative approaches to the classification of tasks were considered (Wheaton, 1968), it seemed desirable to approach classification on quantitative rather than on qualitative grounds. One technique available for this purpose was the similarity coefficient described by Cattell and Coulter (1966). This coefficient was designed to describe the similarity between pairs of profiles in terms of a distance function. Therefore, if descriptive profiles could be generated for tasks, it would be possible to mathematically express the similarity among them in terms of a matrix of similarity coefficients. These data could then be analyzed by cluster analytical techniques to define clusters or classes of highly similar tasks (Silverman, 1967). Although this type of analysis was not of primary concern in the present research, it did emphasize the need for a descriptive system which treated tasks in terms of quantitative profiles.

Prediction - Another use to which descriptive data could be put was in predicting learning rates or proficiency levels on tasks for which performance data were not already available. Emphasis in this approach was not on classifying tasks but rather on identifying those characteristics of tasks which were correlates of performance. It was this latter approach which was pursued in the present study.

A multiple-regression model was developed in which task characteristic descriptors were treated as predictor variables. The model was based on the premise that descriptive terms could be selected which represented correlates of performance and, as such, could be used to predict average learning rates or proficiency levels on different tasks. The rationale underlying the regression approach was as follows. Suppose a single group of operators performed two different tasks yielding the same type of performance measures. If individuals' scores were averaged on each task and if these two means differed, then, since identical subjects are involved, the difference between means could only be attributed to differences between the tasks themselves (assuming "environmental" variables to be identical in both situations). The difference between tasks would be specified in terms of task descriptors.

If the concept of differences between tasks and consequent differences between performance means were extended to a larger set of tasks, performed by the same operators under the same conditions, then a variable ($\bar{P}_m$) would be created. A given value on this variable would represent the mean performance score associated with a particular task (m) within the set of tasks. It was hypothesized, therefore, that specific values for this variable could be predicted in terms of task characteristic scale values. The multiple regression equation required for that purpose would have the following form:

$$\bar{P}'_m = a_o + a_1 X_{m_1} + a_2 X_{m_2} + \ldots + a_n X_{m_n}$$

where

$\bar{P}'_m$ = predicted mean performance score on task "m"

a_n = regression weight for the nth task descriptor, and

X_{m_n} = the value for task "m" on task descriptor "n".

To accomplish these ends, however, it was necessary to impose a major restriction on the model. The tasks under investigation at any one time had to share a common response measure (e.g., reaction time, time on target, percent correct, etc.). This restriction had profound consequences for it implied that different regression equations would be required to handle different types of performance measures. Such would not have been the case had it been possible to describe different measures of performance in terms of a single common metric. The absence of this universal metric, however, made it necessary to categorize tasks in terms of the measures employed to describe performance on them. The categories of performance described by Teichner and Olson (1969) were considered for this purpose. Separate regressions were anticipated for tasks yielding such diverse performance measures as probability of detection, reaction time, percent correct, and percent time on target.

The consequences of the regression model for the descriptive system were readily determined. The system had to contain multiple dimensions, each of which could be applied to any selected task. The dimensions had to be quantitative in nature and had to possess a reasonably high reliability. Finally, if the model were to aid in predicting parameters of performance, the descriptive dimensions had to represent correlates of performance.

Objectives

Based upon these background considerations, the present research attempted to accomplish the following objectives. A series of generically applicable quantitative rating scales was to be developed for description of various task characteristics. The reliability with which these scales could be used to describe tasks was to be determined. Finally, the feasibility of using the descriptive data as predictors of mean levels of performance on different tasks needed to be determined. The remainder of this report describes the activities conducted in pursuit of these objectives.

SCALE DEVELOPMENT

Task Definition

The development of task characteristics received initial guidance from a definition of the term "task" which was devised early in the project. Given that interest lay in predicting performance, a task was defined as a potential means of eliciting performance. More specifically, it referred to a complex situation capable of eliciting goal-directed performance from an operator. Given this orientation, a task was conceived of as having several components with each component possessing certain salient characteristics. These components were: an explicit goal, procedures, input stimuli, responses, and stimulus-response relationships.

An explicit goal was a specification of the "state" or "condition" to be achieved by the operator. By "explicit" was meant that the goal was indicated to at least the operator and one independent observer, and that some objective procedure existed whereby the observer could verify whether or not the goal had been achieved. A task also had to include a statement of the "means" by which the goal was to be attained. The "means" consisted of procedures which were statements specifying the types of stimulus-response relationships to be formed, and their sequencing. Then, too, the task had to contain a set of relevant input stimuli attended to by the operator. Finally, the statement of the task had to describe a set of responses contributing to goal attainment.

Task Characteristics

Given the arbitrary requirement that a task possess these components, it followed that if a potential "task" did not possess all of these components, then by definition it was not a task under the present system, and if an operator failed to perform in accordance with the specified procedures, the question of goal attainment for that task could not be raised. The operator, by definition, would not have performed the task

in question; in fact, he would have performed a different task. This latter point led to a direct consideration of what it was that served to make tasks different. That is, given that all tasks had the above components, what distinctions could be made within these common components? What were, for example, characteristics of a task goal which, when measured in some fashion, would serve to differentiate among various task goals?

In order to differentiate among tasks, therefore, the components of a task were treated as categories within which to devise task characteristics or descriptors. As previously mentioned, additional requirements were set forth regarding these characteristics. Each had to be applicable to most, if not all, types of tasks so as to avoid the problem of not being able to rate or measure all tasks on a comparable set of dimensions. Each characteristic had to be expressed quantitatively, being scaled in at least an ordinal fashion. Each had to possess an acceptable degree of reliability. Finally, to achieve economy of use, it was desirable that the characteristics require a minimum of training time and application time on the part of the user.

Figure 1 clarifies the relationship among the terms "task", "task components", and "characteristics". Each characteristic was cast into a rating scale format which presented a definition of the characteristic, and provided a seven-point scale with defined anchor- and mid-points along with examples for each point (Smith & Kendall, 1963). A sample rating scale is shown in Table 1. The complete set of 19 scales originally developed is shown in Appendix 1.

The original set of scales has undergone changes due to refinement, additions, and deletions. Consequently, the appendix section contains three separate sets of task characteristic scales, each having been used in a separate reliability study*. This evolutionary process is still

* Three sets of task characteristic scales rather than one final set are presented since there is no "final" set in the sense that a reader could rate a task on it and then apply appropriate Beta weights to gain an estimate of performance on that task. The research is still in its early stages where a demonstration of its feasibility is the issue being addressed. In addition, the results of the various reliability and postdiction studies require the inclusion of the specific scales and tasks used.

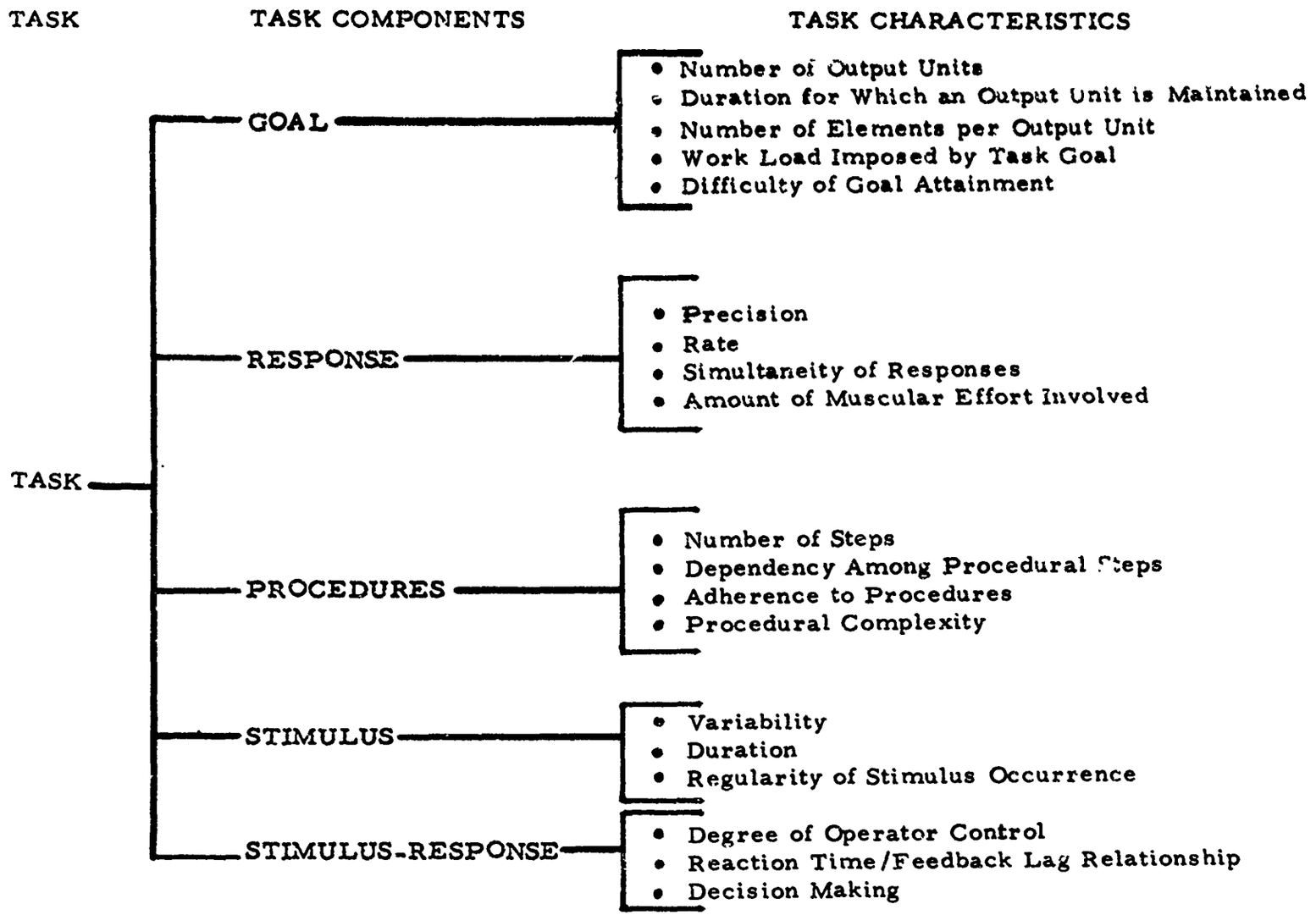


Figure 1. Relationship among the terms "task," "components," and "characteristics"

Table 1

SAMPLE TASK CHARACTERISTIC RATING SCALE

VARIABILITY OF STIMULUS LOCATION

Judge the degree to which the physical location of the stimulus or stimulus complex is predictable over task time.

Definitions		Examples
High predictability - stimulus location remains basically unchanged.	7	• Stimulus is a red light located on a display panel.
	6	
	5	
Medium predictability - location changes but in a known manner or pattern.	4	• Visually following an arrow in flight toward a target.
	3	
	2	
Low-predictability - location changes in an almost random fashion.	1	• Predicting which leaf will fall from a tree next.

not complete, but it has progressed far enough to provide a demonstration of the basic approach. During this developmental phase the task characteristics were viewed as critical independent variables which, if manipulated, would influence task performance. While an indirect test of this view was attempted in the "post-diction" studies to be discussed later, the ultimate test would entail actual manipulation of these characteristics within an experimental task and observation of concomitant changes in performance.

RELIABILITY STUDIES

First Reliability Study

Following development of the original set of rating scales a series of reliability studies was conducted. In the first such study the task characteristic rating scales were employed in their original form. Three research assistants were trained in the use of the scales and were then asked to rate 37 rather simple psychomotor tasks on each of 19 scales. The task descriptions with which the raters worked are referenced in Appendix 2.

The obtained ratings were cast into analyses of variance to determine intraclass correlation coefficients for each scale. Following the method described by Winer (1962, p. 124), two coefficients (r_k and r_1) were calculated. The r_k coefficient provided an estimate of the reliability of the mean of the three judges' (k = 3) ratings. The r_1 coefficient estimated the reliability of a single rating. The obtained coefficients, together with the variance components used in the calculation of r_k, are shown in Table 2.

Winer (1962, p. 128) suggested an interesting interpretation of the intraclass correlations. Each r_k coefficient was an estimate of the correlation which would be obtained were the mean ratings of the present three judges correlated with the mean ratings from another random sample of three judges rating the same tasks. Using an r_k equal to or greater than 0.70 as an arbitrary index of acceptable reliability, seven of the 19 original scales appeared to be adequate.

Three of the scales (7, 12, 18) shown in Table 2 possessed r_k's with negative values. Theoretically, r_k may range in value from zero (o) to plus one (+1). In practice, however, it can be demonstrated that r_k will assume a negative value in those cases where the mean-square within term is greater than the mean-square between term (e.g., $r_k = 1 - \frac{MSw}{MSb}$). Interpretation of such negative r_k coefficients is difficult.

Table 2

RELIABILITY ESTIMATES FOR THREE JUDGES USING ORIGINAL SCALES TO RATE 37 TASKS

Rating Scale	Between Task Variance	Within Task Variance	Reliabilities Average (r_k)	Reliabilities Individual (r_1)
1. # Output Units	15.097	2.342	0.84	0.64
2. Duration	7.882	0.775	0.90	0.75
3. # Elements	3.151	1.243	0.60	0.33
4. Work Load	9.293	2.757	0.70	0.44
5. Difficulty	2.060	1.234	0.40	0.18
6. Precision	3.048	1.081	0.64	0.37
7. Rate	6.245	7.874	-0.26	-0.07
8. Effort	1.009	0.441	0.56	0.30
9. Simultaneity	2.280	0.631	0.72	0.46
10. # Steps	1.674	0.495	0.70	0.44
11. Dependency	3.046	1.189	0.61	0.34
12. Adherence	0.721	1.649	-1.28	-0.23
13. Complexity	2.414	1.568	0.35	0.15
14. Variability	4.678	1.207	0.74	0.48
15. Stim. Dur.	2.646	0.378	0.85	0.66
16. Regularity	2.645	1.234	0.53	0.27
17. Control	2.247	1.748	0.22	0.08
18. Reaction	1.868	2.027	-0.08	-0.02
19. Decision	1.207	1.099	0.08	0.03

Inspection of the rating data showed that the three judges were actually in strong agreement on Scale # 12. However, the judges were not able to differentiate among tasks very effectively, as shown by the relatively small between-task variance component for this scale. Evaluating this scale on another and more heterogeneous sample of tasks would either raise its estimated reliability or confirm its insensitivity. Scales # 7 and # 18 had relatively large within-task variances suggesting a lack of consistency among judges. Inspection of the actual ratings confirmed this impression, particularly in the case of Scale # 7 where judges were in confusion about the end-points (1 or 7) of the scale.

