8

Development of a Taxonomy of Human Performance:

Evaluation of a Task Classification System For Generalizing Research Findings from a Data Base

(1)

Warren H. Teichner

Joan Whitehead

Technical Report 8

APRIL 1971

AIR
AMERICAN INSTITUTES FOR RESEARCH
WASHINGTON OFFICE
Address: 8555 Sixteenth Street, Silver Spring, Maryland 20910
Telephone: (301) 587-8201

R71-11

43

AMERICAN INSTITUTES FOR RESEARCH

WASHINGTON, D.C.

EDWIN A. FLEISHMAN, PhD, DIRECTOR

Albert S. Glickman, PhD, 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.

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: EVALUATION OF A TASK CLASSIFICATION SYSTEM FOR GENERALIZING RESEARCH FINDINGS FROM A DATA BASE

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

Interim Technical Report

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

Warren H. Teichner and Joan Whitehead

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1971	40	11

8a. CONTRACT OR GRANT NO. F44620-67-C-0116 (AFOSR) DAHC19-71-C-0004 (Army-BESRL) b. PROJECT NO. c. ARPA Order Number 1032 ARPA Order Number 1623 d.	9a. ORIGINATOR'S REPORT NUMBER(S) AIR-726/2035-4/71-TR8 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) BESRL Research Study 71-8

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

This publication reports on a research effort to assess the feasibility of constructing a data base for improving generalization of research results about human performance.

A "criterion measure" task classification system was applied to a portion of the existing literature on learning and environmental variables. "Optimum distribution of practice" and "knowledge of results" were the two learning variables investigated. The environmental factor investigated was "the effects of different noise intensities."

It was shown that for certain variables and certain task conditions the categorization system was effective in predicting human performance across a variety of tasks. Implications for developing a data base are described.

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Data Base						
Taxonomy						
Task Classification						
Human Performance						

AIR-726/2035-4/71-TR8

AD

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: EVALUATION OF A TASK CLASSIFICATION SYSTEM FOR GENERALIZING RESEARCH FINDINGS FROM A DATA BASE

Warren H. Teichner
Joan Whitehead

TECHNICAL REPORT 8

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

April 1971

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.

The present publication reports on an effort to evaluate the usefulness of a system for improving the extent to which research findings about task performance can be generalized. A "criterion measure" classification system was applied to existing data concerned with selected training and environmental variables. It was shown that for certain variables and certain task conditions the categorization system was effective in predicting human performance across a variety of tasks. Implications for developing a data base are described.

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

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: EVALUATION OF A TASK CLASSIFICATION SYSTEM FOR GENERALIZING RESEARCH FINDINGS FROM A DATA BASE

BRIEF

Requirement:

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 research was undertaken to assess the feasibility of constructing a data base founded on a "criterion measure" task classification system, which could improve generalizations of research results about human performance.

Procedure:

The purpose of the present report is to present some early findings in applying one task classification system to a portion of the existing literature on learning and environmental effects. The two learning variables investigated were "optimum distribution of practice," and "knowledge of results"; the environmental factor investigated was "the effects of different noise intensities."

Segments of the literature on human performance were collected and evaluated for their adequacy as data sources. A particular task classification system was then applied and the data within each class collated and expressed in terms of the functional relationships identified. To the degree that these steps can be taken, the feasibility of a human performance data base may be said to be established for the literature used and the classification system employed, and encouragement provided for more extensive and more complex efforts.

The task classification system used was that provided by the approach of Teichner and Olson (1969) to the establishment of functional relationships between task variables and dependent measures of performance. In the present project, this approach has been called the "Criterion Measure" approach to task classification. In general, Teichner and Olson (1969) defined classes of task performance by dependent measures. For example, one class of performance, called switching, was defined by measures indicating the latency of the operator's response; another, called coding, was defined by the percent of correct responses made by the operator in task performance.

The approach used was ideally suited to the present purpose since it provided a small set of operationally-defined task classes, it required a minimum of qualifications in order to classify the tasks used in the literature, and because the approach was designed for expression in terms of relationships between variables known to have received considerable study.

The literature base to which the classification system was applied consisted of three sets of experimental reports from the scientific literature included in the human performance data base developed in the project. In each case it was necessary to evaluate the paper for (a) sufficient precision of description of tasks and procedures, and (b) experimental adequacy. If the paper was found adequate on these counts, it was classified into the "Criterion Measure" categories.

Findings:

Of those literatures sampled, two ("knowledge of results" and "effects of noise") did not appear to contain enough studies of a reliability sufficient for the purposes to which a data base might be put. This conclusion is quite independent of the task classification system. The only one of the three literatures which does appear to be useful, after evaluation of individual studies, is that concerned with massed and distributed practice.

The task classification system was applicable to the studies surveyed regardless of area. The system appeared to be a feasible one. This is a general conclusion based upon ease of application. With the system it was possible to organize the literature on distributed practice in terms of: (a) functional relationships and (b) different functions for different task categories. In fact, some hitherto unreported relationships were strongly suggested. It is important to note that these "principles" are general to operationally-defined task categories where each category contains a variety of different tasks.

Utilization of Findings:

Both the method and the distributed practice literature are useful for data base purposes. Other segments of the human performance literature are probably also useful and amenable to this classification method. How far its utility will go remains to be determined empirically. On the other hand, other classification systems can now be applied to the distributed practice literature and can now be evaluated against this one. It is possible that other systems will not survive the test of application, or they might be even more successful, or they might serve to reveal still other kinds of relationships. Regardless, important results of the present study are (a) the identification of a usable literature, (b) the reduction of its studies to those that are reasonably acceptable on scientific grounds, and (c) the identification of principles of learning relating practice schedules and performance change for a variety of human tasks.

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: EVALUATION OF A TASK CLASSIFICATION SYSTEM FOR GENERALIZING RESEARCH FINDINGS FROM A DATA BASE

CONTENTS

	Page
INTRODUCTION	1
OBJECTIVES	3
METHOD	4
Selection of the Task Classification System	4
The Literature Base	6
RESULTS	8
Application of the Taxonomy to Massed Versus Distributed Practice	8
Application of the Taxonomy to Knowledge of Results Studies	18
Application of the Taxonomy to Studies on the Effects of Noise	21
DISCUSSION AND CONCLUSIONS	27
REFERENCES	31
DD Form 1473 (Document Control Data - R&D)	3[illegible]

FIGURES

Figure 1.	Distribution of qualitative effects of distributed practice reported for 111 experimental comparisons	9
2.	Percentage change in performance as a function of intertrial interval for simple coding tasks	11
3.	Percentage change in performance as a function of mean intertrial interval (a) and ratio of intertrial intervals (b) for 61 experimental comparisons among simple coding tasks	12
4.	Percentage change in performance as a function of intertrial interval for successive coding tasks	14
5.	Percentage change in performance as a function of mean intertrial interval for 23 experimental comparisons among successive coding tasks	15
6.	Percentage change in performance as a function of intertrial interval for tracking tasks	17

FIGURES | Page

Figure 7. Percentage change in performance as a function of the ratio of massed to distributed inter-trial intervals for tracking tasks 19

8. Distribution of the qualitative effects of knowledge of results for 60 experimental comparisons 22

9. Distribution of the qualitative effect of noise reported for 44 experimental comparisons 24

10. The effects of noise on performance 26

DEVELOPMENT OF A TAXONOMY OF HUMAN PERFORMANCE: EVALUATION OF A TASK CLASSIFICATION SYSTEM FOR GENERALIZING RESEARCH FINDINGS FROM A DATA BASE

INTRODUCTION

For many years, the effects of training and environmental conditions on human performance have been studied with a great variety of tasks. Vast quantities of data have been accumulated. Yet, as we have pointed out elsewhere (e.g., Fleishman, 1967), when new systems are conceived for defense, exploration of space, etc., it appears very difficult to apply the accumulated data and experience of the past. The problems of skill identification training, and performance for the new tasks must frequently be restudied.

The problem is not only one of generalizing principles from one operational system to another. It also involves the generalization of findings from laboratory tasks to operational tasks. It is difficult to integrate results from several laboratory studies investigating the same learning and environmental factors due to differences in the tasks involved in these studies. Tasks selected in laboratory research are not often based on any clear rationale about the class of task or skill represented. One reason why much of current research on learning in the experimental laboratory is difficult to apply to real-life training situations is the absence of information on the relevant common task dimensions. This is also true, of course, for laboratory studies of the effects of environmental factors, drugs,and other variables. What is needed is a learning and performance theory which ascribes task dimensions a central role (Fleishman, 1967).

The current project has as one objective the development and evaluation of descriptive systems which could improve generalizations of research results about human performance. It is hoped that a common task descriptive language could be developed which would (a) help integrate much of the human performance information in the current literature, and (b) allow better communication between researchers and individuals who need to apply research to applied problems. The assumption is that the world of human tasks is not impossibly diverse and that common task dimensions can be identified which will allow improved predictions of human performance on these tasks.

Earlier reviews (Fleishman, 1967; Wheaton, 1968; Farina, 1969) have indicated a variety of task descriptive systems varying from the highly detailed and specific task descriptions of the job and system to the general categories frequently seen in the experimental literature (e.g., motor vs. cognitive skills). It was concluded that such highly specific or highly general categories are not likely to be useful in generalizing principles across tasks. Also, it was found that no empirical evaluations had actually been made of the extent to which various descriptive systems have been useful in improving predictions and generalizations about factors affecting human performance.

The present project has proceeded along several lines. First, a number of taxonomic systems are under development, based on some rationale about common factors in task performance. Examples are the "ability requirements approach" (Fleishman, 1967; Theologus, Romashko, and Fleishman, 1970; Theologus and Fleishman, 1971), the "task characteristics approach" (Farina and Wheaton, 1971), the "information-theoretic approach" (Levine and Teichner, 1971), and the "task strategies approach" (Miller, 1971).