Second Reliability Study

After the first study, many of the original scales were examined in an attempt to improve their reliabilities. Some scales were deleted and others underwent minor or major revision to clarify the exact nature of the dimension being rated and the meaning of the scale anchor points. The resulting instrument consisted of 16 scales (Appendix 3). In an attempt to estimate the reliability of the revised scales, 28 judges rated 20 tasks on each scale. The 28 judges were college students recruited from a local university. Prior to the actual study, the judges were thoroughly familiarized with the meaning of each scale and with the rating procedure. The judges were paid for their participation.

Reliability estimates were obtained for each of the 16 scales. These data were based on only 15 of the 20 tasks which were actually rated. The five tasks which were eliminated were cognitive, paper-and-pencil tasks. They were originally included to determine whether or not the judges could describe them reliably in terms of the task characteristics. The judges were largely unsuccessful in this effort. Consequently, it was decided to limit use of the scales, at least initially, to psychomotor tasks. Descriptions of the 15 tasks which were finally analyzed in terms of r_k and r_1 are shown in Appendix 4.

The reliability estimates are shown in Table 3 together with the relevant variance components. The striking feature of these data was the relatively low reliability for an individual rater (r_1). Were only one judge of the type employed in this study to assign ratings, he would be fairly reliable only on one scale (# 18). More reliable ratings could be obtained, however, were the mean ratings of either three or five judges utilized. Using the Spearman-Brown Prophecy Formula (Winer, 1962, p. 127) it can be shown that if $r_1 \geq .33$, then $r_3 \geq .60$ and $r_5 \geq .71$. On this basis, adequate reliability could be expected on at least seven scales. The remaining scales appeared to need additional revision.

Third Reliability Study

Finally, additional reliability data were obtained during an analysis of 21 tracking tasks (see Appendix 6) under a contract with the U. S. Naval Training Device Center. In this effort two judges evaluated the tasks in terms of many different measures, including 18 task characteristic scales. The 18 scales (Appendix 5) represented revised versions of many of the earlier scales. In this study both judges were highly familiar with the scales and the procedures for their use.

As shown in Table 4, the rating data from this study were evaluated in several ways. First, as in the preceding studies, analyses of variance were conducted which permitted calculation of the intraclass correlation coefficients (r_k and r_1). Second, similarity coefficients (r_p) were computed which expressed how similar the two judges were in evaluating the tasks on each scale. The technique was essentially one of profile analysis. The r_p statistic (Cattell & Coulter, 1966) could range in value from -1.0 to 1.0 being asymptotic with respect to -1.0. An r_p value of 1.0 meant that the two profiles fell on exactly the same point in multi-dimensional space. An r_p of -1.0 meant that the two profiles were maximally dissimilar. Finally, for each scale the number of times the two judges were within plus or minus one scale unit of each other was determined and expressed as a percentage of 21 cases.

Table 3

RELIABILITY ESTIMATES FOR TWENTY-EIGHT JUDGES USING REVISED SCALES TO RATE 15 TASKS

Rating Scale	Between Task Variance	Within Task Variance	Reliabilities Average (r_k)	Individual (r_1)
1. # Output Units	95.121	3.294	.97	.50
2. Duration	64.630	2.626	.96	.46
3. # Elements	18.507	1.578	.91	.28
4. Work Load	66.005	3.401	.95	.40
5. Precision	46.829	1.633	.97	.50
6. Rate	36.104	3.119	.91	.27
7. Effort	3.081	1.276	.58	.05
8. Simultaneity	54.983	1.278	.98	.60
9. # Steps	25.815	1.330	.95	.40
10. Dependency	56.520	3.943	.93	.32
11. Variability	70.555	3.110	.96	.44
12. Stim. Dur.	16.744	2.663	.84	.16
13. Regularity	28.520	3.463	.88	.21
14. Stim. Con.	48.331	4.271	.91	.27
15. Resp. Con.	29.936	5.441	.82	.14
16. Feedback	11.717	2.547	.78	.11

Table 4

RELIABILITY ESTIMATES FOR TWO JUDGES USING
EIGHTEEN SCALES TO RATE 21 TASKS

	Rating Scale	Between Task Variance	Within Task Variance	Reliabilities Average (r_k)	Reliabilities Individual (r_1)	Similarity Coefficient (r_p)	Percentage Agreement (% (± One Unit)
1.	# Units	17.095	0.000	1.00	1.00	1.000**	100%
2.	Duration	0.407	0.429	-0.05	-0.02	0.319*	90%
3.	# Elements	1.579	0.810	0.48	0.32	0.410**	76%
4.	Workload	3.481	2.095	0.40	0.25	0.126	62%
5.	Precision	1.595	1.333	0.16	0.09	0.436**	62%
6.	Rate	0.329	0.381	-0.16	-0.07	-0.063	95%
7.	T-dependency	0.574	1.000	-0.74	-0.27	-0.167	90%
8.	N-dependency	1.545	0.905	0.41	0.26	0.169	90%
9.	Resp. Con.	0.229	0.286	-0.25	-0.11	-0.048	100%
10.	Simultaneity	2.045	0.143	0.93	0.87	0.770**	100%
11.	# Responses	3.181	0.952	0.70	0.54	0.433**	86%
12.	# Steps	2.131	0.524	0.75	0.60	0.418**	86%
13.	Feedback	4.314	3.738	0.13	0.07	0.251	62%
14.	Effort	1.081	0.667	0.38	0.24	0.367*	81%
15.	Stim. Con.	0.217	0.238	-0.10	-0.05	0.046	100%
16.	Regularity	1.774	2.452	-0.38	-0.16	-0.180	71%
17.	S. Dura.	3.267	1.905	0.42	0.26	0.496**	19%
18.	Variability	4.457	4.405	0.01	0.01	0.312*	29%

*$p < .05$
**$p < .01$

Interpretation of the intraclass coefficients shown in Table 4 was again difficult. Four r_k coefficients were above 0.70 and appeared to represent reasonably reliable scales. In terms of the similarity coefficients (r_p) ten were significant, implying agreement between judges' profiles. Finally, on eight scales the judges were in agreement at least 90% of the time. Only three scales (# 4, # 13, and # 16) failed to exhibit either a high r_k ($r_k \geq .70$), a significant r_p, or a high percentage (90%) of agreement.

Discussion

Our experience in assessing the reliability of the task characteristic scales indicated that the statistical methods used often tended to preclude a definitive answer to the question of scale reliability. Interpretation of the intraclass correlation technique proved troublesome when a small but consistent bias existed among raters in the use of a scale, and each rater assigned but one scale value to all tasks. In these instances the question was whether the tasks were truly homogeneous with respect to those scales or whether the scales were insensitive to differences among tasks.

The similarity coefficient technique (r_p) also yielded cases where an inspection of the raw ratings was required before an interpretation could be made. Finally, the percent agreement data, while intuitively appealing in their logic, lacked any formal status as a statistic.

The entire issue of reliability as it applied to the rating data was not clear-cut. Test-retest reliability, for example, would assess how consistent an average rater was in applying a particular scale. It would not address itself to the equally important question of how well the raters would agree among themselves in their collective use of a scale. Similarly, the intraclass correlation coefficient did shed some light on inter-rater agreement, but it appeared to require some unknown amount of heterogeneity among the tasks being rated to do so. Ideally, one would want each rater to be highly consistent in his use of a scale

on a test-retest basis, and also to have raters in high agreement on a scale's use across tasks. Unfortunately, no one statistical technique seemed applicable to assessing both of these aspects.

Regarding the scales themselves, it appeared that a subset of scales consistently emerged which had adequate reliability in all three studies. Table 5 shows the sets of scales for each study which were most reliable. There was a high degree of consistency between the reli e scales emerging from the three-judge and 28-judge studies. Comparing this common subset to the reliable scales of the two-judge study, four of the six were again reliable. Additional scales were also reliable but these were employed only in the two-judge study.

In general, consideration of these three reliability studies led to the following recommendations:

(a) the raters should have a background in psychology or human factors, or a good awareness of such concepts as stimulus and response;

(b) at least three raters should be used in applying the scales in their present form, with an average of their ratings being used as the value to be assigned to the characteristic in question;

(c) further development of the scales should go in the direction of enumeration (counting) rather than rating; and

(d) further efforts should include an assessment of test-retest reliability.

Table 5

LISTING OF THE MOST RELIABLE SCALES WITHIN EACH OF THE THREE RELIABILITY STUDIES

3-Judge Study[1]	28-Judge Study[2]	2-Judge Study[3]
1. Number of output units	1. Number of output units	1. Number of output units
2. Duration for which an output unit is maintained	2. Duration for which an output unit is maintained	2. Duration for which an output unit is maintained
3. Work load imposed by task goal	3. Work load imposed by task goal	3. Simultaneity of responses
4. Simultaneity of responses	4. Simultaneity of responses	4. Number of procedural steps
5. Number of procedural steps	5. Number of procedural steps	5. Response Rate
6. Variability of stimulus location	6. Variability of stimulus location	6. Tutorial Dependency
7. Stimulus or stimulus-complex duration	7. Precision of responses	7. Operator control over response
		8. Number of responses
		9. Operator control over stimulus

[1] Reliability data for these scales are shown in Table 2 where they are listed as Scales #1, 2, 4, 9, 10, 14 and 15, respectively.

[2] Reliability data for these scales are shown in Table 3 where they are listed as Scales #1, 2, 4, 8, 9, 11 and 5, respectively.

[3] Reliability data for these scales are shown in Table 4 where they are listed as Scales #1, 2, 10, 12, 6, 7, 9, 11 and 15, respectively.

POST-DICTION STUDIES

The paradigm used to determine whether the task characteristics were correlates of performance upon which predictive relationships might be established was that of "post-diction". Post-diction simply refers to the fact that existing criterion data were used, whereas in prediction, arrangements are made to collect data in accordance with some specific experimental design. Post-diction sacrifices precise control over many variables in order to rapidly acquire a relevant set of data for analysis. Ratings were made of the tasks used in these studies and then these task characteristic ratings were entered into a multiple regression analysis to establish the extent to which they were related to or predictive of the performance in question. The task descriptions in the literature were often too brief to use, but it was possible to obtain detailed descriptions from either a study's author (e.g., Fleishman in the first post-diction study), or by acquiring the references an author made to more detailed descriptions of the task/apparatus. Through these means it was possible to provide the judges with explicit description of the tasks to be rated. Employing the post-diction paradigm, two studies were conducted.

Both studies shared a number of common restrictions. First, in selecting studies for the two post-diction efforts, there was the need to have a common metric of performance within each. That is, the studies used for any one regression analysis had to be comparable in terms of the unit of performance. Thus, for the first post-diction the performance measures of all studies was expressed in terms of "the number of output units produced per unit time". The second post-diction used studies in which the common performance metric was "percent time on target". In general, this need for a common metric served to reduce the number of studies available for analysis. The relatively small number of studies in both post-diction efforts created, in turn, the following problems:

1. For a regression analysis the number of predictors should not approach, let alone exceed, the number of cases sampled. As the number of of predictors (i.e., characteristic scales) approaches the number of cases sampled (i.e., studies or tasks), the multiple regression coefficient becomes spuriously large and uninterpretable. Since this was the case initially in both post-dictions, the decision was made to use only a selected set of the task characteristic indices as opposed to the full set. For example, instead of using 19 indices and 26 tasks in the first regression study, a smaller set of six indices was used.

2. The small number of studies sampled precluded any meaningful attempt to perform the important step of cross-validating the resultant regression equations.

First Post-Diction Study

The first post-diction study was based on a portion of the data (Fleishman, 1954) used to conduct the reliability study described earlier in which three judges rated 37 tasks on 19 scales. Applying the requirement for a common performance measure, the 37 tasks were carefully screened in order to determine the types of performance measures associated with them. Although several different measures were represented (e.g., reaction time, percent time on target, or percent correct), 26 of the tasks had one measure in common which was designated as the "number of units produced per unit time". The "units" varied and included such things as: number of blocks moved; number of assemblies completed; number of taps made; and number of correct discriminations given. Common to these 26 tasks was the requirement that as many "units" as possible be produced during specified time periods. Since different amounts of time were allowed for completion of the various tasks (e.g., 25 to 900 seconds), a common time frame was needed to provide a standard basis for comparison. The "unit time" chosen for this purpose was one second. Therefore, the performance score reported for each task was prorated to obtain the average number of units produced per second (i.e., 98.5 units produced in 80 seconds equalled 1.231 units per second). (The 26 tasks are indicated by asterisks in Appendix 2.)

Since the entire set of 19 rating scales (Appendix 1) could not be employed, a smaller subset was selected. The six most reliable scales were chosen for analysis (see Table 2). For each of these scales the ratings provided by three judges were averaged to obtain a single value on each scale for each of the tasks. The specific scales employed in the study were:

1. Stimulus duration (scale # 15),
2. Number of output units (scale # 1),
3. Duration for which an output unit is maintained (scale # 2),
4. Simultaneity of responses (scale # 9),
5. Number of procedural steps (scale # 10), and
6. Variability of stimulus location (scale # 14).

Table 6 presents the data on which the first post-diction study was based. A Wherry-Doolittle stepwise regression analysis was carried out by computer. Six predictor variables were entered into the analysis, but only five were processed. The order in which the scales are listed above represents their order of extraction based upon the percent variance accounted for in the criterion measure (R^2). Although five scales emerged from the analysis, a point of diminishing returns in terms of percent variance accounted for was reached after extraction of the fourth scale. Consequently, a regression equation was written using only the first four scales listed above. The half-diagonal intercorrelation matrix for all seven variables (six predictors, one criterion) is presented in Table 7.

The multiple correlation coefficient for this analysis (based on four predictors) was $R = 0.85$ which accounted for 72% of the variance (R^2) in the criterion measure. This correlation was significant ($F(4, 21) = 13:75$, $p < .01$). It was felt, however, that the small sample ($n = 26$) used in this analysis yielded an inflated multiple R relative to what might have been obtained had a larger sample ($n = \geq 100$) been used. Accordingly, a correction in R for small sample bias (Guilford, 1956, p. 399) was applied. The corrected correlation ($_cR$) was 0.82, which was still significant ($F(4,21) = 10.78$, $p < .01$).

Table 6

BASIC DATA FOR THE FIRST REGRESSION ANALYSIS

Tasks	Avg. No. Units Produced Per Second	Average Rating on Six Scales *					
		1	2	3	4	5	6
1. Two-Plate Tapping	3.98	7	7	1	1	1	4
2. Key Tapping	6.24	7	7	1	1	1	7
3. Ten-Target Aiming	2.02	4	6	1	1	1	4
4. Rotary Aiming	2.49	4	7	1	1	2	4
5. Hand-Precision Aiming	1.87	4	7	1	1	1	4
6. Visual Reaction Time	2.71	4	1	1	1	1	7
7. Auditory Reaction Time	2.86	4	1	1	1	1	7
8. Minnesota - Placing	1.23	4	7	1	1	1	4
9. Minnesota - Turning	1.49	4	7	1	5	3	4
10. Purdue Pegboard - Right Hand	0.56	4	7	1	1	1	4
11. Purdue Pegboard - Both Hands	0.87	4	7	1	4	1	4
12. Purdue Pegboard - Assembly	0.62	4	7	1	3	4	5
13. O'Connor Finger Dexterity	0.53	4	7	1	1	1	4
14. Santa Ana Finger Dexterity	1.80	4	7	1	1	1	4
15. Pin Stick	1.26	4	7	1	1	1	4
16. Dynamic Balance	0.04	4	7	2	4	1	2
17. Medium Tapping	1.34	4	7	1	1	1	4
18. Large Tapping	1.26	4	7	1	1	1	4
19. Aiming	1.31	4	7	1	1	1	4
20. Pursuit Aiming I	2.32	4	7	1	1	1	4
21. Pursuit Aiming II	1.76	4	7	1	1	1	4
22. Square Marking	1.16	4	7	1	1	1	4
23. Tracing	1.89	4	7	1	1	1	3
24. Discrimination Reaction Time-Printed	0.38	4	7	1	1	1	3
25. Marking Accuracy	1.37	4	7	1	1	1	4
26. Verbal Addition Task	0.19	4	5	1	1	1	7

* The six scales were: 1. Stimulus duration; 2. Number of output units; 3. Duration for which an output unit is maintained; 4. Simultaneity of responses; 5. Number of procedural steps; and 6. Variability of stimulus location.

Table 7

INTERCORRELATION MATRIX FOR THE FIRST REGRESSION ANALYSIS

	1	2	3	4	5	6	7*
1	1.00	.01	-.06	-.12	-.10	.27	.78
2		1.00	.07	.15	.12	-.70	-.19
3			1.00	.45	-.07	-.38	-.26
4				1.00	.55	-.23	-.28
5					1.00	.04	-.12
6						1.00	.47
7*							1.00

* Criterion measure

An index of forecasting efficiency (Guilford, 1956, p. 398) which indicated the degree to which predictions made by means of the regression equation were better (more accurate) than those made merely from a knowledge of the mean of the criterion measures was computed. The index for the corrected R was 42.6%, which indicated that use of the regression equation would be superior to using the mean alone.

The regression equation was:

$$\overline{P}'_m = -1.064 + 1.245X_1 - 0.197X_2 - 1.072X_3 - 0.089X_4$$

where

$\overline{P}'_m$ = Predicted mean number of output units produced per second; and

$X_1 - X_4$ = Task characteristic scales # 1 through # 4 listed above.

Second Post-Diction Study

The second post-diction study was based on data from the third reliability study described earlier in which two judges rated 21 tasks on 18 scales. The criterion measure common to the 20 tasks ultimately used was the mean percent time on target achieved after five minutes of practice on the tasks in question. These tasks and their associated performance data were obtained from studies reported in the experimental literature. (See Appendix 6 for references to these studies.)

The need to reduce the set of predictors existed here as in the first post-diction study. Accordingly, the same reductive procedure was followed. This involved ranking the 18 scales (Appendix 5) in terms of their reliability and then selecting the final subset on the basis of high reliability. This operation resulted in the selection of the following scales:

1. Number of procedural steps,
2. Precision of responses,
3. Number of responses,
4. Number of output units,
5. Simultaneity of responses, and
6. Number of elements/output unit.

Table 8 presents the data on which the second post-diction study was based.

A multiple correlation (R) was computed using a stepwise procedure. The order of the scales in the above list paralleled the order in which the predictor variables emerged from the regression analysis. A point of diminishing returns, in terms of percent variance accounted for (R^2), was reached after the fourth predictor emerged. Consequently, a regression equation was written using only the first four scales listed above. The half-diagonal intercorrelation matrix for all seven variables (six predictors, one criterion) is presented in Table 9.

Table 8

BASIC DATA FOR THE SECOND REGRESSION ANALYSIS

Task	Avg. % TOT After 5-Min.	Average Rating on Six Scales* 1	2	3	4	5	6
1. 3-D Pursuit Tracking	15	3	6	2	1	3	3
2. Two-Hand Coordination	24	2	6	2	1	2	3
3. Rudder Control	68	1.5	6.5	2	9	2	2
4. Turret Pursuit	24	3	6	4	1	2	3
5. Wheel Turning	59	1	7	1	1	0	1
6. Two-Hand Coordination	49	2.5	6.5	2	1	2	3
7. Pursuit Rotor	44	1.5	4	1	1	0	2
8. Rudder Control	88	1.5	6	2	1	2	2
9. Iowa Pursuit	20	3.5	5.5	4	1	2	3
10. Two-Hand Coordination	23	2.5	5	2	1	2	3
11. Pursuit Rotor	18	1	4.5	1	1	0	2
12. Two-Hand Coordination	35	3	5.5	2	1	2	3
13. Continuous Compensatory Tracking	32	2	5.5	2	1	0	2
14. Koerth Pursuit Rotor	20	1	4	1	1	0	2
15. Koerth Pursuit Rotor	30	1	4	1	1	0	2
16. Koerth Pursuit Rotor	33	1	4	1	1	0	2
17. Compensatory Tracking	13	2.5	3.5	4	1	0	2
18. Pedestal Gunnery Task	19	5	6	7	8	2	3
19. Compensatory Pursuit	50	4.5	5.5	6	1	3	3
20. Pedestal Sight Manipulation	51	3	5.5	6	8	2	3

*The six scales were: 1. Number of procedural steps; 2. Precision of responses; 3. Number of responses; 4. Number of output units; 5. Simultaneity of responses; and 6. Number of elements/output unit.

Table 3

INTERCORRELATION MATRIX FOR THE SECOND REGRESSION ANALYSIS

	1	2	3	4	5	6	7*
1	1.00	.34	.90	.44	.75	.76	-.54
2		1.00	.25	.34	.61	.26	.30
3			1.00	.59	.60	.60	-.41
4				1.00	.38	.22	.07
5					1.00	.81	-.18
6						1.00	-.46
7							1.00

* Criterion measure

The multiple R achieved for this post-diction study was 0.79, which accounted for 63% of the variance (R^2). This coefficient was significant [$F(4, 15) = 6.42$, $p < .01$]. Correction for small sample bias yielded a $_cR = 0.73$, which was also significant [$F(4, 15) = 4.28$, $p < .05$]. The index of forecasting efficiency for this corrected R was 31.7%. This figure indicated that prediction using the regression equation would be superior to that made on the basis of knowledge of the mean of the criterion measures alone.

The regression equation was:

$$\bar{P}'_m = -1.484 - 19.056X_1 + 12.102X_2 + 4.213X_3 + 1.251X_4$$

where

$\bar{P}'_m$ = Predicted mean percent time on targer after 5 minutes of practice; and

$X_1 - X_4$ = Task characteristic scales #1 through #4 listed above.

Discussion

The results of both post-diction studies are presented for comparison in Table 10.

Table 10

COMPARISON OF POST-DICTION STUDIES 1 AND 2

	Uncorrected		Corrected		Forecasting Efficiency	$p({}_cR)$
	R	R^2	${}_cR$	${}_cR^2$		
Study 1	.85	.72	.82	.67	43%	.01
Study 2	.79	.63	.73	.53	32%	.05

It is apparent that the post-diction efforts were successful in both cases. The critical question of whether these results would hold up in the face of cross-validation remains an open issue. Both studies provide a predictive mechanism which had adequate merit when compared to predicting performance on the basis of knowledge of only the means of the respective samples.

Consideration of these results was interesting in light of the model of performance cited earlier in the report. There, performance was viewed as a function of the operator, the task, and the environment. Given that the operator and the environment components were essentially "uncontrolled" or, at least, were unknown quantities in the studies used here, it was not anticipated that the task component alone would account for as much of the variance (67% and 53%) as it seemingly did.

The model, for instance, suggested that uncontrolled variations in the operator and environmental components might well mask the relationship between task characteristics and performance. This masking may

indeed have been present. That it was not as pronounced as expected, however, may have been due to the fact that the operator and environmental components were being indirectly controlled or almost held constant. For example, with regard to the environment, it could be assumed that any experimenter would attempt to ensure that such conditions as room temperature, noise level, level of illumination, etc., were at least within some "subjective zone of acceptance" when setting up his experiment unless these variables were actually part of his design. Since the studies chosen were picked so as to avoid the presence of such independent variables as stress, drugs, etc., it is reasonably safe to assume that the "environment component" was essentially constant across studies. Furthermore, the use of mean performance scores on each task (obtained by averaging across individuals) tended to minimize the influence of individual difference variables.

Given the limitations inherent in the post-diction approach, these studies nevertheless showed that selected task characteristics were correlates of performance. Use of the task definition described earlier and of the descriptive indices derived from it appeared to provide a basis for systematically relating differences among tasks to variations in performance.

CONCLUSIONS AND RECOMMENDATIONS

The work described in this report has focused on but one of several possible approaches which might be pursued in better organizing information about human performance. In the present approach "tasks" were viewed as more than merely convenient vehicles to be used when assessing the effects of selected experimental treatments on performance. Instead, tasks were treated as complexes of independent variables which, in their own right, were capable of influencing performance. To better understand their influence, therefore, a language was developed to permit objective and direct description of different tasks and to provide a basis for comparing and contrasting various "task treatments".

This effort has tentatively demonstrated that it is possible to describe tasks in terms of a task-characteristics language which is relatively free of the subjective and indirect descriptors found in many other systems. It has further demonstrated that the task characteristics may represent important correlates of performance. Although more convincing proof of this point must await cross-validation exercises, it was possible to describe subtle differences among tasks and to relate such differences systematically to variations in performance.

While successful in many respects, the study also encountered a number of difficulties. First, although several scales proved reasonably reliable, many others did not. Substantial improvement in this area is required and might result from more intensive training of judges, better definition of characteristics, and/or improved methods of quantification. Until higher overall reliabilities can be obtained, continued use of panels of judges will be necessary. This procedure is less attractive than the use of a single rater.

Second, the current language was designed so as to be applicable to all tasks, given our definition of a "task". The study indicated, however, that the scales in their present form were less suitable for the description of "cognitive" paper-and-pencil tasks. It may be

necessary to develop additional descriptors for this type of task, or to treat such tasks separately within an entirely different descriptive system.

Third and finally, one use of the descriptive system was in predicting the mean level of performance expected on different tasks. It became apparent, however, that meaningful regression equations could be developed only when the tasks in question shared the same response measure. In other words, different equations would be required for tasks on which different response measures were employed.

One general consequence of this situation, therefore, is the need for research which attempts to identify the smallest set of distinct response measures which can be used to represent all possible measures. Teichner and Olson (1964) have suggested four such measures (probability of detection, percent error, percentage decrement in time on target, and reaction time) which, if they encompassed a large proportion of all possible tasks, would be worth pursuing. A second consequence bears directly on the language developed in the present study. The possibility exists of tailoring separate descriptive systems for use with different categories of tasks (defined in terms of response measures). While this approach is certainly feasible, and might actually be superior were one only interested in a particular category of tasks, it was not adopted in the present study because of more catholic interests. A language was desired which not only would provide a basis for predicting performance (within categories) but which would also provide for comparisons of tasks across different categories.

Much additional research is required if the approach is to be developed to the fullest extent possible. Two efforts in particular are required. The first would center on the type of application emphasized in the present effort, while the second would attempt to broaden the scope of the approach.