A second line of work has been the development of evaluative systems for testing the reliability and utility of these approaches. For example, observer ratings using scales based on abilities have had some success in predicting empirical factor loadings as well as in predicting performance levels on tasks in various categories (Theologus and Fleishman, 1971). In addition, the task characteristic approach has had some success in predicting performance levels on a variety of tasks (Farina and Wheaton, 1971).

A third line of work has involved the development of a human performance data base for evaluating the effect of provisional taxonomic systems in integrating the experimental literature. The basic notion here is that a taxonomic system should be translatable into an indexing system which allows entry into the available literature in such a way that the tasks used in a large variety of studies can be classified (Chambers, 1969; Korotkin and Chambers, 1969). The data with respect to these task categories can then be examined for consistencies between and within classes. Do alternate systems improve the kinds of generalizations that can be made about the performance effects of certain variables of interest? If such systems could be developed, especially if they are made computer compatible, there would be important implications for retrieving principles of human performance applicable to current and future tasks.

OBJECTIVES

The purpose of the present report is to present some early findings in applying one task classification system to a portion of the existing literature on learning and environmental effects contained in our human performance data base. The two learning variables investigated were "optimum distribution of practice," and "knowledge of results"; the environmental factor investigated was "the effects of different noise intensities".

As implied earlier, the success of computer technology in the organization of data for use in complex management systems suggests application to the use of the enormous available store of scientific information concerning human performance. If those data were available in appropriately coded form, the data base so formed might serve as a primary source of management decisions concerning personnel selection, training, and equipment design. Such a data base might also provide a means for the discovery of previously unknown relationships fundamental to those decisions since, once available, the data could be collated in novel ways and entered into complex mathematical models.

Whether or not such a system is feasible depends upon: (a) the relevance of the literature for the purposes indicated, (b) the amenability of the literature to quantification of its data, (c) the consistency of the results reported, and (d) the utility of the system used for classifying or coding the data for entry into decision-making systems. The present study was a joint test of all of these aspects of feasibility.

Segments of the literature on human performance were collected and evaluated for their adequacy as data sources. A particular task classification system was then applied and the data within each class collated and expressed in terms of functional relationships identified. To the degree that these steps can be taken, the feasibility of a human performance data base may be said to be established for the literature used and the classification system employed, and encouragement provided for more extensive and more complex efforts.

METHOD

Selection of the Task Classification System

Many of the problems and possibilities for a taxonomy of human performance have been reviewed and evaluated in earlier project reports. In the present study our primary concern was not directed to the ultimate value of any particular classification system, but rather to the selection of that one of a variety of possibilities which might provide the greatest ease of application to the research literature. Such a system appeared to have been provided by the approach of Teichner and Olson (1969) to the establishment of functional relationships between task variables and dependent measures of performance. In the present project, this approach has been called the "Criterion Measure" approach to task classification.

In general, Teichner and Olson (1969) defined classes of task performance by dependent measures. For example, one class of performance, called switching, was defined by measures indicating the latency of the operator's response; another, called coding, was defined by the percent of correct responses made by the operator in task performance. In each case, a small number of tentative subclassifications were defined by differences in operational conditions. It was assumed that further subclassifications would develop empirically as the result of attempts to collate the results of studies into a single class, i.e., those studies within a class which could be expressed by the same relationships would be defined as the same in kind; those that required different relationships would be defined as a different subclass.

The approach used by Teichner and Olson was ideally suited to the present purpose since it provided a small set of operationally-defined task classes, it required a minimum of qualifications in order to classify the tasks used in the literature, and because the approach was designed for expression in terms of relationships between variables known to have received considerable study. In the present study, the tasks in the literature selected for study were classified according to the "Criterion Measure" classification system described by Teichner and Olson (1969).

Specifically, each study reviewed was classified into one or the other of the following four of their six primary categories:

"Searching:" The exposure of a sensor to positionally different signal sources or to one source at different times. Searching is receptor orienting or signal seeking. It may be <u>simple orienting</u> as when the ears are positioned to enhance reception of a novel stimulus, or <u>successive orienting</u>, also called scanning. Examples are monitoring, reconnaissance, target seeking. The descriptive measure that will be employed is the <u>probability of detection</u>.

"Switching:" A discrete action which changes the state of the next component in a system. Examples are turning anything on or off, go or no-go, or, in general, making a discrete, selective action involving categorical choices. In a system sense, switching should be described as the time between the initiation of the signal and the completion of the switching response. However, this time will depend critically on the characteristics of the switch that is used. Thus, movement time will be longer the longer the required switch movement, the greater the required torque, etc. Since these factors cannot be anticipated, they must be estimated from specific analysis of the system of interest. Aside from these factors, switching responses vary in the time from the initiation of the signal to the <u>initiation</u> of the response, that is, in <u>reaction time</u>. Therefore, the reaction time or latency is the descriptive measure that will be used to describe switching.

"Coding:" The naming or identifying of a detected signal. <u>Simple coding</u> involves the attachment of a name to characteristics of a stimulus such as color, pitch, direction of movement, position, etc. <u>Group coding</u> refers to the grouping of stimulus characteristics into a single classification such as silverware for knives, spoons, and forks, or "John" for a person, or "attack" for a battle procedure, etc. <u>Successive coding</u> implies a syntax or set of rules which is used to relate or transform names or codes. Examples are translating language and computing. The descriptive measure to be used is the percent of correctly coded responses or equivalent, such as the percent of error.

"Tracking:" Alignment of a response with a changing input. Tracking may be pursuit or compensatory as conventionally used. Examples of tracking are steering, aiming, walking, tuning. The measure to be used will be the percentage decrement in time on target. The use of a relative measure is dictated by the fact, as with switching, that actual time on target will depend on target width, etc., and, therefore, must be determined uniquely."

The "Criterion Measure" approach can be said to be useful if it can be applied to previously unclassified sets of data representing the work of different laboratories and if, in so doing, it is possible to show that data falling into the same classification depend upon the same independent variables. To use the different studies in the literature for this purpose, it is necessary to assume that non-systematic differences between studies at common levels of an independent variable are random error. With this assumption one may average across studies in an attempt to find a systematic relationship between averaged dependent measures and the levels of the independent variable at which the averages fall. That is, relationships should be revealed as a result of these procedures if the following conditions hold:

1. The independent variable has a systematic effect.

2. The independent variable can be or is dimensionalized on a quantitative scale having at least rank order properties.

3. The descriptions of the independent and dependent variables are precise enough for inter-study comparisons.

4. The test or experimental procedures are an adequate basis for drawing conclusions from the results.

Even if none of the above conditions held except the third one, the application of a useful classification system to a set of performance results would provide important information. If a sufficient number of studies were available for use, and if they extended over a reasonable range of the independent variable, classification would indicate whether the variable has a systematic effect and, possibly, its nature. If no functional relationship could be determined, it would provide an organization of the data with which one could determine where the weight of evidence falls. At the very least, if the range of the studies were very limited, it would indicate this and point to where more testing or research is needed.

The Literature Base

The literature base to which the classification system was applied consisted of three sets of experimental reports from the scientific literature included in the human performance data base developed in the project:

1. Eighty-seven studies of the effects of massed and distributed practice carried out between 1914 and 1968 inclusive.

2. One Hundred forty-eight studies of the effects of knowledge of results carried out between 1938 and 1968 inclusive.

3. Seventy studies of the effects of acoustic noise carried out between 1929 and 1968 inclusive.

In each case it was necessary to evaluate the paper for (a) sufficient precision of description of tasks and procedures, and (b) experimental adequacy. If the paper was found adequate on these counts, it was classified into the "Criterion Measure" categories. Sensory studies and studies involving complex tasks, i.e., those that were combinations of classes were not used. Finally, because the experimental conditions varied widely among studies with respect to other factors, no study was accepted unless it provided a controller comparison. With a control group available, it was possible to make decisions about the effect of the experimental conditions that were used.

This "quality filter" phase of the study cannot be over-emphasized. One approach would have been to index all studies, as is the case in many current bibliographic and "human engineering" data files. However, it became readily apparent that quality control of studies was essential to afford any meaningful test of our taxonomic system. The details of each effort, with different parts of this data base, follow.

RESULTS

Application of the Taxonomy to Massed Versus Distributed Practice

Of the eighty-seven studies available on massed versus distributed practice, thirty-five were eliminated for one or the other reason given above. The remaining fifty-two were classified according to "Criterion Measure" system applied to the tasks utilized in these studies.

Since the studies varied widely in the amount of practice given, and in the number of data points on learning curves made available, single measures were developed from each as a data reduction step. Specifically, the arithmetic mean was calculated for the last four trials of each comparison condition regardless of the number of trials employed. All further discussion, except where noted otherwise, is based on these values as basic data.

As a first step toward finding effects, the results were coded according to whether distributed practice produced an increment (+) in performance, no effect (0), or a decrement (-) compared to the massed control condition of the experiment. Each distributed practice comparison condition was treated as a separate result. Since many studies had more than one distributed condition, a total of 111 experimental comparisons were available.

Figure 1 presents a distribution of the results. For this figure, studies were included which did not actually present data, but which instead, provided the results of statistical analysis. The figure shows that most of the tasks were classified as of the "simple coding" type. In fact, most of them were studies of verbal learning. No studies fell into either "searching" or "group coding", and very few into "switching."

For each of the three remaining task categories, simple coding, successive coding, and tracking, it is clear that the weight of the evidence favors distributed practice as the learning condition which produces improved performance. This conclusion is consistent with the general understanding of the field.