First, the predictive methodology should be assessed using a much larger sample of tasks. Resultant regression equations must then be evaluated in formal cross-validation exercises. Given that these efforts were successful, how would the predictive methodology be applied? Ideally, the user -- an equipment design engineer, a training specialist, etc. -- would first identify the type of performance measure most appropriate for the new task in question. He would then refer to a document containing a number of regression equations, each of which was specific to a particular type of performance measure. Here he would see which scales were involved in assessing the type of performance relevant to his interest. He would then rate the new task on these scales and enter these values into the equation which would contain the appropriate weights. The output would be a predicted mean level of performance on that task at some specified point in the learning curve. This estimated level of performance could then be compared to some desired criterion level of performance. If the predicted performance were inadequate relative to the desired level, the user would receive guidance regarding remedial actions, i.e., redesigning certain aspects of the task. For example, beta-weights for each of the terms in the equation would indicate the relative contribution made by each task characteristic to the predicted performance. The user would be in a position to change certain features of the task by assessing which features were most potent versus those which were amenable to change. Having made these changes conceptually, he could rerun the regression equation using the new task values and see if a sufficient improvement in performance had occurred. This iterative process could be accomplished without physically changing the equipment until a change was warranted.

A second, rather different application of the descriptive language should also be studied. The situation here would be represented by the case where a review of the literature was conducted to determine the general effect of a specific "environmental" variable, i.e., massed versus distributed practice, levels of noise, etc., on performance. Typically, the findings of such a survey could be used to define a subset

of studies in which, for example: massed practice proved superior; a subset in which spaced practice proved superior; and, possibly, a third subset which yielded no difference between the variables. Having categorized the studies in terms of which treatment was superior (i.e., massed, distributed, neither), the tasks used in those studies would then be rated, and discriminant function analyses would be conducted to determine whether different task profiles were associated with the various criterion groups. If such were the case, additional studies would be selected, tasks within those studies would be rated, and the obtained profiles would be analyzed in order to predict which distribution of practice should be superior for a given task. The predictions would be checked against actual learning data. If successful, these efforts would have identified those aspects of tasks which were beneficially conducive to the application of, in the case of the example cited, either massed or distributed practice. These findings would be of importance to researchers in both the applied and theoretical fields. Such suggestions were made earlier by Fleishman (1967) and Fleishman, Teichner, and Stephenson (1970); beginnings in this direction have been made under this project by Teichner and Whitehead (1971).

In summary, in addition to pursuing the two major applications cited above, the following activities should also be considered:

(a) the development of descriptive systems for the operator and environmental components;

(b) the development of a response taxonomy or classification system to reduce the number of potential performance measures to a manageable set;

(c) a mathematical procedure for allowing the characteristics of all three of the model's components to enter into a full test of the model's predictive efficiency;

(d) the further development of the task characteristics themselves in the direction of greater quantification;

(e) a more adequate means of assessing the types of "reliability" of interest to the rating situation encountered here;

(f) the development of a collection of suitable tasks both adequate in number and type to permit cross-validation; or

(g) programmatic experimental efforts in which tasks, operators, and the environment can be systematically varied.

The need for further development notwithstanding, the present study has served a valuable purpose. It has demonstrated the essential validity and utility of a rather different method of task description. The characteristics themselves are not the only ones, nor necessarily the best ones, which might be developed. Similarly, only one of several possible uses of the descriptive data was evaluated. Although the specifics of the system may eventually assume a very different form, the present study has demonstrated the soundness of the underlying approach.

REFERENCES

Cattell, R. B. & Coulter, M. A. Principles of behavioral taxonomy and the mathematical basis of the taxonomy computer program. British Journal of Mathematical and Statistical Psychology, 1966, 19 (Part 2), 237-269.

Cotterman, T. E. Task classification: An approach to partially ordering information on human learning. Report No. WADC TN 58-374, January 1959. Wright Patterson Air Force Base, Ohio.

Farina, A. J., Jr. Development of a taxonomy of human performance: Descriptive schemes for human task behavior. Report No. AIR-726-1/69-TR-2, January 1969. American Institutes for Research, Washington, D. C.

Fine, S. A. A functional approach to a broad scale map of work behaviors. Report No. HSR-RM-63/2, September 1963. Human Sciences Research, McLean, Virginia.

Fitts, P. M. Factors in complex skill training. In R. Glasser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962.

Fleishman, E. A. Performance assessment based on an empirically derived task taxonomy. Human Factors, 1967, 9(4), 349-366.

Fleishman, E. A. The description and prediction of perceptual-motor skill learning. In R. Glasser (Ed.), Training research and education. Pittsburgh: University of Pittsburgh Press, 1962.

Fleishman, E. A. & Stephenson, R. W. Development of a taxonomy of human performance: A review of the third year's progress. Report No. AIR-726-9/70-TPR-3, September 1970. American Institutes for Research, Washington, D. C.

Fleishman, E. A., Kinkade, R. G., & Chambers, A. N. Development of a taxonomy of human performance: A review of the first year's progress. Report No. AIR-726-11/68-TPR-1, November 1968. American Institutes for Research, Washington, D. C.

Fleishman, E. A., Teichner, W. H., & Stephenson, R. W. Development of a taxonomy of human performance: A review of the second year's progress. Report No. AIR-726-1/70-TPR-2, January 1970. American Institutes for Research, Washington, D. C.

Folley, J. D., Jr. Development of an improved method of task analysis and beginning of a theory of training. Report No. NAVTRADEVCEN 1218-1, June 1964. U. S. Naval Training Device Center, Port Washington, New York.

Gagne, R. M. Human functions in systems. In R. M. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

Ginsburg, R., McCullers, J. C., Merryman, J. J., Thomson, C. W., & Whitte, R. S. A review of efforts to organize information about human learning, transfer, and retention. San Jose, California: San Jose State College, 1966.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1956.

Hackman, J. R. Tasks and task performance in research on stress. In J. E. McGrach (Ed.), Social and psychological factors on stress. New York: Holt, Rinehart, and Winston, 1968.

McCormick, E. J. Job dimensions: Their nature and possible uses. Paper presented at the International Congress of Applied Psychology, Amsterdam, August 1968.

Melton, A. W. & Briggs, G. E. Engineering psychology. Annual Review of Psychology, 1960, 11, 71-98.

Miller, R. B. Task taxonomy: Science or technology? International Business Machines, Poughkeepsie, New York, 1966.

Miller, R. B. Task description and analysis. In R. M. Gagne (Ed.), Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

Reed, L. E. Advances in the use of computers for handling human factors task data. Report No. AMRL-TR-67-16, April 1967. Wright Patterson Air Force Base, Ohio.

Silverman, J. New techniques in task analysis. Report No. SRM 68-12, 1967. U. S. Naval Personnel Research Activity, San Diego, Calif.

Smith, P. C. & Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47(2), 149-155.

Stolurow, L. M. A taxonomy of learning task characteristics. Report No. AMRL-TDR-64-2, 1964. Wright Patterson Air Force Base, Ohio.

Teichner, W. H. & Olson, D. Predicting human performance in space environments. NASA Report No. CR-1370, 1969. National Aeronautics and Space Administration, Washington, D. C.

Teichner, W. H. & Whitehead, J. Development of a taxonomy of human performance: Evaluation of a task classification system for generalizing research findings from a data base. Report No. AIR-726/2035-4/71-TR-8, April 1971. American Institutes for Research, Washington, D.C. (U. S. Army Behavior and Systems Research Laboratory Research Study 71-8.)

Theologus, G. C., Romashko, R., & Fleishman, E. A. Development of a taxonomy of human performance: A feasibility study of ability dimensions for classifying human tasks. Report No. AIR-726-1/70-TR-5, January 1970. American Institutes for Research, Washington, D. C.

Wheaton, G. R. Development of a taxonomy of human performance: A review of classificatory systems relating to tasks and performance. Report No. AIR-726-12/68-TR-1, December 1968. American Institutes for Research, Washington, D. C.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

APPENDICES

		Page
Appendix 1.	Scales Used in the 3-Judge Study	45
2.	37 Tasks Used in the 3-Judge Study	67
3.	Scales Used in the 28-Judge Reliability Study	73
4.	Tasks Used in the 28-Judge Reliability Study	91
5.	Scales Used in the 2-Judge Study	97
6.	Tasks Used in the 2-Judge Study	115

 Preceding page blank

APPENDIX 1

SCALES USED IN THE 3-JUDGE STUDY

This section contains the 19 scales used in the 3-judge study. Asterisks identify the subset of these scales which were ultimately entered into the multiple regression analysis.

Preceding page blank

TASK CHARACTERISTICS ANSWER SHEET

Rater's Name _____

Date Rating Performed _____

Name and Number of Task Rated _____

Instructions

There are 19 rating scales. Each task should be rated on all 19 scales. As you assigned a scale value to the task, write down the scale value on the line for that rating scale as listed below. There is space at the bottom for you to describe any problems you had in applying the scales to the task.

*1. Number of output units ____

*2. Duration for which an output unit is maintained ____

3. Number of elements per output unit ____

4. Work load imposed by task goal ____

5. Difficulty of goal attainment ____

6. Precision of responses ____

7. Rate of responding ____

8. Amount of muscular effort involved in responses ____

*9. Simultaneity of response ____

*10. Number of procedural steps ____

11. Dependency of procedural steps ____

12. Adherence to procedures ____

13. Procedural complexity ____

*14. Variability of stimulus location ____

*15. Stimulus or stimulus-complex duration ____

16. Regularity of stimulus occurrence ____

17. Degree of operator control ____

18. Reaction time/feedback lag ____

19. Decision-making ____

Problems/Comments

Preceding page blank

* 1. NUMBER OF OUTPUT UNITS

An output unit is specified or implied in the statement of the task goal. Output units are often: an assembly of objects, a stimulus-control relationship, or a specifiable end-product (e. g., arrival at B in the task, run from A to B). You are to judge the number of output units specified or implied by the task goal relative to other quotas which could be established for the same type of task.

Definition		Examples
<u>As many as possible</u> - as many output units as possible are to be produced, usually during a fixed period of time.	7	Insert as many plugs into the connectors as possible in five minutes Do 200 push-ups in five minutes
	6	Do 200 push-ups.
	5	
<u>Moderate number</u> - relative to other possible quotas for the same type of task, a moderate number of output units is to be produced.	4	Do twenty push-ups in five minutes. Do twenty push-ups.
	3	
	2	
<u>One output unit</u> - one output unit is to be produced. It is either maintained or signals the termination of performance.	1	Assume a push-up position. Maintain it for five minutes. Do one push-up. Add the following list of integers.

*2. DURATION FOR WHICH AN OUTPUT UNIT IS MAINTAINED

Once the operator has produced an output unit, he may be required to maintain or continue it for one of several time periods. For example, it can be maintained for as long as possible; or, its completion may be a signal to leave it and go on to produce the next output unit; or, finally, having produced it, performance ends.

Decide where the present output unit belongs on the below scale.

Definition		Examples
Maintenance for as long as possible - an output unit (body position, stimulus-control relationship, etc.) is to be maintained for as long as possible.	7	• Hang in a bent-arm position for as long as possible.
	6	• Maintain a stimulus-control relationship (target and cursor) for 20 minutes.
	5	
Moderate maintenance - relative to other possible periods of maintenance, an output unit is to be maintained for a moderate period of time.	4	• Maintain a stimulus-control relationship for five minutes.
	3	
	2	• Do as many push-ups as possible in ten minutes holding each "down" position for 30 seconds.
		• Do as many push-ups as possible in two minutes.
Short maintenance - production of an output unit signals termination of performance or production of additional units. Maintenance, therefore, is minimal.	1	• Solve the following trigonometric problems.

3. NUMBER OF ELEMENTS PER OUTPUT UNIT

One way of describing an output unit is in terms of the number of elements involved in its production. By elements we mean the objects or components which, when assembled, comprise the output unit. In an addition problem, for example, the numbers to be added are the elements which comprise the output unit.

Rate the present task in terms of the number of elements forming an output unit on the scale below.

Definition		Examples
Many elements: each output unit contains many constituent elements.	7	• Assemble a radio from the components in this kit.
	6	
	5	
Moderate number of elements: each output unit contains several constituent elements.	4	• Change a flat tire. • Rank order these 20 items.
	3	
	2	
One element: each output unit contains only one constituent element.	1	• Push the button when the light comes on.

4. WORK LOAD IMPOSED BY TASK GOAL

Work load is judged in terms of the number of output units to be produced relative to the amount of time allowed for their protection, i.e., output units per time.

There are those tasks in which the goal is to maintain a situation, e.g., stay within 40 feet of the vehicle ahead of you, rather than produce multiple output units. For those tasks, the degree of work load is directly related to the length of time for which maintenance is required.

Rate the present task on the scale below.

Definition		Examples
High work load - as many output units as possible are to be produced in a fixed period of time; a relatively large number of output units is to be produced in a relatively short period of time; an output unit is to be maintained for a relatively long time or for as long as possible.	7	• Drive as many nails as possible in five minutes. • Maintain a stimulus-control relationship for one hour
	6	• Maintain a stimulus-control relationship as long as possible.
	5	
Moderate work load - a moderate number of output units is to be produced in a reasonable period of time; an output unit is to be maintained for a moderate period of time relative to other possible periods.	4	• Drive ten nails in five minutes. • Maintain a stimulus-control relationship for three minutes.
	3	
	2	• Drive these two nails in the next five minutes. • Sum the following five integers. • Maintain a stimulus-control relationship for 30 seconds.
Low work load - a small number of output units is to be produced in a relatively long period of time; an output unit is to be maintained for a relatively short period of time.	1	

5. DIFFICULTY OF GOAL ATTAINMENT

Difficulty of goal attainment is a function of two things: 1) the number of elements in an output unit, and 2) the degree of work load (both these terms have been previously defined). The greater the work load and the higher the number of elements, the more difficult is the goal.

Definition		Examples
<u>High difficulty</u> - not only is the work load high, but the number of elements in an output unit is also high.	7	•You have two days in which you are to assemble as many radio kits as possible.
	6	
	5	
<u>Moderate difficulty</u> - both the work load and the number of elements in an output unit are moderate; this combination results in a task of average difficulty. Or, one measure is high and the other is low, thus yielding a moderate average.	4	•Build two small shipping crates today.
	3	
	2	•Paint the walls of this room. Take as much time as you need.
<u>Low difficulty</u> - relative to other possible values, work load and element number are both very low.	1	

6. PRECISION OF RESPONSES

Tasks may be differentiated with respect to the degree of precision associated with overt observable responses. Degree of precision or motor control required will increase as target size decreases, lag in controls increases, rate of change in stimulus increases, etc. You are to judge the degree of precision required in overt responses.