It cannot be determined from Figure 1 whether or not the instances of no effect and of decrement are the result of a poor choice of comparison between massed and distributed conditions. That is, if the function of distributed practice reaches a limit, and if

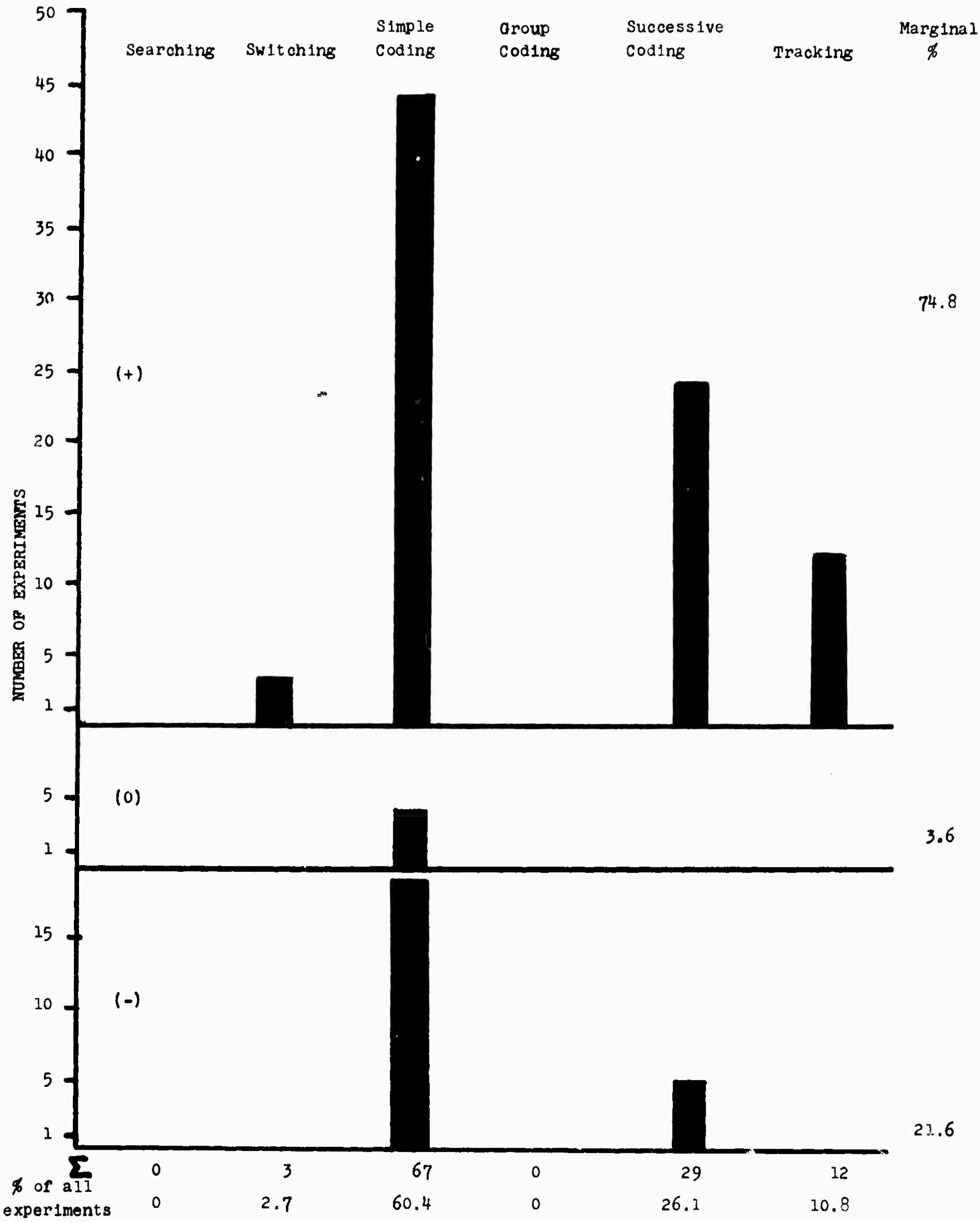


Figure 1. Distribution of qualitative effects of distributed practice reported for 111 experimental comparisons

the control and experimental groups were both selected near the limit, no difference might occur. Similarly, it is possible that beyond some limit of inter-trial interval, distribution might be decremental compared to a particular control condition. Aside from this, since the time between trials is a dimensionalized variable, it was desirable to analyze the data in a way that might test the classification system's ability to show trends and, hopefully, functional relations.

To achieve this, several steps had to be taken. First, studies not providing quantitative data were eliminated. For the remaining studies a common metric had to be developed to deal with the problems of different measurement units in different studies. The common measure used was percent change of each experimental comparison from its control condition. Finally, the decision was made to exclude those few studies which used massed control conditions longer than ten seconds between trials.

In reviewing the studies, it was found that studies varied markedly with regard to selection of a control condition so that what was treated as a "distributed practice condition" in one study was used as a "massed practice condition" in another. To handle this problem, the studies were grouped into class intervals of the massed control condition, viz. 0-3 seconds, 4-7 seconds, 8-10 seconds.

Simple coding task results. The results for simple coding are shown in Figure 2. The plot in Figure 2, of course, represents an enormous variety of confounding. Nevertheless, inspection of the figure shows that the weight of the evidence favors distribution and, for the shortest massing interval (0-3 seconds), that the amount of improvement, on the average and without regard to any other consideration, may increase with increasing distribution.

To investigate this further, the values of Figure 2 at fixed conditions of distributed practice were averaged and plotted in Figure 3-a as a function of distributed interval. The straight line in the figure was drawn by eye.

Figure 3-a shows very clearly that on the average the percentage improvement with practice is proportional to the length of the interval used for the distributed condition. However, the figure also shows

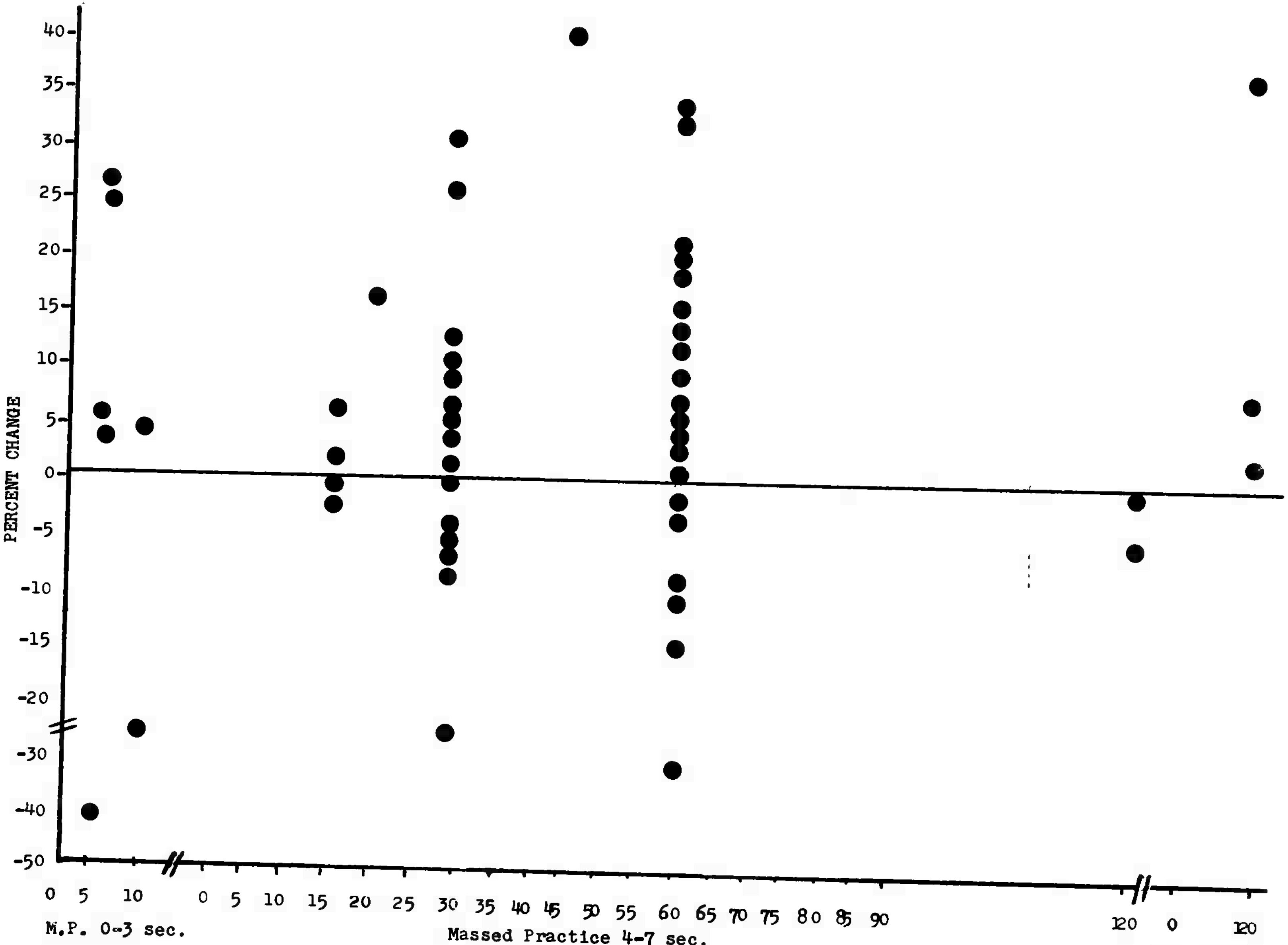


Figure 2. Percentage change in performance as a function of intertrial interval for simple coding tasks

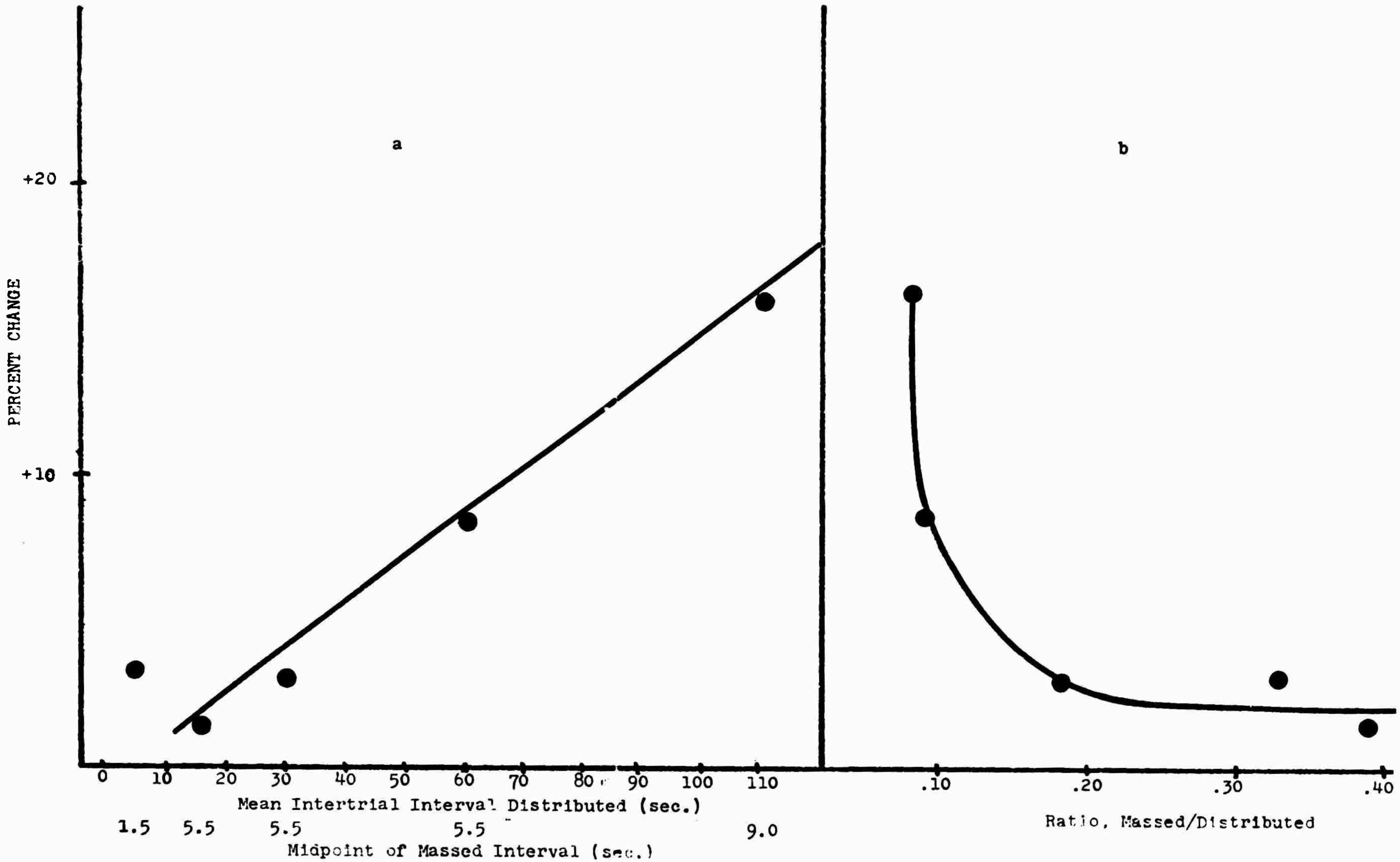


Figure 3. Percentage change in performance as a function of mean intertrial interval (a) and ratio of intertrial intervals (b) for 61 experimental comparisons among simple coding tasks

that as the distributed interval increased in the studies, the control interval also increased. Since the measure is increasing, it follows that the percent change can be expressed as a function of the ratio of the two conditions. Furthermore, since the change is linear, the ratio function must be non-linear. To examine this, the values in Figure 3-a were plotted as a function of the ratio of the massed condition to the distributed condition as shown in Figure 3-b. It is clear from this result that the greater the difference between the two (massed-distributed) conditions, the greater was the improvement. The function, drawn by eye, is reasonably smooth. Its greatest value is that it confirms the linearity suggested by Figure 3-a.

Successive coding task results. Figure 4 provides the percent change in studies classified in terms of another task category, "successive coding." Intervals were not used here since the studies available tended to use either 0, 2, or 4 seconds as a control condition. Inspection of this figure suggests a trend which increases to a limit within the 0-second studies and which may, in fact, continue over the figure or decrease again without regard to the control conditions.

Trial plots of the mean percent change suggested that the relationships are not the same across studies with different control conditions as was the case for simple coding. Therefore, means were plotted separately, for studies having a 0-second control and a 2-second control, as shown in Figure 5. Since only one experiment was available at the 4-second control condition, it was dropped at this point of analysis.

Figure 5, for successive coding tasks is much more complex than was Figure 4, for simple coding tasks. The lines, drawn by eye, represent an attempt to express the trends that are suggested. That is, both sets of data represent an increase in percentage improvement in performance with increasing distribution followed by a decrease in percentage improvement. The fits are reasonably good, but clearly, more work is needed to determine what functions really hold. Meanwhile, the trends of Figure 5 may serve as hypotheses. The hypotheses, in fact, are reasonable if one considers the nature of the successive coding task. This is a task in which successive responses depend upon previous responses, i.e., there is a contingent probability holding between successive stimuli as opposed to simple coding

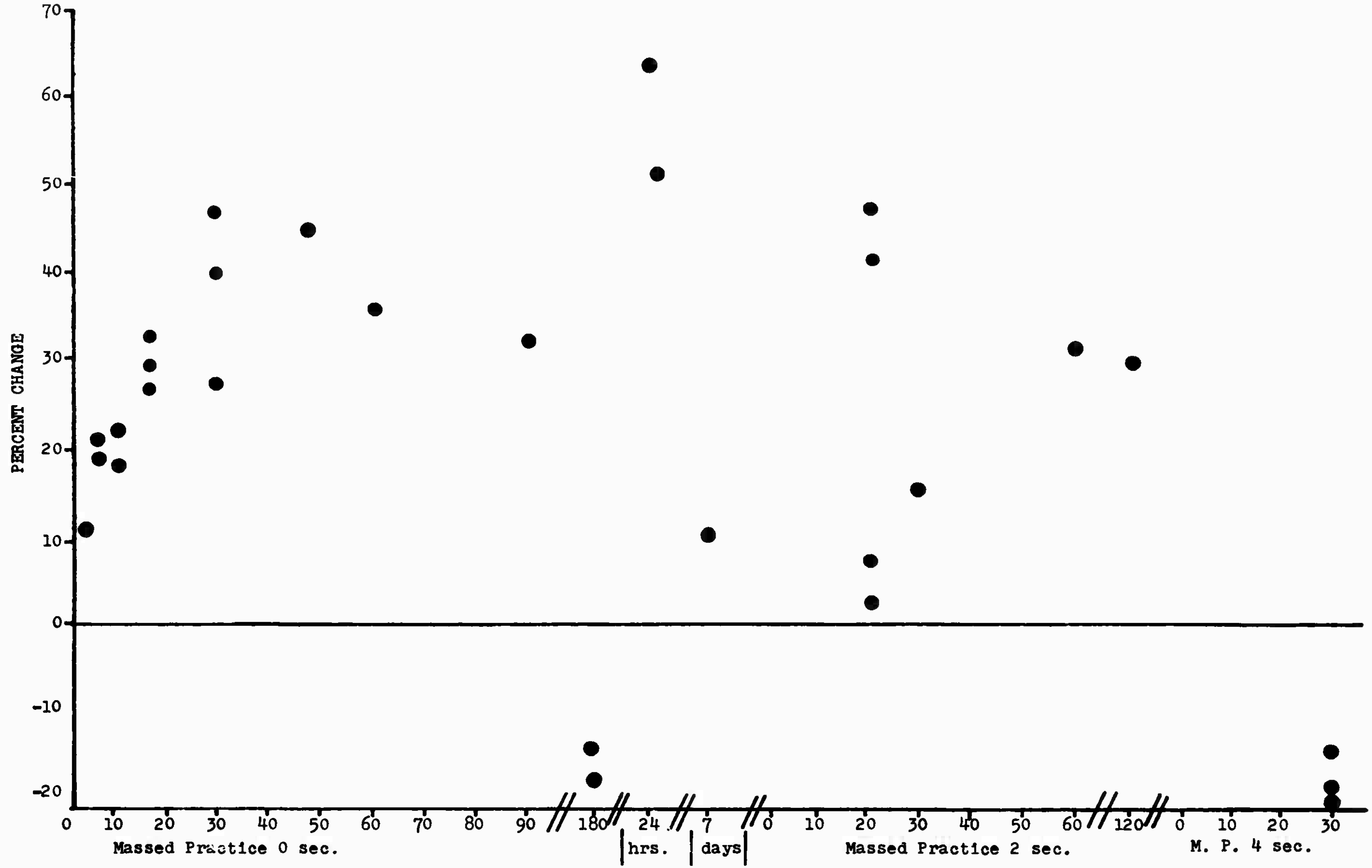


Figure 4. Percentage change in performance as a function of intertrial interval for successive coding tasks