Definition		Examples
High degree of precision - because of small targets, fine scales, sensitive controls, etc. the subject must make responses which are extremely precise.	7	• Using a chemical balance (scales) determine the weight of the following objects to the nearest microgram. • Replace the mainspring in this w[illegible]
	6	
	5	
Moderate precision - relative to the definitions above or below, a moderate degree of precision must accompany subject's responses.	4	• Solder these two wires together. Using your pencil, trace this maze.
	3	
	2	• Do twenty push-ups. • Sort the oranges and lemons into two piles.
Low degree of precision - because of large targets, gross scales, insensitive controls, etc. the subject can make responses which are gross or imprecise.	1	

7. RATE OF RESPONDING

Goal-directed responses can be emitted at different rates. You are to judge the rate of responding in a particular task by considering other rates which are possible for that same task.

Definition		Examples
<u>High rate of responding</u> - many responses are required per unit time relative to other rates which could be employed for the same task. Responses are often repetitive or serial. In the extreme, they are continuous.	7	• Fire 20 rounds as quickly as possible. • Complete this jig-saw puzzle as fast as you can. • Track this target.
	6	
	5	
<u>Moderate rate of responding</u> - a moderate number of responses are required per unit time.	4	• Fire 20 rounds. Fire rapidly but also be as accurate as you can. • You have half an hour to complete this 20 item "True-False" quiz.
	3	
	2	
<u>Low rate of responding</u> - few responses are emitted per unit time. Responses are often singular.	1	• Add the following numbers. Take all the time you need.

8. AMOUNT OF MUSCULAR EFFORT INVOLVED IN RESPONSES

This dimension reflects the degree of muscular effort required in performing the task. It ranges from physical efforts such as weight-lifting to simple verbal responses.

Definition		Examples
High amount of muscular effort-response(s) require a high degree of muscular involvement.	7	• Do 20 pushups. • Lift the heaviest weight possible.
	6	
	5	
Moderate amount of muscular effort required for the response(s)	4	• Tighten nuts on bolts securely.
	3	
	2	
Low amount of muscular effort required	1	• Solder two wires together • Add numbers and report the sum aloud.

*9. SIMULTANEITY OF RESPONSES

An overt response or sequence of responses leading to the production of an output unit may involve one or more effectors (hands, arms, legs, feet, voice, etc.). These effectors may or may not be used simultaneously.

You are to rate the degree of simultaneity involved in using the effectors needed in the response(s) leading to production of an output unit.

Definition		Examples
<u>High simultaneity</u> - responses involve the simultaneous use of several effectors on a fairly continuous basis.	7	•You are to fly this plane at 400 knots and an altitude of 5,000 feet, banking to the left and to the right. •Play this song on the piano.
	6	
	5	
<u>Moderate simultaneity</u> - responses involve the simultaneous use of at least two effectors on a continuous or periodic basis.	4	•Pat your head and rub your stomach. •Hit that target by firing your rifle.
	3	
	2	
<u>Low simultaneity</u> - responses involve the use of only one effector at a time. If other effectors are employed, they are employed sequentially.	1	•Push the button when the light comes on.

*10. NUMBER OF PROCEDURAL STEPS

Earlier we were concerned about the number of elements, i.e., objects or components, involved in the production of one output unit. Now we want to consider the number of procedural steps (responses) needed to produce one output unit. There isn't a necessary one-to-one relationship between objects and responses.

Consider the number of responses or steps involved in producing one output unit for the present task. Rate this task on the scale below.

Definition		Examples
Large number of steps - the procedure consists of a large number of constituent steps.	7	•Build a crystal receiver set following the enclosed instructions.
	6	
	5	
Medium number of steps - the procedure contains a medium number of steps relative to other procedures.	4	•Solve the equation $X^2 - 4X + 4 = 0$. •Type the following business letter.
	3	
	2	
Small number of steps - the procedure consists of few steps. At a minimum, only one step may be necessary.	1	•Open this combination lock (32L - 43R - 10L). •Press the button whenever the light comes on.

II. DEPENDENCY OF PROCEDURAL STEPS

Consider again the number of steps involved in producing one output unit. The steps may be described in terms of the dependency among them; dependency concerns the extent to which the steps must be done in some specified order. For example, dependency exists between steps A and B if step B cannot be accomplished without step A being done first. Note: Procedures which have only one step are automatically low in dependency.

Definition		Examples
High dependency among steps - each step in the procedure is completely dependent upon the preceding procedural step. Systematic ordering of steps is at a maximum.	7	•Using the combination you've been given, open the safe. •Dial this telephone number.
	6	
	5	
Moderate dependency among steps - in the total number of steps comprising the procedure, approximately 50% are dependent upon preceding steps.	4	•Using colored blocks, stack them into columns four blocks high. Do this in the order red and green for the first two blocks. The remaining blocks may be of any color.
	3	
	2	•Using colored blocks, stack them into columns four blocks high. Order of color is unimportant.
Low dependency among steps - procedural steps are not organized in any particular sequence. Step "A" may precede "B" or "B" may precede "A". Procedures having one step are low in dependency.	1	

12. ADHERENCE TO PROCEDURES

Tasks may vary in the extent to which the operator must faithfully adhere to the procedures set forth. In some types of tasks strict adherence is critical; in others, the operator may depart somewhat from stated procedures without jeopardy to the performance.

Judge the degree of adherence to stated procedures for the present task.

Definition		Examples
<u>High</u> - strict adherence to procedures with even small departures being discouraged or having detrimental results.	7	Firing an M-1 rifle according to procedures given by a Marine D.I.
	6	
	5	
<u>Moderate</u> - Some departures from the stated procedures are tolerated.	4	Given conventional procedures to solve a trigonometric problem; alternative procedures exist and can be employed.
	3	
	2	
<u>Low</u> - fairly large departures from stated procedures are tolerated.	1	Type a letter using whatever procedures you are most comfortable with.

13. PROCEDURAL COMPLEXITY

Procedural complexity is a function of the number of steps or responses leading to an output unit and the degree of dependency among these steps.

Rate the present task in terms of its procedural complexity.

Definition		Examples
High complexity - the procedure contains many steps. Each step is dependent upon execution of the step which precedes it.	7	Without referencing any notes, perform a B-52 pre-flight check-list task.
	6	
	5	
Moderate complexity - the procedure contains several steps. Not all steps are dependent upon preceding steps, however.	4	Check and if necessary replace the following ten tubes ($T_1 \ldots T_{10}$) in these 10 radio sets.
	3	
	2	
Low complexity - the procedures consists of few steps and there is little if any dependency among steps.	1	When the light comes on, press this button as fast as you can. Bolt this bracket to that frame.

* 14. VARIABILITY OF STIMULUS LOCATION

Judge the degree to which the physical location of the stimulus or stimulus complex is predictable over task time.

Definition		Examples
<u>High predictability</u> - stimulus location remains basically unchanged.	7	• Stimulus is a red light located on a display panel.
	6	
	5	
<u>Medium predictability</u> - location changes but in a known manner or pattern.	4	• Visually following an arrow in flight toward a target.
	3	
	2	
<u>Low predictability</u> - location changes in an almost random fashion.	1	• Predicting which leaf will fall from a tree next.

*15. STIMULUS OR STIMULUS-COMPLEX DURATION

Consider the critical stimulus or stimulus complex to which the operator must attend in performing the task. Relative to the total task time, for how long a duration is the stimulus or stimulus complex present during the task?

Definition		Examples
Long duration - stimulus would remain indefinitely.	7	•Drawing a picture by observing a model of the object being drawn.
	6	
	5	
Medium duration - stimulus remains present until changed (spatially, temporally, etc.) by the response made to it.	4	•Red light goes out when operator pushes a button.
	3	
	2	
Short duration - stimulus ceases prior to response being made to it.	1	•Operator must identify words or targets presented tachistoscopically.

16. REGULARITY OF STIMULUS OCCURRENCE

Consider the critical stimulus or stimulus-complex to which the operator must attend. Does it occur at regular (i. e., equal) intervals or at irregular intervals. Treat equal intervals and constant presence of the stimulus as equivalent conditions.

Rate the present task on this dimension.

Definition		Examples
High regularity - regular intervals, periodic occurrence. Also refers to stimulus which is constantly present.	7	• Responding to units on an assembly line. • Looking at a picture on a wall.
	6	
	5	
Medium regularity - Irregular intervals but a perceivable pattern of occurrence.	4	• Receiving morse code.
	3	
	2	
Low regularity - Very irregular intervals; stimulus occurrence is aperiodic.	1	• Detecting random signals on a CRT display.

17. DEGREE OF OPERATOR CONTROL OVER THE OCCURRENCE OF THE STIMULUS AND THE RESPONSE

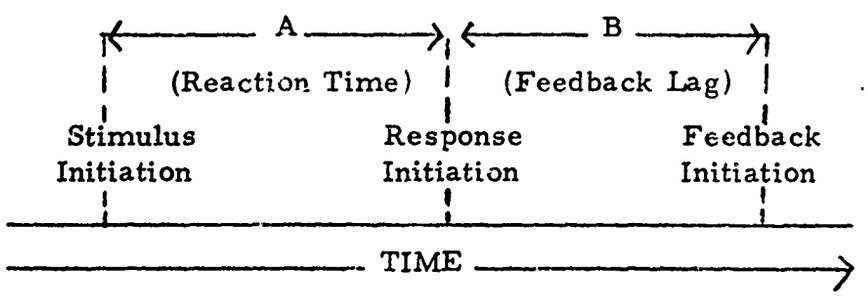
Does the operator determine when the stimulus appears (e.g., self-controlled) or is the occurrence of the stimulus externally-controlled? Given the occurrence of the stimulus, must the operator respond immediately (externally-controlled) or may he respond at will (self-controlled)?

Based on these two decisions, rate the task in question on the following scale.

Definition		Examples
<u>High subject control</u> - (both stimuli and responses are self-paced).	7	• Reading aloud to oneself.
	6	
	5	
<u>Medium subject control</u> - (either the stimuli or the responses are self-paced).	4	• Shooting skeet (shooter determined when 'bird' appears).
	3	
	2	
<u>Low subject control</u> - (both stimuli and responses are externally paced).	1	• Typical reaction time task.

18. REACTION TIME/FEEDBACK LAG RELATIONSHIP

What relationship exists between the operator's reaction time interval (i.e., the time between stimulus appearance and initiation of the operator's response) and the time lag interval occurring before feedback (i.e., knowledge of the effects of the response) begins? Note carefully that the two intervals of interest are formed by the initiation of the stimulus, response, and feedback, e.g.,



Definition		Examples
A > B - Reaction time (A) exceeds feedback lag (B)	7	• Subject listens to a series of digits and repeats them after a 20-second delay.
	6	
	5	
A = B - Reaction time (A) equal to feedback lag (B)	4	• Subject presses button to turn off red light when it comes on. Light goes out when button is pressed.
	3	
A < B - Reaction time (A) is shorter than feedback lag (B)	2	• Subject answers a question on a paper-and-pencil test; gets results at end of test.
	1	

19. DECISION-MAKING

The task instructions guide the operator in producing an output unit. Frequently, the steps leading to the output unit are not of an "A—B—C" nature, but instead they involve choice-points where the operator must decide which of several potential steps should be done next. He bases his choice on the outcome of the last step. For example, the instructions might say, "Press button A and observe the outcome; if a red light comes on, throw the switch. If the blue light comes on, throw the blue switch." The key feature of this situation is that the operator must decide what to do next on the basis of the feedback or outcome of his last response.

Rate the present task on the extent to which it contains choice-points in the steps leading to an output unit.

Definition		Examples
High decision-making - more than 75% of the steps involved in the production of an output unit consist of choice-points.	7	• Trouble shooting a piece of electronic gear
	6	• Diagnosing an illness
	5	
Moderate decision-making- approximately half of the steps involved in the production of an output unit consist of consist of choice-points	4	
	3	
Low decision-making - fewer than 25% of the steps involved in the production at an output unit consist of choice-points.	2	
	1	• Reciting a short verse by memory

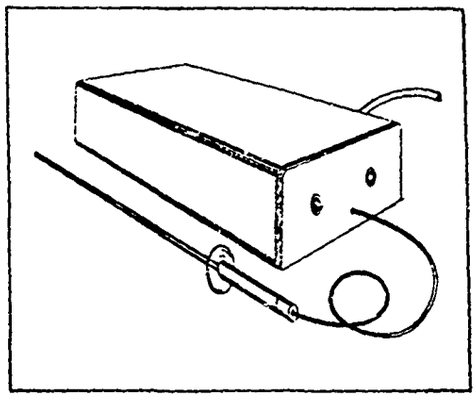
APPENDIX 2

37 TASKS USED IN THE 3-JUDGE STUDY

These tasks were drawn primarily from a study by Fleishman (1954). The raters were provided with a two-page description of each task which contained (a) a picture of the apparatus; (b) a verbal description of the basic task; and (c) the actual instruction read to the subject. Two examples of such tasks are presented in their entirety in this appendix, along with a listing of all 37 tasks by name and source. Double asterisks (**) indicate the subset of 26 tasks which ultimately entered the multiple regression analysis.

TASK 1

Apparatus



Description

The S is seated before a long rectangular boxlike apparatus containing two openings. Each opening is the entrance to a straight passageway which S must negotiate with a long stylus. He moves the stylus forward at slightly below shoulder height and at arm's length. He must move the stylus slowly and steadily away from his body, trying not to hit the sides of the cylindrical passage. As he reaches the end of the passage he strikes a contact point and withdraws the stylus, again trying to avoid hitting any part of the passageway. He then negotiates the second passageway. Two complete negotiations constitute a trial. Counters record the number of contacts and clocks record the amount of time in contact. Six trials, no time limit.