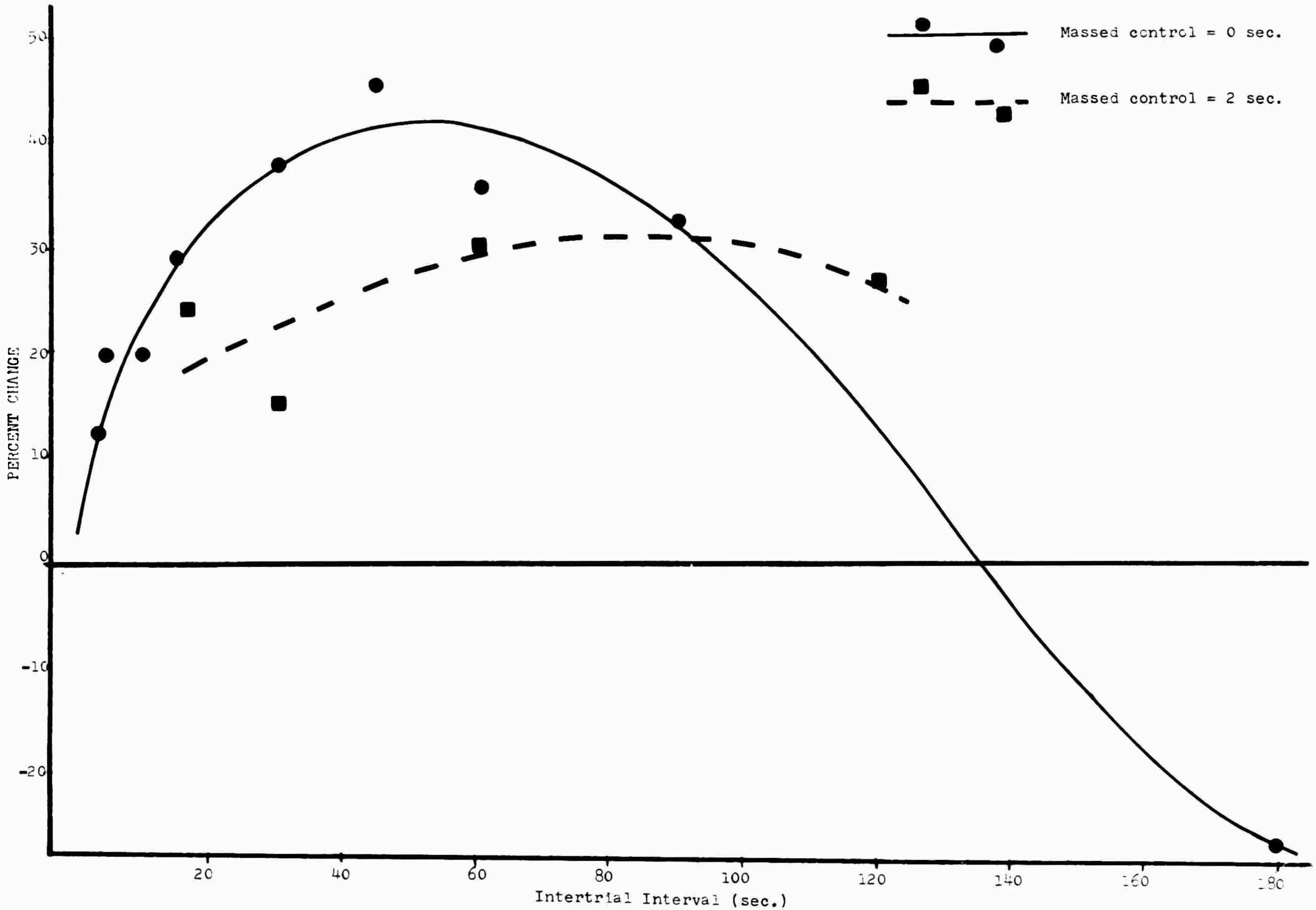


Figure 5. Percentage change in performance as a function of mean intertrial interval for 23 experimental comparisons among successive coding tasks

where each stimulus is an independent event. Under the conditions of successive coding, short-term memory would be expected to be a very important cognitive process, as postulated by Teichner and Olson (1969). The longer the intertrial interval the greater the risk of decrement due to forgetting. On this basis the decreasing incremental effect of distribution would be overcome by the increasing decremental effect of forgetting. The result would be a curve which first increased and then decreased as, in fact, is shown in Figure 5.

Tracking task results. The effects of the intertrial interval on tracking are shown in Figure 6. The data are those from ten studies which were considered to have produced acceptable quantitative results, or which used massed control groups with not more than twenty seconds between trials. The figure shows the effects of comparisons made against control conditions having zero time between trials (i.e., continuous practice), two seconds between trials, ten seconds between trials, and twenty seconds between trials. These four conditions are arranged from left to right according to the number of studies available for each rather than in any other systematic way. The smooth line, fitted by eye to the 0-second control comparisons, ignores the higher of the two 30-second distributed conditions on the assumption that, since it is out of the range of all other studies, it is unrepresentative.

Figure 6 shows that distributed practice produces better performance than massed practice under all conditions in which comparisons were made. The results also suggest that the gain to be expected with the more distributed condition decreases as the intertrial interval associated with it increases. The smooth line provides a general statement of that relationship. The curve suggests that the effect of increasing distributed condition intervals decreases to a limit. However, it is possible that with intervals longer than those studied, the curve might continue its drop to some point where, relative to a smaller interval, the distributed condition would be deleterious.

The remaining portions of Figure 6 are difficult to interpret beyond what has already been said, i.e., the gain in performance attributable to the more distributed condition is less the longer the distributed interval.

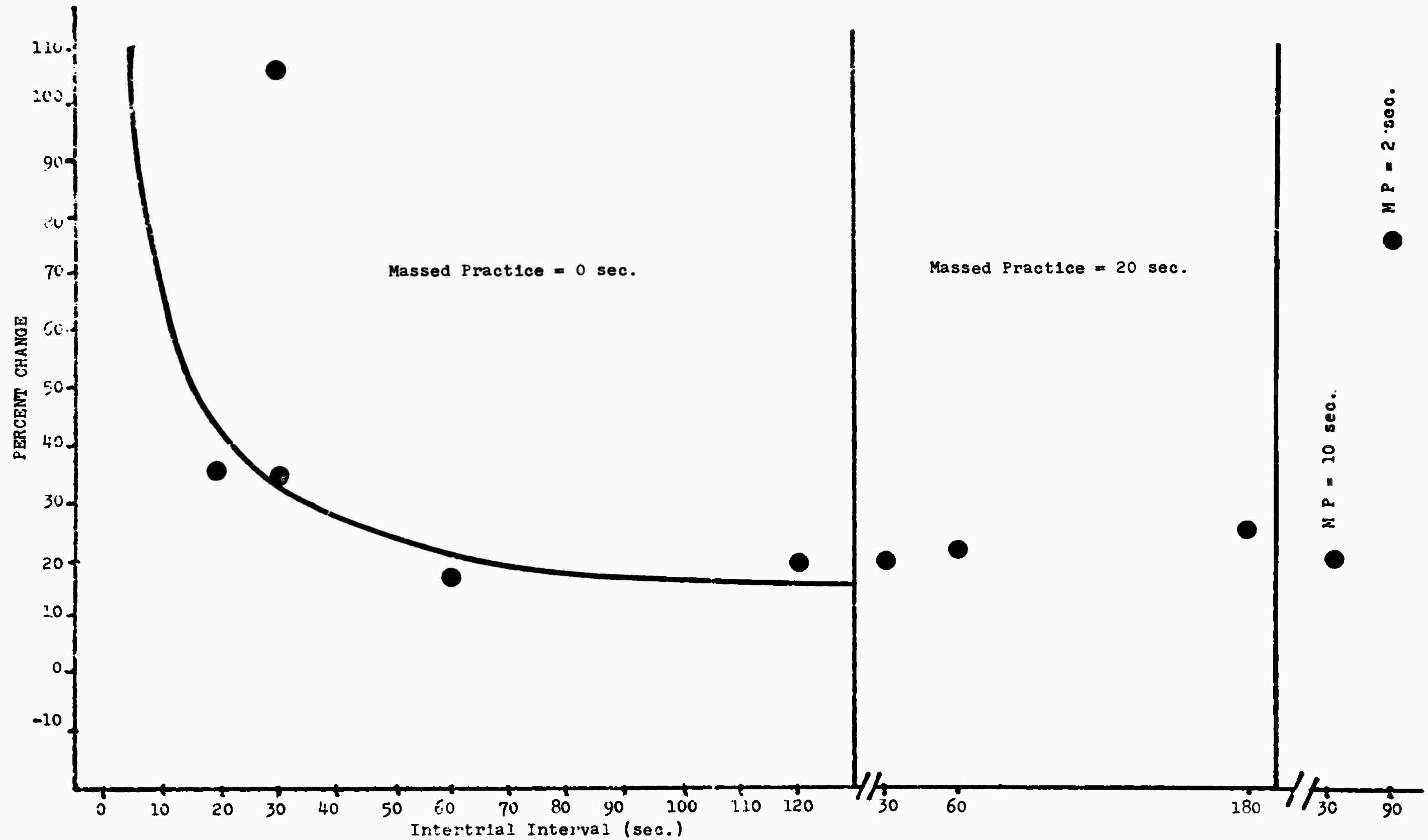


Figure 6. Percentage change performance as a function of intertrial interval for tracking tasks

To investigate this further, as well as to seek a single dimension on which to put the various studies, the data of five studies were plotted as a function of the ratio of the massed interval to the distributed interval. The results of this operation are shown in Figure 7 where it may be seen that all five studies are ordered systematically regardless of the length of the intervals used. The function, drawn by eye, drops rapidly and is flat between .10 and .20 after which the gain is small and constant. We may conclude that the greater the difference between the massed and distributed intervals, the greater the gain to be associated with the distributed condition until the ratio of the two approaches .20. After that value the gain is approximately twenty percent regardless of the difference. The conclusion holds for the continuous practice comparison as well, as was shown in Figure 6. That is, the greater the distributed interval, the less the gain up to about 80 seconds between trials. After that the distributed condition is associated with a gain of about fifteen percent.

The suggestions indicated by our organization of the data must be qualified, of course, by the procedures that we used to develop a comparison measure. In particular, variations due to the different amounts of practice used are confounded in the measure. Our means, based on the last four practice trials, are necessarily sensitive to the steepness of the learning curve at these trials. Thus, studies which provided extensive practice are likely to show smaller differences between the massed and distributed conditions than are studies with fewer trials because the latter are more likely to be at a steep part of the learning curve. Our use of the percentage difference equalizes this factor only in part. On the other hand, the systematic nature of the results suggests that these other considerations were not enough to obscure the effects of the intertrial interval.

Application of the Taxonomy to Knowledge of Results Studies

The second learning research area investigated by means of the "Criterion Measure" taxonomy was that of the effects of "knowledge of results." Although it is generally accepted that learning reaches a higher level when the learner is provided with knowledge of results, great difficulty

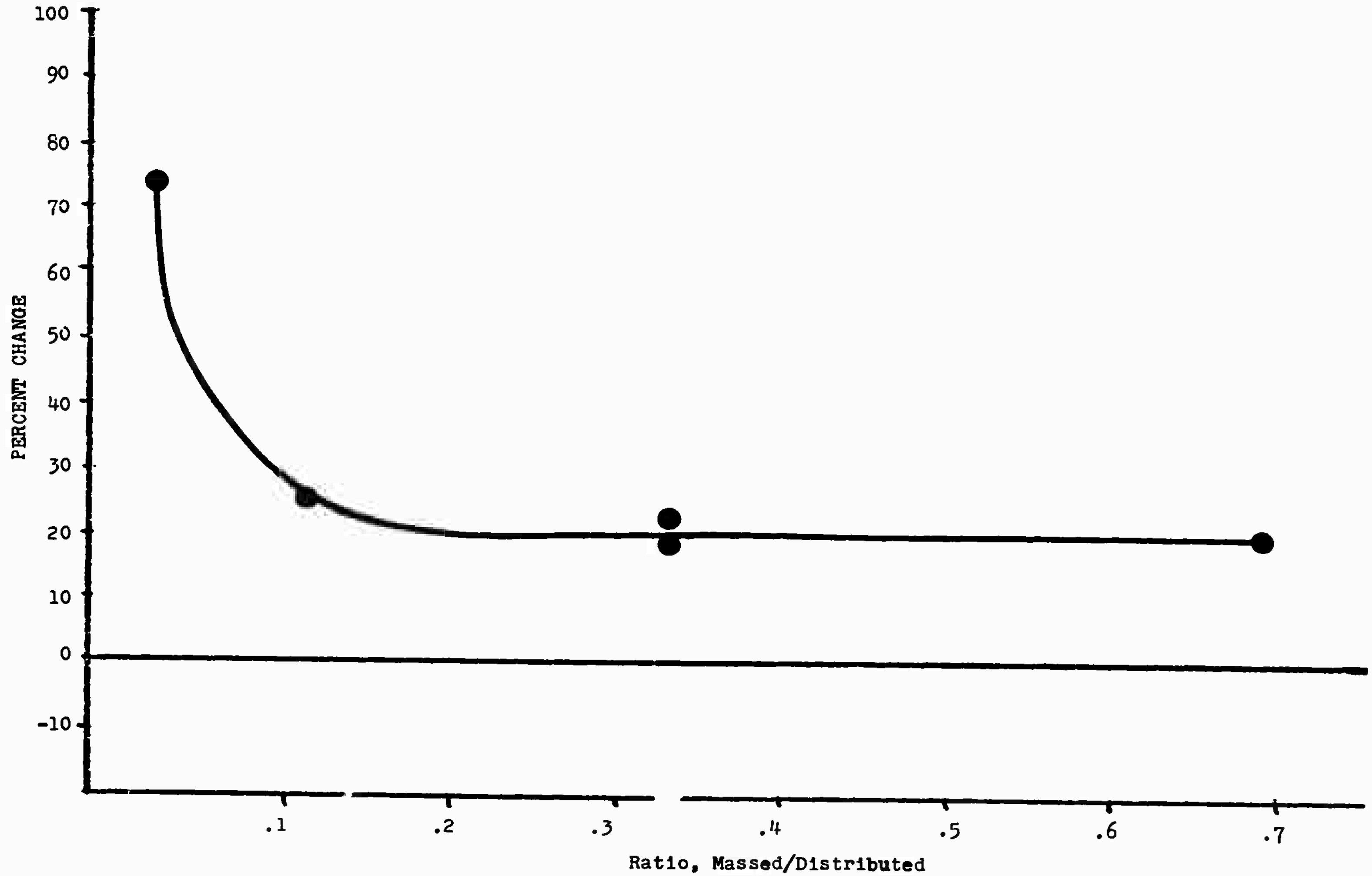


Figure 7. Percentage change in performance as a function of the ratio of massed to distributed intertrial intervals for tracking tasks

was encountered in accepting many of the studies which have purported to have verified that principle. One problem that arose had to do with the distinction between a signal or stimulus which provides knowledge of performance and one which provides information without which the subject is unable to perform the task. In the first case, a signal which provides knowledge of results (KOR) is simply a redundantly informing signal. One example arises in those tracking studies in which a signal tells the subject that he is on-target. In actuality it merely tells him what he already knows. Sometimes such a signal has been called augmented feedback. Regardless of what it may be called, it is difficult to accept such a study as having shown that performance is poor without knowledge of results. On the other hand, studies which have hidden the visual target, and thereby not provided necessary information have produced such poor performance that a KOR signal acts simply as a delayed informing signal.

Another example arises in search studies in which the subject is given a signal to indicate that he detected a target. In most cases, it was not necessary to do this since the subject could tell that he had detected it. Telling the subject that he has missed a target seems to be a clearer instance of KOR. Perhaps, the redundant signal should be thought of as a reward rather than KOR. In any case, it is logically possible to conceive of a variety of ways in which what has been called KOR might be provided. For example, the subject might be informed only when he is right in some sense, e.g., on target. Or he might be informed only when he is wrong in some way. Or he might be provided with both right and wrong information. There are still other possibilities which include the direction and the amount of error. Because performance might depend differentially upon these various KOR conditions, we felt the necessity of partitioning the studies available in terms of them.

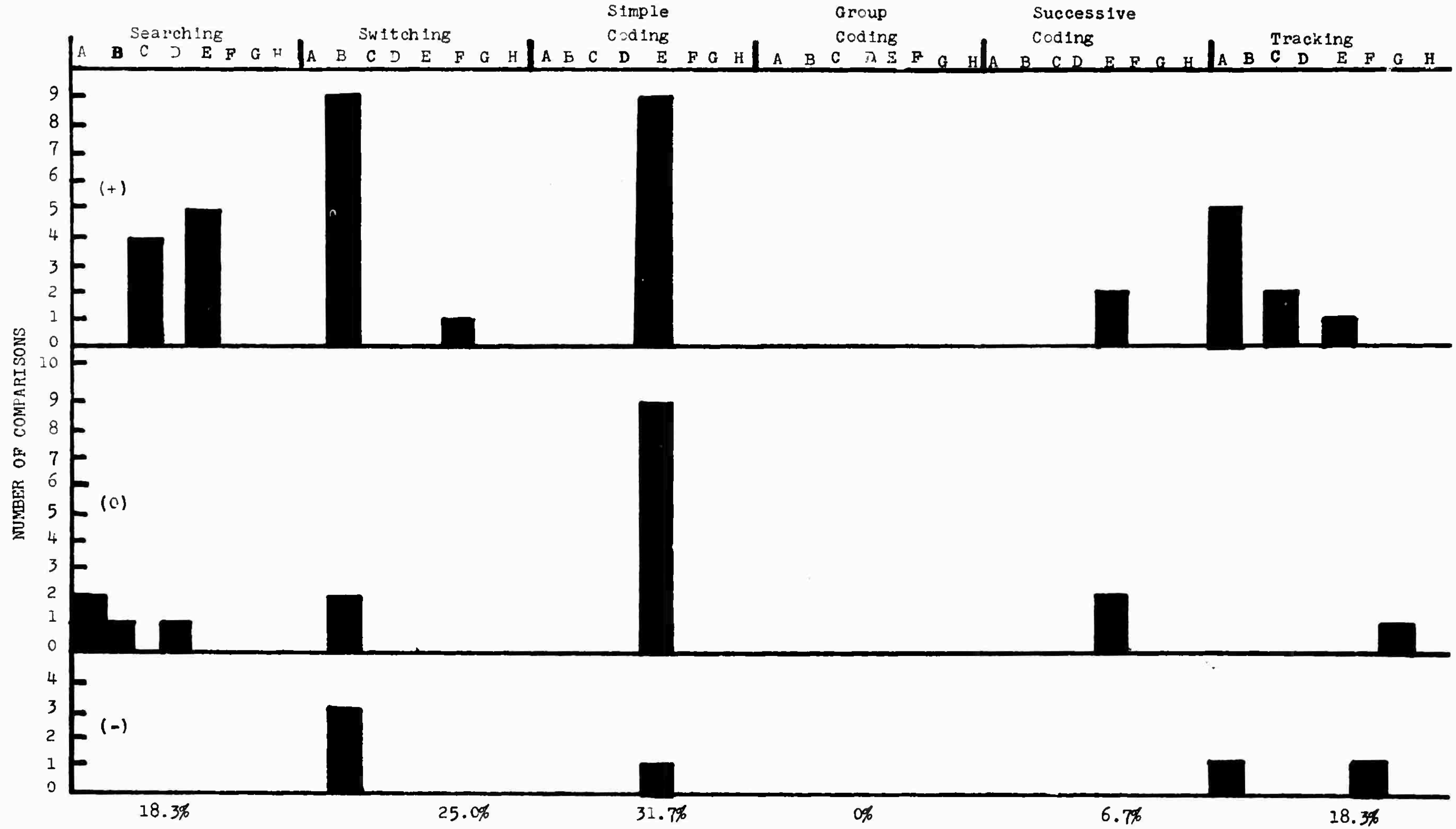
A second kind of problem arose because KOR has not often enough been studied in a way which provides a dimension of amount of KOR. The literature allows only qualitative comparisons. In counting the comparisons we ignored the manner or providing KOR, whether verbally, with signal lights or buzzers, etc. As before, studies failing to provide a control group or those which appeared to be based on inseparable experimental confoundings, etc., were rejected. Those studies which did provide more or less acceptable conditions yielded sixty experimental comparisons. Figure 8 summarizes the results in terms of whether KOR produced a relative gain, no effect, or a decrement for each of eight possible KOR parameter combinations.