Instructions

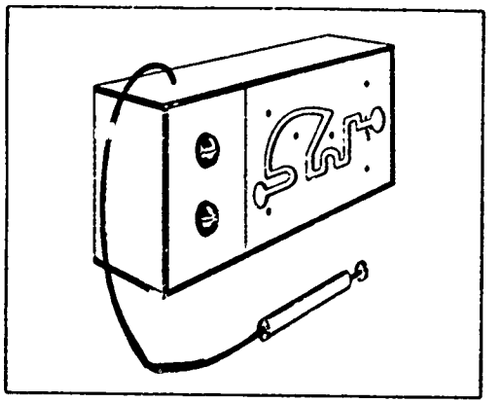
Your task is to move this stylus slowly and carefully arms length through the openings. You are to do this without touching the sides of the passageway with the stylus. When the stylus makes contact with the end of the passageway, withdraw it carefully and slowly without touching the sides. When you have moved the stylus in and out of opening No. 1, move to opening No. 2 and repeat the procedure. After moving in and out of the second passageway, place the stylus beside the machine and rest until told to continue. You will repeat the procedure. Are there any questions?

Remember to keep the stylus at arms length at all times and to move as carefully as possible to reduce errors which is time you contact the sides of the passageway. Begin when I say 'Start'.

 Preceding page blank

TASK 3

Apparatus



Description

The S is required to negotiate an irregular slot pattern with a T-shaped stylus. He sits at arm's length from the apparatus box and moves slowly and steadily through the pattern from right to left, depresses a plunger at the end of the pattern with his stylus, and then returns through the pattern. This constitutes one trial.

Errors are recorded each time any part of the stylus touches the top, bottom, or back of the slot. Four trials, no time limit.

Instructions

Your task is to move the stylus at arms length slowly and carefully through this slot. You are to do this without allowing the stylus to touch the top, bottom, or inside of the slot. Any time the stylus touches any part of the metal plate around the slot, errors will be automatically counted against you. The red light tells you when you are making errors. When you get to the end of the slot, push in on the little plunger with your stylus, and then retrace the pattern without removing the stylus from the slot. When you have completed tracing back through the slot, put your stylus down and place your hand in your lap. Rest until told to begin.

Remember, it is important that you move slow enough so that you may avoid hitting any part of the slot.

Are there any questions?

Pick up the stylus and begin when the green light goes on.

TASK LISTING*

1. Precision Steadiness
2. Steadiness Aiming
3. Tracking Tracing
** 4. Two-Plate Tapping
** 5. Ten-Target Aiming
** 6. Visual Reaction Time
** 7. Minnesota Rate of Manipulation-Turning
** 8. Purdue Pegboard-Right Hand
9. Rotary Pursuit
10. Complex Coordination
**11. Key Tapping
**12. Rotary Aiming
**13. Hand-Precision Aiming
**14. Auditory Reaction Time
**15. Minnesota Rate of Manipulation-Placing
**16. Purdue Pegboard-Two Hands
**17. Purdue Pegboard-Assembly Test
**18. O'Connor Finger Dexterity
**19. Santa Ana Finger Dexterity
**20. Pin Stick
**21. Dynamic Balance
22. Postural Discrimination
23. Postural Discrimination
24. Discrimination Reaction Time
25. Rudder Control
**26. Medium Tapping
**27. Large Tapping
**28. Pursuit Aiming I
**29. Pursuit Aiming II

**30. Aiming

**31. Square Marking

**32. Tracing

33. Steadiness

**34. Discrimination Reaction Time-Printed

**35. Marking Accuracy

**36. Verbal Addition Task[1]

37. Silent Reading Task[2]

---

*Tasks numbered 1 through 35 were abstracted from:

Fleishman, E. A. Dimensional analysis of psychomotor abilities. Journal of Experimental Psychology, 48, 6, 1954, 437-454.

Certain of these tasks (7, 15; 8, 16, 17; and 22, 23) were used more than once as there were different aspects of the tasks which could be scored. This had the net effect of changing the nature and number of the output units and certain of the other characteristics.

**Indicates the 26 tasks which ultimately entered the multiple regression analysis.

[1]This task was abstracted from:

Mech, E. V. Factors influencing routine performance under noise: 1. The influence of "set". Journal of Psychology, 1953, 35, 283-298.

[2]This task was abstracted from:

McGuigan, F. J., & Rodier, W. I. Effects of auditory stimulation on covert oral behavior during silent reading. Journal of Experimental Psychology, 1968, 76, 4, 649-655.

APPENDIX 3

SCALES USED IN THE 28-JUDGE RELIABILITY STUDY

This section contains the 16 scales used in the 28-judge reliability study.

Revision 2. 3/70

TASK CHARACTERISTICS ANSWER SHEET

Rater's Name _____

Date Rating Performed _____

Task Number _____

Instructions

There are 16 rating scales. Each task should be rated on all 16 scales. As you assigned a scale value to the task, write down the scale value on the line for that rating scale as listed below. There is space at the bottom for you to describe any problems you had in applying the scales to the task.

1. Number of output units _____
2. Duration for which an output unit is maintained _____
3. Number of elements per output unit _____
4. Work load _____
5. Precision of responses _____
6. Response rate _____
7. Degree of muscular effort involved _____
8. Simultaneity of responses _____
9. Number of procedural steps _____
10. Dependency of procedural steps _____
11. Variability of stimulus location _____
12. Stimulus or stimulus complex duration _____
13. Regularity of stimulus occurrence _____
14. Operator control of the stimulus _____
15. Operator control of the response _____
16. Rapidness of feedback _____

Problems/Comments

1. NUMBER OF OUTPUT UNITS

The entire purpose of the task is to create output units. An output unit is the end product resulting from the task. Output units can take different forms. For example, sometimes the output unit is a physical object assembled from several parts. It may also take the form of a relationship between two or more things, e.g., drive three car-lengths behind the car in front of you. An output unit might also be a destination, e.g., run from here to the corner, with the corner being the destination.

First, identify what the output unit(s) is in the present task. Now, judge the number of such output units that someone performing this task is supposed to produce.

Definition		Examples
As many as possible - as many output units as possible are to be produced, usually during a fixed period of time.	7	• Insert as many plugs into the connectors as possible in five minutes.
	6	
	5	
Moderate number - a moderate number of output units is to be produced.	4	• Do twenty push-ups in five minutes.
	3	
	2	• Assume a push-up position and maintain it for five minutes. • Do one push-up. • Add the following list of numbers
One output unit - one output unit is to be produced. It is either maintained or it signals the termination of performance.	1	

2. DURATION FOR WHICH AN OUTPUT UNIT IS MAINTAINED

Once the operator has produced an output unit he may be required to maintain or continue it for one of several time periods. For example, it can be maintained for as long as possible. Another alternative is that completing one output unit is a signal to leave it and go on to produce the next output unit. Or, having produced the output unit, performance ends.

Decide where the present output units belong on the below scale.

Definition		Examples
Maintenance for as long as possible - an output unit (body position, stimulus-control relationship, etc.) is to be maintained for as long as possible.	7	• Hang in a bent-arm position for as long as possible.
	6	• Maintain a stimulus-control relationship for 20 minutes.
	5	
Moderate maintenance - relative to other possible periods of maintenance, an output unit is to be maintained for a moderate period of time.	4	• Maintain a stimulus-control relationship for five minutes.
	3	
	2	• Do as many push-ups as possible in ten minutes holding each "down" position for 30 seconds.
Short maintenance - production of an output unit signals the end of performance or the production of additional units. Maintenance, therefore in minimal time.	1	• Solve the following trigonometric problems.

3. NUMBER OF ELEMENTS PER OUTPUT UNIT

One way of describing an output unit is in terms of the number of elements involved in its production. By elements we mean the parts or components which comprise the output unit. In an addition problem, for example, the numbers to be added are the elements which comprise the output unit. In a more physical task, the elements could be parts to be assembled or apparatus to be manipulated.

Rate the present task on the scale below in terms of the number of elements entering into a single output unit.

Definition		Examples
Many elements: each output unit contains many elements.	7	• Assembly a radio from the components in this kit.
	6	
	5	
Moderate number of elements: each output unit contains several elements.	4	• Change a flat tire. • Rank order these 20 items.
	3	
	2	
One element: each output unit contains only one element.	1	• Push the button when the light comes on

4. WORK LOAD

Work load refers to the number of output units to be produced relative to the time allowed for their production. We are interested in the ratio of the number of output units per unit time, e.g., make 5 widgets in 10 minutes = 1 widget produced every two minutes.

However, there are those tasks in which the goal is to maintain a situation rather than to produce multiple output units. For example, a driving task where you are to stay within 40 feet of the vehicle ahead of you. For these types of tasks, work load refers to the length of time for which maintenance is required. The longer the maintenance period, the higher the work load.

Therefore, rating a task in terms of work load resolves to answering one of two questions:

1) How much has to be produced in what amount of time; or
2) How long does this situation have to be maintained or continued?

Definition		Examples
High work load - as many output units as possible are to be produced in a fixed period of time; a relatively large number of output units is to be produced in a relatively short period of time; an output unit is to be maintained for relatively long time or for as long as possible.	7	• Drive as many nails as possible in five minutes. • Maintain a stimulus-control relationship as long as possible.
	6	
	5	
Moderate work load - a moderate number of output units is to be produced in a reasonable period of time; an output unit is to be maintained for a moderate period of time relative to other possible periods.	4	• Drive ten nails in five minutes. • Maintain a stimulus-control relationship for three minutes.
	3	
	2	• Drive these two nails in the next five minutes. • Sum the following five numbers. • Maintain a stimulus-control relationship for 30 seconds.
Low work load - a small number of output units is to be produced in a relatively long period of time; an output unit is to be maintained for a relatively short period of time.	1	

5. PRECISION OF RESPONSES

Tasks may differ in terms of how precise or exact the operator's responses must be. Judge the degree of precision involved in the present task.

Definition		Examples
<u>High degree of precision</u> - because of small targets, fine scales, sensitive controls, etc. the subject must make responses which are extremely precise.	7	• Using a chemical balance (scales) determine the weight of the following objects to the nearest microgram. • Replace the mainspring in this wrist-watch.
	6	
	5	
<u>Moderate precision</u> - relative to the definitions above or below, a moderate degree of precision must accompany subject's responses.	4	• Solder these two wires together. • Using your pencil, trace this maze.
	3	
	2	• Do twenty push-ups. • Sort the oranges and lemons into two piles.
<u>Low degree of precision</u>-because of large targets, gross scales, insensitive controls, etc. the subject can make responses which are gross or imprecise.	1	

6. RESPONSE RATE

Responses can be made at different rates. That is, the frequency with which responses must be made can vary from task to task. For example, you would have a higher rate of responding if you were playing a singles game of tennis than if you were playing chess. The responses would come more frequently in the first case than in the second. You are to judge what rate of responding is called for in the task being judged.

Definition		Examples
<u>High rate of responding</u> - many responses are required per unit time. In the extreme case responses become continuous.	7	• Fire 20 rounds as quickly as possible. • Complete this jig-saw puzzle as fast as you can. • Track this target.
	6	
	5	
<u>Moderate rate of responding</u> - a moderate number of responses are required per unit time.	4	• Fire 20 rounds. Fire rapidly but also be as accurate as you can.
	3	
	2	
<u>Low rate of responding</u> - few responses are emitted per unit time. Responses are often singular.	1	• Add the following numbers. Take all the time you need.

7. DEGREE OF MUSCULAR EFFORT INVOLVED

This dimension considers the amount of muscular effort required to perform the task. Examine the task and identify the most physically strenuous part of it. Rate this part on the scale below.

Definition		Examples
High amount of muscular effort-response(s) require a high degree of muscular involvement.	7	• Do 40 push ups. • Lift the heaviest weight possible.
	6	
	5	
Moderate amount of muscular effort required for the response(s)	4	• Tighten nuts on bolts securely with a wrench.
	3	
	2	
Low amount of muscular effort required	1	• Solder two wires together • Add numbers and report the sum aloud.

8. SIMULTANEITY OF RESPONSES

The responses which the operator makes in producing an output may involve one or more effectors (e.g., hand, foot, arm, voice, etc.). Depending upon the task, these effectors may or may not be used simultaneously. For example, both hands (two effectors) are used simultaneously in playing a piano.

You are to rate the degree of simultaneity involved in using the effectors needed for the response(s).

Definition		Examples
High simultaneity - responses involve the simultaneous use of several effectors.	7	• You are to fly this plane at 400 knots and an altitude of 5,000 feet, banking to the left and to the right. • Play this song on the piano.
	6	
	5	
Moderate simultaneity - responses involve the simultaneous use of at least two effectors.	4	• Pat your head and rub your stomach. • Hit that target by firing your rifle.
	3	
	2	
Low simultaneity - responses involve the use of only one effector at a time. If other effectors are employed, they are employed sequentially.	1	• Push the button when the light comes on.

9. NUMBER OF PROCEDURAL STEPS

Earlier we were concerned about the number of elements, i.e., objects or components, involved in the production of one output unit. Now we want to consider the number of procedural steps (responses) needed to produce one output unit. There isn't a necessary one-to-one relationship between objects and responses.

Consider the number of responses or steps involved in producing one output unit for the present task. Rate this task on the scale below.

Definition		Examples
Large number of steps - the procedure consists of a large number of constituent steps.	7	• Build a color TV kit following the enclosed instructions.
	6	
	5	
Medium number of steps - the procedure contains a medium number of steps relative to other procedures.	4	• Solve the equation $X^2 - 4X \quad 4 = 0$ • Type the following business letter.
	3	
	2	• Open this combination lock (32L-43R-10L). • Press the button whenever the light comes on.
Small number of steps - the procedure consists of few steps. At a minimum, only one step may be necessary.	1	

10. DEPENDENCY OF PROCEDURAL STEPS

Consider again the number of steps (responses) involved in producing one output unit. The steps may be described in terms of the dependency among them; dependency concerns the extent to which the steps must be done in some specified order. For example, dependency exists between steps A and B if step B cannot be accomplished without step A being done first. Note: Procedures which have only one step are automatically low in dependency.