Figure 8 shows that the most frequently-made comparisons involved the task classification of "simple coding." No "group coding" studies were found at all. The figure also shows that the nature of KOR provided varied with the task. Most tracking studies provided only "correct" information, whereas both simple and successive coding studies were restricted to the provision of "correct" and "error" information. Searching studies used "correct and error" slightly more frequently than any other kind, with "error" a close second

Figure 8 shows that KOR aided learning in nine comparisons of "search" performance and had no effect in four comparisons. On the other hand, none of the nine comparisons used the same KOR conditions as the four which had no effect. It appears, therefore, that a conclusion favoring KOR for "searching" must be limited to the "error" only or the "correct and error" kinds of KOR information.

KOR was beneficial in nine out of fifteen comparisons of "switching" in which KOR was expressed as "correct and direction" and one case of "correct, error, and direction." Some form of augmented KOR or signal information, as the case may be, did aid "tracking," but three of the eleven comparisons did not favor KOR. Beyond that, for the one form of KOR used, it cannot be concluded that KOR aided learning for either "simple" or "successive coding."

Figure 8 demonstrates that the weight of the evidence favors KOR slightly, but whether it really aids performance depends on the task and the form of KOR employed. Since the data reported do not lend themselves to a meaningful quantitative analysis, these conclusions must be restricted to the presence or absence of KOR rather than the amount of KOR.



A	Right	13.3%	E	Right & Wrong	48.3%
B	Right & Direction	23.3%	F	Right, Wrong & Direction	3.3%
C	Wrong	10.0%	G	Right, Wrong & Amount	1.7%
D	Wrong & Direction	0.0%	H	Right, Wrong, Direction & Amount	0.0%

Figure 8. Distribution of the qualitative effects of knowledge of results for 60 experimental comparisons

Application of the Taxonomy to Studies on the Effects of Noise

Our review of the scientific literature on the effects of noise on human performance led to the conclusion that this literature is one of the poorest in terms of scientific rigor. Aside from studies which were rejected because of poor or ambiguous procedures, a large number of studies were rejected for a failure to specify the noise levels used! These included studies with limited descriptions of control conditions (e.g., "quiet") as well as those which presented noise at a specified level from a speaker to a subject, but at an unspecified, or undetermined, or variable distance and position from the source. Some did not specify whether the level was measured at the source or at the subject.

It also appears that one investigator's quiet is another's noise. Thus, the "quiet" control condition in many studies was a more intense acoustic exposure than the experimental noise condition in other studies. Finally, as a major criticism, it should be noted that the noises used included continuous, intermittent, pure tone, broad band sound, etc., sometimes unspecified and often passed through unspecified impedances before reaching the subject.

The first step taken to organize the noise literature was to plot the frequency of occurrence of reported improvements, decrements, and "no effects." This was done without regard to whether the study provided data which could be used for quantitative purposes. The results for each of the task classes are shown in Figure 9.

It is apparent from Figure 9 that none of the studies fell into the "group coding" class. It is also apparent that the most frequent result was a failure to show an effect of noise. Beyond that, improvements were essentially as frequent as decrements. Since all three possible results were actually very similar in frequency of occurrence, Figure 9 suggests that acoustic noise has no significant effect on performance. The conclusion appears warranted regardless of how the tasks might have been classified. This conclusion is based upon the marginal frequencies and is upheld by the frequencies plotted within task classes as well.

Figure 9 was based upon the general results reported. It is possible that the effects of noise are dependent on the nature of the exposure. For example, the initial effect of noise might be a decrement or an

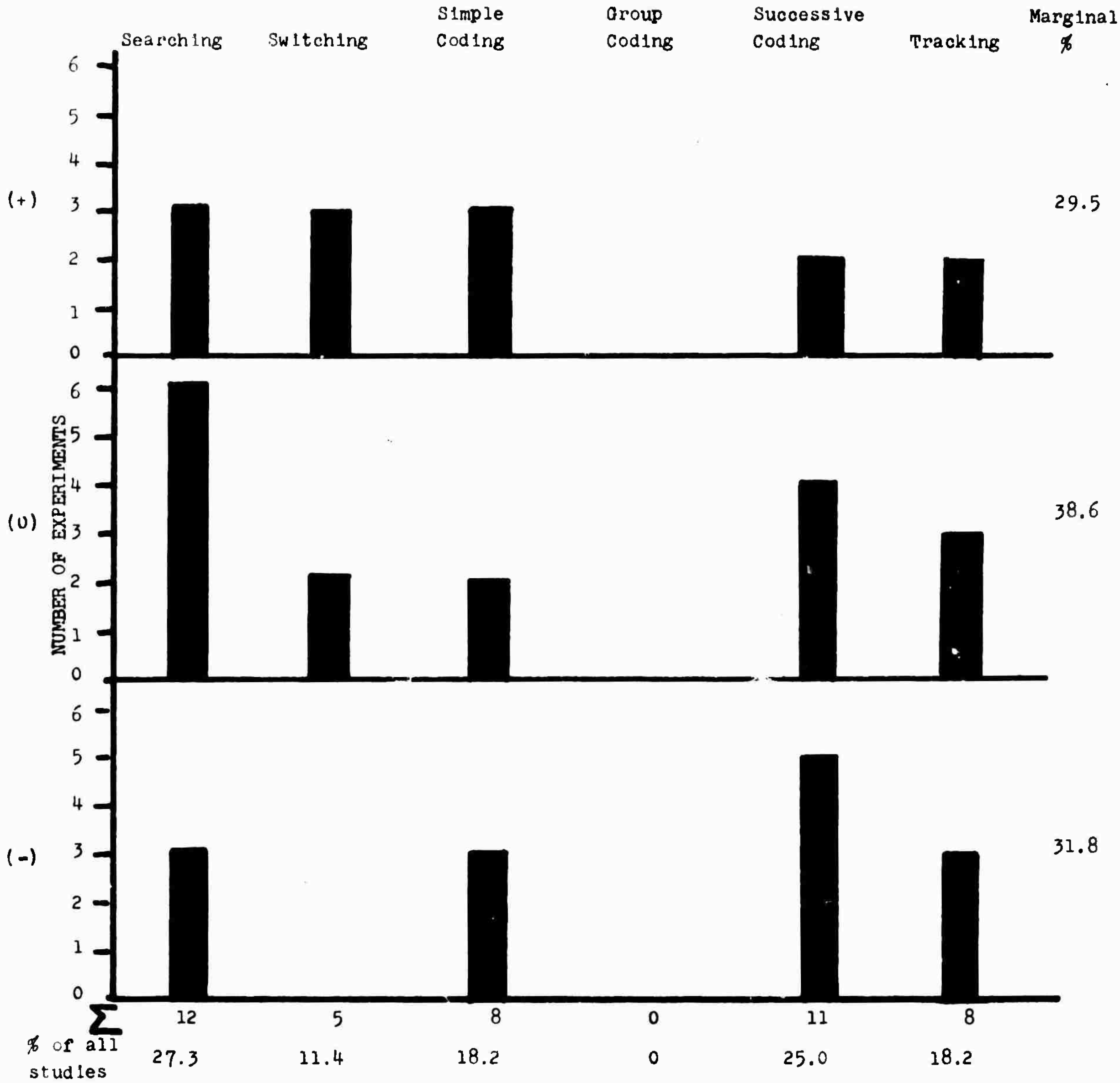


Figure 9. Distribution of qualitative effect of noise reported for 44 experimental comparisons

increment of performance. With continued exposure, that effect might be altered. To investigate this possibility the studies on which Figure 9 was based were coded according to a more detailed analysis of the results.

The effect of this operation is shown in Figure 10 for the five task classes into which the literature fell. Unfortunately, as the figure shows, the refinements employed do not permit changing the conclusions drawn from the previous figure.

An attempt was made to investigate the possibility that the quantitative results might have information not revealed qualitatively. The number of studies available for this purpose totaled eighteen. For each, a percent change measure was determined exactly as for the massed vs. distributed practice literature described earlier. The studies were then further subdivided according to the level of quiet control condition, e.g., 30-45 dB or 75-90 dB re .0002 dyne/cm^2, etc., depending upon the conditions used with each task in the literature, and plotted as a function of the noise levels of the experimental groupings.

Only one study remained available for "simple coding" and two for "successive coding." All three studies reported decrements. Plots of five "searching" studies suggested either no effect (one study) or an improvement (three studies) or a decrement (one study). The decrement was, interestingly, reported with the most complex (3-clock monitoring) of the five search tasks.

Plots of five tracking studies showed no effect when the experimental condition was 100 dB compared to a control between 60-75 dB. The three remaining studies did show decrements, but not exceeding ten percent.

Regardless of possible quantitative effects, it is not logically sound to draw a different general conclusion from the more quantitative analysis than from the qualitative one. They differ, in one sense only, in that the former is based upon the discarding of relevant information. In any case, the plots made were not considered to present anything reliable. For this reason they have not been presented. We conclude that the effects of noise are either not demonstrated or that they are not there to be demonstrated.

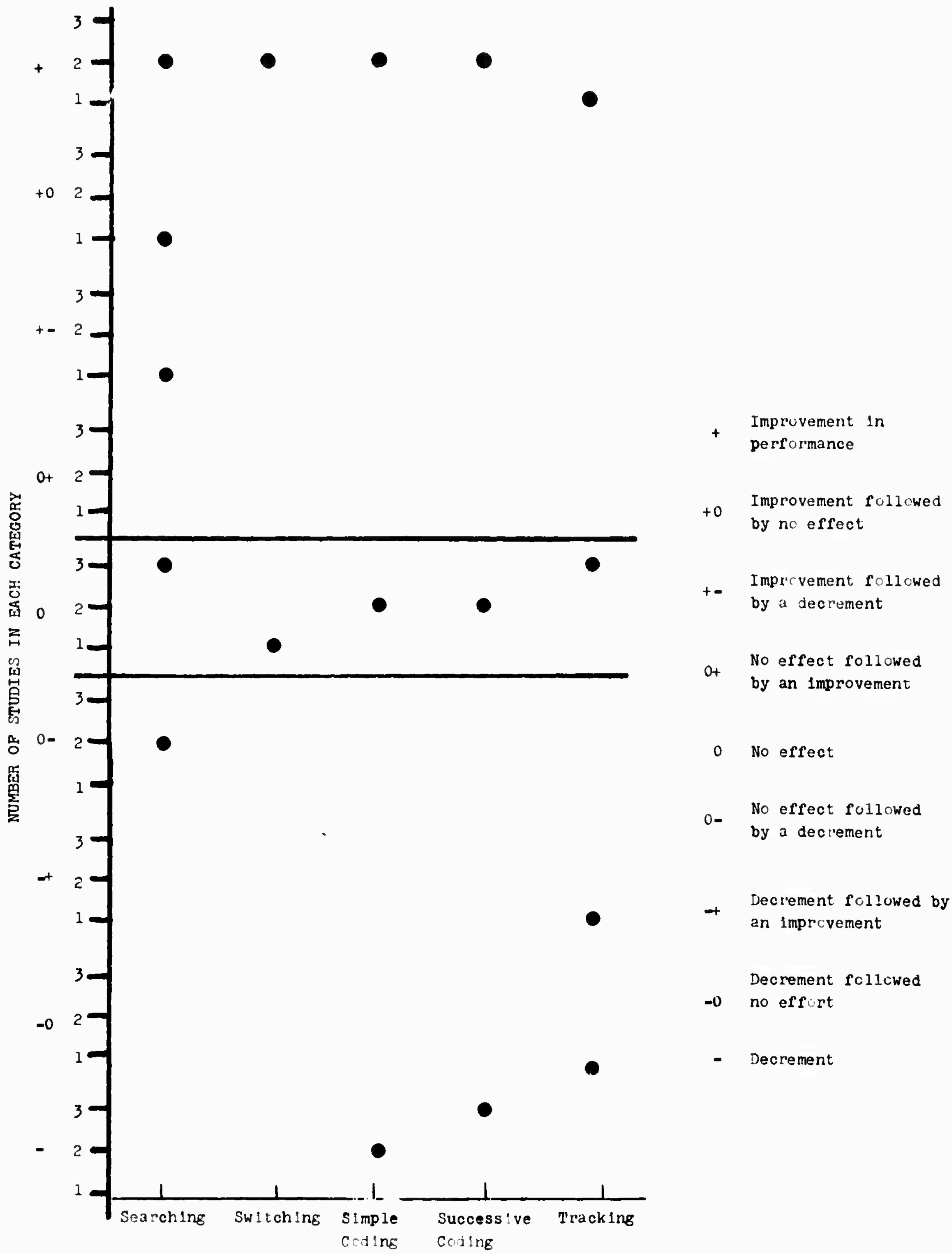


Figure 10. The effects of noise on performance

DISCUSSION AND CONCLUSIONS

An attempt was made to evaluate the feasibility of a human performance data base using the method of Teichner and Olson (1969) to classify the tasks found in the literature. We have called this the "Criterion Measure" approach to task classification since the classification is operationally defined by the measure itself. There is no additional inference about function or process involved required.

Of those literatures sampled, two ("knowledge of results" and "effects of noise") did not appear to contain enough studies of a reliability sufficient for the purposes to which a data base might be put. This conclusion is quite independent of the task classification system. The only one of the three literatures which does appear to be useful, after evaluation of individual studies, is that concerned with massed and distributed practice.

The task classification system was applicable to the studies surveyed regardless of area. This is a general conclusion based upon ease of application. The ease of application of the method decreases for those tasks which Teichner and Olson defined as combinations of the simpler tasks. For that reason we have not presented the results obtained with that classification, although it was used.

It was noted earlier that the study was intended as a joint test of the classification system and the literature. As it turned out, the literature could be evaluated independently in terms of marginal frequencies and numbers of available acceptable studies. Since the classification system was internally consistent with those overall evaluations, it would appear to be supported as a feasible system. Even more convincing, however, was the finding that with the system it was possible to organize the literature on distributed practice in terms of: (a) functional relationships and (b) different functions for different task categories. In fact, some hitherto unreported relationships were strongly suggested. It is important to note that these "principles" are general to operationally defined task categories where each category contains a variety of different tasks.

The application of the taxonomy to studies of massed vs. distributed practice led to several interesting functional relationships. For <u>simple</u>

coding tasks, performance change was a linear function of intertrial interval in the range of 10 to 110 seconds, with massed practice periods of 1.5 to 9 seconds. When these results were plotted as a function of the ratio of the massed condition to the distributed condition, it was indicated that the greater the difference between the two conditions the greater was the improvement in performance. For successive coding tasks, on the other hand, it was determined that there was an increase in percentage improvement in performance with increasing distribution followed by a decrease in percentage improvement.

The tracking task results suggested that distributed practice produces better performance than massed practice and that the gain to be expected with the more distributed condition decreases as the intertrial interval associated with it increases. This result is true, however, only for comparisons made against control conditions having zero time between trials (that is, continuous practice). When performance was plotted as a function of the ratio of massed to distributed practice it was apparent that the greater the difference between the massed and distributed intervals, the greater the gain that was associated with the distributed condition until the ratio of the two approached .20.

The application of the taxonomic system to knowledge of results studies and noise studies did not provide as clear a set of relationships as was the case for massed vs. distributed practice. While the weight of the evidence indicated that knowledge of results did result in improved performance, whether or not it really aided performance depended upon the task and the form of knowledge of results employed. The data did not lend themselves to a meaningful quantitative analysis so these conclusions must be restricted to the presence or absence of knowledge of results rather than the amount.

In terms of our task categories, it was apparent that switching tasks provided the most consistent results. For these tasks, knowledge of results aided performance. For the other types of tasks, that is searching, simple coding, successive coding and tracking, the data did not indicate any systematic increment or decrement in performance as a result of providing knowledge of results.

With respect to the noise literature, it was apparent that the most frequent result was a failure to show consistent effects of noise on tasks in any category. Improvements were essentially as frequent as decrements under high noise conditions. The suggestion was that acoustic noise in the ranges previously investigated has no significant effect on performance. It was concluded that the effects of noise are either not demonstrated or that they are not there to be demonstrated. Other task classification systems may be more useful in illuminating whatever effects are there.

It should be pointed out that the above relationships are illustrative of the types capable of being developed with such systems. It is also important to note that, had the tasks been grouped without regard to the separate taxonomic categories, these functional relationships would have been obscured and few generalizations about performance would have been possible.

We conclude that both the method and the distributed practice literature are useful for data base purposes. Other segments of the human performance literature are probably also useful and amenable to this classification method. How far its utility will go remains to be determined empirically. On the other hand, other classification systems can now be applied to the distributed practice literature and can now be evaluated against this one. It is possible that other systems will not survive the test of application, or they might be even more successful, or they might serve to reveal still other kinds of relationships. Regardless, one important result of the present study is the identification of a usable literature and the reduction of its studies to those that are reasonably acceptable on scientific grounds.

Chambers, A. N. Development of a taxonomy of human performance: A heuristic model for the development of classification systems. Technical Report AIR-726-3/69-TR4. Washington, D. C.: American Institutes for Research, March 1969.

Farina, A. J., Jr. Development of a taxonomy of human performance: A review of descriptive schemes for human task behavior. Technical Report AIR-726-2/69-TR2. Washington, D. C.: American Institutes for Research, February 1969.

Farina, A. J., Jr., & Wheaton, G. R. Development of a taxonomy of human performance: The task characteristics approach to performance prediction. Technical Report AIR-726/2035-2/71-TR7. Washington, D. C.: American Institutes for Research, February 1971. (U. S. Army Behavior and Systems Research Laboratory Research Study 71-7.)

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

Korotkin, A. L., & Chambers, A. N. A human performance data base for evaluation of taxonomies. Paper presented at the annual meeting of the American Psychological Association, September 1969.

Levine, J. M., & Teichner, W. H. Development of a taxonomy of human performance: An information-theoretic approach. Technical Report AIR-726/2035-2/71-TR9. Washington, D. C.: American Institutes for Research, February 1971. (U. S. Army Behavior and Systems Research Laboratory Research Study 71-6.)

Miller, R. B. Development of a taxonomy of human performance: A user-oriented approach. Technical Report AIR-726/2035-3/71-TR6. Washington, D. C.: American Institutes for Research, March 1971. (U. S. Army Behavior and Systems Research Laboratory Research Study 71-5.)

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

Theologus, G. C., & Fleishman, E. A. Development of a taxonomy of human performance: Validation study of ability scales for classifying human tasks. Technical Report AIR-726/2035-4/71-TR10. Washington, D. C.: American Institutes for Research, April 1971. (U. S. Army Behavior and Systems Research Laboratory Research Study 71-9.)

Theologus, G. C., Romashko, T., & Fleishman, E. A. Development of a taxonomy of human performance: Feasibility study of ability dimensions. Technical Report AIR-726-1/70-TR5. Washington, D. C.: American Institutes for Research, January 1970.

Wheaton, G. R. Development of a taxonomy of human performance: A review of classification systems relating to tasks and performance. Technical Report AIR-726-12/68-TR1. Washington, D. C.: American Institutes for Research, December 1968.

AMERICAN INSTITUTES FOR RESEARCH