Definition		Examples
High dependency among steps - each step in the procedure is completely dependent upon the preceding procedural step. Systematic ordering of steps is at a maximum.	7	• Using the combination you've been given, open the safe. • Dial this telephone number.
	6	
	5	
Moderate dependency among steps - in the total number of steps comprising the procedure, approximately 50% are dependent upon preceding steps.	4	• Using colored blocks, stack them into columns four blocks high. Do this in the order red and green for the first two blocks. The remaining blocks may be of any color.
	3	
	2	
Low dependency among steps - procedural steps are not organized in any particular sequence. Step "A" may precede "B" or "B" may precede "A". Procedures having one step are low in dependency.	1	• Using colored blocks, stack them into columns four blocks high. Order of color is unimportant.

11. VARIABILITY OF STIMULUS LOCATION

Judge the degree to which the physical location of the stimulus or stimulus complex is predictable over task time.

Definition		Examples
<u>High predictability</u> - stimulus location remains basically unchanged.	7	• Stimulus is a red light located on a display panel.
	6	
	5	
<u>Medium predictability</u> - location changes but in a known manner or pattern.	4	• Visually following an arrow in flight toward a target.
	3	
	2	
<u>Low-predictability</u> - location changes in an almost random fashion.	1	• Predicting which leaf will fall from a tree next.

12. STIMULUS OR STIMULUS COMPLEX DURATION

Consider the critical stimulus or stimulus-complex to which the operator must attend in performing the task. Relative to the total task time, for how long a duration is the stimulus or stimulus-complex present during the task?

Definition		Examples
Long duration - stimulus would remain indefinitely.	7	• Drawing a picture by observing a model of the object being drawn.
	6	
	5	
Medium duration - stimulus remains present until changed (spatially, temporally, etc.) by the response made to it.	4	• Red light goes out when operator pushes a button.
	3	
	2	
Short duration - stimulus ceases prior to response being made to it.	1	• Operator must identify words or targets presented tachistoscopically.

13. REGULARITY OF STIMULUS OCCURRENCE

Consider the critical stimulus or stimulus complex to which the operator must attend. Does it occur at regular (i.e., equal) intervals or at irregular intervals. Treat regular intervals and constant presence of the stimulus as equivalent conditions.

Rate the present task on this dimension.

Definition		Examples
High regularity - stimulus occurs at regular intervals or is constantly present.	7	• Cars coming along an assembly line. • Looking at a photograph of an object.
	6	
	5	
Medium regularity - stimulus occurs at irregular (unequal) intervals but there is a pattern of occurrence.	4	• Receiving morse code.
	3	
	2	
Low regularity - stimulus occurs at very irregular (almost random) intervals.	1	• Detecting random signals on a CRT display.

14. OPERATOR CONTROL OF THE STIMULUS

What degree of control does the operator have over either the occurrence or relevance of the stimulus?

Definition		Examples
Full operator control - the operator is the sole determiner of when the stimulus occurs or when it becomes relevant.	7	• Shooting skeet; shooter determines when "bird" appears.
	6	
	5	
Partial operator control - the operator has some control over when the stimulus either occurs or becomes relevant.	4	• Controlling the speed of your car in approaching a traffic light in order to have a green light when you get to the intersection.
	3	
	2	
No operator control - the operator has no control over when the stimulus occurs or when it becomes relevant.	1	• Waiting for the telephone to ring.

15. OPERATOR CONTROL OF THE RESPONSE

Given the occurrence of the stimulus, what degree of control does the operator have over when he must initiate response?

Definition		Examples
Full operator control - the operator is the sole determiner of when the response will be made.	7	• Playing a game of chess by yourself where you play both sides and there is no time limit for responding.
	6	
	5	
Partial operator control - the response must be made within a reasonable time after the stimulus occurs but the operator determines when within the interval the response will take place.	4	• The traffic light turns red when you are 500 yards from it; you have options as to when you will hit the brake.
	3	
	2	
No operator control - the operator must respond as soon as the stimulus occurs.	1	• Typical reaction time task. When the light comes on, push this button as fast as you can.

16. RAPIDNESS OF FEEDBACK

For present purposes the term FEEDBACK refers to information which an operator may get about the correctness of a response. In this scale we are interested in how quickly feedback occurs once the response is made.

Definition		Examples
Immediate feedback - Operator knows whether the response was correct as soon as it was completed.	7	• Finding the correct switch to turn on a light.
	6	
	5	
Delayed feedback - operator receives feedback regarding his responses after entire task is completed.	4	• Opening a combination lock having five numbers.
	3	
	2	
No feedback provided - Operator never receives feedback	1	• Student takes a mid-term exam but is not told what grade he got.

APPENDIX 4

TASKS USED IN THE 28-JUDGE RELIABILITY STUDY

This section contains the 15 tasks* used by 28 judges in an assessment of the reliability of 16 scales. The information provided on each task consisted of: (a) a picture of the apparatus; (b) a verbal description of the basic task; and (c) the actual instructions read to the subject. Two examples of these tasks are presented in their entirety in this section; the remainder are listed by name along with a reference to the study from which they are abstracted.

---

* The original reliability study employed 20 tasks: 15 psychomotor and 5 paper-and-pencil (cognitive) tasks. The scales proved entirely unreliable for the latter tasks and, hence, these five descriptions are omitted from this section.

TASK LISTING*

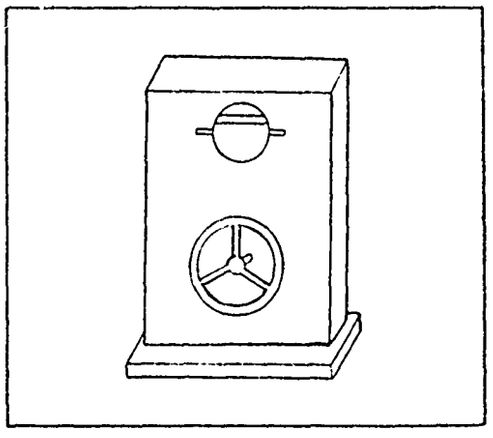
1. Two-Plate Striking
2. Ten Target Aiming
3. Purdue Pegboard
4. Control Sensitivity
5. Two-Hand Coordination
6. Pursuit Confusion
7. Bimanual Matching
8. Visual Reaction Time Test
9. Steadiness Aiming
10. Single Dimension Pursuit
11. Complex Coordination Test
12. Tracking Tracing
13. Rotary Pursuit
14. Precision Steadiness
15. Minnesota Rate of Manipulation

*Descriptions and illustrations of these tasks were abstracted from:
Parker, J. R., Jr., & Fleishman, E. A. Ability factors and component performance measures as predictors of complex tracking behavior. Psychological Monograph, 1960, 74, No. 503.

Preceding page blank

TASK 10

Apparatus



Description

The subject makes compensatory adjustments (in and out movements) of a control wheel in order to keep a horizontal line in a null position as it deviates from center in irregular fashion. The control wheel is damped pneumatically, introducing a lag into the system. Score is the time the horizontal line is held in a null position during the four 1-minute trials.

Instructions

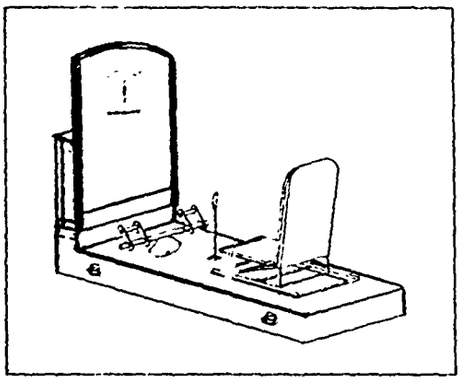
In this test your job is to keep this white line inside the circle centered between these two points. When the test starts, the line will start to move out of position. Your task is to keep the line centered as it deviates from the center. You can move the line up by pulling out on this wheel and you can move it down by pushing in on the wheel. Rotating the wheel has no effect. Your score will be the total time you are able to keep the white line centered.

READY?

BEGIN?

TASK 11

Apparatus



Description

The S is required to make complex motor adjustment of stick and pedal controls in response to successively presented patterns of visual signals.

A correct response (movement of stick and rudder controls to proper positions) is not accomplished until both the hands and feet have completed and maintained the appropriate adjustment. A new pattern appears as each correct response is completed. Score is the number of completed matchings. Four 2-minute test periods.

Instructions

Your task will be to line up a green light with each of the three red lights. Moving the stick from side to side moves the top green light. Moving the stick forward and backward moves the middle green light; and moving the rudder bar moves the bottom green light. Move the stick sideways to match the top green light with the top red light. Get it directly underneath. If it is off to one side like this it will not work. Then hold the stick in position to keep the top lights matched while you move it forward or backward to match the middle lights. Then hold the stick steady while you match the bottom lights with the rudder bar.

When you have matched all three lights, a new setting of red lights will appear. Go right ahead and match the new setting of red lights without bothering to come back to neutral.

TASK 11 (Continued)

If you move any of the controls as far as it will go there will be no green light. You must ease back a bit to find the end green light.

When the test starts, you may use either your right or left hand on the stick, but use only one hand throughout the test. Keep your heels off the floor. Match as many settings of the lights as you can until go out. If the red lights ever fail to come on, let me know immediately.

Your score will be the number of matchings you can make in the time allowed. Work as rapidly as you can. When the buzzer sounds, the test period begins. When all the lights go out again, the test will be over.

APPENDIX 5

SCALES USED IN THE 2-JUDGE STUDY

This section contains the 18 scales used in the 2-judge study. Asterisks identify the subset of these scales which were ultimately entered into the multiple regression analysis.

TASK CHARACTERISTICS ANSWER SHEET

Rater_____

Study No._____ Author:_____

Date_____

Type Task_____

*1. Number of output units_____

2. Duration_____

*3. Number of elements/output unit_____

4. Work load_____

*5. Precision of responses_____

6. Response rate_____

7. Tutorial dependency_____

8. Natural dependency_____

9. Operator control over response_____

*10. Simultaneity of responses_____

*11. Number of responses_____

*12. Number of procedural steps_____

13. Feedback_____

14. Degree of muscular effort_____

15. Operator control over stimulus_____

16. Regularity of stimulus occurrence_____

17. Stimulus duration_____

18. Variability of stimulus location_____

 Preceding page blank

*1. NUMBER OF OUTPUT UNITS (UNIT)

The entire purpose of the task is to create output units. An output unit is the end product resulting from the task. Output units can take different forms. For example, sometimes the output unit is a physical object assembled from several parts. It may also take the form of a relationship between two or more things, e.g., drive three car-lengths behind the car in front of you. An output unit might also be a destination, e.g., run from here to the corner, with the corner being the destination.

First, identify what the output unit(s) is in the present task. Now, count the number of such output units that someone performing this task is supposed to produce. Use the designation AMAP (As many as possible) where no actual limit exists.

2. DURATION FOR WHICH AN OUTPUT UNIT IS MAINTAINED (DURA)

Once the operator has produced an output unit he may be required to maintain or continue it for one of several time periods. For example, it can be maintained for as long as possible. Another alternative is that completing one output unit is a signal to leave it and go on to produce the next output unit. Or, having produced the output unit, performance ends.

Choose which of the following alternatives applies here:

1) Maintain unit as long as possible.

2) Maintain unit as long as possible but continue to produce additional units.

3) Leave unit and go on to produce next unit.

4) Production of unit signals end of task.

3. NUMBER OF ELEMENTS PER OUTPUT UNIT (ELEM)

One way of describing an output unit is in terms of the number of elements involved in its production. By elements we mean the parts or components which comprise the output unit. In an addition problem, for example, the numbers to be added are the elements which comprise the output unit. In a more physical task, the elements could be parts to be assembled or apparatus to be manipulated.

Count the number of different displays and controls which are manipulated in producing a single output unit.

4. WORK LOAD (LOAD)

Work load refers to the number of output units to be produced relative to the time allowed for their production. We are interested in the ratio of the number of output units per unit time, e.g., make 5 widgets in 10 minutes = 1 widget produced every two minutes.

However, there are those tasks in which the goal is to maintain a situation rather than to produce multiple output units. For example, a driving task where you are to stay within 40 feet of the vehicle ahead of you. For these types of tasks, work load refers to the length of time for which maintenance is required. The longer the maintenance period, the higher the work load.

Therefore, rating a task in terms of work load resolves to answering one of two questions:

1) How much has to be produced in what amount of time; or
2) How long does this situation have to be maintained or continued?

Definitions		Examples
High work load - as many output units as possible are to be produced in a fixed period of time; a relatively large number of output units is to be produced in a relatively short period of time; an output unit is to be maintained for a relatively long time or for as long as possible.	7	• Drive as many nails as possible in five minutes. • Maintain a stimulus-control relationship as long as possible.
	6	
	5	
Moderate work load - a moderate number of output units is to be produced in a reasonable period of time; an output unit is to be maintained for a moderate period of time relative to other possible periods.	4	• Drive ten nails in five minutes. • Maintain a stimulus-control relationship for three minutes.
	3	
	2	• Drive these two nails in the next five minutes. • Sum the following five numbers. • Maintain a stimulus-control relationship for 30 seconds.
Low work load - a small number of output units is to be produced in a relatively long period of time; an output unit is to be maintained for a relatively short period of time.	1	

5. PRECISION OF RESPONSES (PREC)

Tasks may differ in terms of how precise or exact the operator's responses must be. Judge the degree of precision involved in the present task by considering the most precise response made in producing an output unit.

Definitions		Examples
High degree of precision - because of small targets, fine scales, sensitive controls, etc. the subject must make responses which are extremely precise.	7	• Using a chemical balance (scales) determine the weight of the following objects to the nearest microgram. • Replace the mainspring in this wrist-watch.
	6	
	5	
Moderate precision - relative to the definitions above or below, a moderate degree of precision must accompany subject's responses.	4	• Using your pencil, trace this maze.
	3	
	2	• Do twenty push-ups. • Sort the oranges and lemons into two piles.
Low degree of precision-because of large targets, gross scales, insensitive controls, etc. the subject can make responses which are gross or imprecise.	1	

R. RESPONSE RATE (RATE)

Responses can be made at different rates. That is, the frequency with which responses must be made can vary from task to task. For example, you would have a higher rate of responding if you were playing a singles game of tennis than if you were playing chess. The responses would come more frequently in the first case than in the second. You are to judge what rate of responding is called for in producing one output unit in the task being judged.

Definitions		Examples
High rate of responding - many responses are required per unit time. In the extreme case responses become continuous.	7	• Fire 20 rounds for effect as quickly as possible. • Complete this jig-saw puzzle as fast as you can. • Track this target.
	6	
	5	
Moderate rate of responding - a moderate number of responses are required per unit time.	4	• Fire 20 rounds. Fire rapidly but also be as accurate as you can.
	3	
	2	
Low rate of responding - few responses are emitted per unit time. Responses are often singular.	1	• Add the following numbers. Take all the time you need.

7. TUTORIAL DEPENDENCY OF RESPONSES (TUDE)

Consider again the number of steps (responses) involved in producing one output unit. The steps may be described in terms of the dependency among them; dependency concerns the extent to which the steps must be done in some specified order. For example, dependency exists between steps A and B if step B cannot be accomplished without step A being done first. Note Procedures which have only one step are automatically low in dependency. Tutorial dependency refers to a dependency imposed as part of the training in an effort to standardize trainee operations.

Definitions		Examples
High dependency among steps - each step in the procedure is completely dependent upon the preceding procedural step. Systematic ordering of steps is at a maximum.	7	• Using the combination you've been given, open the safe. • Dial this telephone number.
	6	
	5	
Moderate dependency among steps - in the total number of steps comprising the procedure, approximately 50% are dependent upon preceding steps.	4	• Using colored blocks, stack them into columns four blocks high. Do this in the order red and green for the first two blocks. The remaining blocks may be of any color.
	3	
	2	
Low dependency among steps - procedural steps are not organized in any particular sequence. Step "A" may precede "B" or "B" may precede "A". Procedures having one step are low in dependency.	1	• Using colored blocks, stack them into columns four blocks high. Order of color is unimportant.

8. NATURAL DEPENDENCY OF RESPONSES (NADE)

Consider again the number of steps (responses) involved in producing one output unit. The steps may be described in terms of the dependency among them; dependency concerns the extent to which the steps must be done in some specified order. For example, dependency exists between Steps A and B if step B cannot be accomplished without step A being done first. Note: Procedures which have only one step are automatically low in dependency. Natural dependency refers to dependency that is inherent in the operation of the equipment.

Definitions		Examples
High dependency among steps - each step in the procedure is completely dependent upon the preceding procedural step. Systematic ordering of steps is at a maximum.	7	• Using the combination you've been given, open the safe. • Dial this telephone number.
	6	
	5	
Moderate dependency among steps - in the total number of steps comprising the procedure, approximately 50% are dependent upon preceding steps.	4	• Using colored blocks, stack them into columns four blocks high. Do this in the order red and green for the first two blocks. The remaining blocks may be of any color.
	3	
	2	
Low dependency among steps - procedural steps are not organized in any particular sequence. Step "A" may precede "B" or "B" may precede "A". Procedures having one step are low in dependency.	1	• Using colored blocks, stack them into columns four blocks high. Order of color is unimportant.

9. OPERATOR CONTROL OF THE RESPONSE (OCOR)

Given the occurrence of the stimulus, what degree of control does the operator have over when he must initiate his response.

Definitions		Examples
Full operator control - the operator is the sole determiner of when the response will be made.	7	• Playing a game of chess by yourself where you play both sides and there is no time limit for responding.
	6	
	5	
Partial operator control - the response must be made within a reasonable time after the stimulus occurs but the operator determines when within the interval the response will take place.	4	• The traffic light turns red when you are 500 yards from it; you have options as to when you will hit the brake.
	3	
	2	
No operator control - the operator must respond as soon as the stimulus occurs.	1	• Typical reaction time task. When the light comes on, push this button as fast as you can.

*10. SIMULTANEITY OF RESPONSES (SIMU)

The responses which the operator makes in producing one output unit may involve one or more effectors (e.g., hand, foot, arm, voice, etc.). Depending upon the task, these effectors may or may not be used simultaneously. For example, both hands (two effectors) are used simultaneously in playing a piano.

How many effectors are being used simultaneously during the present task?

zero_____ two_____ three_____ four_____

*11. NUMBER OF RESPONSES (NO. R)

Earlier we were concerned about the number of elements, i.e., objects or components, involved in the production of one output unit. Now we want to consider the number of responses needed to produce one output unit. There isn't a necessary one-to-one relationship between objects and responses.

Count the number of responses or steps involved in producing one output unit for the present task. Enter this number on the answer sheet.

12. NUMBER OF PROCEDURAL STEPS

Earlier we were concerned about the number of elements, i.e., objects or components, involved in the production of one output unit. Now we want to consider the number of procedural steps (responses) needed to produce one output unit. There isn't a necessary one-to-one relationship between objects and responses.

Consider the number of responses or steps involved in producing one output unit for the present task. Rate this task on the scale below.

Definitions		Examples
Large number of steps - the procedure consists of a large number of constituent steps.	7	• Build a color TV kit following the enclosed instructions.
	6	
	5	
Medium number of steps - the procedure contains a medium number of steps relative to other procedures.	4	• Solve the equation $X^2 - 4X \quad 4 = 0$ • Type the following business letter.
	3	
	2	• Open this combination lock (32L-43R-10L). • Press the button whenever the light comes on.
Small number of steps - the procedure consists of few steps. At a minimum, only one step may be necessary.	1	

13. FEEDBACK (FEED)

For present purposes the term FEEDBACK refers to information which an operator may get about the correctness of a response. In this scale we are interested in how quickly feedback occurs once the response is made.

Definitions		Examples
Immediate feedback - Operator knows whether the response was correct as soon as it was completed.	7	• Finding the correct switch to turn on a light.
	6	
	5	
Delayed feedback - operator receives feedback regarding his responses after entire task is completed.	4	• Opening a combination lock having five numbers.
	3	
	2	
No feedback provided - Operator never receives feedback	1	• Student takes a mid-term exam but is not told what grade he got.

14. DEGREE OF MUSCULAR EFFORT INVOLVED (MUSC)

This dimension considers the amount of muscular effort required to perform the task. Examine the task and identify the most physically strenuous part of it. Rate this part on the scale below.

Definitions		Examples
High amount of muscular effort-response(s) require a high degree of muscular involvement.	7	• Do 40 push ups. • Lift the heaviest weight possible.
	6	
	5	
Moderate amount of muscular effort required for the response(s)	4	• Tighten nuts on bolts securely with a wrench.
	3	
	2	
Low amount of muscular effort required	1	• Solder two wires together • Add numbers and report the sum aloud.

15. OPERATOR CONTROL OF THE STIMULUS (OCOS)

What degree of control does the operator have over either the occurrence or relevance of the stimulus?

Definitions		Examples
Full operator control - the operator is the sole determiner of when the stimulus occurs or when it becomes relevant.	7	• Shooting skeet; shooter determines when "bird" appears.
	6	
	5	
Partial operator control - the operator has some control over when the stimulus either occurs or becomes relevant.	4	• Controlling the speed of your car in approaching a traffic light in order to have a green light when you get to the intersection.
	3	
	2	
No operator control - the operator has no control over when the stimulus occurs or when it becomes relevant.	1	• Waiting for the telephone to ring.

16. REGULARITY OF STIMULUS OCCURRENCE (ROSO)

Consider the critical stimulus or stimulus complex to which the operator must attend. Does it occur at regular (i.e., equal) intervals or at irregular intervals. Treat regular intervals and constant presence of the stimulus as equivalent conditions.

Rate the present task on this dimension.

Definitions		Examples
High regularity - stimulus occurs at regular intervals or is constantly present.	7	• Cars coming along an assembly line. • Looking at a photograph of an object.
	6	
	5	
Medium regularity - stimulus occurs at irregular (unequal) intervals but there is a pattern of occurrence.	4	• Receiving morse code.
	3	
	2	
Low regularity - stimulus occurs at very irregular (almost random) intervals.	1	• Detecting random signals on a CRT display.

17. STIMULUS OR STIMULUS COMPLEX DURATION (SDUR)

Consider the critical stimulus or stimulus-complex to which the operator must attend in performing the task. Relative to the total task time, for how long a duration is the stimulus or stimulus-complex present during the task?

Definitions		Examples
Long duration - stimulus would remain indefinitely.	7	• Drawing a picture by observing a model of the object being drawn.
	6	
	5	
Medium duration - stimulus remains present until changed (spatially, temporally, etc.) by the response made to it.	4	• Red light goes out when operator pushes a button.
	3	
	2	
Short duration - stimulus ceases prior to response being made to it.	1	• Operator must identify words or targets presented tachistoscopically.

18. VARIABILITY OF STIMULUS LOCATION (VARS)

Judge the degree to which the physical location of the stimulus or stimulus complex is predictable over task time.

Definitions		Examples
High predictability - stimulus location remains basically unchanged.	7	• Stimulus is a red light located on a display panel.
	6	
	5	
Medium predictability - location changes but in a known manner or pattern.	4	• Visually following an arrow in flight toward a target.
	3	
	2	
Low-predictability - location changes in an almost random fashion.	1	• Predicting which leaf will fall from a tree next.

APPENDIX 6

TASKS USED IN 2-JUDGE STUDY

The judges in this study rated tasks appearing in a number of published articles. In each case, their attention was directed toward the method section, focusing on the apparatus and instructions.

A list of the references so viewed is provided in lieu of descriptions of the tasks themselves.

REFERENCES USED IN SECOND POST-DICTION STUDY

Adams, J. A. Psychomotor performance as a function of intertrial rest interval. Journal of Experimental Psychology, 1954, 48, 131-133.

Archer, E. J., Kent, G. W., & Mote, F. A. Effect of long-term practice and time-on-target information feedback on a complex tracking task. Journal of Experimental Psychology, 1956, 51, 103-112.

Bilodeau, E. A. Some effects of various degrees of supplemental information given at two levels of practice upon the acquisition of a complex motor skill. Research Bulletin 52-15, April 1952, Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas.

Birren, J. E., & Fisher, M. B. Standardization of two tests of hand-eye coordination: A two-hand complex tapping test and a rotary pursuit test. Research Project X-293, Report No. 6, 1945, NMRI, Bethesda, Maryland.

Briggs, G. E., Fitts, P. M., & Bahrick, H. P. Learning and performance in a complex tracking task as a function of visual noise. Research Report AFPTRC-TN-56-67, June 1956, Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas.

Brown, C. W., Ghiselli, E. E., Jarrett, R. F., Minium, E. W., & U'Ren, R. M. Comparison of aircraft controls for prone and seated position in three-dimensional pursuit task. AF Technical Report No. 5956, October 1949, U. S. Air Force Air Materiel Command, Wright-Patterson Air Force Base, Dayton, Ohio.

Cook, B. S., & Hilgard, E. R. Distributed practice in motor learning: Progressively increasing and decreasing rests. Journal of Experimental Psychology, 1949, 39, 169-172.

Dore, L. R., & Hilgard, E. R. Spaced practice and maturation hypothesis. Journal of Psychology, 1937, 4, 245-259.

Fleishman, E. A. Unpublished data on two-hand coordinator.

Fleishman, E. A., & Rich, S. Role of kinesthetic and special-visual abilities in perceptual-motor learning. Journal of Experimental Psychology, 1963, 66, 6-11.

Preceding page blank

Gagne, R. M., & Bilodeau, E. A. The effects of target size variation on skill acquisition. Research Bulletin AFPTRC-TR-54-5, April 1954, Air Force Personnel and Training Research Center, Lackland Air Force Base, San Antonio, Texas.

Goldstein, M., & Rittenhouse, C. H. The effects of practice with triggering omitted on performance of the total pedestal sight gunnery task. Technical Report 53-9, May 1953, Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas.

Howland, D., & Merrill, E. N. The effect of physical constants of a control on tracking performance. Journal of Experimental Psychology, 1953, 46, 353-360.

Lewis, D., & Shephard, A. H. Devices for studying associative interference in psychomotor performance. IV. The turret pursuit apparatus. Journal of Psychology, 1950, 29, 173-182.

Lincoln, R. S. Learning and retaining a rate of movement with the aid of kinesthetic and verbal cues. Journal of Experimental Psychology, 1956, 51, 199-204.

Noble, C. E. An attempt to manipulate incentive motivation in a continuous tracking task. Research Bulletin AFPTRC-TR-54-43, October 1954, Air Force Personnel and Training Research Center, Lackland Air Force Base, San Antonio, Texas.

Reynolds, B., & Adams, J. A. Effect of distribution and shift in distribution of practice within a single training session. Journal of Experimental Psychology, 1953, 46, 137-145.

Reynolds, B., & Bilodeau, I. M. Acquisition and retention of three psychomotor tests as a function of distribution of practice during acquisition. USAF Human Resources Research and Development, Lackland Air Force Base, San Antonio, Texas.

Spieth, W. An investigation of individual susceptibility to interference in the performance of three psychomotor tasks. Research Bulletin 53-8, April 1953, Human Resources Research Center, Lackland Air Force Base, San Antonio, Texas.*

*This study yielded two groups and, hence, two sets of learning data for the post-diction study.

END